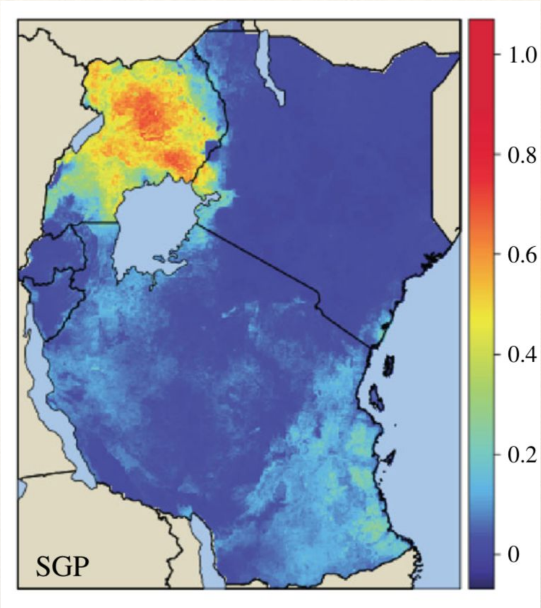
Gaussian Processes: Basics

Markus Michael Rau

Questions? Good!

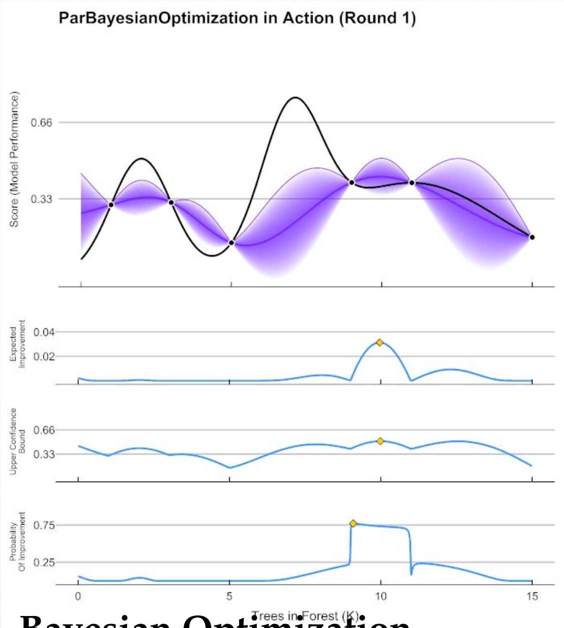
markusmichael.rau@googlemail.com

Desease Risk Mapping



Bhatt, et al. 2021

http://dx.doi.org/10.1098/rsif.2017.0520



Bayesian Optimization

Credit: AnotherSamWilson

Time Series Forecasting

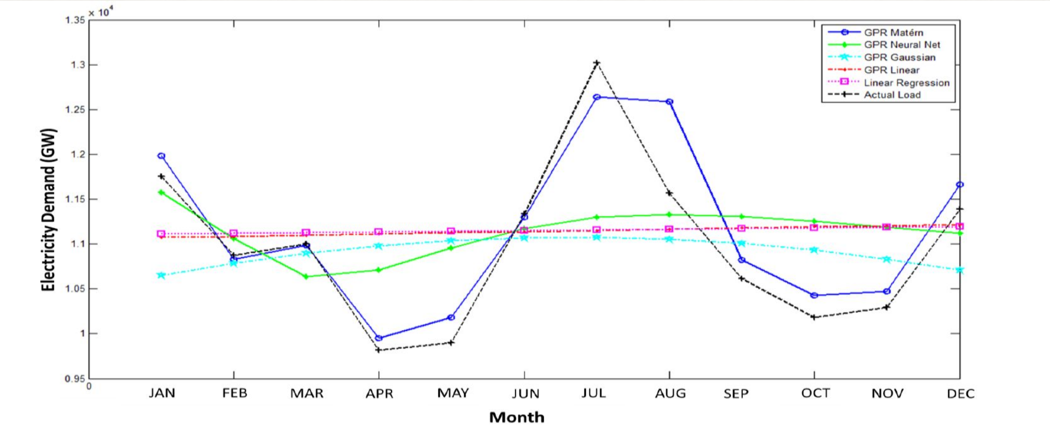
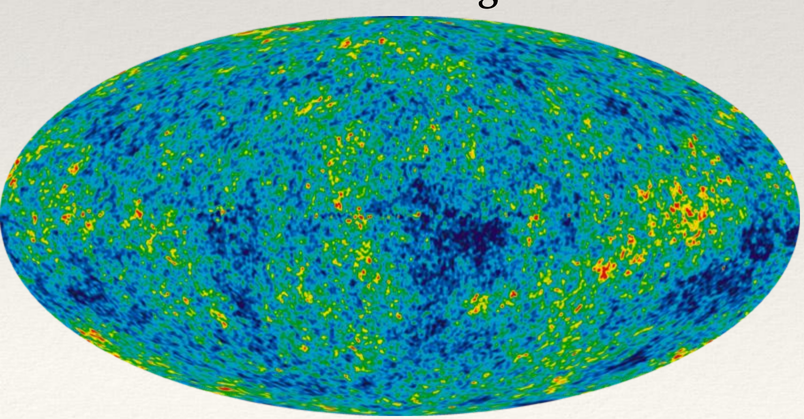


