Statistics April 14, 2021

# Gaussian Processes

We say that $f$ is a Gaussian process:

$$f \sim GP(m, k)$$

$\Rightarrow$ Marginal distributions $f(x) \sim N(m(x), k(x,x))$

Finite dimensional joint distributions:

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left( \begin{pmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & \ddots & \\ k(x_3,x_1) & & \\ \vdots & & k(x_n,x_n) \end{pmatrix} \right)$$

If we denote by $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, then $\vec{f}(\vec{x}) \sim N(\vec{m}(\vec{x}), K(\vec{x},\vec{x}))$

What are the properties of a covariance kernel:

① symmetric $k(x,y) = k(y,x)$.

② positive definite: for any $\vec{x}$, $\vec{x}^T K(\vec{x},\vec{x})\,\vec{x} > 0$.

Recall the density function for a multivariate normal distribution:

$$f(x_1,..,x_n) = \frac{1}{2\pi^{n/2}\sqrt{\det(K)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T K^{-1}(\vec{x}-\vec{\mu})}$$

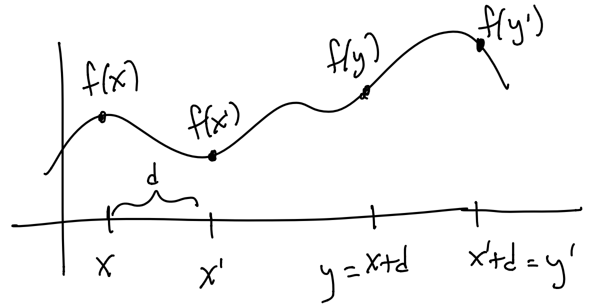
Continuous interpretation:

$$\iint f(x)\, K(x,y)\, f(y)\, dx\, dy > 0$$

If we take the continuous covariance kernel $k$ and form the matrix $K(\vec{x},\vec{x})$ with entries $K_{ij} = k(x_i, x_j)$, the $K$ = Gram Matrix.

Other properties that are useful in modeling:

① Translation invariance: $K(x, x') = k(x - x')$ (stationary)




$$\mathrm{cov}(f(x), f(x')) = \mathrm{cov}(f(y), f(y'))$$

② Isotropic: $k(x, x') = k(|x - x'|)$

Ex: $k(x,x') = e^{-|x-x'|}$

$$k(x,x') = a\, e^{-(x-x')^2/b}$$

Positive definite translation invariant covariance kernels have a nice 1-1 correspondence with "spectral densities"

Thm A stationary covariance kernel $k = k(x-x') = k(\tau)$ can be written as

$$\underbrace{k(\tau) = \int \overbrace{S(s)}\, e^{2\pi i s \tau}\, ds}_{\text{Inverse Fourier Transform}}$$

spectral density, or power spectrum of $k$.

where $S$ is a positive function, i.e. $S(s) > 0$ for each $s$. (Bochner's Theorem).

and therefore by Fourier inversion,

$$S(s) = \underbrace{\int k(\tau)\, e^{-2\pi i s \tau}\, d\tau}_{\text{Fourier Transform.}}$$

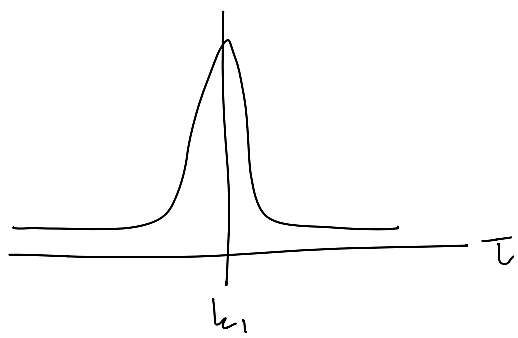
<u>Ex:</u> $k(0) = \mathrm{Var}(x,x)$.

