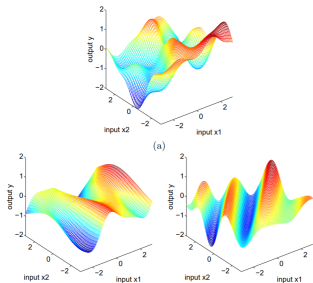


Introduction to Machine Learning

Gaussian Processes

Covariance functions for GPs



Learning goals

- Covariance functions encode key assumptions about the GP
- Common covariance functions like squared exponential and Matérn

VALID COVARIANCE FUNCTIONS

- Recall marginalization property of GPs: for any $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$,

$$\mathbf{f} = f(\mathbf{X}) \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

with $\mathbf{m} = m(\mathbf{X})$, $\mathbf{K} = k(\mathbf{X}, \mathbf{X})$

- Cov. function (or kernel) determines cov / kernel / Gram matrix:
 $\mathbf{K} = k(\mathbf{X}, \mathbf{X})$
- For \mathbf{K} to be a valid cov matrix it needs to be positive semi-definite (PSD) for any choice of inputs \mathbf{X}
- Implication: only **PSD functions** (i.e., those that induce PSD \mathbf{K}) are valid cov functions
- Also look at SVM chapter for background info on kernels, many further details in e.g. [Duvenaud 2014](#)



STATIONARY COVARIANCE FUNCTIONS

- Recall concept of spatial correlation

$\mathbf{x}, \tilde{\mathbf{x}}$ close in $\mathcal{X} \Rightarrow f(\mathbf{x}), f(\tilde{\mathbf{x}})$ close / more correlated in \mathcal{Y}

- Measure “closeness” via $\mathbf{d} = \mathbf{x} - \tilde{\mathbf{x}}$
- $k(\cdot, \cdot)$ called **stationary** \Leftrightarrow function of \mathbf{d}

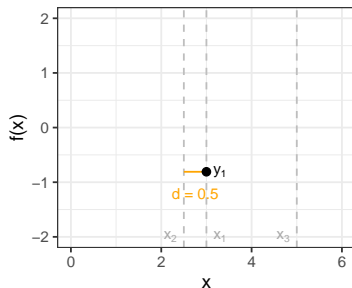
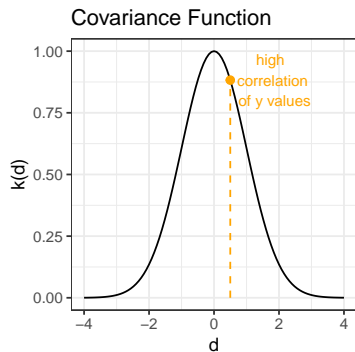
$$k(\mathbf{x}, \tilde{\mathbf{x}}) = k(\mathbf{d})$$

- Intuition: stationary $k(\cdot, \cdot)$ implies functions that do not depend on where the points lie in input space but only their difference



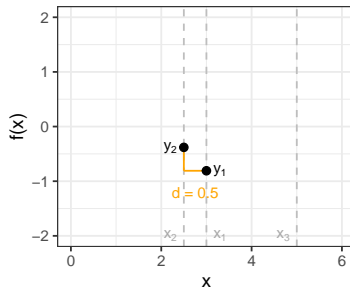
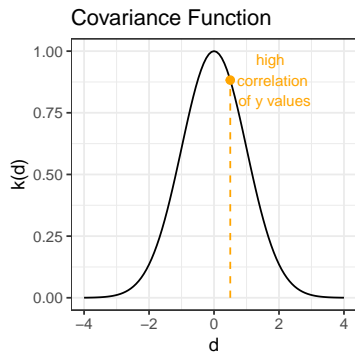
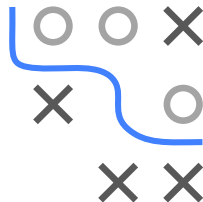
EXAMPLE: STATIONARY COVARIANCE

- Let $f \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$ with $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\frac{1}{2}\|\mathbf{d}\|^2)$
- Consider two points $\mathbf{x}^{(1)} = 3$ and $\mathbf{x}^{(2)} = 2.5$
- To get corr. between $f(\mathbf{x}^{(1)})$ and $f(\mathbf{x}^{(2)})$ look at $\mathbf{d}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$



