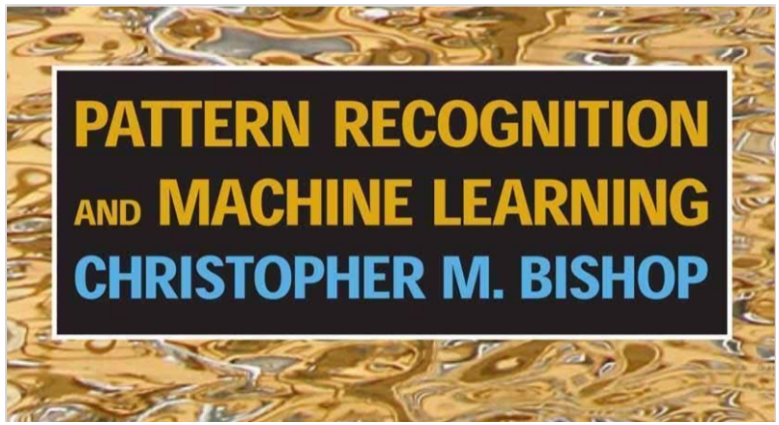
GP recap

Rodrigo A. Vargas-Hernández

7/Feb/2023

References for today's lecture

Chapter 3 and 6



- Week 3 and 4 slides.
- Gaussian Processes for ML
- Online material

Gaussian Process

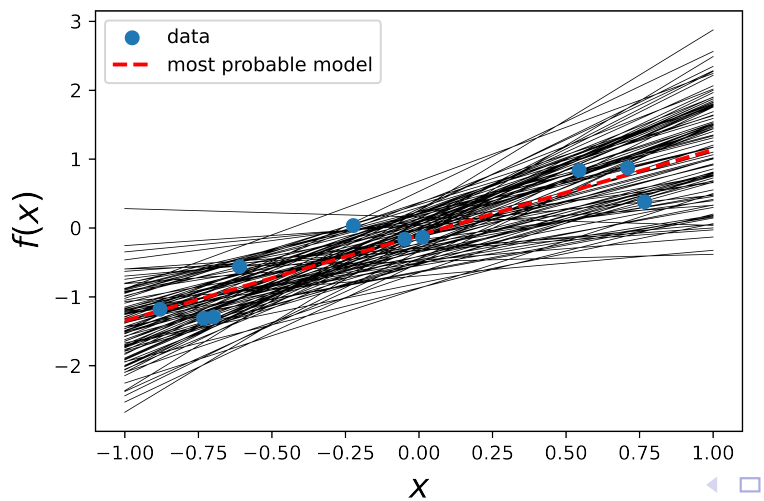
$$[\theta_1, \theta_0] \sim p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \underbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}_{\textit{likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\textit{prior}}$$

Predictions with Gaussian Process

$$[\theta_1, \theta_0] \sim p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \underbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}_{\textit{likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\textit{prior}}$$

$\boldsymbol{\theta}$ is not deterministic!!

$$p(y_*|\mathbf{x}_*\mathbf{X}, \mathbf{y}) = \underbrace{\int \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_* \; p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\theta}}_{\text{average over different models}} \approx \frac{1}{M}\sum_{\ell}^{M} \boldsymbol{\theta}_{\ell}^{\mathrm{T}}\mathbf{x}_*$$



Predictions with Gaussian Process

$$p(y_*|\mathbf{x}_*\mathbf{X},\mathbf{y}) = \overbrace{\int \underbrace{p(y_*|\mathbf{x}_*,\boldsymbol{\theta},\mathbf{X},\mathbf{y})}_{\text{Gaussian}} \underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{Gaussian}} d\boldsymbol{\theta}}^{\text{Gaussian}}$$

Prediction,

$$p(y_*|\mathbf{x}_*\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\underbrace{\mathbf{x}_*^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbb{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}_{\mu(\mathbf{x}_*)}, \underbrace{\mathbf{x}_*^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbb{I}\right)^{-1}\mathbf{x}_*}_{\sigma(\mathbf{x}_*)}\right)$$

- Apply the kernel trick!!

