Gaussian processes
-
Refresher and some more insights

Marcel Lüthi

Graphics and Vision Research Group
Department of Mathematics and Computer Science
University of Basel

Outline

- Gaussian process – refresher
- Vector-valued and scalar valued Gaussian processes
- The space of samples
- Gaussian process regression

Gaussian process: Formal definition

A Gaussian process $p(u) = GP(\mu, k)$

is a probability distribution over functions

$$u : \mathcal{X} \to \mathbb{R}^d$$

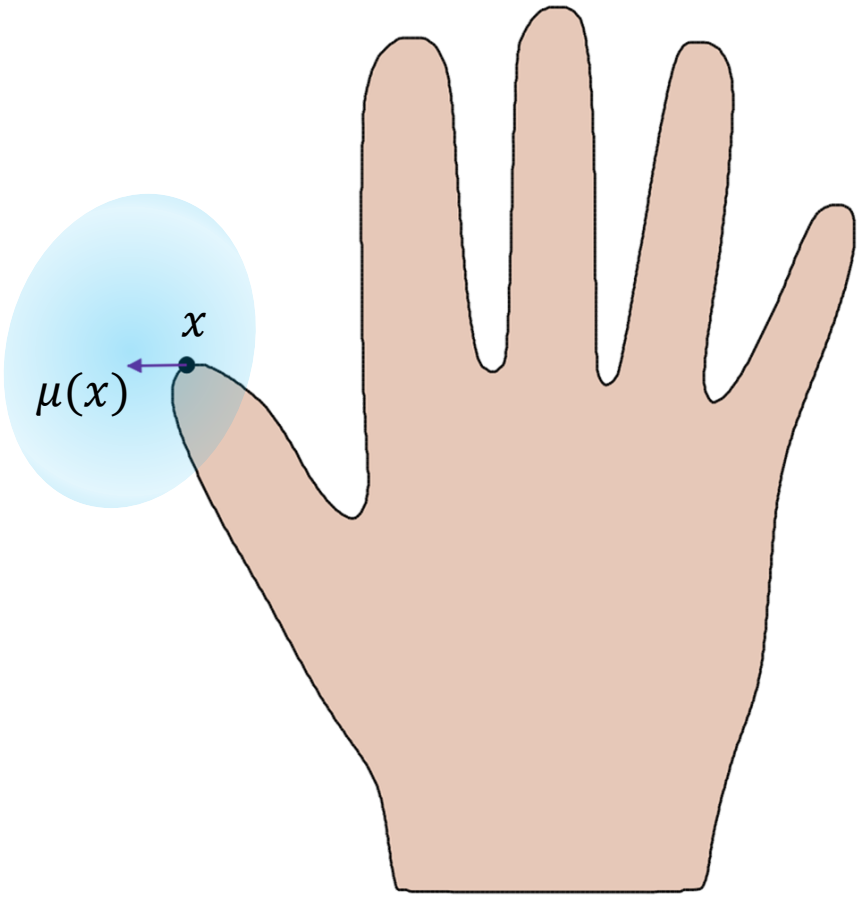
such that every finite restriction to function values

$$u_X = (u(x_1), \ldots, u(x_n))$$

is a multivariate normal distribution

$$p(u_X) = N(\mu_X, k_{XX}).$$

Gaussian process: Illustration

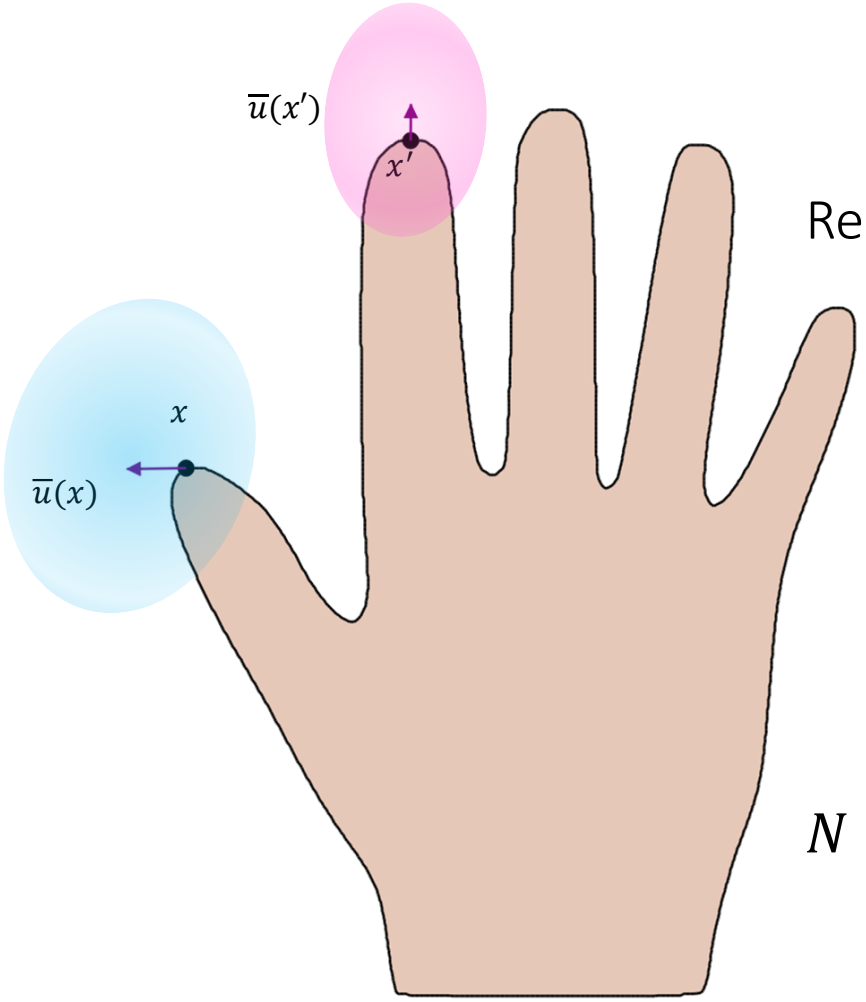


Restriction to values at points $X = \{x\}$

$$u(x) = \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix} \sim N(\mu_X, k_{XX})$$

$$= N\left(\begin{pmatrix} \mu_1(x) \\ \mu_2(x) \end{pmatrix}, \begin{pmatrix} k_{11}(x,x) & k_{12}(x,x) \\ k_{21}(x,x) & k_{22}(x,x) \end{pmatrix}\right)$$

Gaussian process: Illustration



Restriction to values at points $X = \{x, x'\}$

$$\begin{pmatrix} u(x) \\ u(x') \end{pmatrix} = \begin{pmatrix} u_1(x) \\ u_2(x) \\ u_1(x') \\ u_2(x') \end{pmatrix} \sim N(\mu_X, k_{XX}) =$$

$$N\left(\begin{pmatrix} \mu_1(x) \\ \mu_2(x) \\ \mu_1(x') \\ \mu_2(x') \end{pmatrix}, \begin{pmatrix} \mathrm{k}_{11}(x,x) & \mathrm{k}_{12}(x,x) & \mathrm{k}_{11}(x,x') & \mathrm{k}_{12}(x,x') \\ \mathrm{k}_{21}(x,x) & \mathrm{k}_{22}(x,x) & \mathrm{k}_{21}(x,x') & \mathrm{k}_{22}(x,x') \\ \mathrm{k}_{11}(x',x) & \mathrm{k}_{12}(x',x) & \mathrm{k}_{11}(x',x') & \mathrm{k}_{12}(x',x') \\ \mathrm{k}_{21}(x',x) & \mathrm{k}_{22}(x',x) & \mathrm{k}_{21}(x',x') & \mathrm{k}_{22}(x',x') \end{pmatrix}\right)$$

Defining a Gaussian process

A Gaussian process

$$GP(\mu, k)$$

is completely specified by a mean function μ and covariance function (or kernel) k.

