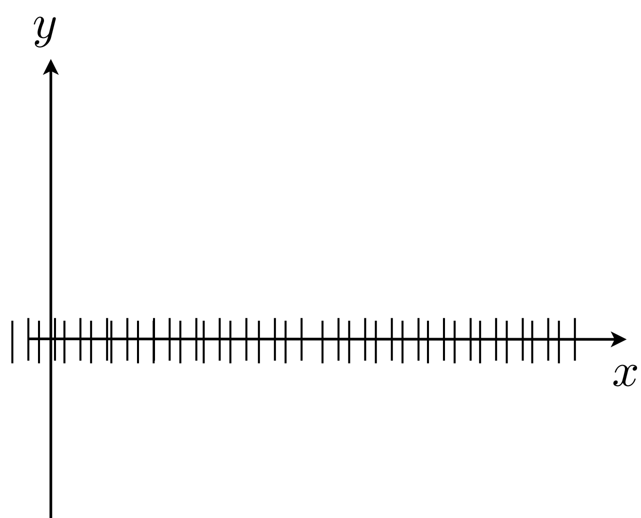
What is Gaussian process?

From distribution to process

Distribution: probability defined in a finite dimensional space.

Process: probability defined in an infinite dimensional space (e.g., function).



<u>Gaussian distribution</u>

$$p(\boldsymbol{y}) = \mathrm{Gauss}_{100000}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{y} \in \mathbb{R}^{100000},$$
$$\boldsymbol{\mu} \in \mathbb{R}^{100000},$$
$$\boldsymbol{\Sigma} \in \mathbb{R}^{100000 \times 100000},$$

<u>Gaussian process</u>

$$p(y) = \mathrm{GP}\left(y; \nu, k\right)$$

$$y(x) \in \mathbb{R} \text{ for } x \in \mathbb{R},$$
$$\nu(x) \in \mathbb{R} \text{ for } x \in \mathbb{R},$$
$$k(x, x') \in \mathbb{R} \text{ for } (x, x') \in \mathbb{R}^2,$$

Formal definition

A Gaussian process is a collection of random variables,
any of finite number of which have a joint Gaussian distribution.

[Rasmussen&Williams:2016]

Practical property 1 (Marginalization property)

Random functions $f(\cdot) : \mathbb{R}^L \mapsto \mathbb{R}$ follow the Gaussian process $f \sim \mathrm{GP}(\nu(\cdot), k(\cdot, \cdot))$ with a mean function $\nu(\cdot)$ and a kernel function $k(\cdot, \cdot)$.

For any given input set $\{\boldsymbol{x}^{(n)} \in \mathbb{R}^L\}_{n=1}^N$, it holds that

$$p(\boldsymbol{f}|\boldsymbol{\nu}, \boldsymbol{K}) \propto \exp\left(-\frac{(\boldsymbol{f}-\boldsymbol{\nu})^\top \boldsymbol{K}^{-1} (\boldsymbol{f}-\boldsymbol{\nu})}{2}\right),$$

where

$$\boldsymbol{f} = \left(f(\boldsymbol{x}^{(1)}), \ldots, f(\boldsymbol{x}^{(N)})\right)^\top,$$

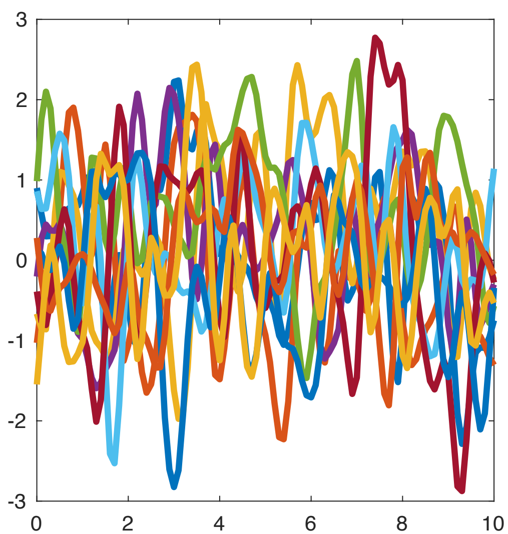
$$\boldsymbol{\nu} = \left(\nu(\boldsymbol{x}^{(1)}), \ldots, \nu(\boldsymbol{x}^{(N)})\right)^\top,$$