Predictions with Gaussian Process

$$
\begin{aligned}
y_* &\sim \mathcal{N}\Big(\boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})+\lambda\mathbb{I}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})\mathbf{y}, \\
&\qquad \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})+\lambda\mathbb{I}\right)^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\Big) \\
y_* &\sim \mathcal{N}\Big(\boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})\left(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})+\lambda\mathbb{I}\right)^{-1}\mathbf{y}, \\
&\qquad \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\left(\mathbb{I}-\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})\left(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}}+\lambda\mathbb{I}\right)^{-1}\boldsymbol{\Phi}(\mathbf{X})\right)\boldsymbol{\phi}(\mathbf{x}_*)\Big) \\
y_* &\sim \mathcal{N}\Big(\underbrace{\boldsymbol{\kappa}(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}}\left(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})+\lambda\mathbb{I}\right)^{-1}\mathbf{y}}_{\mu(\mathbf{x}_*)}, \\
&\qquad \underbrace{\boldsymbol{\kappa}(\mathbf{x}_*,\mathbf{x}_*)-\boldsymbol{\kappa}(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}}\left(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{X})+\lambda\mathbb{I}\right)^{-1}\boldsymbol{\kappa}(\mathbf{x}_*,\mathbf{X})}_{\sigma(\mathbf{x}_*)}\Big)
\end{aligned}
$$

- NO SAMPLING!!!

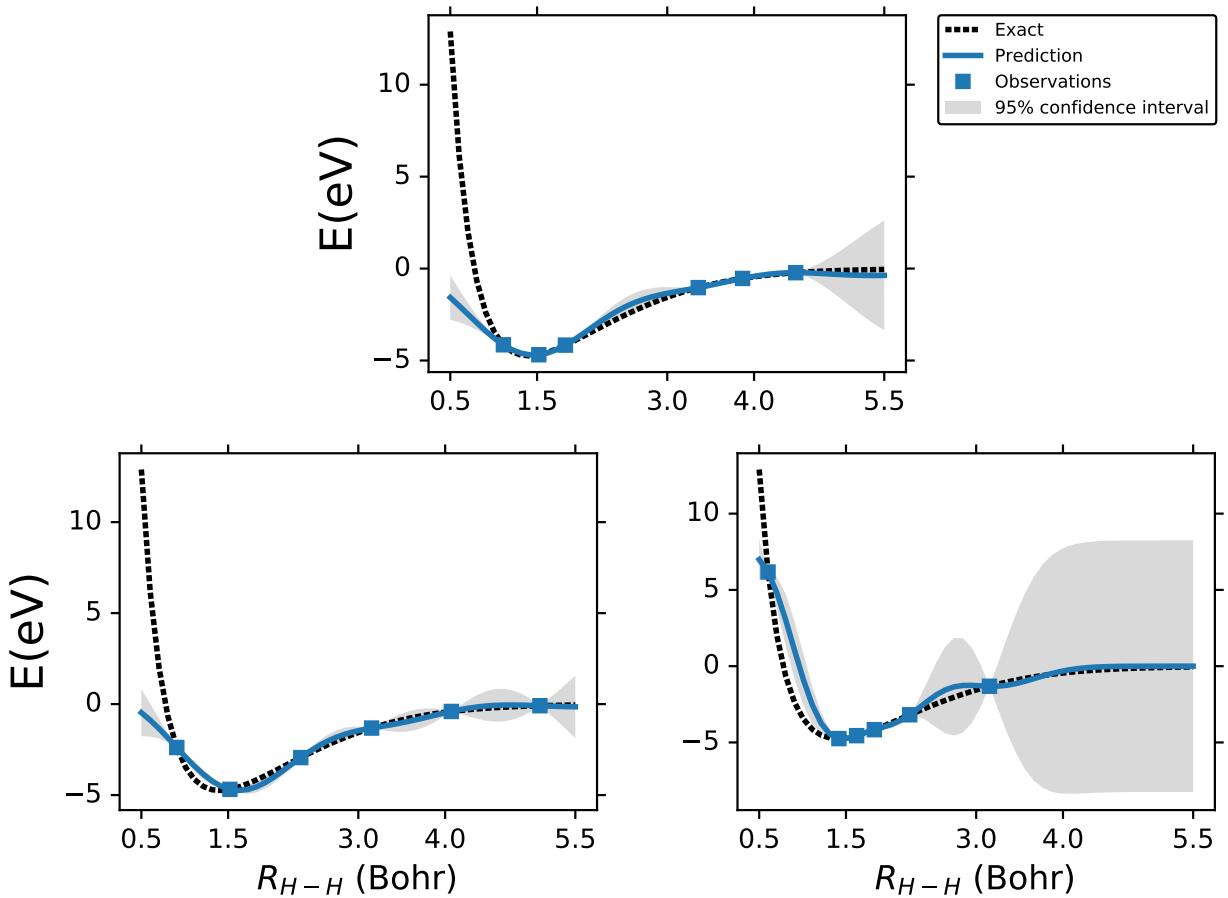
Gaussian process are defined as, $f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')\right))$

