DS-GA 3001.009
Modeling Time Series Data
Lab 11

Artie Shen | Center for Data Science

NYU

Agenda

- Recap
  - Common Kernels
  - Combining Kernels
- Programming
  - Kernel Exporation
  - Combining Kernels
  - Application: CO2 Proportion Trend

Definition

A kernel is a positive-definite function of two inputs x and x'. Gaussian Process uses kernels to model values of two values of a function evaluated at each points:

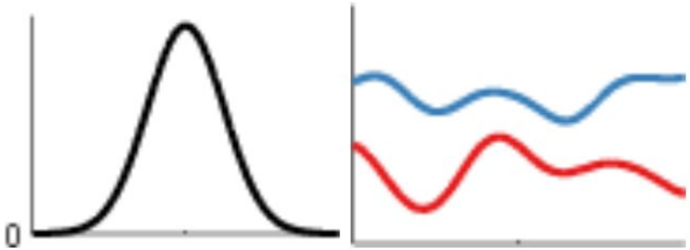
$$\mathbf{cov}[\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')] = \mathbf{k}(\mathbf{x}, \mathbf{x}')$$

From another perspective, kernels determine which functions are likely under the GP prior.

Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Common Kernels

Squared Exponential Kernel



A.K.A. the Radial Basis Function kernel, the Gaussian kernel. It has the form:

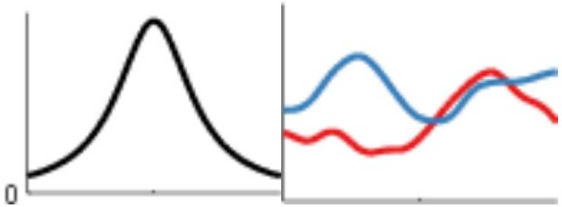
$$k_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

- The lengthscale ℓ determines the length of the 'wiggles' in your function. In general, you won't be able to extrapolate more than ℓ units away from your data.
- The output variance σ^2 determines the average distance of your function away from its mean. Every kernel has this parameter out in front; it's just a scale factor.

Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Common Kernels

Rational Quadratic Kernel



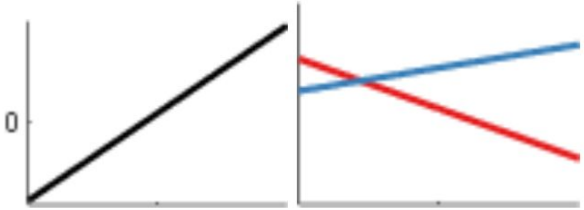
$$k_{\mathrm{RQ}}(x, x') = \sigma^2 \left(1 + \frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

This kernel is equivalent to adding together many SE kernels with different lengthscales. So, GP priors with this kernel expect to see functions which vary smoothly across many lengthscales. The parameter α determines the relative weighting of large-scale and small-scale variations. When $\alpha \to \infty$, the RQ is identical to the SE.

Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Common Kernels

Linear Kernel

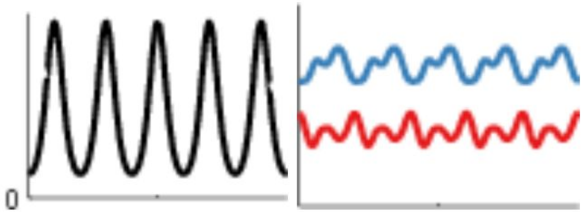


$$k_{\text{Lin}}(x, x') = \sigma_b^2 + \sigma_v^2 (x - c)(x' - c)$$

If you use just a linear kernel in a GP, you're simply doing Bayesian linear regression, and good news! You can do this in time $\mathcal{O}(N)$ instead of $\mathcal{O}(N^3)$, so you should probably go use software specifically designed for that.

Common Kernels

Periodic Kernel



$$k_{\text{Per}}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|x-x'|/p)}{\ell^2}\right)$$

The periodic kernel (derived by David Mackay) allows one to model functions which repeat themselves exactly. Its parameters are easily interpretable:

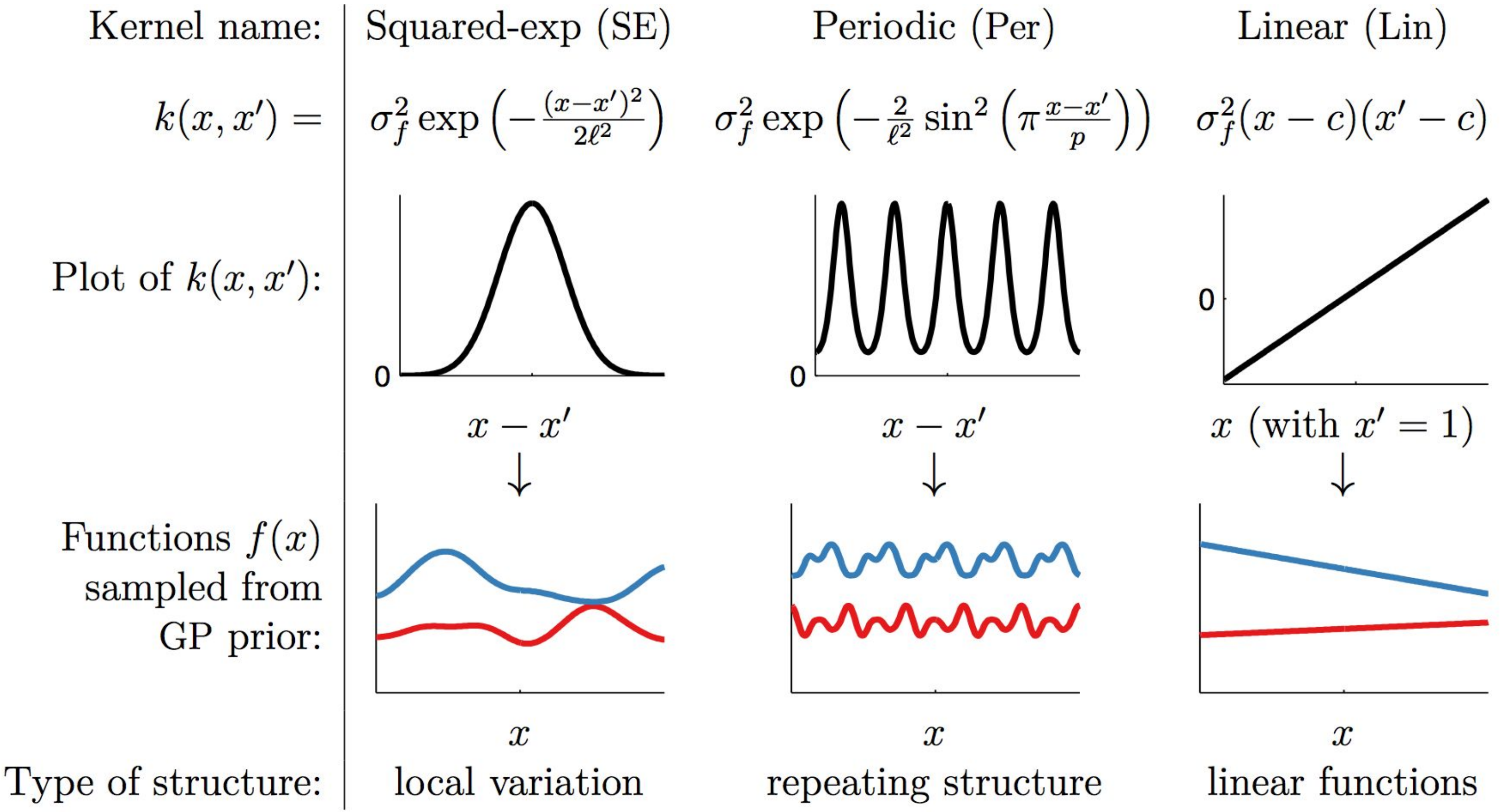
- The period p simply determines the distnace between repititions of the function.
- The lengthscale ℓ determines the lengthscale function in the same way as in the SE kernel.

Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Common Kernels

Stationary and Non-stationary The SE and Per kernels are *stationary*, meaning that their value only depends on the difference $x - x'$. This implies that the probability of observing a particular dataset remains the same even if we move all the $\mathbf{x}$ values by the same amount. In contrast, the linear kernel (Lin) is non-stationary, meaning that the corresponding GP model will produce different predictions if the data were moved while the kernel parameters were kept fixed.