- $\mu: \mathcal{X} \to \mathbb{R}^d$ defines how the average deformation looks like
- $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$ defines how it can deviate from the mean
  - Must be positive semi-definite

Marginalization property

Let $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_m)$

$$\mathrm{p}(X, Y) = N\left(\begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}\right)$$

The marginal distribution $p(X) = \int p(X, Y) dY$ is given by

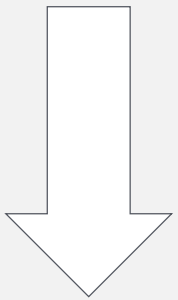
$$p(X) = N(\mu_X, \Sigma_{XX}).$$

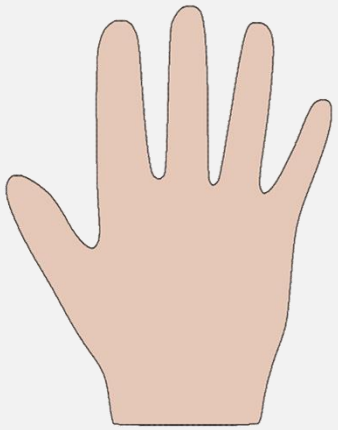
- Evaluating the Gaussian process $GP(\mu, k)$ defined on domain $\mathcal{X}$ at the points $X = (x_1, \dots, x_n)$ is marginalizing out (ignoring) all random variables $\mathcal{X} \setminus X$

From continuous to discrete

Conceptual formulation:

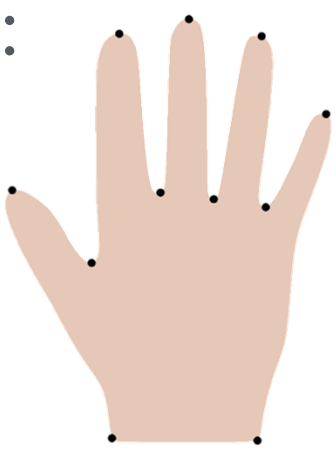
Continuous: $GP(\mu, k)$

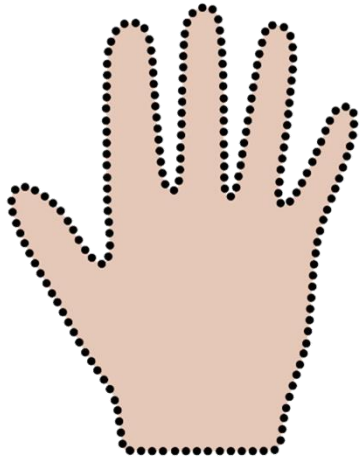


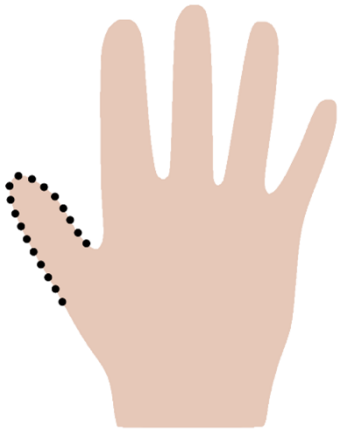


Practical implementation:

Discrete: $N(\mu, K)$







The Karhunen-Loève expansion

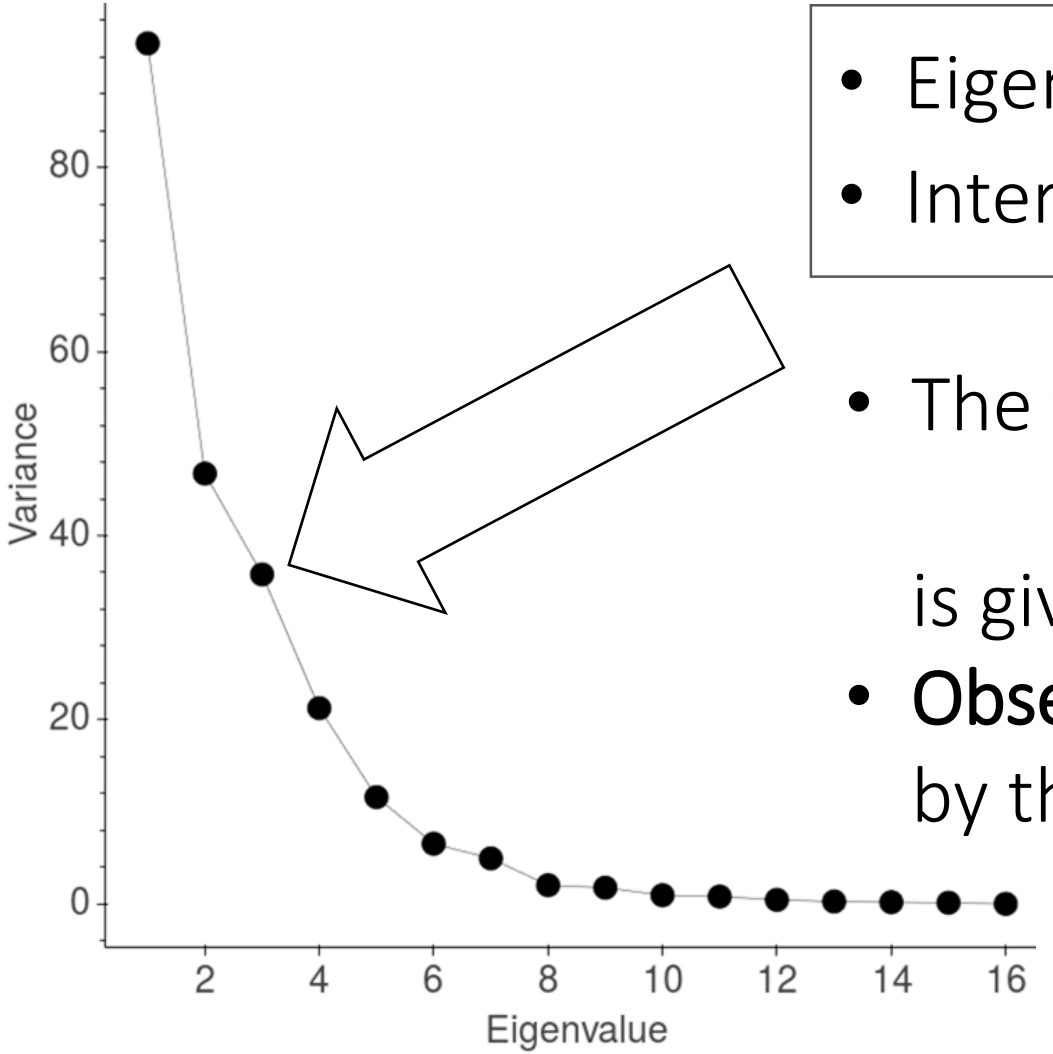
We can write $\mathrm{u} \sim GP(\mu, k)$

as $u \sim \mu + \sum_{i=1}^{\infty} \alpha_i \sqrt{\lambda_i}\, \phi_i, \;\; \alpha_i \sim N(0,1)$

- ϕ_i is the eigenfunction with associated eigenvalue λ_i of the linear operator

$$[T_k u](x) = \int k(x,s)u(s)ds$$

Eigenvalues and variance



- Eigenvalue λ_i
- Interpretation: Variance of $\alpha_i \sqrt{\lambda_i} \phi_i$

- The total variance of the process
$$u \sim GP(\mu, k)$$
is given by $\sum_{i=1}^{\infty} \lambda_i$.
- **Observation**: Most variance is explained by the first eigenfunctions

Low-rank approximation

$$u = \mu + \sum_{i=1}^{r} \alpha_i \sqrt{\lambda_i}\, \phi_i, \qquad \alpha_i \sim N(0,1)$$

Main idea: Represent process using only the first r components

- We have a finite, parametric representation of the process.
- Any deformation u is determined by the coefficients $\alpha = (\alpha_1, \dots, \alpha_r)$

$$p(u) = p(\alpha) = \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi}} \exp(-\alpha_i^2/2)$$

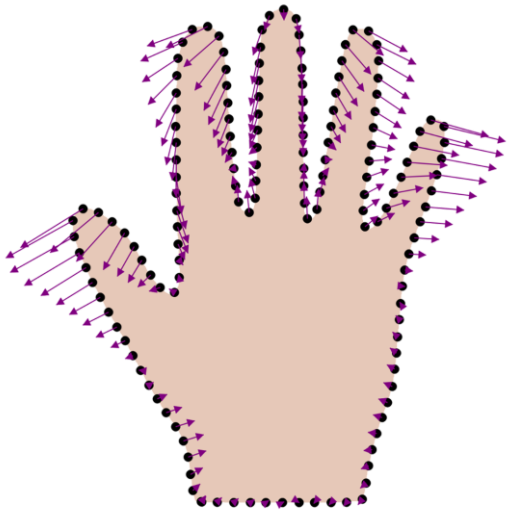
Vector-valued and single valued Gaussian processes

Scalar-valued Gaussian processes

Vector-valued (this course)

- Samples u are deformation fields:

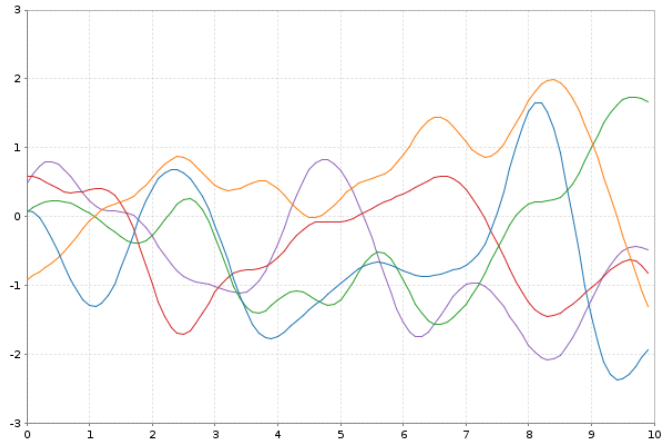
$$u: \mathbb{R}^n \to \mathbb{R}^d$$



Scalar-valued (more common)