Use Matrix cook book

Predictions with Gaussian Process

$$\mu(\mathbf{x}_*) = \kappa(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} \left(\Phi(\mathbf{X}) \Phi^{\mathrm{T}}(\mathbf{X}) + \lambda \mathbb{I} \right)^{-1} \mathbf{y}$$

$$\sigma(\mathbf{x}_*) = \kappa(\mathbf{x}_*, \mathbf{x}_*) - \kappa(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} \left(\Phi(\mathbf{X}) \Phi^{\mathrm{T}}(\mathbf{X}) + \lambda \mathbb{I} \right)^{-1} \kappa(\mathbf{x}_*, \mathbf{X})$$



What about the kernel parameters?

- marginal likelihood: the marginalization over the function values $\mathbf{f}$,

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f}$$

" In Bayesian statistics, it represents the probability of generating the observed sample from a prior and is therefore often referred to as model evidence or simply evidence. [1] "

$$\log p(\mathbf{y}|\mathbf{X}) = \underbrace{-\frac{1}{2}\mathbf{y}^{\mathrm{T}} (K + \lambda\mathbb{I})^{-1} \mathbf{y}}_{\text{data-fit}} \underbrace{-\frac{1}{2}\log |K + \lambda\mathbb{I}|}_{\text{complexity penalty}} - \frac{N}{2}\log(2\pi)$$

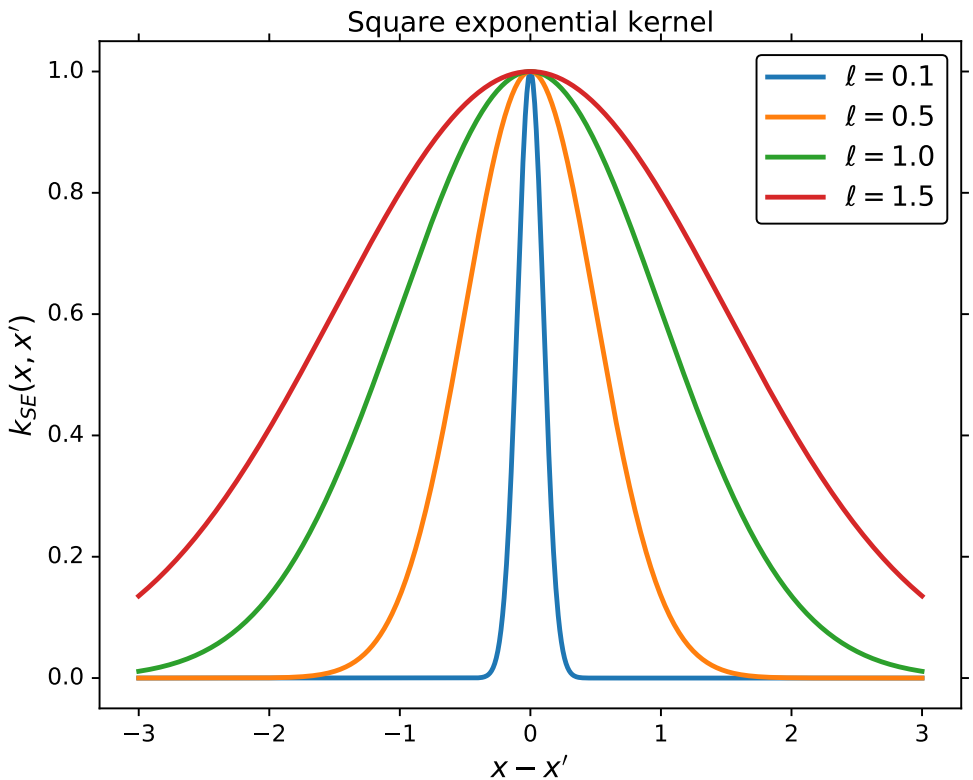
- $\nabla_{\ell} \log p(\mathbf{y}|\mathbf{X})$ has an analytic form ($\boldsymbol{\ell} = \{\ell_i\}_i^m$ kernel parameters)