$$= \int S(s)\, e^{2\pi i s \cdot 0}\, ds$$

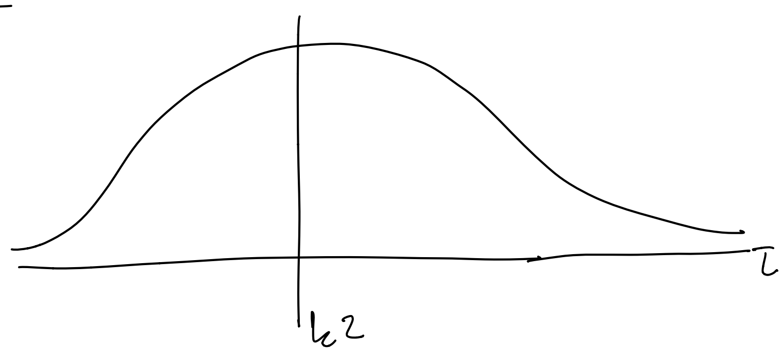
$$= \int S(s)\, ds \implies \text{this must be finite.}$$

( See chapter 4 of Rasmussen & Williams for more, very nice clean theory if you know a little Fourier analysis and stochastic processes.).

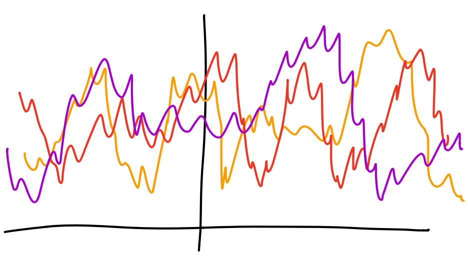
## Effect of the covariance kernel

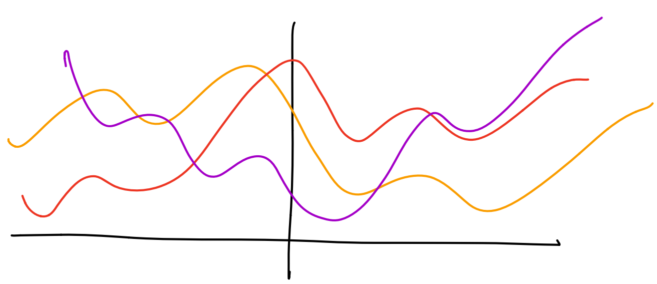



short-range correlations




long-range correlations





# Gaussian Process Regression

**Noise Free**:

Ground truth: $y = f(x)$, a deterministic function

Model: $y = f(x)$, $f \sim GP(0, k)$

Training data: $(x_1, y_1) \ldots (x_n, y_n)$

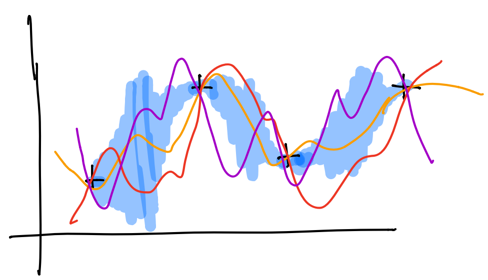
Predictions: $(x_1^*, y_1^*) \ldots (x_m^*, y_m^*)$

Joint distribution of $\vec{y}, \vec{y}_*$ ~ multivariate normal distribution

$$\begin{pmatrix} \vec{y} \\ \vec{y}_* \end{pmatrix} \sim N\left( \vec{0}, \begin{pmatrix} K(\vec{x},\vec{x}) & K(\vec{x},\vec{x}_*) \\ K(\vec{x}_*,\vec{x}) & K(\vec{x}_*,\vec{x}_*) \end{pmatrix} \right)$$

We want to compute the distribution

of $\vec{y}_* \mid \vec{x}_*, \vec{x}, \vec{y}$ ← posterior, conditioned on the observed data.



+ : observed data points.

In this setup, draws from the posterior have to pass through the observed data.