Figure 13. Monthly Electricity Load Values for Year 2008

Alamaniotis, et al. 2014

DOI: 10.1049/cp.2014.1693

Cosmic Microwave Background



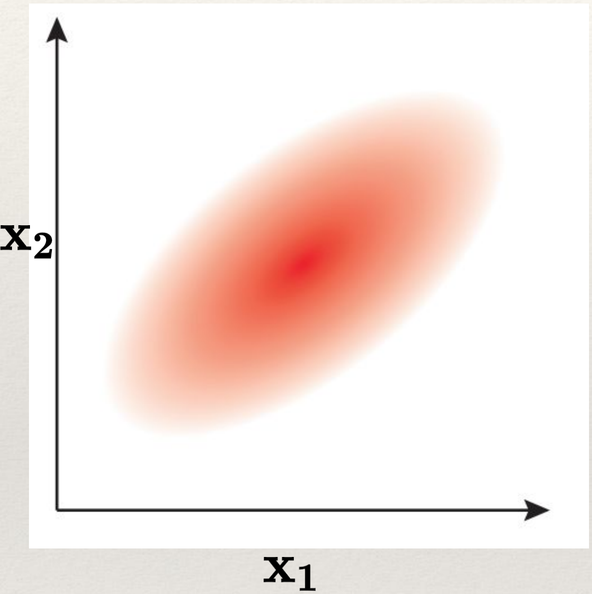
http://wmap.gsfc.nasa.gov/media/101080

Gaussian Process

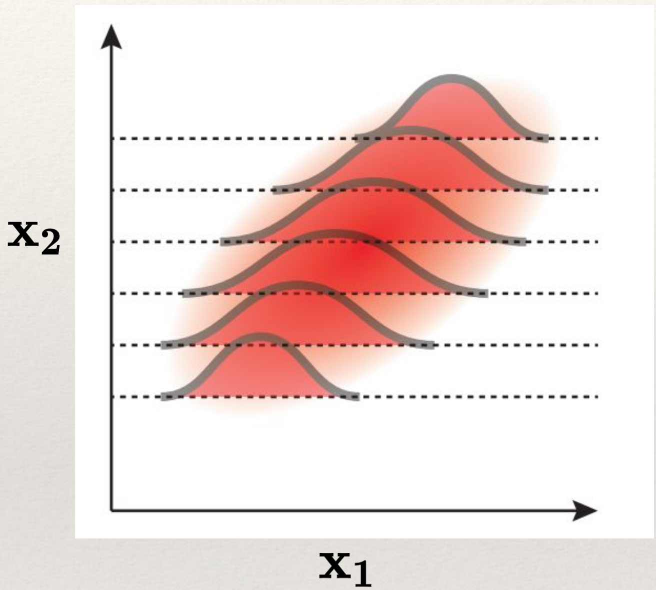
1. The multivariate normal distribution
2. Gaussian Processes: Basics
3. Inference
    a. Before the measurement
    b. Effect of Kernel Parameters
    c. The continuous limit
    d. Gaussian Process Regression
4. Demo
5. Summary
6. Literature

Multivariate Normal Distribution

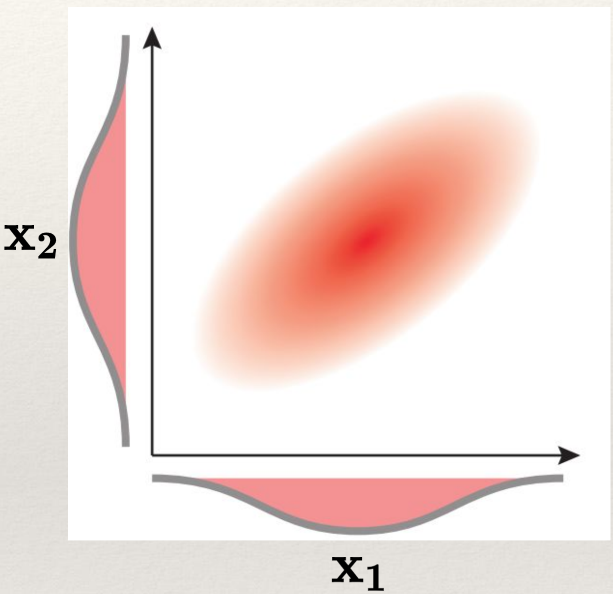
Pictures from https://ericmjl.github.io/blog/2018/8/7/joint-conditional-and-marginal-probability-distributions/