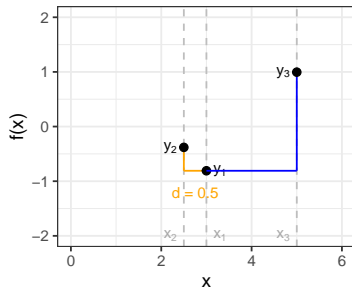
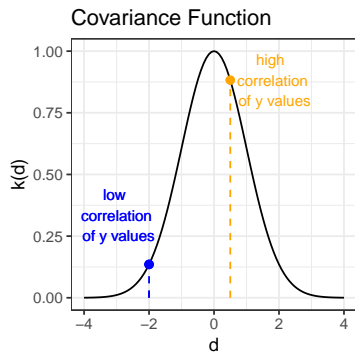
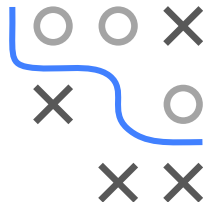
EXAMPLE: STATIONARY COVARIANCE

- Suppose we observe $y^{(1)} = -0.8$
- $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ are close in \mathcal{X} space
- Under the above GP assumption, $y^{(2)}$ should be close to $y^{(1)}$



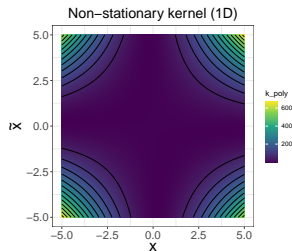
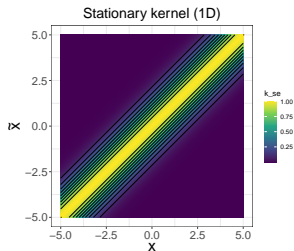
EXAMPLE: STATIONARY COVARIANCE

- Consider now $\mathbf{x}^{(3)} = 5$
- This is now further from $\mathbf{x}^{(1)}$
- \Rightarrow expect lower correlation between $y^{(3)}, y^{(1)}$



PROPERTIES OF COVARIANCE FUNCTIONS I

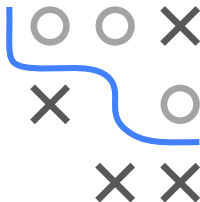
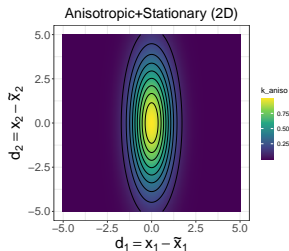
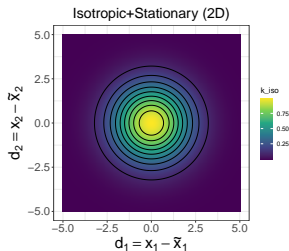
- **Stationary:** $k = k(\mathbf{d})$ with $\mathbf{d} = \mathbf{x} - \tilde{\mathbf{x}}$
 \Rightarrow invariant to translations in \mathcal{X} : $k(\mathbf{x}, \mathbf{x} + \mathbf{d}) = k(\mathbf{0}, \mathbf{d})$
(so we sometimes abuse notation and write $k(\mathbf{d})$)
- Consider $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}$ and plot contours of $k(\mathbf{x}, \tilde{\mathbf{x}})$:



- Stationarity implies: contour set of $\mathbf{x} \rightarrow k(\mathbf{x}, \tilde{\mathbf{x}}_1)$ must be same as for $\mathbf{x} \rightarrow k(\mathbf{x}, \tilde{\mathbf{x}}_2)$ translated by $\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2$
- Left: contours are parallel lines to diagonal ✓
Right: non-lin. contours, not translation-invariant \Rightarrow non-stationary

PROPERTIES OF COVARIANCE FUNCTIONS II

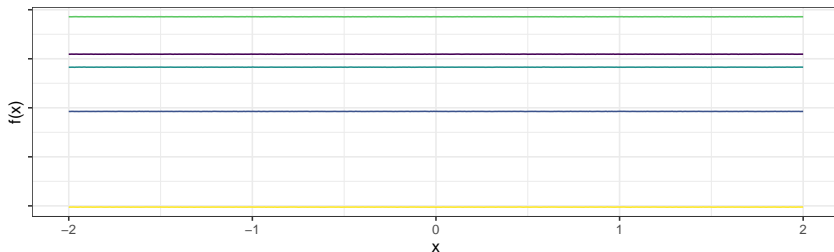
- **Isotropic:** $k = k(r)$ with $r = \|\mathbf{x} - \tilde{\mathbf{x}}\|$
 \Rightarrow invariant to rotations, implies stationarity
(again slight notational abuse)
- Since k is stationary can write $k(\mathbf{d})$ and plot contours for $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^2$