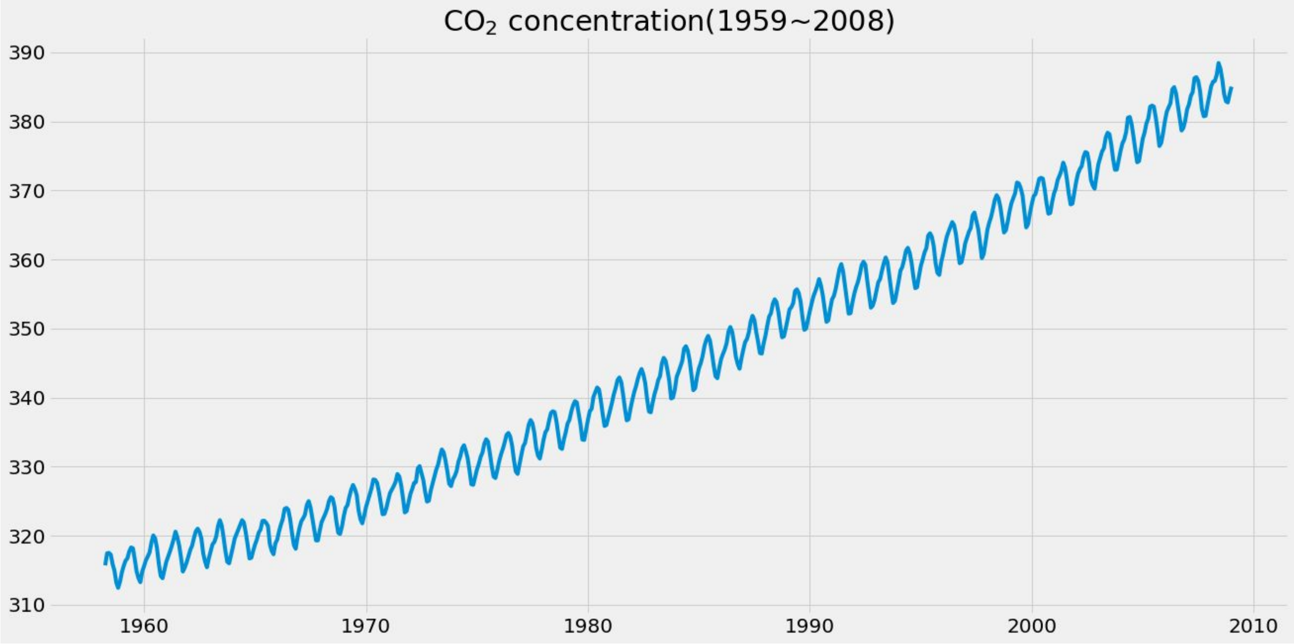
Common Kernels



Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Combining Kernels

- **Motivation**



Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Combining Kernels

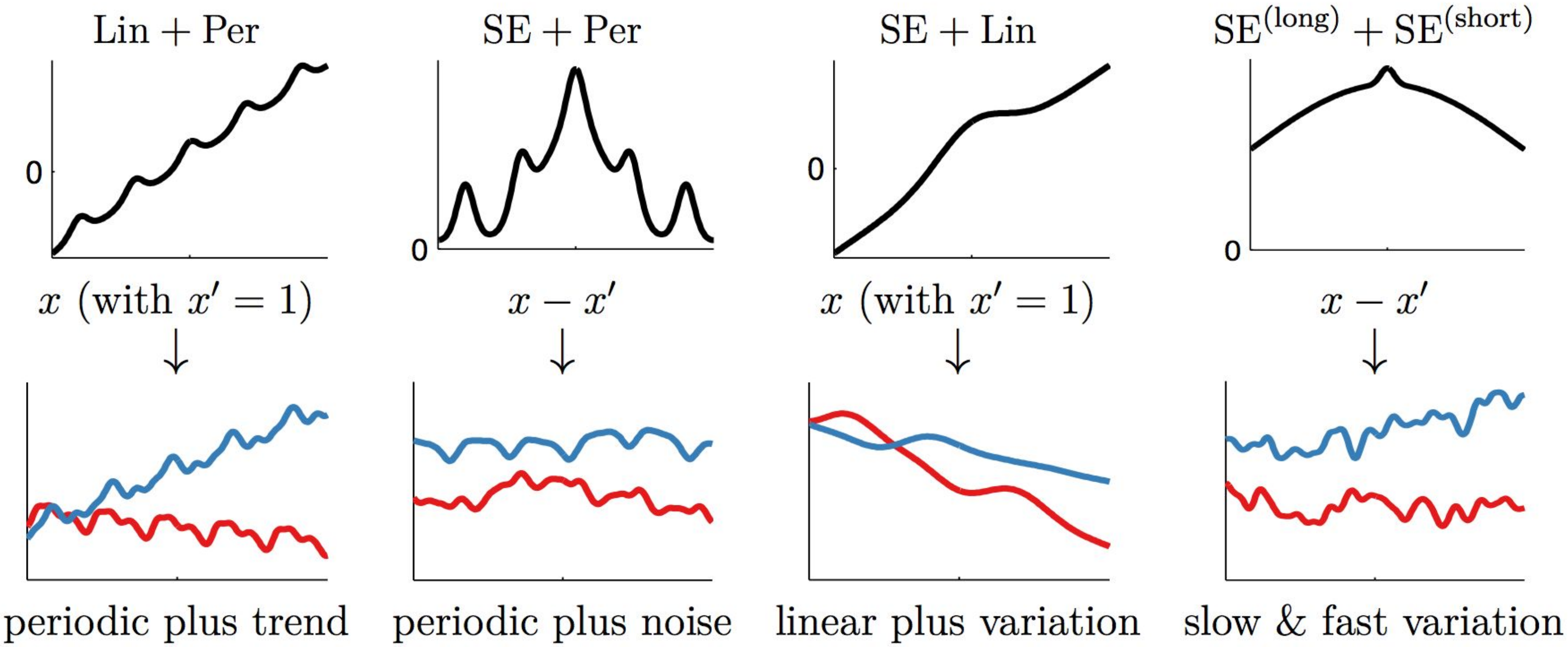
- **Two popular methods:**
  - Summation
  - Multiplication