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(N)}) \\ \vdots & & \vdots \\ k(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(N)}) \end{pmatrix}.$$

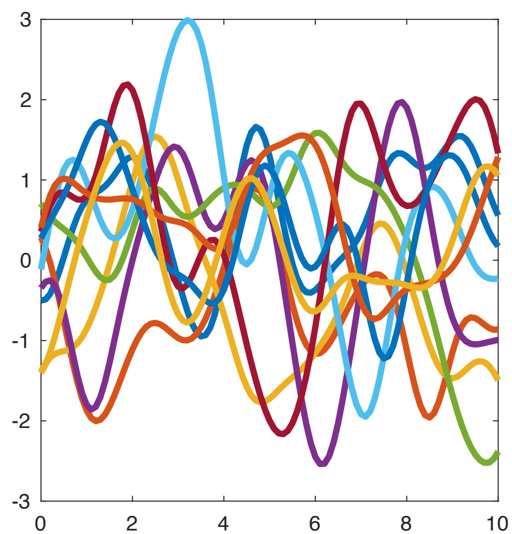
Assume infinite dimension, work in finite dimension.

Examples

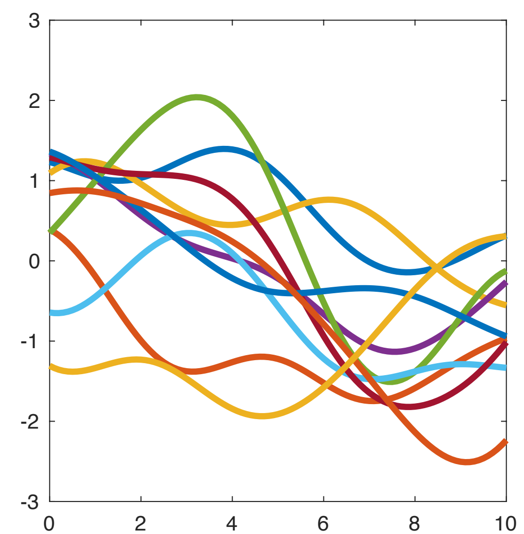
$$f \sim \mathrm{GP}(\nu, k), \qquad \text{where} \quad \nu(\boldsymbol{x}) \equiv 0, \quad k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$$



$\gamma^2 = 0.1$



$\gamma^2 = 1$



$\gamma^2 = 10$

Practical property 2

Given another input set $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$, the joint distribution is

$$p(\boldsymbol{f}, \boldsymbol{f}^*) \propto \exp\left(- \frac{\left(\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f}^* \end{pmatrix} - \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\nu}^* \end{pmatrix} \right)^\top \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}^* \\ \boldsymbol{K}^{*\top} & \boldsymbol{K}^{**} \end{pmatrix}^{-1} \left(\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f}^* \end{pmatrix} - \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\nu}^* \end{pmatrix} \right)}{2} \right),$$

where $\boldsymbol{K}^*$ is the kernel between the two sets, $\{\boldsymbol{x}^{(n)} \in \mathbb{R}^L\}_{n=1}^{N}$ and $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$, and $\boldsymbol{K}^{**}$ is the kernel within the second set, $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$.

The conditional distribution of $\boldsymbol{f}^*$ given $\boldsymbol{f}$ is written as

$$p(\boldsymbol{f}^*|\boldsymbol{f}) \propto \exp\left(- \frac{(\boldsymbol{f}^* - \boldsymbol{\mu}_{\mathrm{c}})^\top \boldsymbol{\Sigma}_{\mathrm{c}}^{-1} (\boldsymbol{f}^* - \boldsymbol{\mu}_{\mathrm{c}})}{2} \right),$$

where

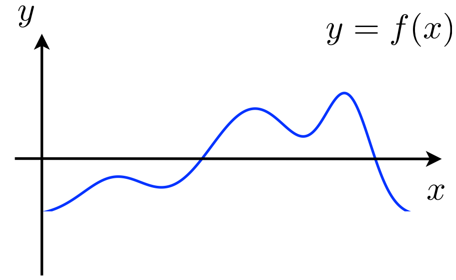
$$\boldsymbol{\mu}_{\mathrm{c}} = \boldsymbol{\nu}^* + \boldsymbol{K}^{*\top} \boldsymbol{K}^{-1} (\boldsymbol{f} - \boldsymbol{\nu}),$$

$$\boldsymbol{\Sigma}_{\mathrm{c}} = \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top} \boldsymbol{K}^{-1} \boldsymbol{K}^*.$$

Information transfer from training locations to test locations!