- Isotropic = circle/hypersphere contours
Anisotropic = non-spherical ellipsoid contours
- **Dot product:** $k = k(\mathbf{x}^T \tilde{\mathbf{x}})$

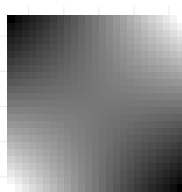
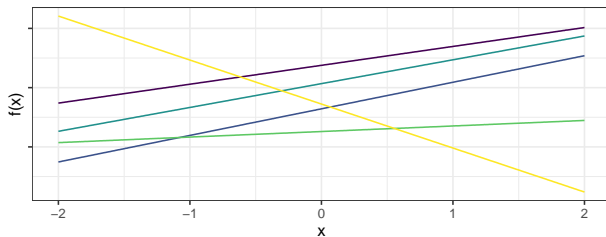
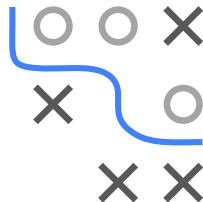
CONSTANT KERNEL

- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \theta_0 > 0$
- Constant function priors
- Global correlation irresp. of concrete inputs $\mathbf{x}, \tilde{\mathbf{x}}$
- Practically pretty useless



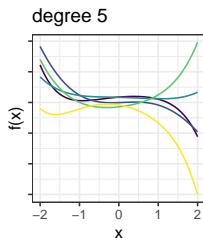
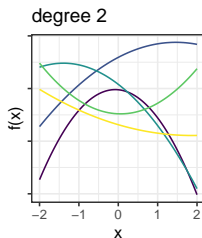
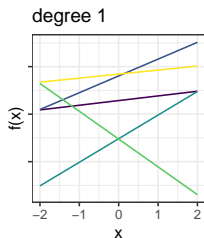
LINEAR KERNEL

- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \theta_0 + \mathbf{x}^T \tilde{\mathbf{x}}$
- Linear function priors
- Measures directional similarity: higher if vectors point in similar dirs
- In general, non-stationary \Rightarrow depends on locations of $\mathbf{x}, \tilde{\mathbf{x}}$
- See Bayesian linear model part
- LHS: Random draws. RHS: covariance matrix of 25 sampled points (black = high pos cor, white = high neg cor)



POLYNOMIAL KERNEL

- $k(\mathbf{x}, \tilde{\mathbf{x}}) = (\theta_0 + \mathbf{x}^T \tilde{\mathbf{x}})^p, \quad p \in \mathbb{N}$
- Polynomial function priors
- Allows for non-linearity through higher-order monomials & interaction terms

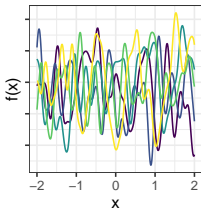


SQUARED EXPONENTIAL KERNEL

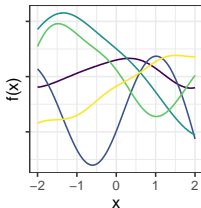
- Aka Gaussian kernel, RBF kernel
- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|^2}{2\ell^2}\right)$
- Very smooth: continuous, ∞ differentiable (not always realistic)
- Cov decays quickly \Rightarrow quadratic in distance



