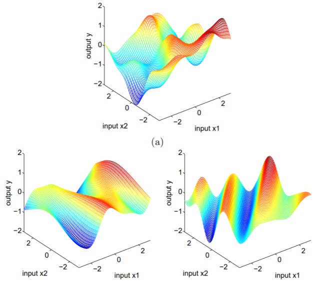
Introduction to Machine Learning

Covariance Functions for GPs



Learning goals

- Know that covariance functions encodes key assumptions about the GP
- Know commonly used covariance functions

COVARIANCE FUNCTION OF A GP

The marginalization property of the Gaussian process implies that for any finite set of input values, the corresponding vector of function values is Gaussian:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), ..., f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right),$$

- The covariance matrix $\boldsymbol{K}$ is constructed based on the chosen inputs $\left\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right\}$.
- Entry $\boldsymbol{K}_{ij}$ is computed by $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$.
- Technically, for **every** choice of inputs $\left\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right\}$, $\boldsymbol{K}$ needs to be positive semi-definite in order to be a valid covariance matrix.
- A function $k(., .)$ satisfying this property is called **positive definite**.

COVARIANCE FUNCTION OF A GP

- Recall, the purpose of the covariance function is to control to which degree the following is fulfilled:

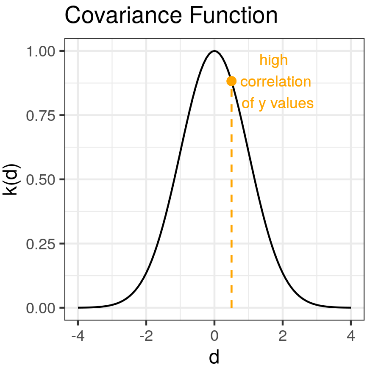
  If two points $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ are close in $\mathcal{X}$-space, their function values $f(\mathbf{x}^{(i)}), f(\mathbf{x}^{(j)})$ should be close (**correlated**!) in $\mathcal{Y}$-space.

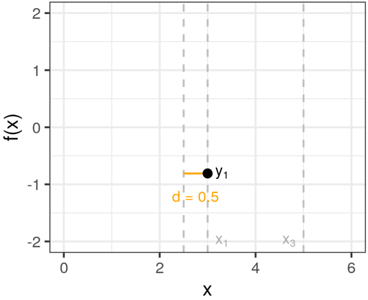
- Closeness of two points $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ in input space $\mathcal{X}$ is measured in terms of $\boldsymbol{d} = \mathbf{x}^{(i)} - \mathbf{x}^{(j)}$:

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = k(\boldsymbol{d})$$

COVARIANCE FUNCTION OF A GP: EXAMPLE

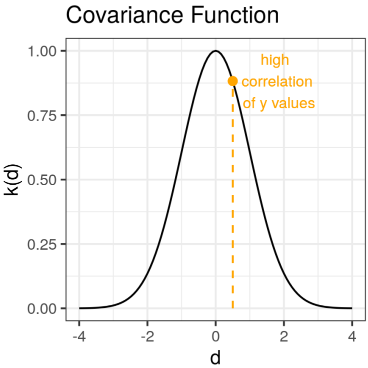
- Let $f(\mathbf{x})$ be a GP with $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}\|\boldsymbol{d}\|^2)$ with $\boldsymbol{d} = \mathbf{x} - \mathbf{x}'$.
- Consider two points $\mathbf{x}^{(1)} = 3$ and $\mathbf{x}^{(2)} = 2.5$.
- If you want to know how correlated their function values are, compute their correlation!