GP regression

GP regression



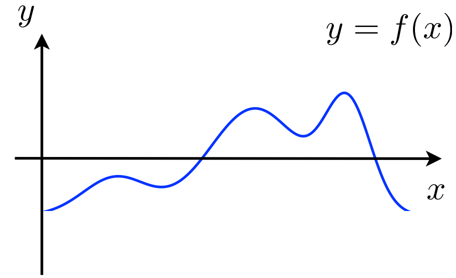
Observed: $\boldsymbol{x} \in \mathbb{R}^L, y \in \mathbb{R}$

Parameter: $f : \mathbb{R}^L \to \mathbb{R}$

Model dist.: $p(y|f) \propto \exp\left(-\frac{\|y - f(\boldsymbol{x})\|^2}{2\sigma^2}\right)$

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$

GP regression



Observed: $\boldsymbol{x} \in \mathbb{R}^L, y \in \mathbb{R}$

Parameter: $f : \mathbb{R}^L \to \mathbb{R}$

Model dist.: $p(y|f) \propto \exp\left(-\frac{\|y - f(\boldsymbol{x})\|^2}{2\sigma^2}\right)$

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$

Data set: $\{\boldsymbol{x}^{(n)}, y^{(n)}\}_{n=1}^N$

$$\boldsymbol{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix}$$

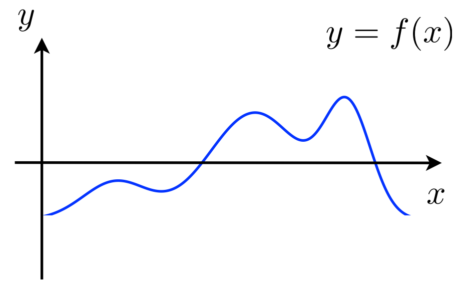
$$p(\boldsymbol{y}|\boldsymbol{f}) \propto \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{f}\|^2}{2\sigma^2}\right)$$

$$\boldsymbol{f} = \begin{pmatrix} f(\boldsymbol{x}^{(1)}) \\ \vdots \\ f(\boldsymbol{x}^{(N)}) \end{pmatrix}$$

$$p(\boldsymbol{f}) \propto \exp\left(-\frac{\boldsymbol{f}^\top \boldsymbol{K}^{-1} \boldsymbol{f}}{2}\right)$$

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(N)}) \\ \vdots & & \vdots \\ k(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(N)}) \end{pmatrix}.$$

GP regression



Posterior on **f**:

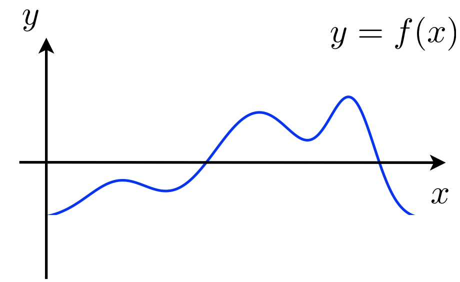
$$
\begin{aligned}
p(\boldsymbol{f}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}) \\
&\propto \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{f}\|^2}{2\sigma^2} - \frac{\boldsymbol{f}^\top \boldsymbol{K}^{-1}\boldsymbol{f}}{2}\right) \\
&\propto \exp\left(-\frac{-2\sigma^{-2}\boldsymbol{y}^\top \boldsymbol{f} + \boldsymbol{f}^\top\left(\boldsymbol{K}^{-1}+\sigma^{-2}\boldsymbol{I}_N\right)\boldsymbol{f}}{2}\right) \\
&\propto \exp\left(-\frac{(\boldsymbol{f}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})}{2}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu} &= \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{y}, \\
\boldsymbol{\Sigma} &= \left(\boldsymbol{K}^{-1} + \sigma^{-2}\boldsymbol{I}_N\right)^{-1}.
\end{aligned}
$$

Posterior at training data points are Gaussian!