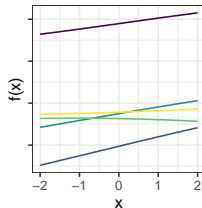
length scale 0.1



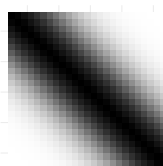
length scale 1.0



length scale 10.0

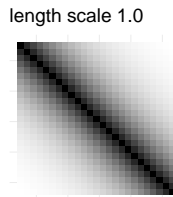
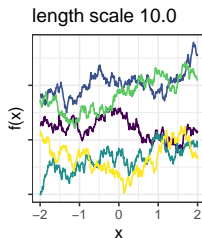
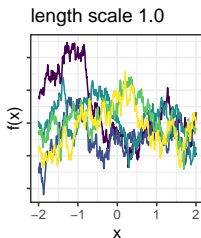
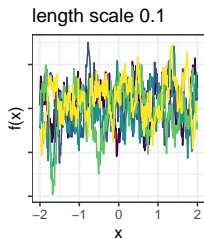


length scale 1.0



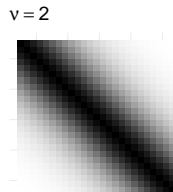
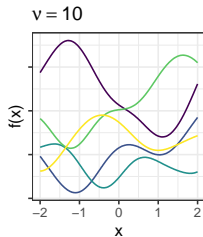
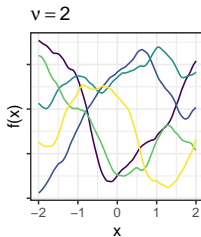
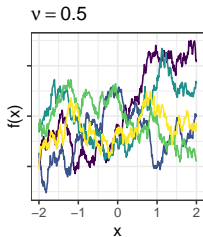
EXPONENTIAL KERNEL

- Aka Ornstein-Uhlenbeck kernel
- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\ell}\right)$
- Non-smooth: continuous but not differentiable, can model functions with abrupt variations
- Cov decays exponentially with distance (modulated by ℓ)



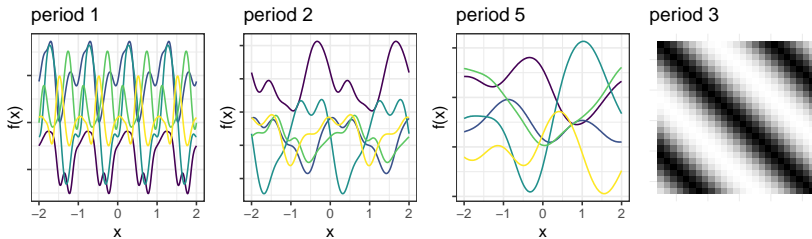
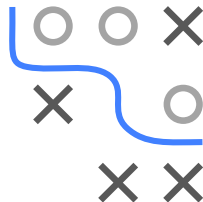
MATÉRN KERNEL

- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{2^\nu \Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\ell} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\ell} \right)$
- ν : smoothness param, Γ : gamma function, ℓ : length scale, K_ν : modified Bessel function
- Stationary & isotropic
- Allows for controlled degree of smoothness via choice of ν
- ν also determines differentiability
- Squared exponential kernel ($\nu = \infty$) and exponential kernel ($\nu = 0.5$) are special cases



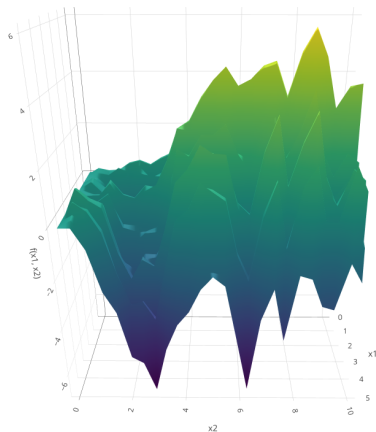
PERIODIC KERNEL

- E.g., radial periodic kernel: $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(\frac{-2 \sin^2(\pi \|\mathbf{x} - \tilde{\mathbf{x}}\|/m)}{\ell^2}\right)$
- m : period, ℓ : length-scale
- $f(\mathbf{x})$ should be periodically similar to points with a distance which is a multiple of m ;
for distances in between, this is modulated by ℓ
- Alternative: Product of 1D periodic kernels, with m_j period in dimension j : $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(\sum_j \frac{-2 \sin^2(\pi |x_j - \tilde{x}_j|/m_j)}{\ell^2}\right)$



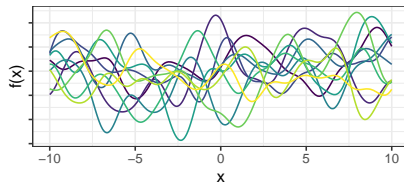
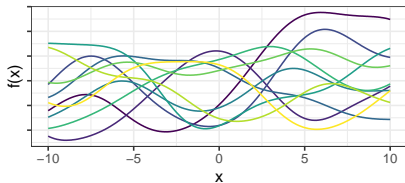
EXAMPLE: BROWNIAN MOTION