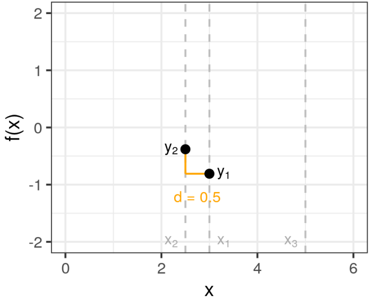




COVARIANCE FUNCTION OF A GP: EXAMPLE

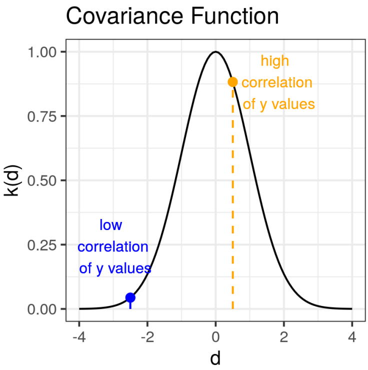
- Assume we observed a value $y^{(1)} = -0.8$, the value of $y^{(2)}$ should be close under the assumption of the above Gaussian process.

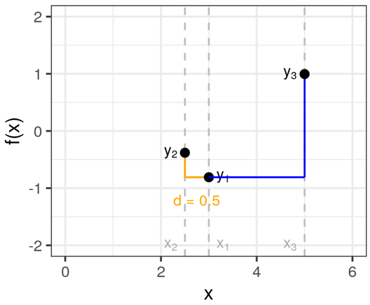




COVARIANCE FUNCTION OF A GP: EXAMPLE

- Let us compare another point $\mathbf{x}^{(3)}$ to the point $\mathbf{x}^{(1)}$
- We again compute their correlation
- Their function values are not very much correlated; $y^{(1)}$ and $y^{(3)}$ might be far away from each other





COVARIANCE FUNCTIONS

There are three types of commonly used covariance functions:

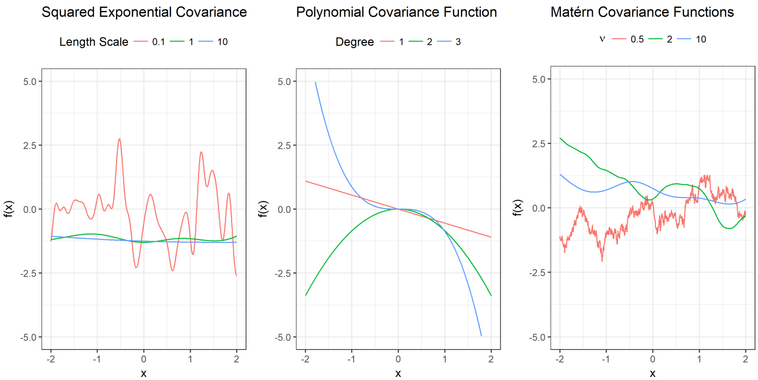
- $k(.,.)$ is called stationary if it is as a function of $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}'$, we write $k(\boldsymbol{d})$.
  Stationarity is invariance to translations in the input space:
  $k(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{d}) = k(\boldsymbol{0}, \boldsymbol{d})$
- $k(.,.)$ is called isotropic if it is a function of $r = \|\boldsymbol{x} - \boldsymbol{x}'\|$, we write $k(r)$.
  Isotropy is invariance to rotations of the input space and implies stationarity.
- $k(.,.)$ is a dot product covariance function if k is a function of $\boldsymbol{x}^T\boldsymbol{x}'$

COMMONLY USED COVARIANCE FUNCTIONS

Name	$k(\boldsymbol{x}, \boldsymbol{x}')$
constant	σ_0^2
linear	$\sigma_0^2 + \boldsymbol{x}^T \boldsymbol{x}'$
polynomial	$(\sigma_0^2 + \boldsymbol{x}^T \boldsymbol{x}')^p$
squared exponential	$\exp(-\frac{\lVert \boldsymbol{x}-\boldsymbol{x}' \rVert^2}{2\ell^2})$
Matérn	$\frac{1}{2^\nu \Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \lVert \boldsymbol{x} - \boldsymbol{x}' \rVert \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{\ell} \lVert \boldsymbol{x} - \boldsymbol{x}' \rVert \right)$
exponential	$\exp\left(-\frac{\lVert \boldsymbol{x}-\boldsymbol{x}' \rVert}{\ell}\right)$

$K_\nu(\cdot)$ is the modified Bessel function of the second kind.