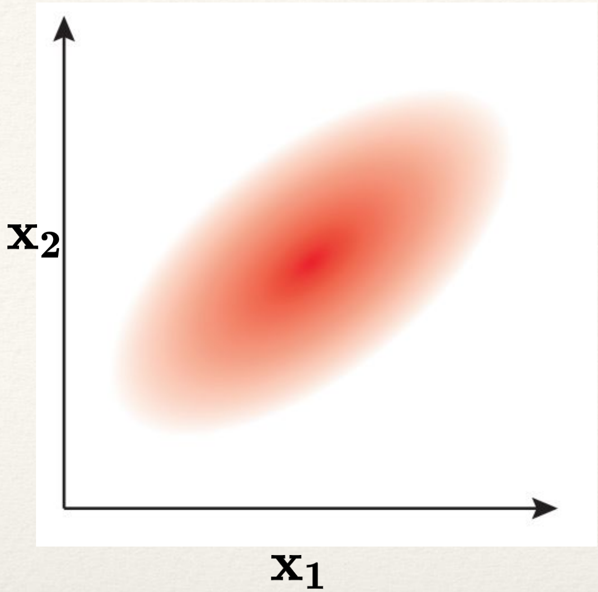
Joint Distribution



Conditional Distribution



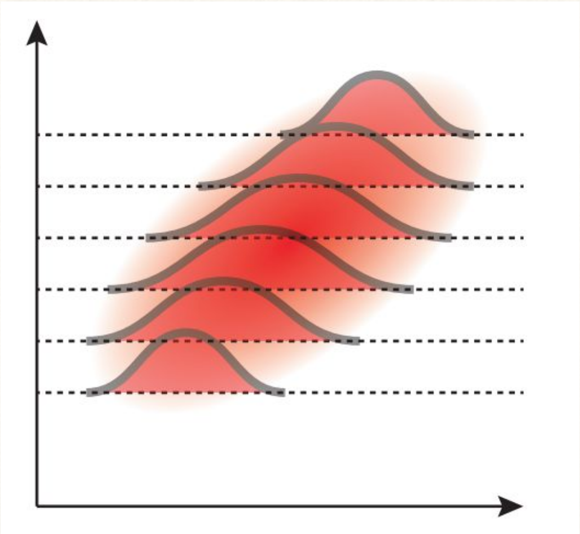
Marginal Distribution



$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$



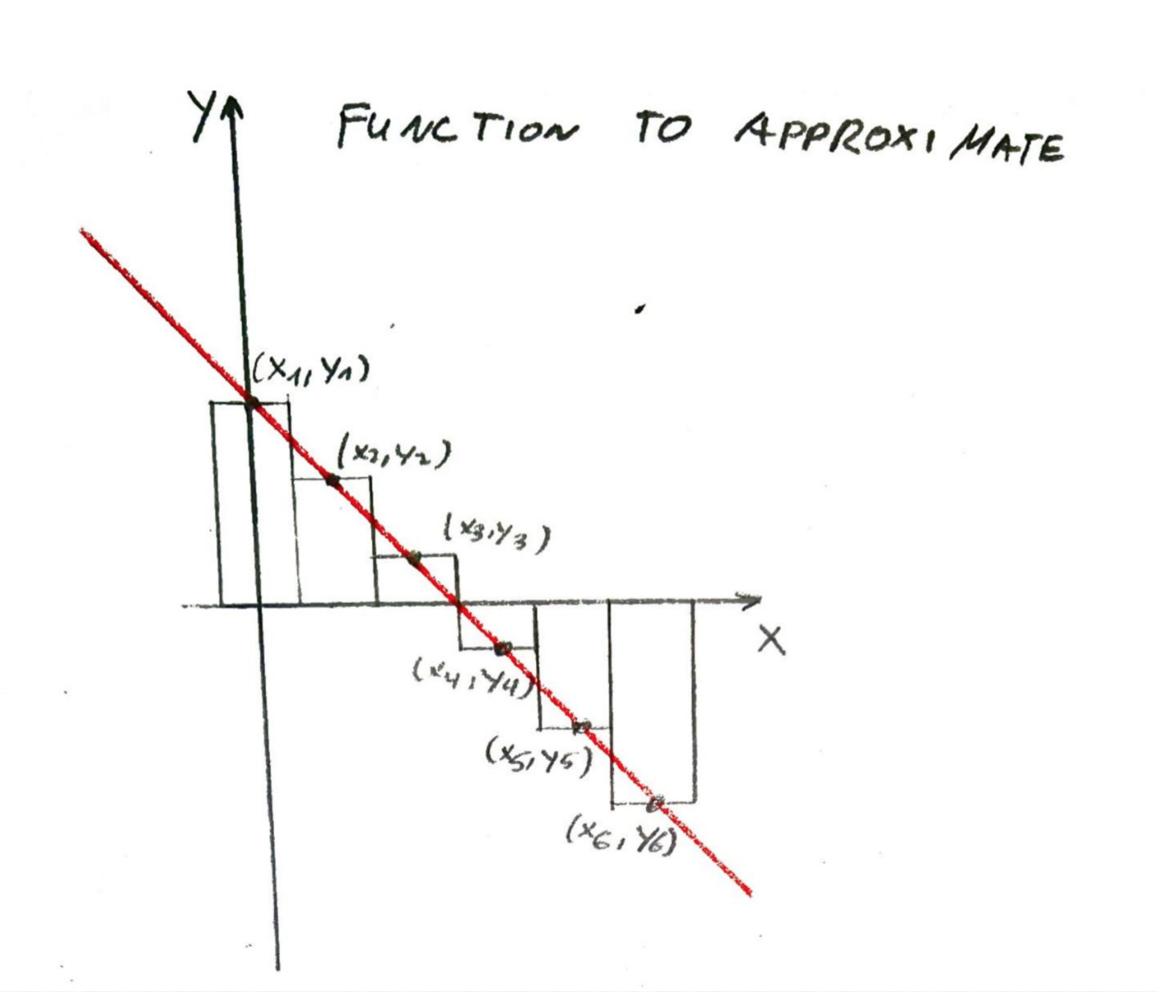
Covariance matrix of joint distribution:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{a} - \boldsymbol{\mu}_2 \right) \qquad \overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}.$$

$$\mathbf{p}(\mathbf{x_1} \mid \mathbf{x_2} = \mathbf{a}) = \mathcal{N}\left(\overline{\mu},\ \overline{\Sigma}\right)$$

Gaussian Processes: Basics

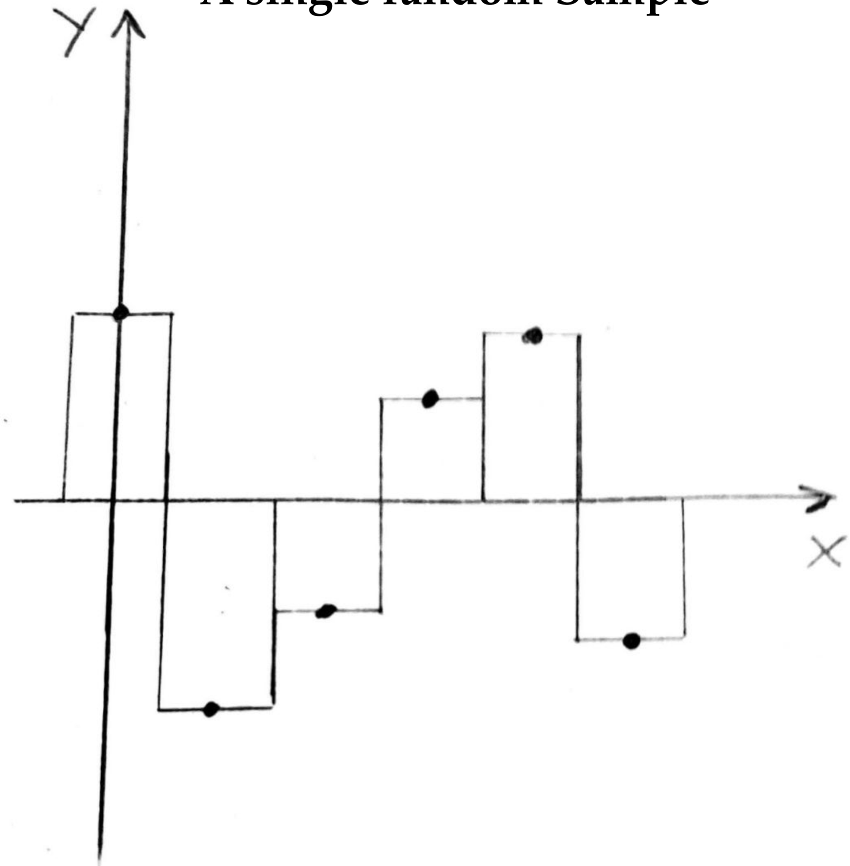


Discretize the red linear function in bins and denote the height of the bins as (x, y)