$$k_a + k_b = k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}')$$

$$k_a \times k_b = k_a(\mathbf{x}, \mathbf{x}') \times k_b(\mathbf{x}, \mathbf{x}')$$

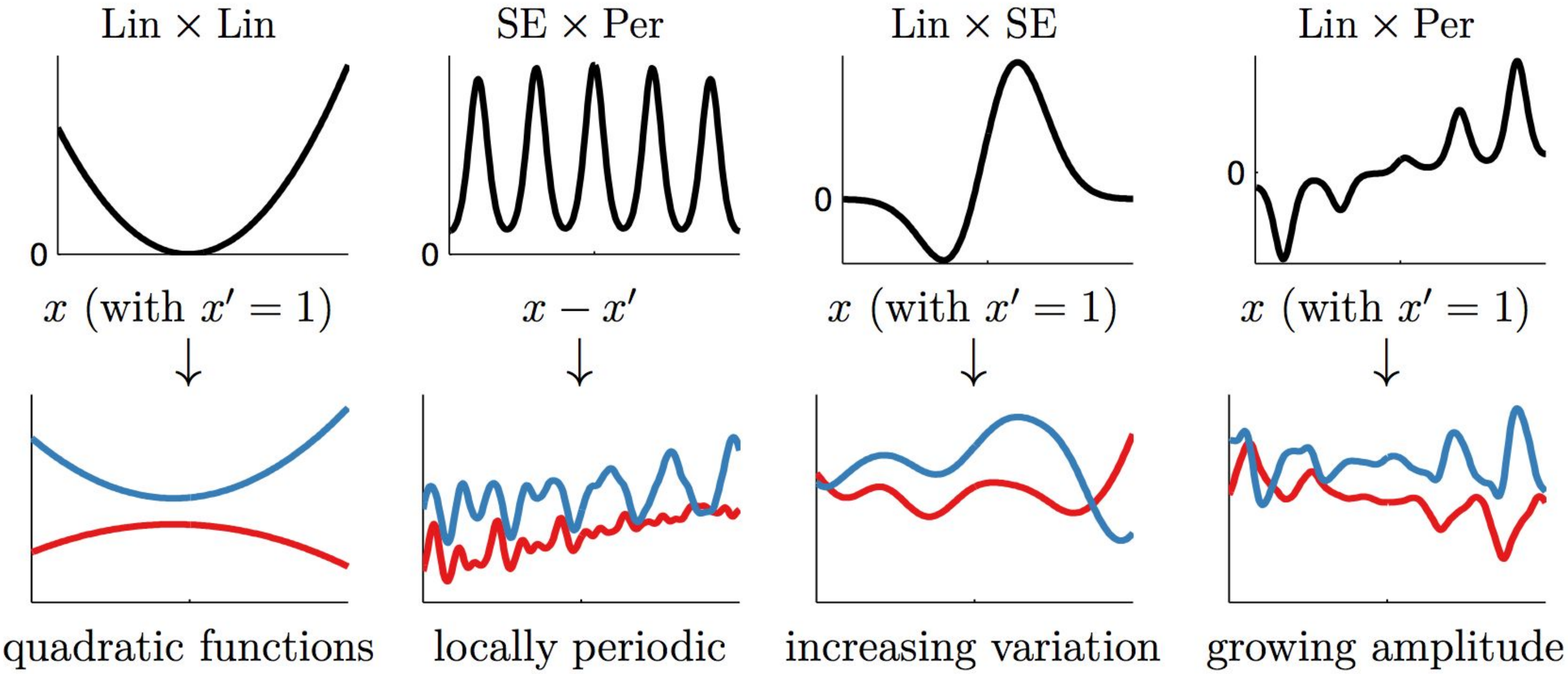
Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Combining Kernels



Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Combining Kernels



Reference: https://www.cs.toronto.edu/~duvenaud/thesis.pdf

- Github:
  - https://github.com/charlieblue17/timeseries2018
- No submission required.