- $k(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_j \min(x_j, \tilde{x}_j)$
- Physics application: random fluctuations of particles
- With non-1D inputs aka Brownian sheet
- Correlation in each dimension is 1D-like Brownian motion



CHARACTERISTIC LS : ISOTROPIC CASE

- Every (isotropic) kernel can be written as $k(r)$ where
- $r = ||\mathbf{x} - \tilde{\mathbf{x}}||$
- E.g. $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{||\mathbf{x} - \tilde{\mathbf{x}}||^2}{2\ell^2}\right)$ or $k(r) = \exp\left(-\frac{1}{2}\left(\frac{r}{\ell}\right)^2\right)$
- Controls how quickly function values become uncorrelated
- High (low) ℓ : smooth (wiggly) functions



- In SVMs we called this bandwidth or (inverse) kernel width

CHARACTERISTIC LS : STATIONARY CASE

- For stationary kernels $k(\mathbf{d})$
- We modulate every distance component d_j by an individual ℓ_j
- We can turn the isotropic examples from above into stationary ones – with individual length scales
- Write $\|\mathbf{d}\|^2 = \sum d_j^2$ and put an $1/\ell_j$ before each d_j
- E.g. for squared exp:

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

- Also note: this is a product of 1D kernels, one for each input dim. The correlation in each dim is described by the 1D kernel and its distance component is modulated by ℓ_j



BENEFITS OF DIM-WISE LENGTH-SCALES

- ℓ_1, \dots, ℓ_p : characteristic length-scales
- Intuition for ℓ_i : how far to move along i -th axis for fun. values to become uncorrelated?
- Implements **automatic relevance determination** (ARD): inverse of ℓ_i determines importance of i -th feature
- Very large $\ell_i \Rightarrow$ cov effectively independent of i -th feature
- For features on different scales: rescale automatically by estimating ℓ_1, \dots, ℓ_p



CHARACTERISTIC LS : WEIGHTED EUCLID DIST

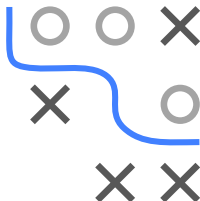
- Can even generalize the above principle
- Move to weighted (squared) Euclidean distance
- E.g. for squared exp again:

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

- This covers the case before
- Possible choices for \mathbf{M} :

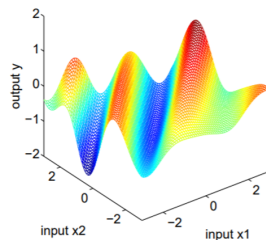
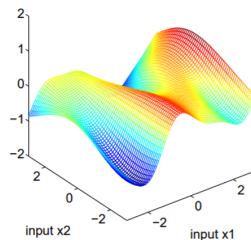
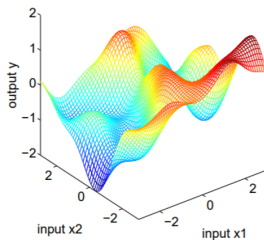
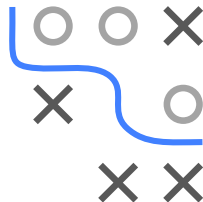
$$\mathbf{M}_1 = \ell^{-2} \mathbf{I}_p \quad \mathbf{M}_2 = \text{diag}(\ell)^{-2} \quad \mathbf{M}_3 = \Gamma \Gamma^T + \text{diag}(\ell)^{-2}$$

where $\ell \in \mathbb{R}_+^p$, $\Gamma \in \mathbb{R}^{p \times k}$



EXAMPLES: CHARACTERISTIC LS

- Left: $\mathbf{M} = \mathbf{I} \Rightarrow$ same variation in all directions
- Middle: $\mathbf{M} = \text{diag}(\ell)^{-2} \Rightarrow$ less variation in x_2 direction ($\ell_2 > \ell_1$)
- Right: $\mathbf{M} = \Gamma\Gamma^T + \text{diag}(\ell)^{-2}$ with $\Gamma = (1, -1)^T$ and $\ell = (6, 6)^T \Rightarrow \Gamma$ determines dir. of most rapid variation



► [Click for source](#)