I can fully specify the discretized function if I measured these 6 points

Before the Measurement: Prior

A single random Sample



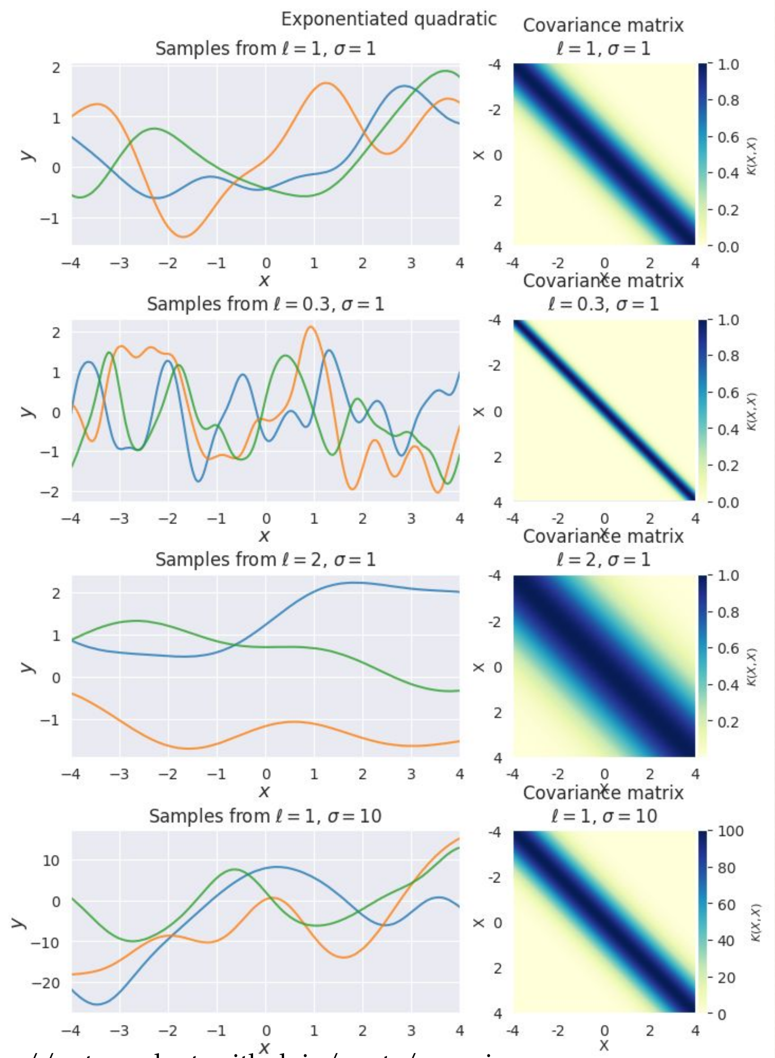
Before I have any information about the height of the histogram discretization I assume that the heights are normally distributed with zero mean

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

We also want the correlation between points scale with their distance

$$k(x_1, x_2) = \sigma^2 \exp(-\frac{(\mathbf{x_1} - \mathbf{x_2})^2}{2l^2})$$

Effect of Kernel Parameters?