- Samples f are real-valued functions

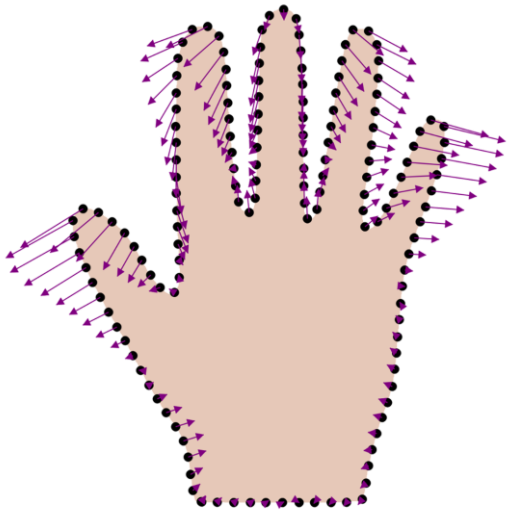
$$f : \mathbb{R}^n \to \mathbb{R}$$



Scalar-valued Gaussian processes

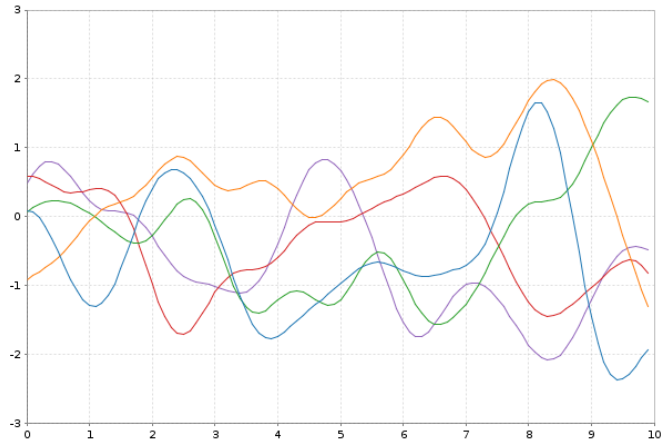
Vector-valued (this course)

$$u \sim GP(\vec{\mu}, \boldsymbol{k})$$
$$\vec{\mu}: \mathcal{X} \to \mathbb{R}^d$$
$$\boldsymbol{k}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$$



Scalar-valued (more common)

$$f \sim GP(\mu, k)$$
$$\mu: \mathcal{X} \to \mathbb{R}$$
$$k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$



A connection

Matrix-valued kernels can be reinterpreted as scalar-valued kernels:

Matrix valued kernel: $\boldsymbol{k}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\boldsymbol{d}\times\boldsymbol{d}}$

Scalar valued kernel: $k\colon \mathcal{X} \times (1..d) \times \mathcal{X} \times (1..d) \to \mathbb{R}$

Bijection: Define

$$k((x,i),\big((x',j)\big) = \boldsymbol{k}(x',x')_{i,j}$$

GP Regression – Vector-valued case

$$\boldsymbol{K} = \begin{pmatrix} k_{11}(x_1, x_1) & k_{12}(x_1, x_1) & \cdots & k_{11}(x_1, x_n) & k_{12}(x_1, x_n) \\ k_{21}(x_1, x_1) & k_{22}(x_1, x_1) & & k_{21}(x_1, x_n) & k_{22}(x_1, x_n) \\ \vdots & & & & \vdots \\ k_{11}(x_n, x_1) & k_{12}(x_n, x_1) & & k_{11}(x_n, x_n) & k_{12}(x_n, x_n) \\ k_{21}(x_n, x_1) & k_{22}(x_n, x_1) & \cdots & k_{21}(x_n, x_n) & k_{22}(x_n, x_n) \end{pmatrix}$$

$$\Updownarrow$$

$$K = \begin{pmatrix} k\big((x_1,1),(x_1,1)\big) & k\big((x_1,1),(x_1,2)\big) & \cdots & k\big((x_1,1),(x_n,1)\big) & k\big((x_1,1),(x_n,2)\big) \\ k\big((x_1,2),(x_1,1)\big) & k\big((x_1,2),(x_1,2)\big) & & k\big((x_1,2),(x_n,1)\big) & k\big((x_1,2),(x_n,2)\big) \\ \vdots & & & & \vdots \\ k\big((x_n,1),(x_1,1)\big) & k\big((x_n,1),(x_1,2)\big) & & k\big((x_n,1),(x_n,1)\big) & k\big((x_n,1),(x_n,2)\big) \\ k\big((x_n,2),(x_1,1)\big) & k\big((x_n,2),(x_1,2)\big) & \cdots & k\big((x_n,2),(x_n,1)\big) & k\big((x_n,2),(x_n,2)\big) \end{pmatrix}$$

A connection

Matrix-valued kernels can be reinterpreted as scalar-valued kernels:

Matrix valued kernel: $\boldsymbol{k}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\boldsymbol{d}\times\boldsymbol{d}}$

Scalar valued kernel: $k: \mathcal{X} \times (1..d) \times \mathcal{X} \times (1..d) \to \mathbb{R}$

Bijection: Define

$$k((x,i),((x',j)) = \boldsymbol{k}(x',x')_{i,j}$$

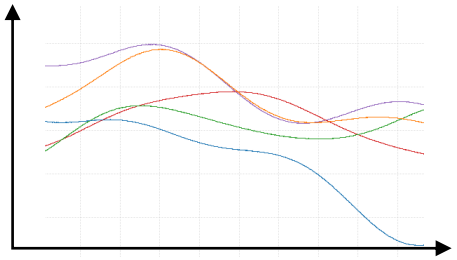
All the theory developed for the scalar-valued GPs holds also for vector-valued GPs!

Sampling revisited

Finite views on infinite objects

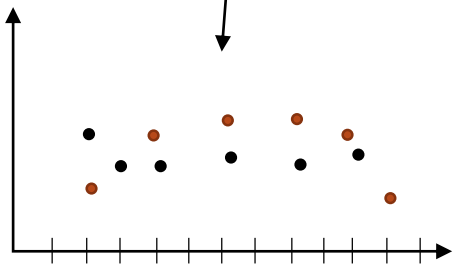
Infinite dimensional

Gaussian process

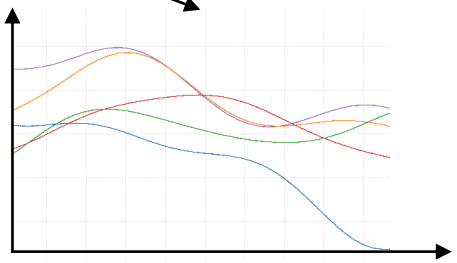


Continuous domain

Finite dimensional



Finite domain (Marginalization)



Finite rank (KL-Expansion)

The space of samples

Sampling from $GP(\mu, k)$ is done using the corresponding normal distribution $N(\vec{\mu}, \mathrm{K})$

Algorithm for sampling (slightly inefficient)

1. Do an SVD: $\mathrm{K} = UD^2U^T$
2. Draw a normal vector $\alpha \sim N(0, I_{n\times n})$
3. Compute $\vec{\mu} + UD\alpha$

The space of samples

- From $\mathrm{K} = UD^2U^T$ (using that $U^TU = I$) we have that
$$\mathrm{K}UD^{-1} = UD$$
- Any sample
$$s = \vec{\mu} + UD\alpha = \vec{\mu} + \mathrm{K}UD^{-1}\alpha = \mu + \mathrm{K}\beta$$
is a linear combinations of the columns of K.

Two ways to represent sample:

1. KL-Expansion: $s = \vec{\mu} + \sum_i d_i \alpha_i u_i$
2. Linear combination of kernels: $s = \vec{\mu} + \sum_j \beta k_j$