GP regression



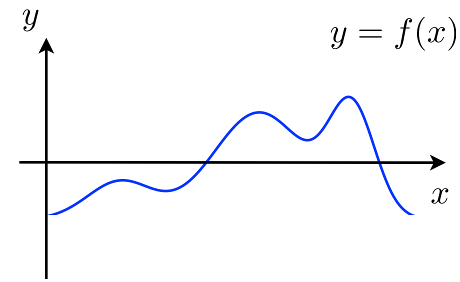
Predictive for test points $\{\boldsymbol{x}^{*(n)}\}_{n=1}^{N^*}$:

$$y^* = \begin{pmatrix} y^{*(1)} \\ \vdots \\ y^{*(N)} \end{pmatrix} \qquad \boldsymbol{f}^* = \begin{pmatrix} f(\boldsymbol{x}^{*(1)}) \\ \vdots \\ f(\boldsymbol{x}^{*(N)}) \end{pmatrix}$$

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \int p(\boldsymbol{y}^*|f)p(f|\boldsymbol{y})df \\ &= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \int p(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{f}^* \\ &= \int \underbrace{p(\boldsymbol{y}^*|\boldsymbol{f}^*)}_{\text{model dist.}} \int \underbrace{p(\boldsymbol{f}^*|\boldsymbol{f})}_{\text{conditional (transfer)}} \underbrace{p(\boldsymbol{f}|\boldsymbol{y})}_{\text{posterior on } \mathbf{f}} d\boldsymbol{f} d\boldsymbol{f}^* \end{aligned}$$

Everything is Gaussian!

GP regression



Predictive for test points $\{\boldsymbol{x}^{*(n)}\}_{n=1}^{N^*}$:

$$\boldsymbol{y}^* = \begin{pmatrix} y^{*(1)} \\ \vdots \\ y^{*(N)} \end{pmatrix}$$

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \int p(\boldsymbol{y}^*|f)p(f|\boldsymbol{y})df$$

$$= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \int p(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{f}^*$$

$$= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \underbrace{\int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y}) d\boldsymbol{f}}_{\text{posterior on } \mathbf{f}^*} d\boldsymbol{f}^*$$

$$\boldsymbol{f}^* = \begin{pmatrix} f(\boldsymbol{x}^{*(1)}) \\ \vdots \\ f(\boldsymbol{x}^{*(N)}) \end{pmatrix}$$

Practical property 2

Given another input set $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$, the joint distribution is

$$p(\boldsymbol{f}, \boldsymbol{f}^*) \propto \exp\left(- \frac{\left(\begin{pmatrix}\boldsymbol{f} \\ \boldsymbol{f}^*\end{pmatrix} - \begin{pmatrix}\boldsymbol{\nu} \\ \boldsymbol{\nu}^*\end{pmatrix}\right)^\top \begin{pmatrix}\boldsymbol{K} & \boldsymbol{K}^* \\ \boldsymbol{K}^{*\top} & \boldsymbol{K}^{**}\end{pmatrix}^{-1} \left(\begin{pmatrix}\boldsymbol{f} \\ \boldsymbol{f}^*\end{pmatrix} - \begin{pmatrix}\boldsymbol{\nu} \\ \boldsymbol{\nu}^*\end{pmatrix}\right)}{2} \right),$$

where $\boldsymbol{K}^*$ is the kernel between the two sets, $\{\boldsymbol{x}^{(n)} \in \mathbb{R}^L\}_{n=1}^{N}$ and $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$, and $\boldsymbol{K}^{**}$ is the kernel within the second set, $\{\boldsymbol{x}^{*(n)} \in \mathbb{R}^L\}_{n=1}^{N^*}$.