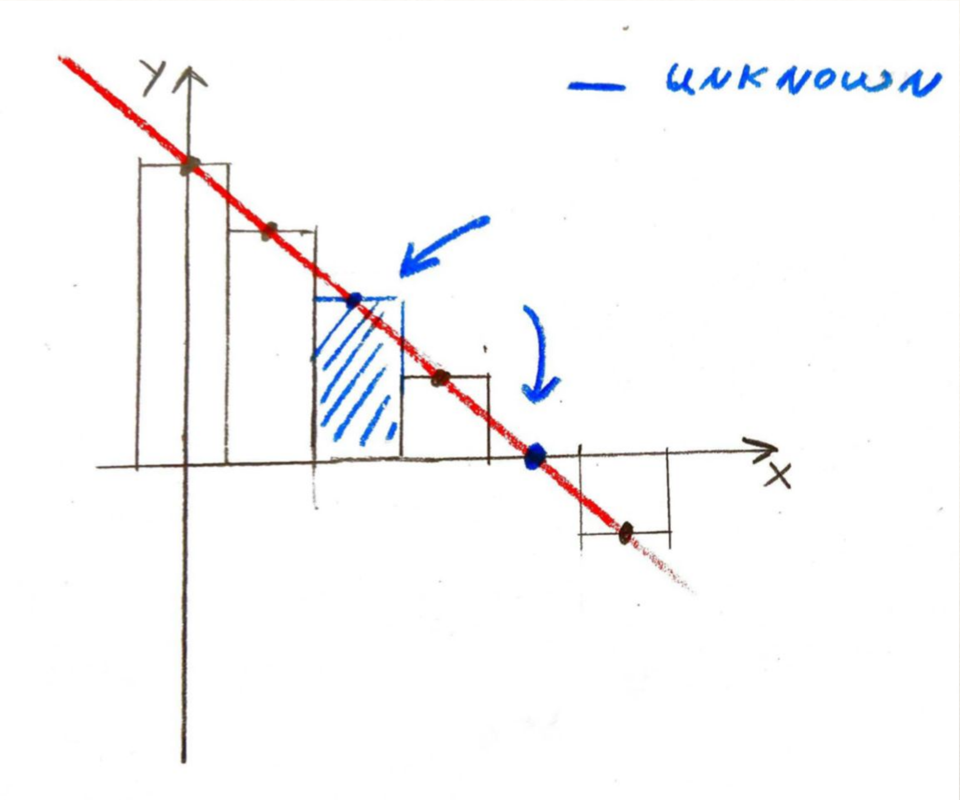
https://peterroelants.github.io/posts/gaussian-process-kernels/

The prefactor scales the fluctuations, the scale length determines the smoothness or correlation

$$k(x_1,\, x_2) \;=\; \sigma^2 \exp(-\frac{(\mathbf{x_1} - \mathbf{x_2})^2}{2l^2})$$

This function is often called a Kernel.

A Kernel is a non-negative real valued integratable function. Often the optional properties of symmetry and normalization (integrates to unity) are imposed.



$$\begin{pmatrix} y_{Known} \\ y_{Unknown} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x_{Known}, x_{Known}) & K(x_{Unknown}, x_{Known}) \\ K(x_{Known}, x_{Unknown}) & K(x_{Unknown}, x_{Unknown}) \end{pmatrix} \right)$$

$$\boxed{y_{Unknown} \,\Big|\, y_{Known} \sim \mathcal{N}\left(\overline{y}, \overline{\Sigma}\right)}$$

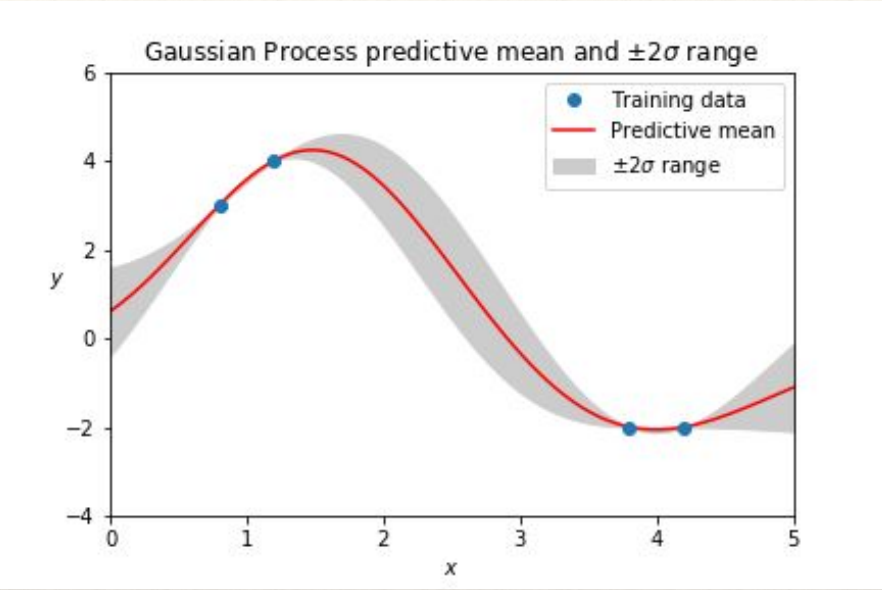
[Here is the Proof]

$$\overline{\mu} = K\left(x^{Unknonw}, x^{Known}\right) K\left(x^{Known}, x^{Known}\right)^{-1} y^{Known}$$

$$\Sigma = K\left(x^{Unknown}, x^{Unknown}\right) - K\left(x^{Unknown}, x^{Known}\right) K\left(x^{Known}, x^{Known}\right)^{-1} K\left(x^{Known}, x^{Unkown}\right)$$