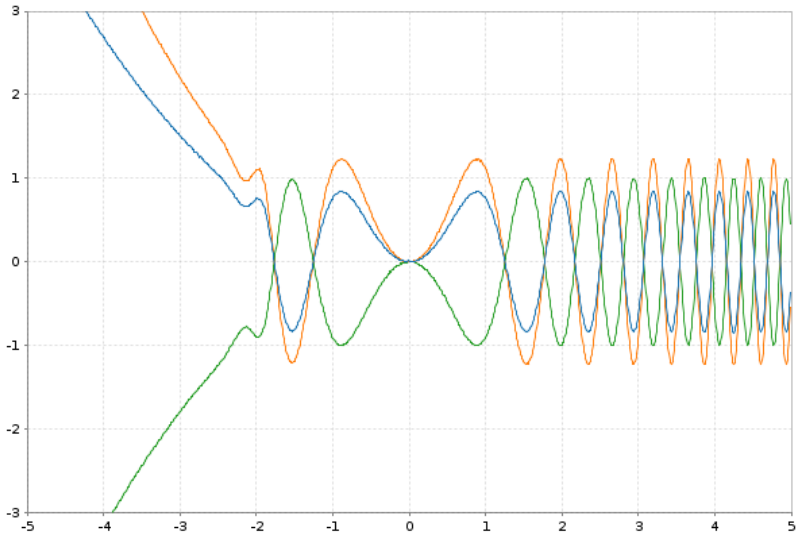
Four examples covariance functions

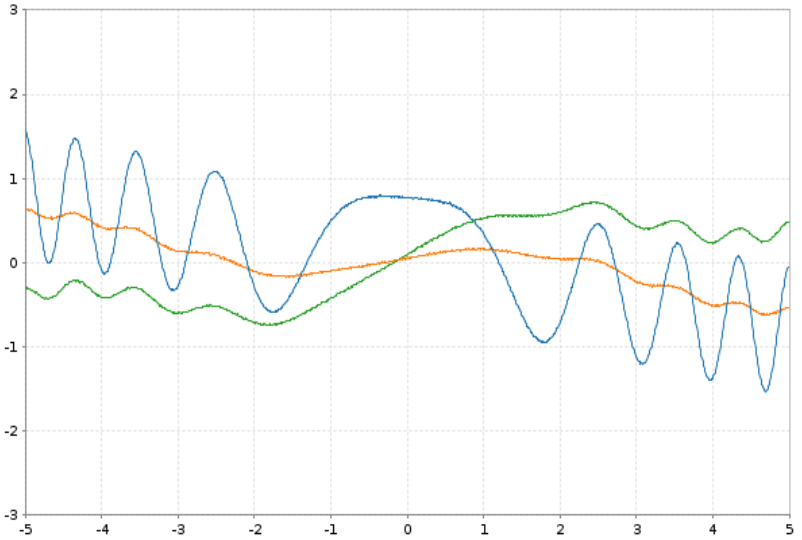
$$k(x, x') = f(x)f(x')$$

$$\mathrm{f(x)} = (1 - s(x))2x^2 + \ s(x)\sin(x^2)$$

$$k(x, x') = \sum_{i=1}^{3} f_i(x)f_i(x')$$

$$f_1(x) = \sin(x)\,, f_2(x) = x, f_3(x) = \cos(x^{\,2})$$

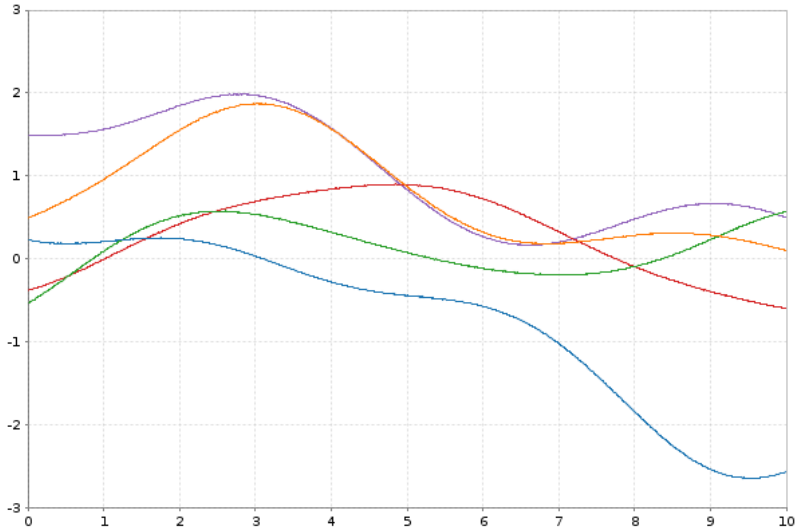


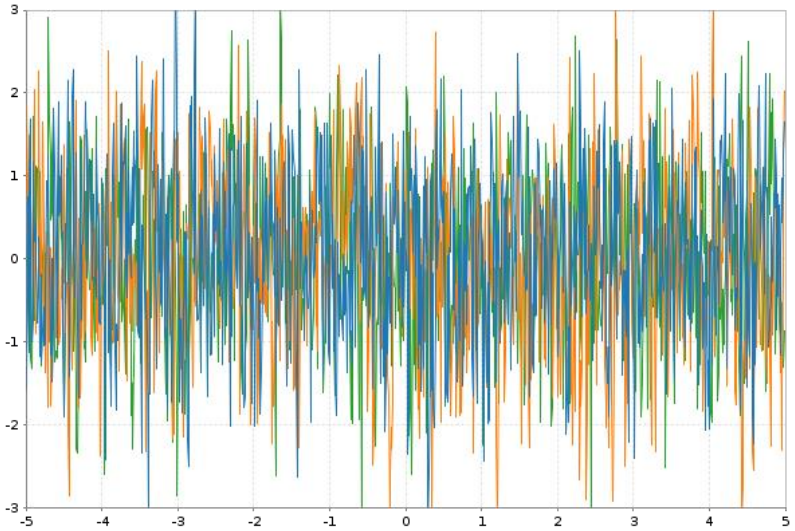