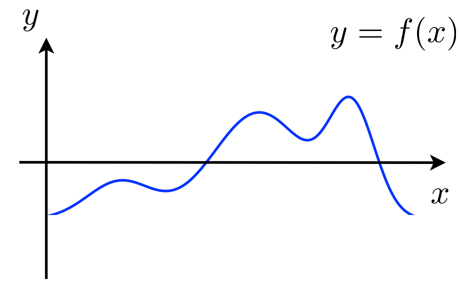
The conditional distribution of $\boldsymbol{f}^*$ given $\boldsymbol{f}$ is written as

$$p(\boldsymbol{f}^*|\boldsymbol{f}) \propto \exp\left(- \frac{(\boldsymbol{f}^* - \boldsymbol{\mu}_{\mathrm{c}})^\top \boldsymbol{\Sigma}_{\mathrm{c}}^{-1} (\boldsymbol{f}^* - \boldsymbol{\mu}_{\mathrm{c}})}{2} \right),$$

where
$$\boldsymbol{\mu}_{\mathrm{c}} = \boldsymbol{\nu}^* + \boldsymbol{K}^{*\top}\boldsymbol{K}^{-1}(\boldsymbol{f} - \boldsymbol{\nu}),$$
$$\boldsymbol{\Sigma}_{\mathrm{c}} = \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}\boldsymbol{K}^{-1}\boldsymbol{K}^{*}.$$

Information transfer from the training locations to the test locations!

GP regression



Predictive for test points $\{\boldsymbol{x}^{*(n)}\}_{n=1}^{N^*}$:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \int p(\boldsymbol{y}^*|f)p(f|\boldsymbol{y})df$$
$$= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \int p(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}^*$$
$$= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \underbrace{\int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}^*}_{\text{posterior on } \mathbf{f}^*}$$

$$\boldsymbol{y}^* = \begin{pmatrix} y^{*(1)} \\ \vdots \\ y^{*(N)} \end{pmatrix}$$

$$\boldsymbol{f}^* = \begin{pmatrix} f(\boldsymbol{x}^{*(1)}) \\ \vdots \\ f(\boldsymbol{x}^{*(N)}) \end{pmatrix}$$

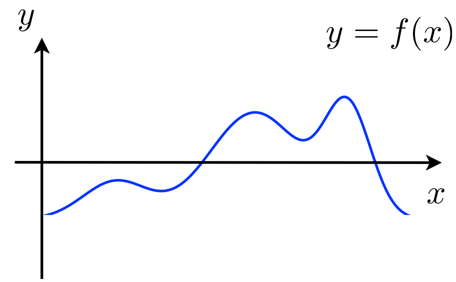
After some algebra…

$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f} \propto \exp\left(-\frac{\left(\boldsymbol{f}^* - \boldsymbol{\mu}_f\right)^\top \boldsymbol{\Sigma}_f^{-1} \left(\boldsymbol{f}^* - \boldsymbol{\mu}_f\right)}{2}\right),$$

where
$$\boldsymbol{\mu}_f = \boldsymbol{K}^{*\top}\left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_N\right)^{-1}\boldsymbol{y},$$
$$\boldsymbol{\Sigma}_f = \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}\left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_N\right)^{-1}\boldsymbol{K}^*.$$

Posterior at test data (any) points are Gaussian!

GP regression



Predictive for test points $\{\boldsymbol{x}^{*(n)}\}_{n=1}^{N^*}$:

$$\boldsymbol{y}^* = \begin{pmatrix} y^{*(1)} \\ \vdots \\ y^{*(N)} \end{pmatrix} \qquad \boldsymbol{f}^* = \begin{pmatrix} f(\boldsymbol{x}^{*(1)}) \\ \vdots \\ f(\boldsymbol{x}^{*(N)}) \end{pmatrix}$$

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \int p(\boldsymbol{y}^*|f)p(f|\boldsymbol{y})df \\ &= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \int p(\boldsymbol{f},\boldsymbol{f}^*|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}^* \\ &= \int p(\boldsymbol{y}^*|\boldsymbol{f}^*) \underbrace{\int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}^*}_{\text{posterior on } \mathbf{f}^*} \end{aligned}$$