The continuous limit

- So far we have parametrized the Gaussian Process using a discretization.
- The considerations presented in the previous slide also hold true for continuous functions (limit of infinitely small histogram bins).



https://planspace.org/20181226-gaussian_processes_are_not_so_fancy/

Gaussian Process Regression

- So far our `known' points have been known exactly.
- In practise the y values often have an error component
- We assume here that this error is independently identically distributed.
- Let the noise be parametrized by a scalar sigma. The rest is the same

$$\begin{pmatrix} y_{Known} \\ y_{Unknown} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x_{Known}, x_{Known}) + 1\sigma^2 & K(x_{Unknown}, x_{Known}) \\ K(x_{Known}, x_{Unknown}) & K(x_{Unknown}, x_{Unknown}) \end{pmatrix} \right)$$

$$\overline{\mu} = K(x^{Unknonw}, x^{Known}) \left(K(x^{Known}, x^{Known}) + 1\sigma^2 \right)^{-1} y^{Known}$$

$$\Sigma = K(x^{Unknown}, x^{Unknown}) - K(x^{Unknown}, x^{Known}) \left(K(x^{Known}, x^{Known}) + 1\sigma^2 \right)^{-1} K(x^{Known}, x^{Unkown})$$

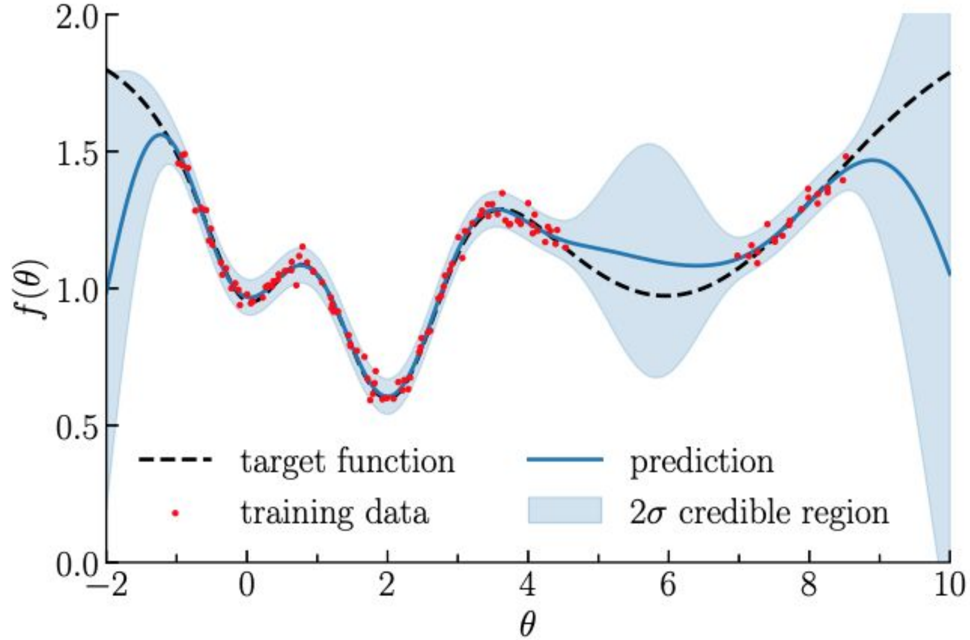


FIG. 3. Illustration of Gaussian process regression in one dimension, for the target test function $f : \theta \mapsto 2 - \exp\left[-(\theta-2)^2\right] - \exp\left[-(\theta-6)^2/10\right] - 1/(\theta^2+1)$ (dashed line). Training data are acquired (red dots); they are subject to a Gaussian observation noise with standard deviation $\sigma_{\mathrm{n}} = 0.03$. The blue line shows the mean prediction $\mu(\theta)$ of the Gaussian process regression, and the shaded region the corresponding $2\sigma(\theta)$ uncertainty. Gaussian processes allow interpolating and extrapolating predictions in regions of parameter space where training data are absent.

Leclercq et al. 2018

Important here:

Same pattern as before but the error around the points is nonzero! (because of the excess variance due to their error

FAQ: How can we select the parameters that describe the Kernel and noise in the observations?