Four examples covariance functions

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{9}\right)$$

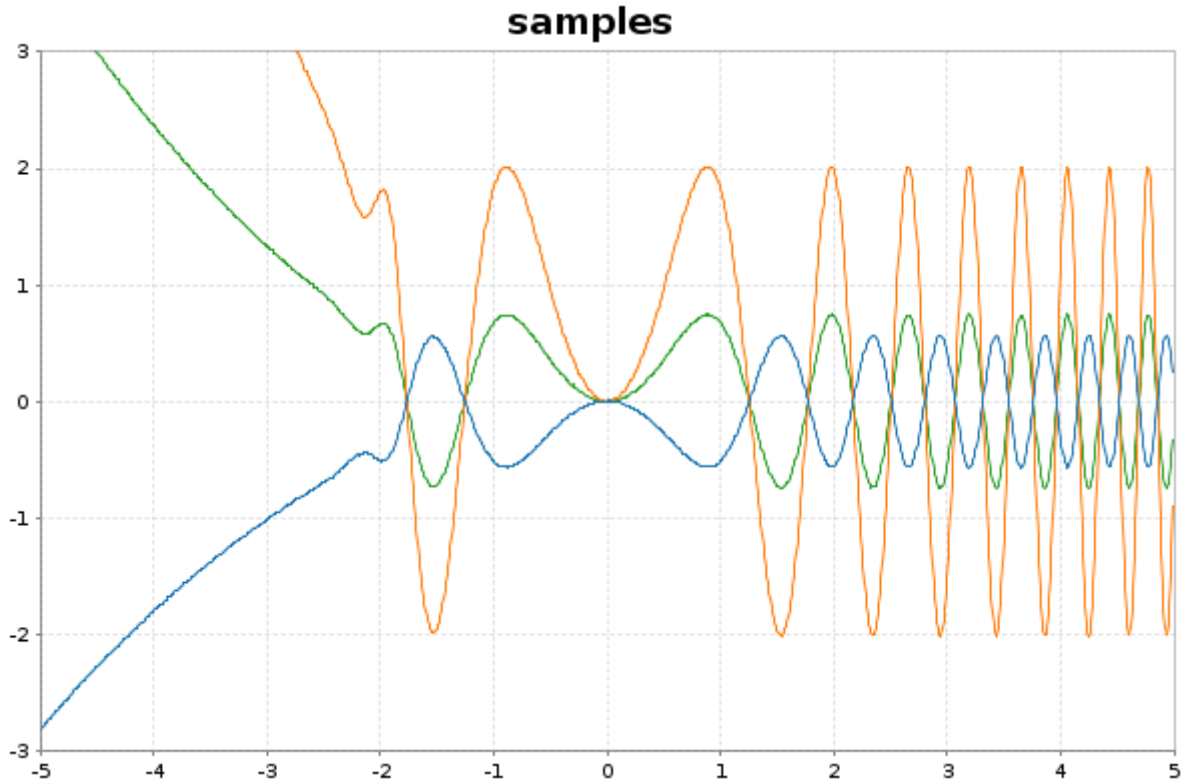
$$k(x, x') = \delta(x, x')$$





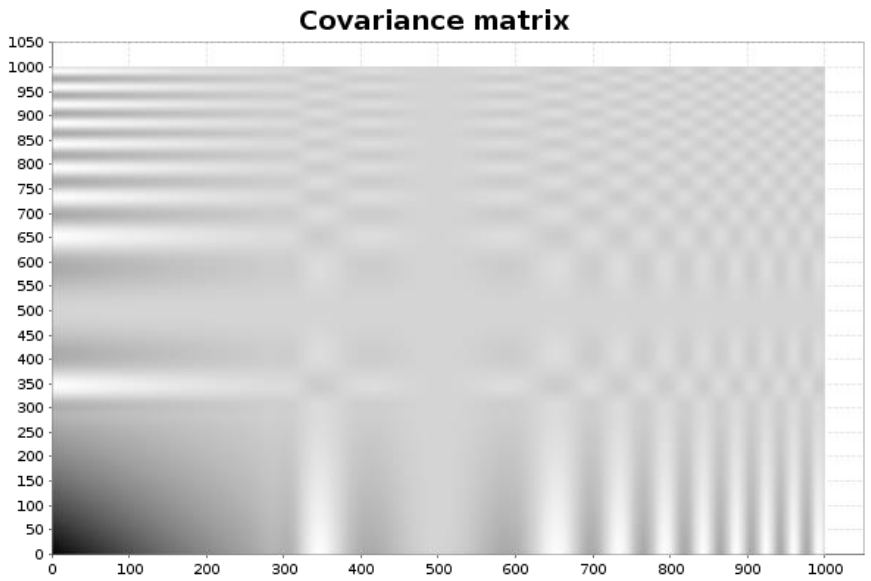
Example 1

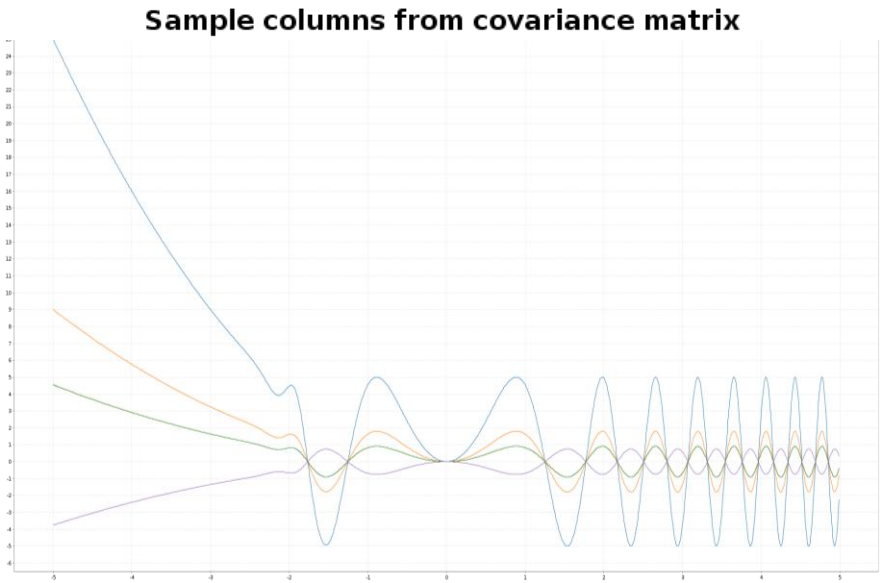
$$k(x, x') = f(x)f(x')$$



Example 1

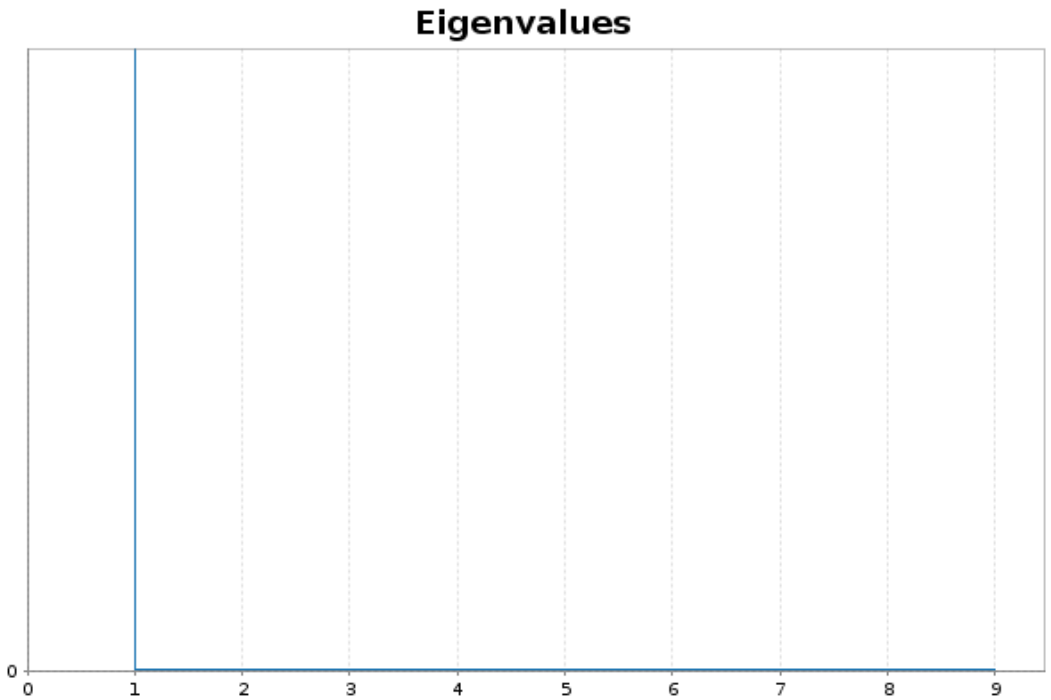
$$k(x, x') = f(x)f(x')$$