It can be shown that the conditional distribution is

$$\vec{y}_* \mid \vec{x}_*, \vec{x}, \vec{y} \sim N\left( K(\vec{x}_*,\vec{x})\, K(\vec{x},\vec{x})^{-1} \vec{y},\ K(\vec{x}_*,\vec{x}_*) - K(\vec{x}_*,\vec{x})\, K(\vec{x},\vec{x})^{-1} K(\vec{x},\vec{x}_*) \right)$$

**Exercise** To prove this to yourself.

An analogous calculation:

Solve $\begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\begin{pmatrix} \vec{x} \\ \vec{y} \end{pmatrix} = \begin{pmatrix} \vec{a} \\ \vec{b} \end{pmatrix}$

$$\Rightarrow \begin{pmatrix} A - CB^{-1}C^T & 0 \\ C^T & B \end{pmatrix}\begin{pmatrix} \vec{x} \\ \vec{y} \end{pmatrix} = \begin{pmatrix} \vec{a} - CB^{-1}\vec{b} \\ \vec{b} \end{pmatrix}$$

With independent noise

deterministic function

Observe $\quad y = f(x) + \epsilon$

$\epsilon \sim N(0, \sigma^2)$

Model: $\quad y = f(x) + \epsilon$

Gaussian process $GP(0, k)$

Normal random variable, also known as white noise, also a Gaussian process.

$\Rightarrow$ $y$ is a Gaussian Process with

$$\operatorname{cov}(y_i, y_j) = k(x_i, x_j) + \sigma^2 \delta_{ij}$$

Kronecker delta function,

$\delta_{ij} = 1$ if $i=j$

$0$ otherwise.

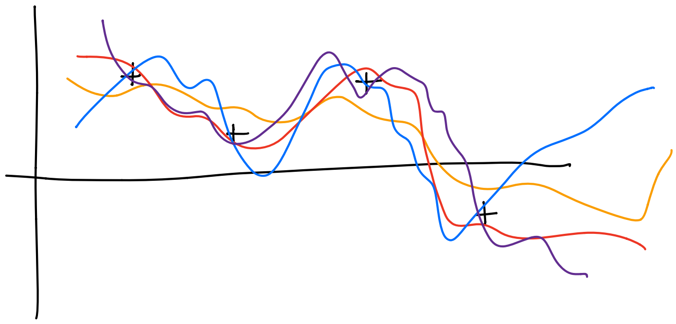
$$\operatorname{Cov}(\vec{y}) = K(\vec{x}, \vec{x}) + \sigma^2 I$$

So the joint distribution of the training data $\vec{x}, \vec{y}$ with the predicted $\vec{x}_*, \vec{y}_*$ is

$$\begin{pmatrix} \vec{y} \\ \vec{y}_* \end{pmatrix} \sim N\left(\vec{0}, \begin{pmatrix} K(\vec{x},\vec{x}) + \sigma^2 I & K(\vec{x},\vec{x}_*) \\ K(\vec{x}_*,\vec{x}) & K(\vec{x}_*,\vec{x}_*) \end{pmatrix}\right)$$

Same calculation to compute posterior distribution:

$$\vec{y}_* \mid \vec{x}_*, \vec{x}, \vec{y} \sim N\Big( K(\vec{x}_*, \vec{x})\big(K(\vec{x},\vec{x}) + \sigma^2 I\big)^{-1} \vec{y},$$
$$K(\vec{x}_*, \vec{x}_*) - K(\vec{x}_*, \vec{x})\big(K(\vec{x},\vec{x}) + \sigma^2 I\big)^{-1} K(\vec{x}, \vec{x}_*)\Big)$$



+ observations

$y = f(x) + \epsilon$

Draws from posterior do not pass through data.

One last comment:

The Bayesian "predictor" or "estimator" at the points $\vec{x}_*$ is:

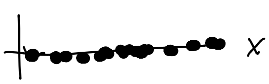
$$\hat{y}_* = K(\vec{x}_*, \vec{x})\big(K(\vec{x},\vec{x}) + \sigma^2 I\big)^{-1} \vec{y}$$

$$= \sum \alpha_i K(\vec{x}_*, x_i)$$ ← a linear combination of covariance kernels.