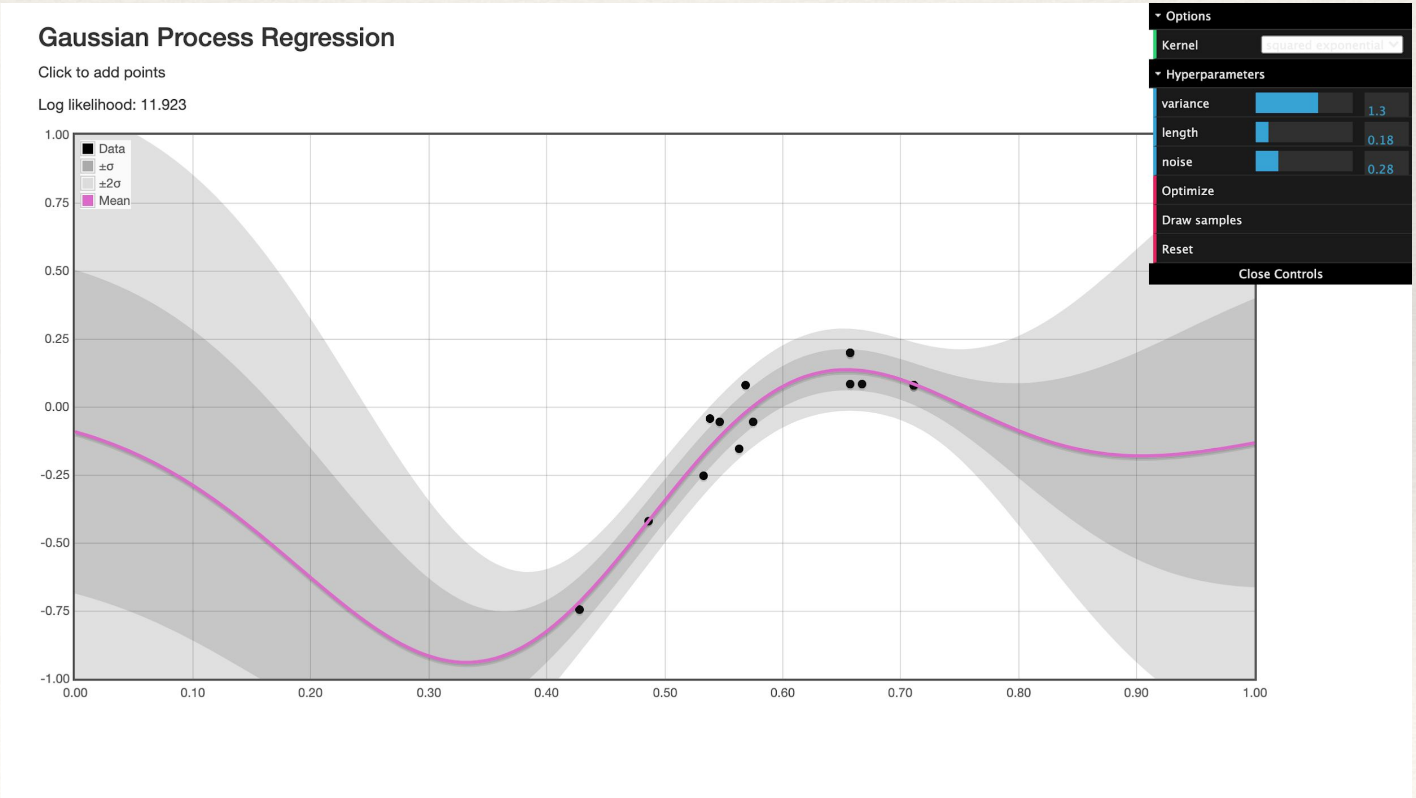
For Gaussian Observations simple scheme available

See Bayesian Data Analysis, Gelman et al. page. 503

Note: Gaussian Assumption often arbitrary. Comparison with other options (t-processes) desirable

Demo

http://chifeng.scripts.mit.edu/stuff/gp-demo/



Summary

Key Idea 1: Prior over functions
Impose a joint distribution over function evaluation points. Impose a covariance to perform function interpolation and regression.

Key Idea 2: The multivariate normal exhibits closed form solutions for marginal and conditional distributions. Ideal to formulate this process!

Exercise: Implement a simple Gaussian Process regression to determine the blue points in slide 10 using our histogram example. To this end implement the formulas on slide 12 using the exponential kernel on slide 9.

Literature

- **Bayesian Data Analysis, Gelman et al.**
- **Gaussian Processes for Machine Learning, Carl Edward Rasmussen**
- **Pattern Recognition and Machine Learning, Bishop**
- **Links in the lecture**