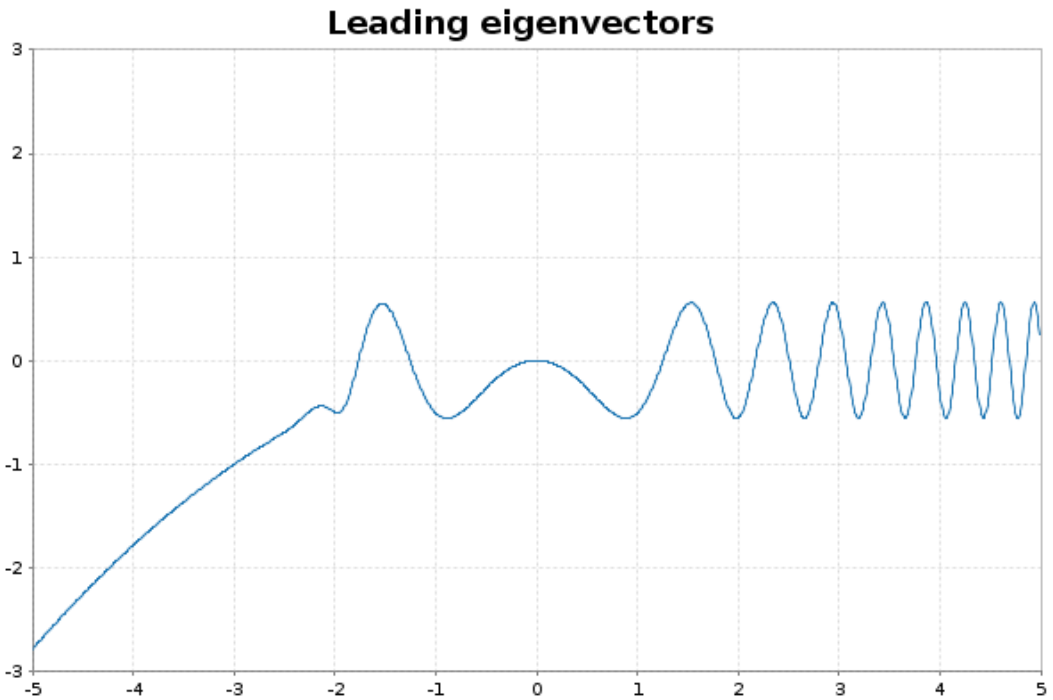




Example 1

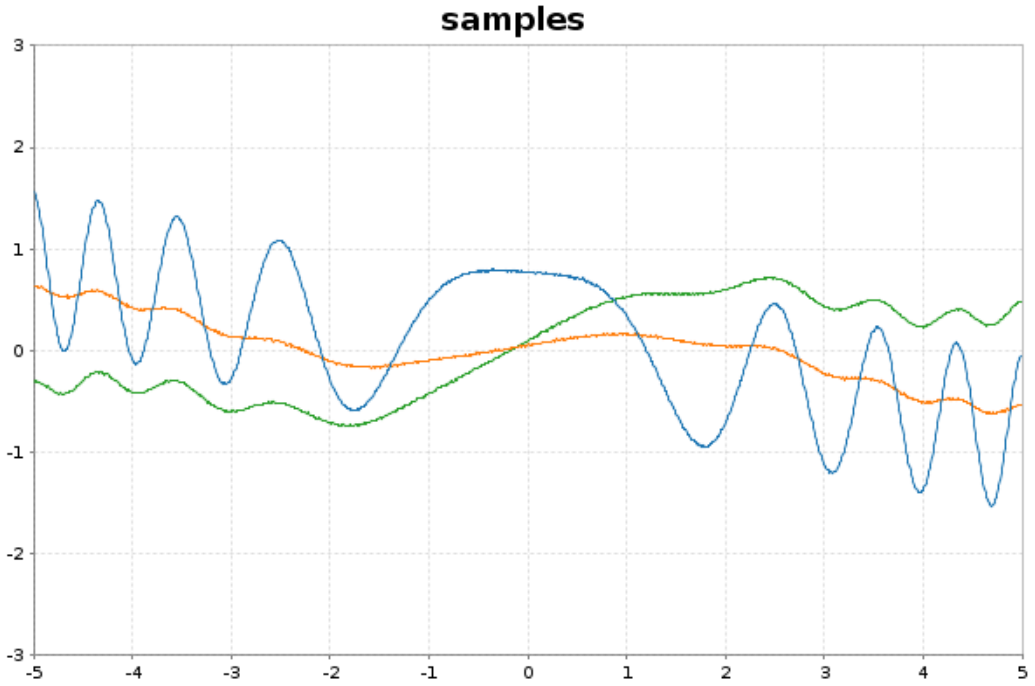
$$k(x, x') = f(x)f(x')$$





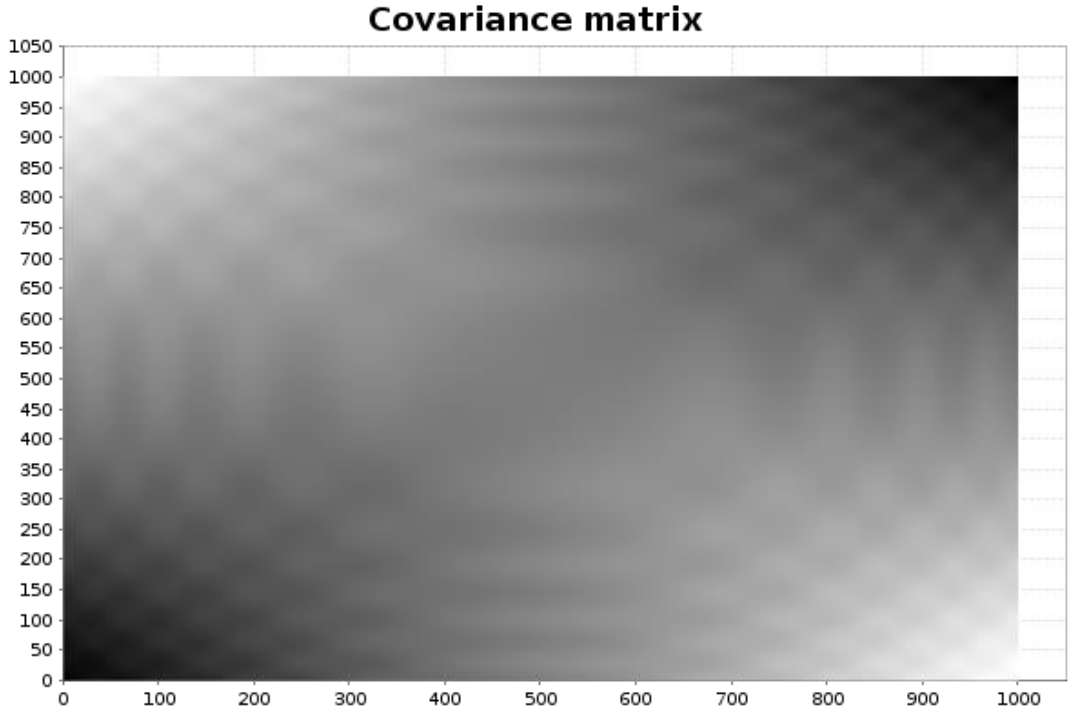
Example 2

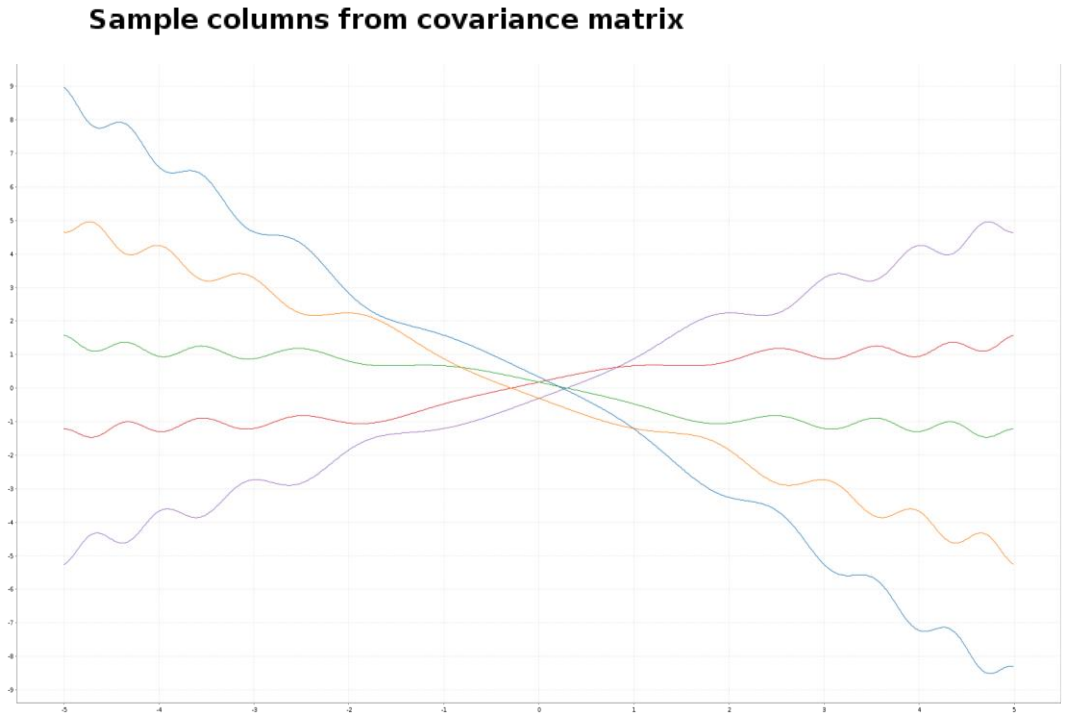
$$k(x, x') = \sum_{i=1}^{3} f_i(x) f_i(x')$$



Example 2

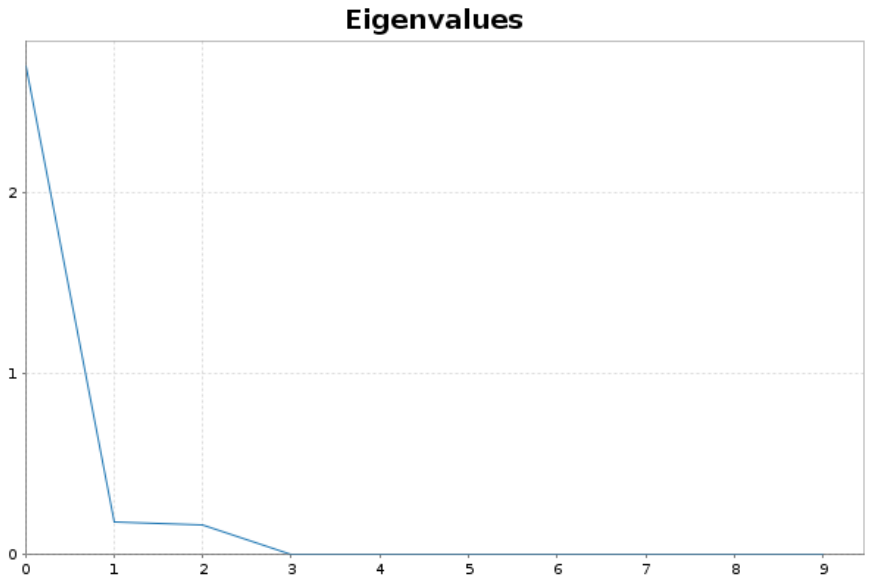
$$k(x, x') = \sum_{i=1}^{3} f_i(x) f_i(x')$$