[1] Wikipedia

Kernel parameters

- Radial basis function kernel

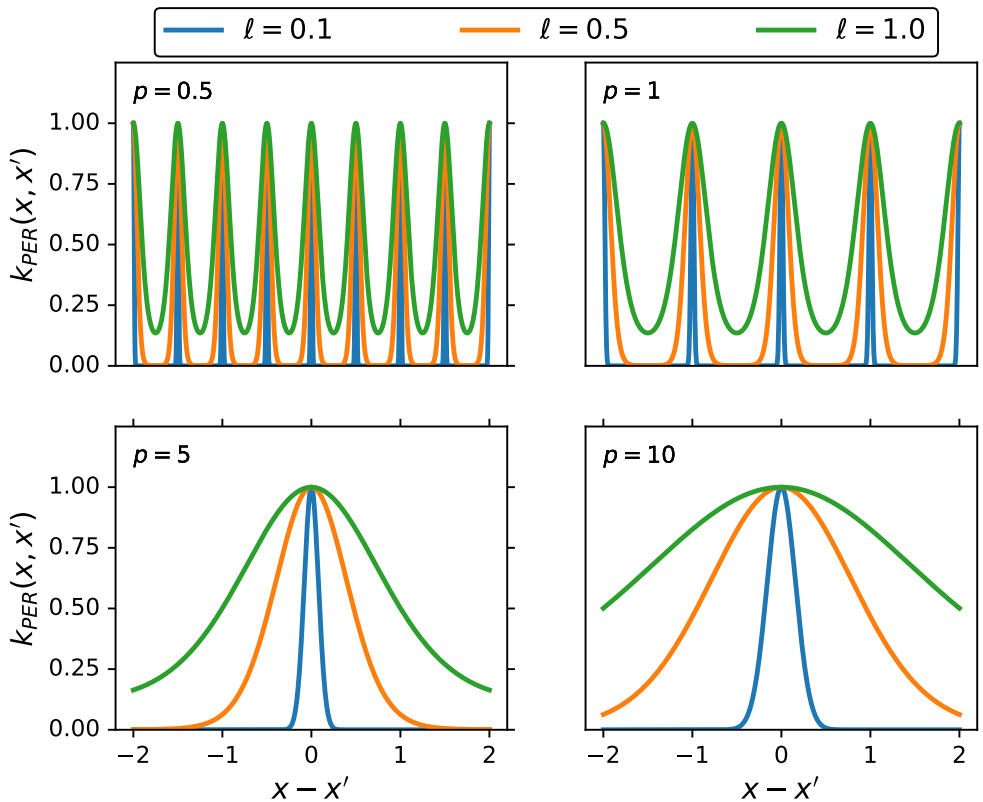
$$k_{RBF}(x, x') = c \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$