COMMONLY USED COVARIANCE FUNCTIONS



- Random functions drawn from Gaussian processes with a Squared Exponential Kernel (left), Polynomial Kernel (middle), and a Matérn Kernel (right, $\ell = 1$).
- The length-scale hyperparameter determines the “wiggliness” of the function.
- For Matérn, the ν parameter determines how differentiable the process is.

SQUARED EXPONENTIAL COVARIANCE FUNCTION

The squared exponential function is one of the most commonly used covariance functions.

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right).$$

Properties:

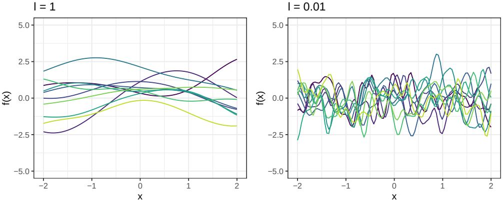
- It depends merely on the distance $r = \|\mathbf{x} - \mathbf{x}'\| \rightarrow$ isotropic and stationary.

- Infinitely differentiable $\rightarrow$ sometimes deemed unrealistic for modeling most of the physical processes.

CHARACTERISTIC LENGTH-SCALE

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\ell^2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$$

ℓ is called **characteristic length-scale**. Loosely speaking, the characteristic length-scale describes how far you need to move in input space for the function values to become uncorrelated. Higher ℓ induces smoother functions, lower ℓ induces more wiggly functions.



CHARACTERISTIC LENGTH-SCALE

For $p \geq 2$ dimensions, the squared exponential can be parameterized:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\left(\mathbf{x} - \mathbf{x}'\right)^\top \boldsymbol{M}\left(\mathbf{x} - \mathbf{x}'\right)\right)$$

Possible choices for the matrix $\boldsymbol{M}$ include

$$\boldsymbol{M}_1 = \ell^{-2}\boldsymbol{I} \qquad \boldsymbol{M}_2 = \operatorname{diag}(\boldsymbol{\ell})^{-2} \qquad \boldsymbol{M}_3 = \Gamma\Gamma^\top + \operatorname{diag}(\boldsymbol{\ell})^{-2}$$

where $\boldsymbol{\ell}$ is a p-vector of positive values and Γ is a $p \times k$ matrix.

The 2nd (and most important) case can also be written as

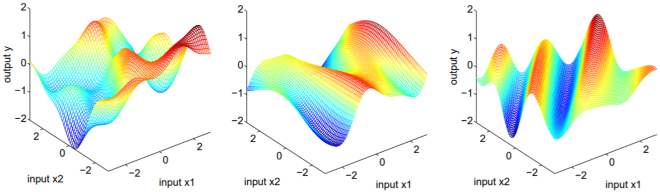
$$k(\mathbf{d}) = \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\frac{d_j^2}{l_j^2}\right)$$

CHARACTERISTIC LENGTH-SCALE

What is the benefit of having an individual hyperparameter ℓ_i for each dimension?

- The $\ell_1, \ldots, \ell_p$ hyperparameters play the role of **characteristic length-scales**.
- Loosely speaking, ℓ_i describes how far you need to move along axis i in input space for the function values to be uncorrelated.
- Such a covariance function implements **automatic relevance determination** (ARD), since the inverse of the length-scale ℓ_i determines the relevancy of input feature i to the regression.
- If ℓ_i is very large, the covariance will become almost independent of that input, effectively removing it from inference.
- If the features are on different scales, the data can be automatically **rescaled** by estimating $\ell_1, \ldots, \ell_p$

CHARACTERISTIC LENGTH-SCALE



For the first plot, we have chosen $\boldsymbol{M} = \boldsymbol{I}$: the function varies the same in all directions. The second plot is for $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{\ell})^{-2}$ and $\boldsymbol{\ell} = (1, 3)$: The function varies less rapidly as a function of x_2 than x_1 as the length-scale for x_1 is less. In the third plot $\boldsymbol{M} = \Gamma\Gamma^T + \mathrm{diag}(\boldsymbol{\ell})^{-2}$ for $\Gamma = (1, -1)^\top$ and $\boldsymbol{\ell} = (6, 6)^\top$. Here Γ gives the direction of the most rapid variation. (Image from Rasmussen & Williams, 2006)