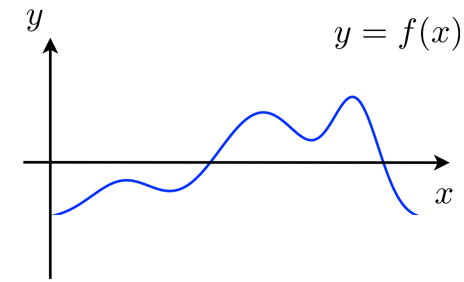
After some algebra…

$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f} \propto \exp\left(-\frac{\left(\boldsymbol{f}^* - \boldsymbol{\mu}_f\right)^\top \boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{f}^* - \boldsymbol{\mu}_f\right)}{2}\right),$$

where
$$\begin{aligned} \boldsymbol{\mu}_f &= \boldsymbol{K}^{*\top}\left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_N\right)^{-1}\boldsymbol{y}, \\ \boldsymbol{\Sigma}_f &= \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}\left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_N\right)^{-1}\boldsymbol{K}^*. \end{aligned}$$

Posterior is GP! $\quad p(f|\boldsymbol{y}) = \mathrm{GP}(f; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$

GP regression



Predictive for test points $\{\boldsymbol{x}^{*(n)}\}_{n=1}^{N^*}$:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) \propto \exp\left(-\frac{\left(\boldsymbol{y}-\boldsymbol{\mu}_{\mathrm{y}}\right)^\top \boldsymbol{\Sigma}_{\mathrm{y}}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}_{\mathrm{y}}\right)}{2}\right),$$

where

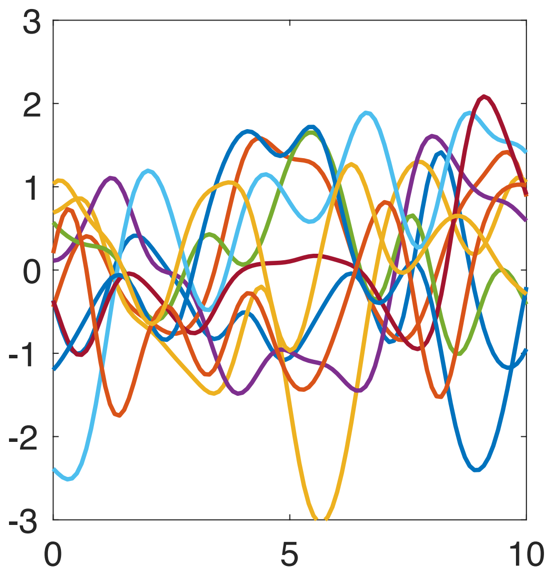
$$\boldsymbol{\mu}_y = \boldsymbol{K}^{*\top}\left(\boldsymbol{K}+\sigma^2\boldsymbol{I}_N\right)^{-1}\boldsymbol{y},$$

$$\boldsymbol{\Sigma}_y = \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}\left(\boldsymbol{K}+\sigma^2\boldsymbol{I}_N\right)^{-1}\boldsymbol{K}^* + \underbrace{\sigma^2\boldsymbol{I}_N}.$$

Bayesian predictive is analytically obtained without any approximation!

GP regression

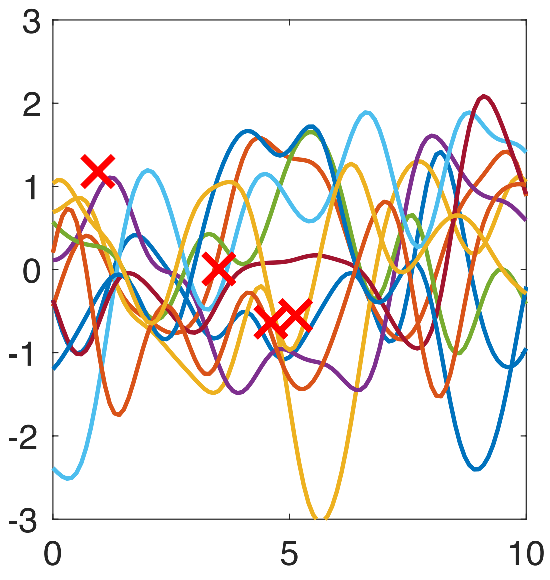
Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right)$.



GP regression

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right)$.

observation!