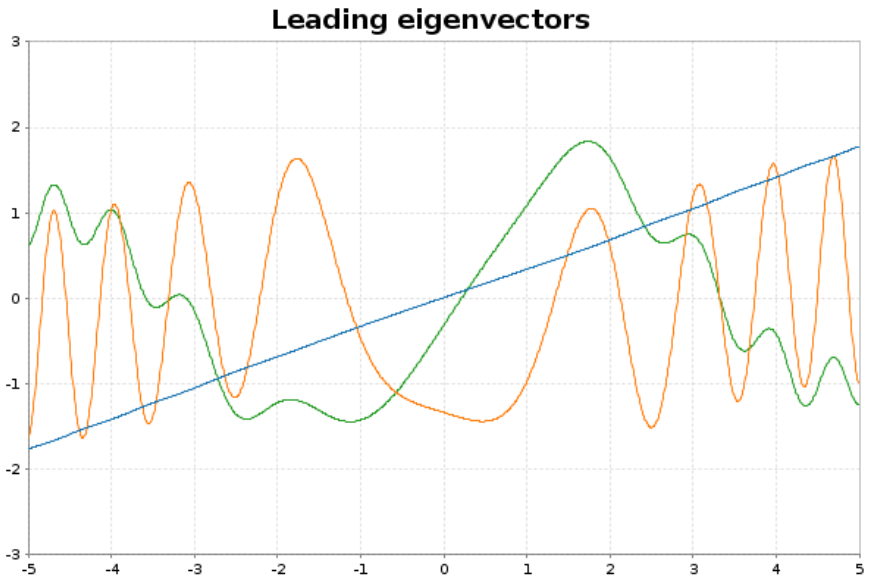




Example 2

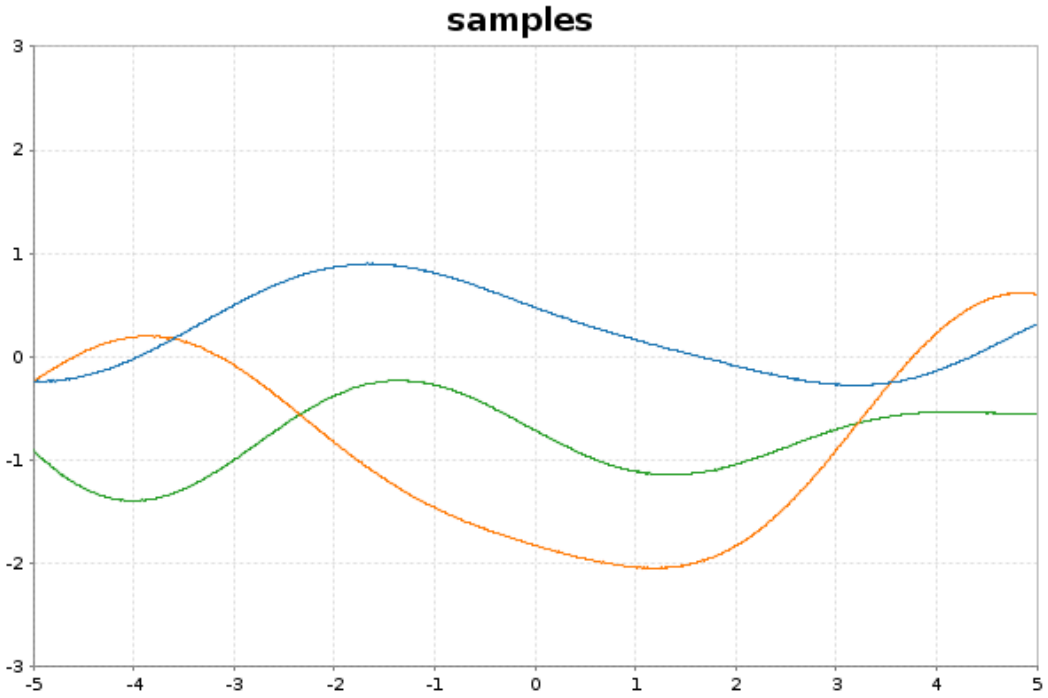
$$k(x, x') = \sum_{i=1}^{3} f_i(x) f_i(x')$$





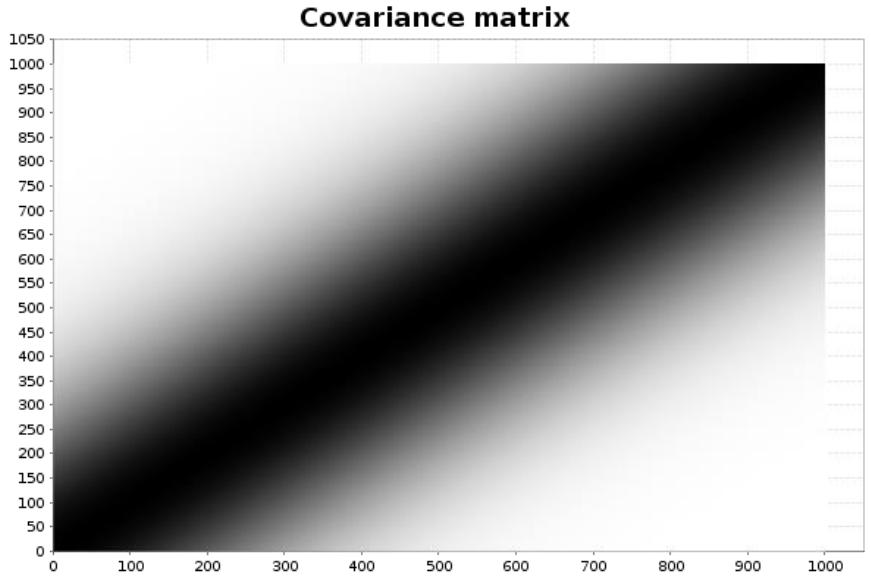
Example 3

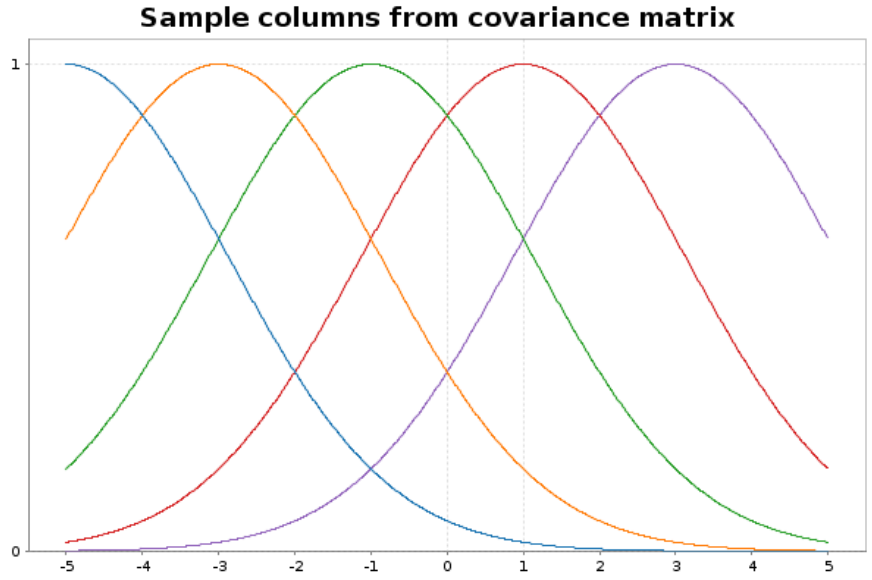
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{9}\right)$$



Example 3

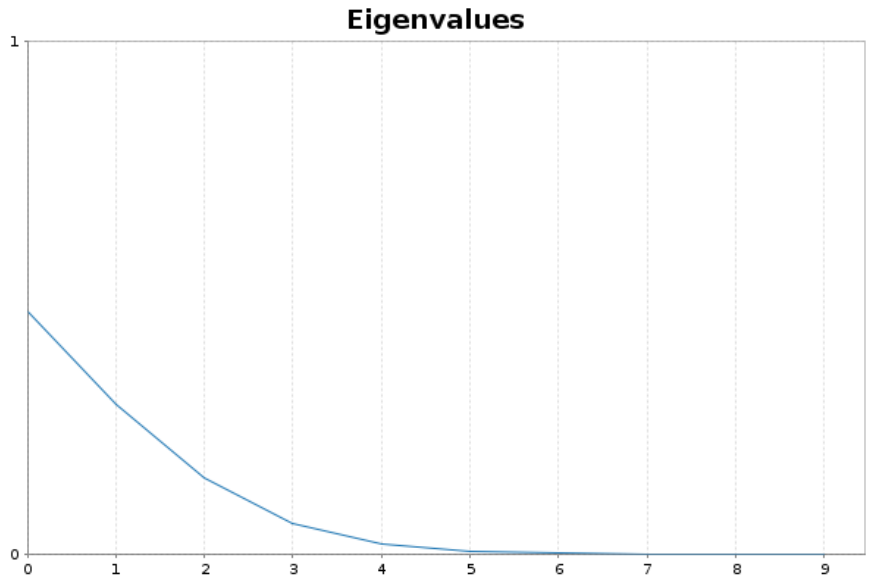
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{9}\right)$$