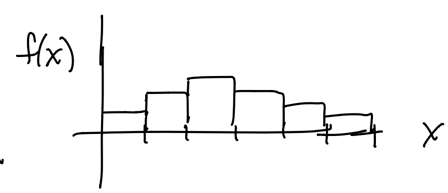
$\alpha_i$ = $i^{th}$ entry of $\big(K(\vec{x},\vec{x}) + \sigma^2 I\big)^{-1} \vec{y}$

## Non-parametric Methods (curve smoothing)

Ex: Estimate a probability density from observed data

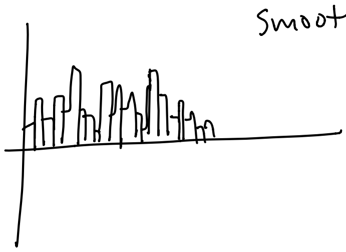


→ bin the data, create a histogram

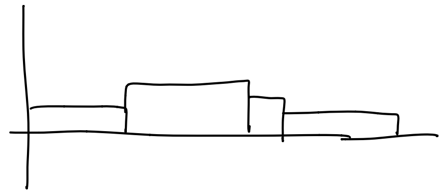


## How do you pick the bin width?

Narrow, not enough smoothing



Wide, oversmoothed



For example: Measure the quality of the histogram (one option) using the mean-squared error ($L_2$ error)

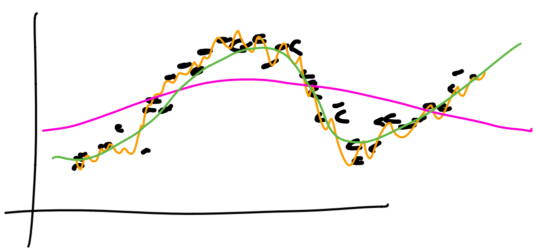
Let $\hat{f}$ be the histogram estimator of $f$

$$MSE = \int \left( f(x) - \hat{f}(x) \right)^2 dx$$

true pdf — histogram


pick bin width around here.

### Example Regression



- observed data
- under smoothed
- over smoothed
- just right

## Bias - Variance Tradeoff

Loss function: pointwise error

$L_2$ loss: $L(f, \hat{f})(x) = \left( f(x) - \hat{f}(x) \right)^2$

$$\underline{\text{Risk}} = \mathbb{E}\left( L(f,\hat{f})(x) \right) = R(f,\hat{f})(x),$$

random variable,
function of the observations

expection taken with respect
to $\hat{f}$

It can be shown that the $L_2$ Risk can be written as :

$$R(f,\hat{f})(x) = \text{bias}_x^2 + \text{Var}_x$$

where $\text{bias}_x = \mathbb{E}\left(\hat{f}(x)\right) - f(x)$

$$\text{Var}_x = \text{Var}\left(\hat{f}(x)\right)$$

To capture an average error, integrate:

Integrated risk
Integrate MSE

$$= \int R(f,\hat{f})(x)\, dx$$

$$= \int \mathbb{E}\left( L(f,\hat{f})(x) \right) dx$$

$$\neq \int \left( f(x) - \hat{f}(x) \right)^2 dx \qquad = R(f,\hat{f})$$

<u>Example</u> In the case of a regression, the average MSE can be used:

$$R(r,\hat{r}) = \frac{1}{n} \sum_i R(r,\hat{r})(x_i)$$

**Example** Regression model : $Y_i = r(x_i) + \epsilon_i$

$\epsilon_i \sim N(0, \sigma^2)$

Compute some estimator $\hat{r}$ for $r$.

Now predict at each of the original $x_i$'s:

$$\hat{Y}_i = \hat{r}(x_i)$$

$\Rightarrow$ squared prediction error $= \left(Y_i - \hat{r}(x_i)\right)^2$

$$= \left(r(x_i) + \epsilon_i - \hat{r}(x_i)\right)^2$$

$\Rightarrow$ prediction risk $= \mathbb{E}\left(\frac{1}{n}\sum\left(Y_i - \hat{r}(x_i)\right)^2\right)$

$$= \frac{1}{n}\sum\left(r(x_i) - \hat{r}(x_i)\right)^2 + \frac{1}{n}\sum\sigma^2$$

$$= R(r, \hat{r}) + \sigma^2$$

The challenge is to balance the bias and the variance

lots of smoothing : $\uparrow$ bias, $\downarrow$ variance

too little smoothing : $\downarrow$ bias, $\uparrow$ variance