Kernel parameters

- Periodic kernel

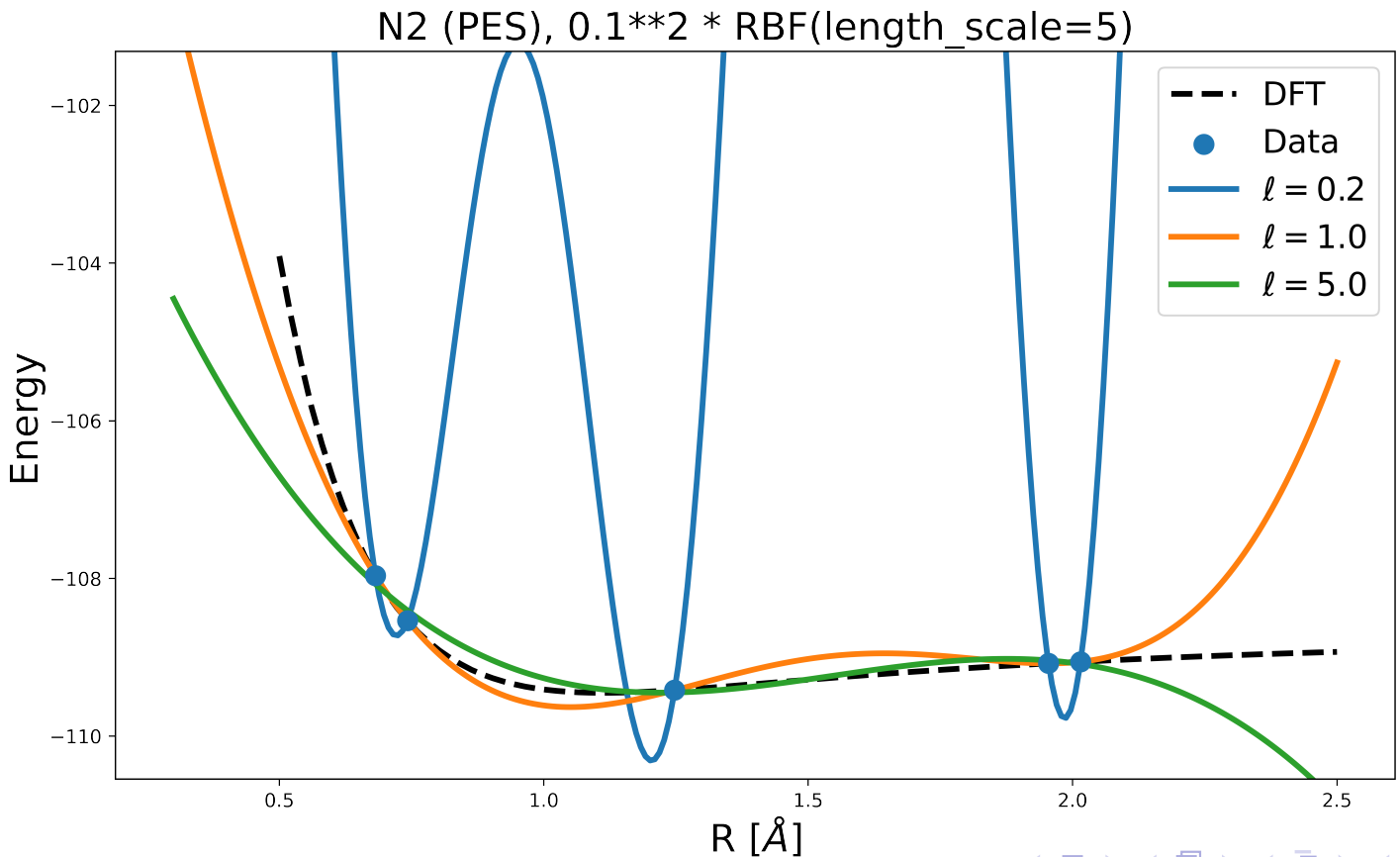
$$k_{PER}(x, x') = c \ \exp\left(-\frac{2\sin^2(\frac{\pi r}{p})}{\ell^2}\right) \quad ; \quad r = \sqrt{(x - x')^2}$$