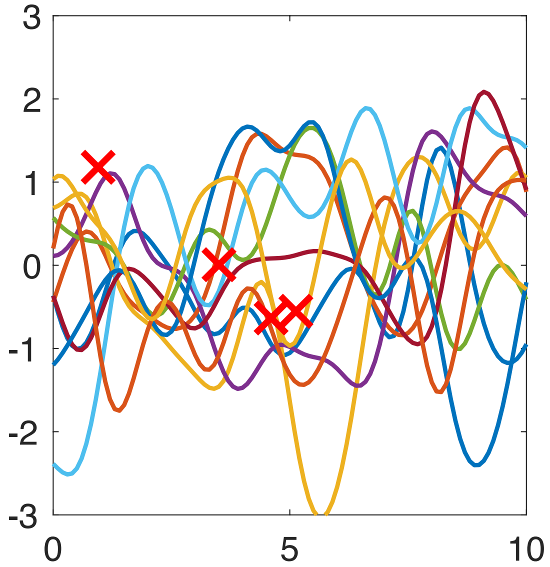


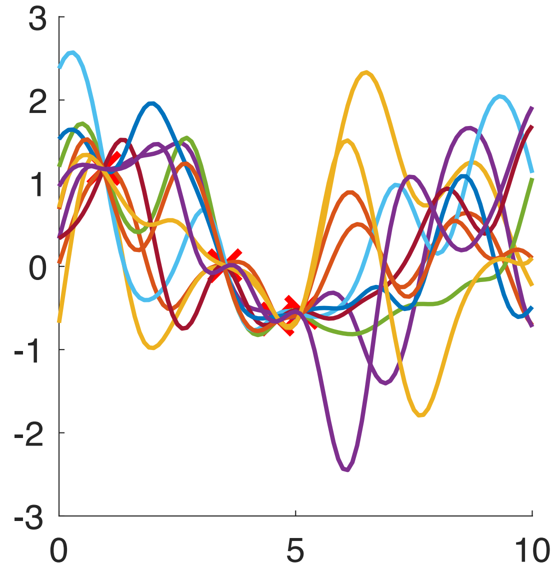
GP regression

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$

Posterior: $p(f|\boldsymbol{y}) = \mathrm{GP}(f; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$

observation!



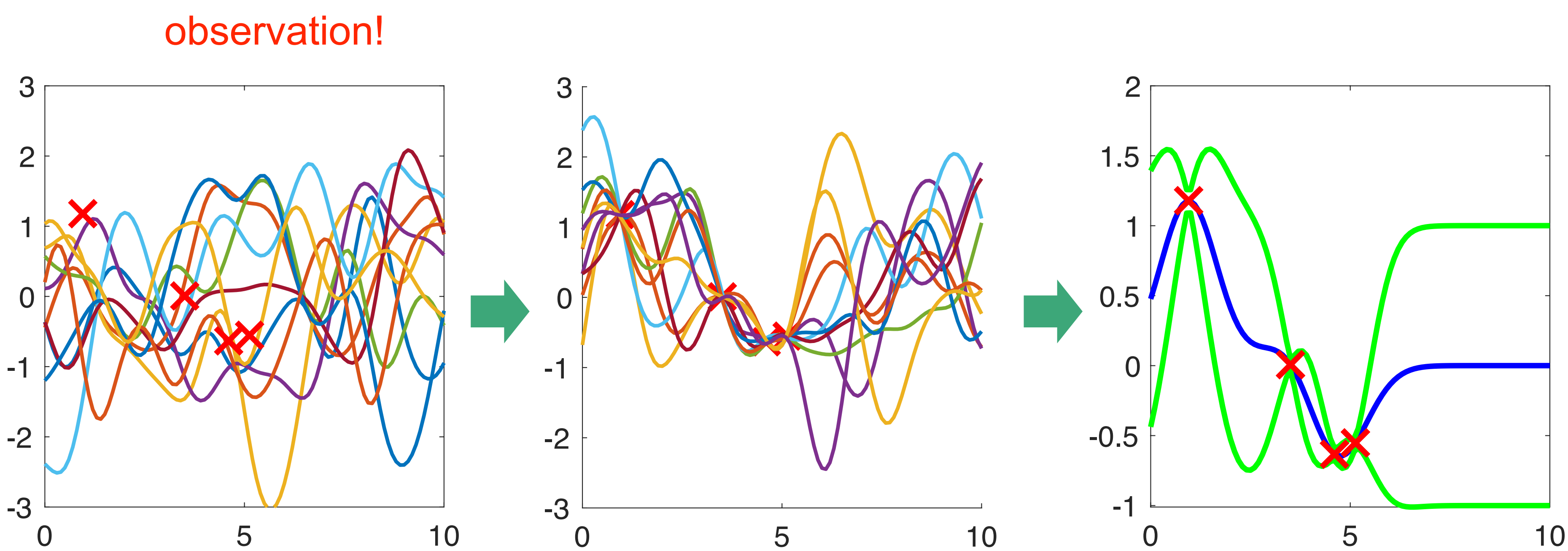


GP regression

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right)$.