Kernel parameters

- Radial basis function kernel

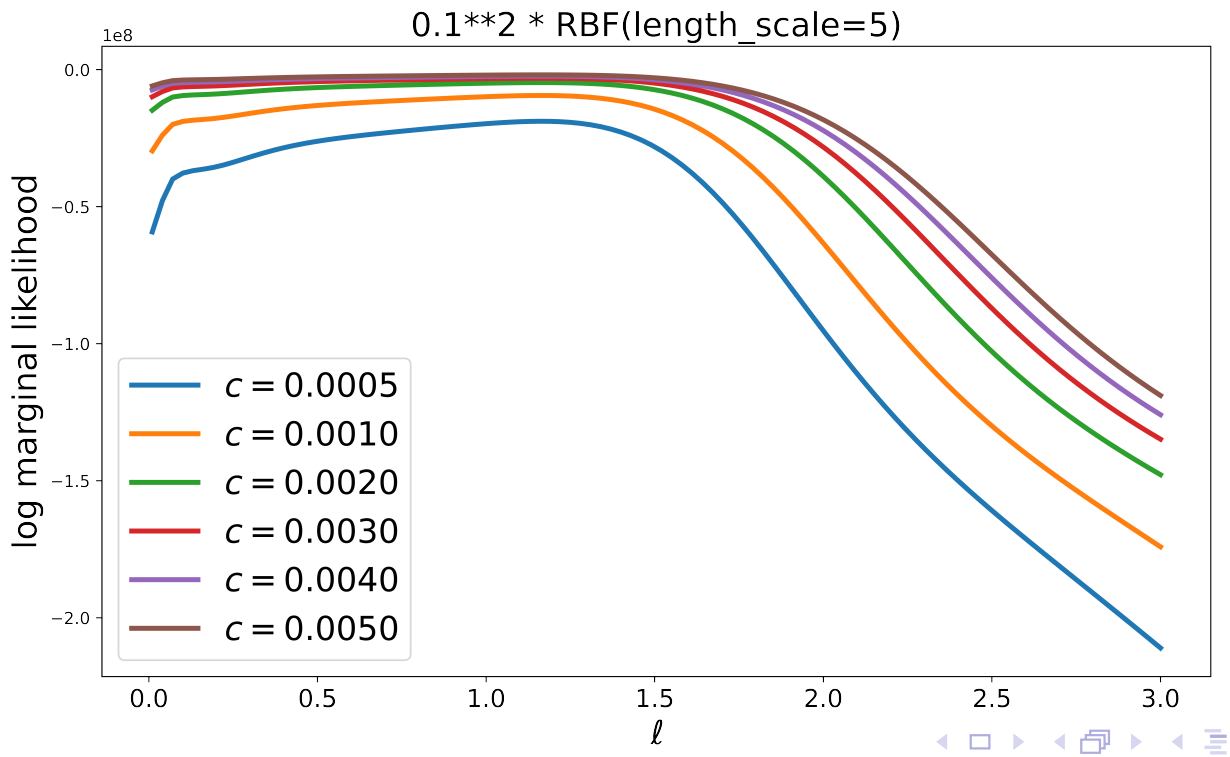
$$k_{RBF}(x_i, x_j) = c \ \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right)$$



Kernel parameters: Radial basis function kerne

$$k_{RBF}(x_i, x_j) = c \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right)$$

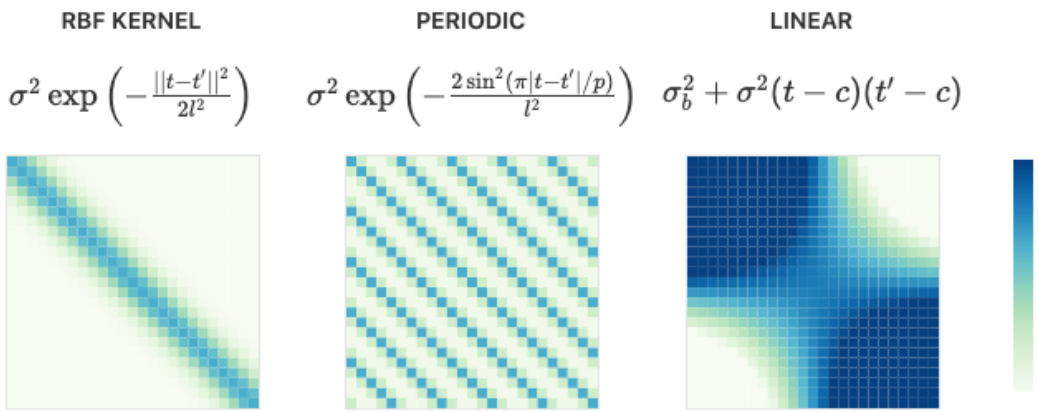
$$\log p(\mathbf{y}|\mathbf{X}) = \underbrace{-\frac{1}{2}\mathbf{y}^{\mathrm{T}}(K + \lambda\mathbb{I})^{-1}\mathbf{y}}_{\text{data-fit}} \underbrace{-\frac{1}{2}\log|K + \lambda\mathbb{I}|}_{\text{complexity penalty}} - \frac{N}{2}\log(2\pi)$$



Covariance matrix

The matrix elements of the Covariance matrix are computed with the kernel function,

$$[K(\mathbf{X}, \mathbf{X})]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$$



- Kernels online tutorial
- The Kernel Cookbook

Search for the best kernel

VA How can we select the best kernel function for machine learning methods?

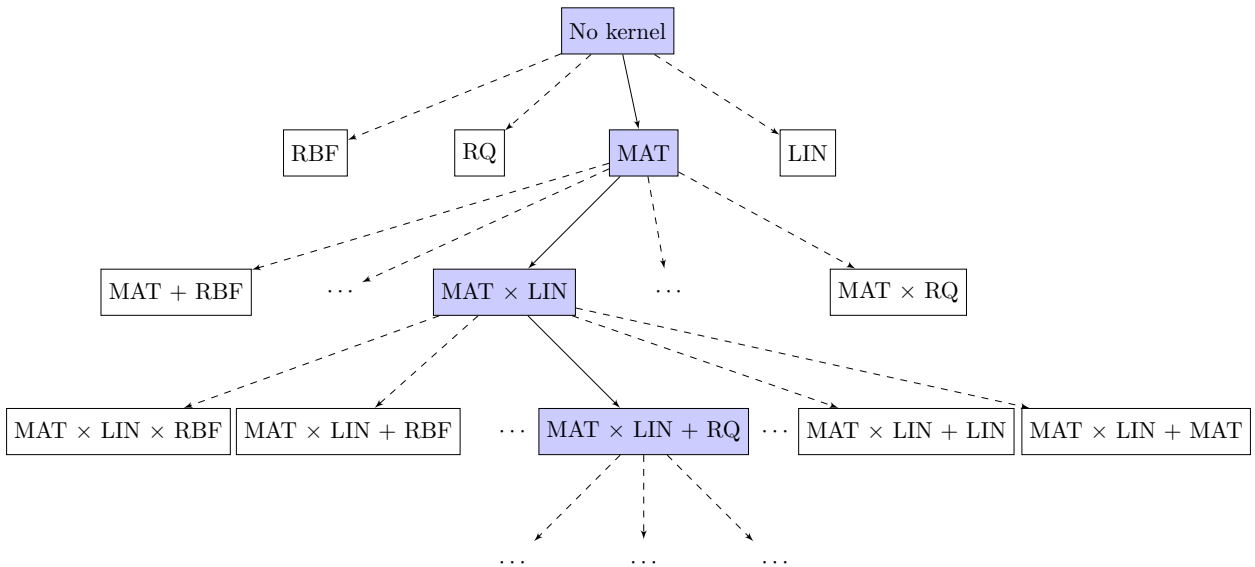
Selecting the best kernel function for a machine learning method depends on several factors, including:

1. The nature of the data: If the data is linearly separable, a linear kernel is usually a good choice, while a non-linear kernel is appropriate for non-linearly separable data.
2. The number of features: A higher-dimensional feature space may require a more complex kernel to capture the relationships between the features.
3. The number of samples:

(Extra) Search for the best kernel

Bayesian information criterion

$$\text{BIC}(\ell) = \underbrace{\log p(\mathbf{y}|\mathbf{X})}_{\text{log marginal like-lihood}} - \underbrace{\frac{1}{2}|\ell| \log N}_{\text{penality}}$$



Papers:

- Structure Discovery in Nonparametric Regression through Compositional Kernel Search
- Automatic Construction and Natural-Language Description of Nonparametric Regression Models