Posterior: $p(f|\boldsymbol{y}) = \mathrm{GP}(f; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$

Predictive: $p(\boldsymbol{y}^*|\boldsymbol{y}) \propto \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^\top \boldsymbol{\Sigma}_{\mathrm{y}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})}{2}\right)$,

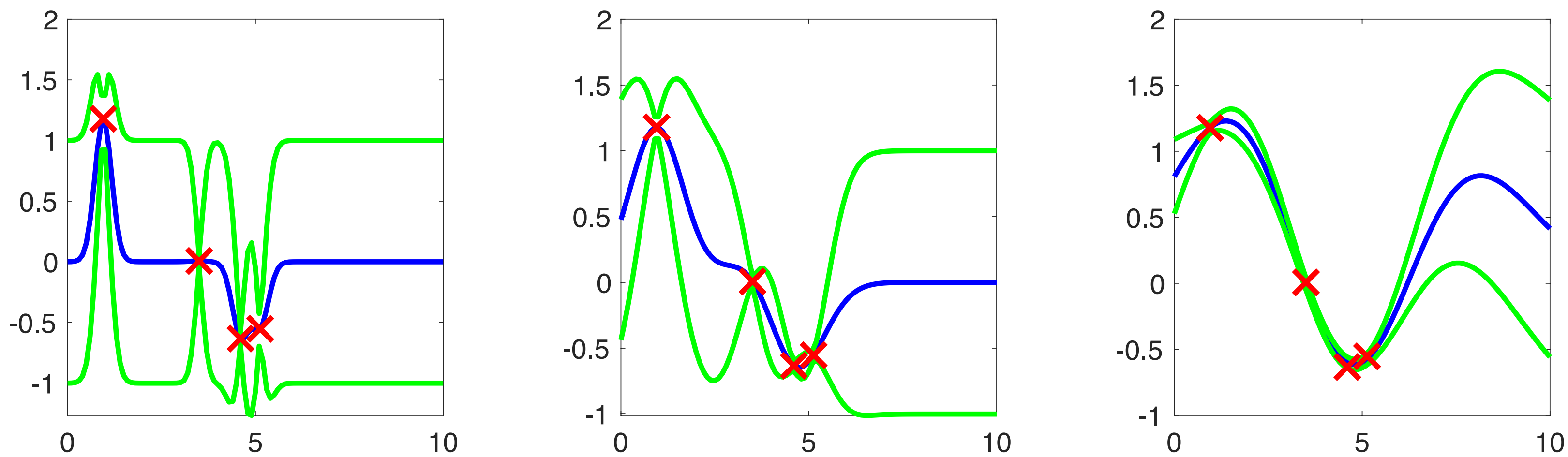


GP regression

Prior dist.: $p(f) = \mathrm{GP}(f; 0, k)$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\gamma^2}\right)$.

Model selection with marginal likelihood or cross validation,
but note that <u>criterion is task dependent</u>!

$$p(\boldsymbol{y}|\boldsymbol{K}) = (2\pi)^{-N/2}|\sigma^2\boldsymbol{I}_N + \boldsymbol{K}|^{-1/2}\exp\left(-\frac{\boldsymbol{y}^\top\left(\sigma^2\boldsymbol{I}_N+\boldsymbol{K}\right)^{-1}\boldsymbol{y}}{2}\right),$$



$\gamma^2 = 0.1$ $\quad$ $\gamma^2 = 1$ $\quad$ $\gamma^2 = 10$

$-\log p(\boldsymbol{y}|\gamma^2) = 1.4544$ $\quad$ $-\log p(\boldsymbol{y}|\gamma^2) = 1.4490$ $\quad$ $-\log p(\boldsymbol{y}|\gamma^2) = 0.6408$