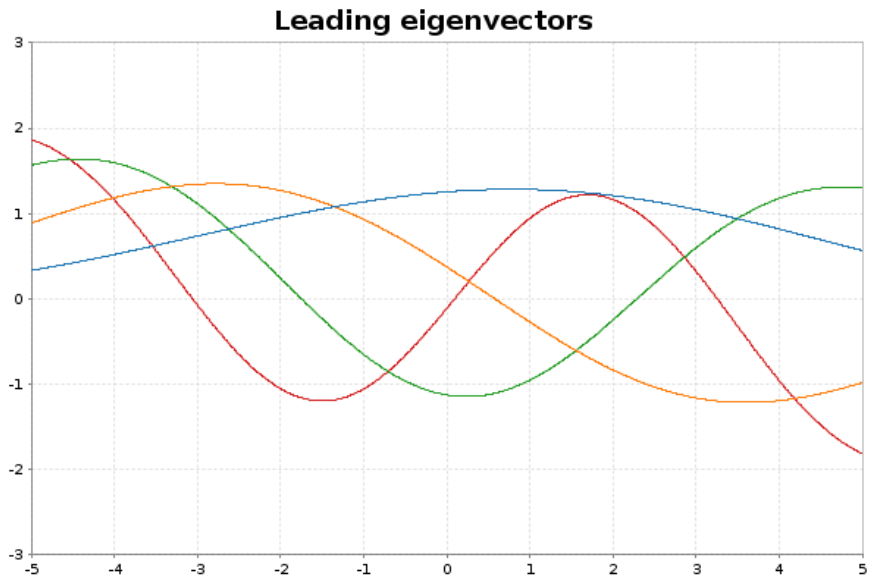




Example 3

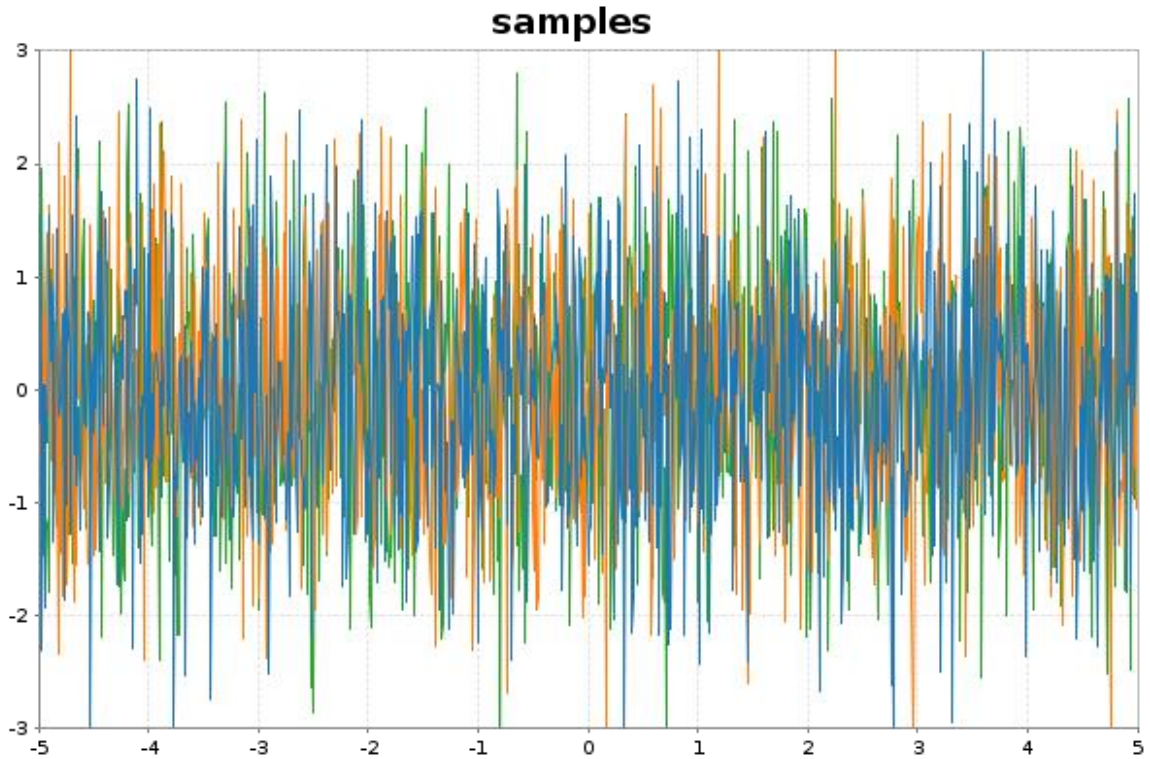
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{9}\right)$$





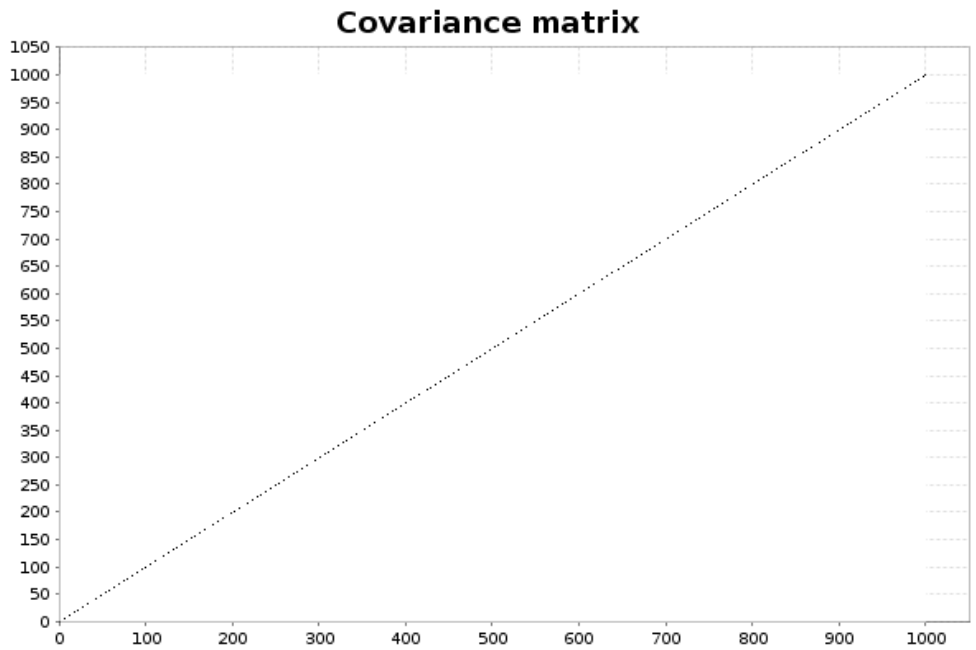
Example 4

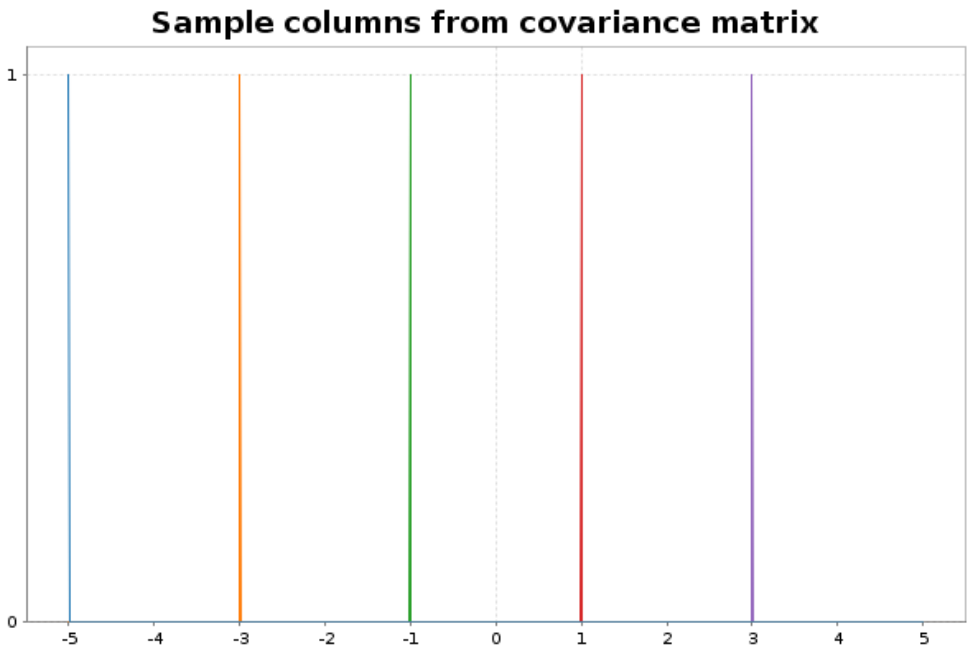
$$k(x, x') = \delta(x, x')$$



Example 4

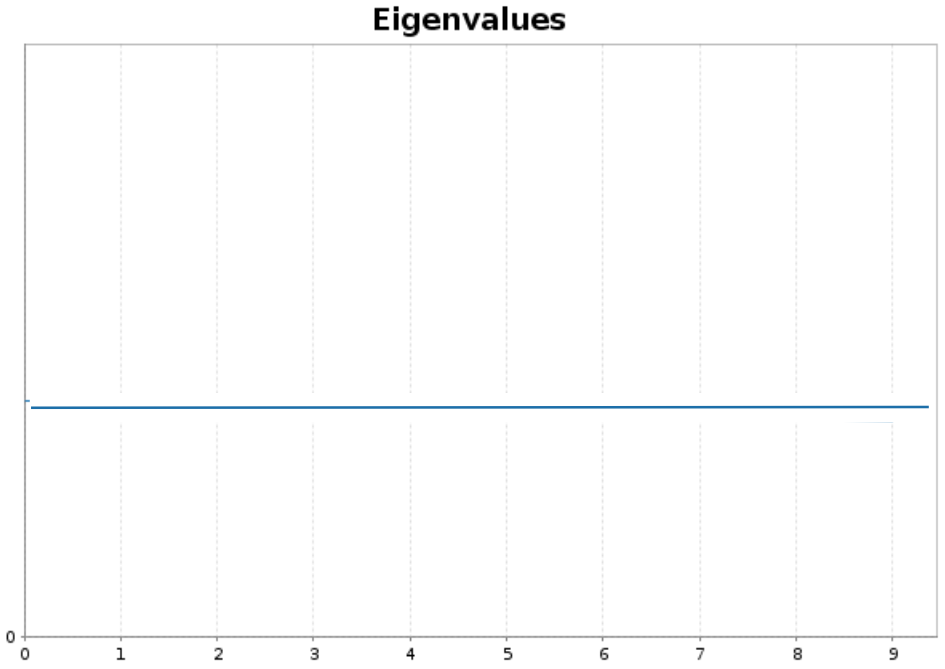
$$k(x, x') = \delta(x, x')$$

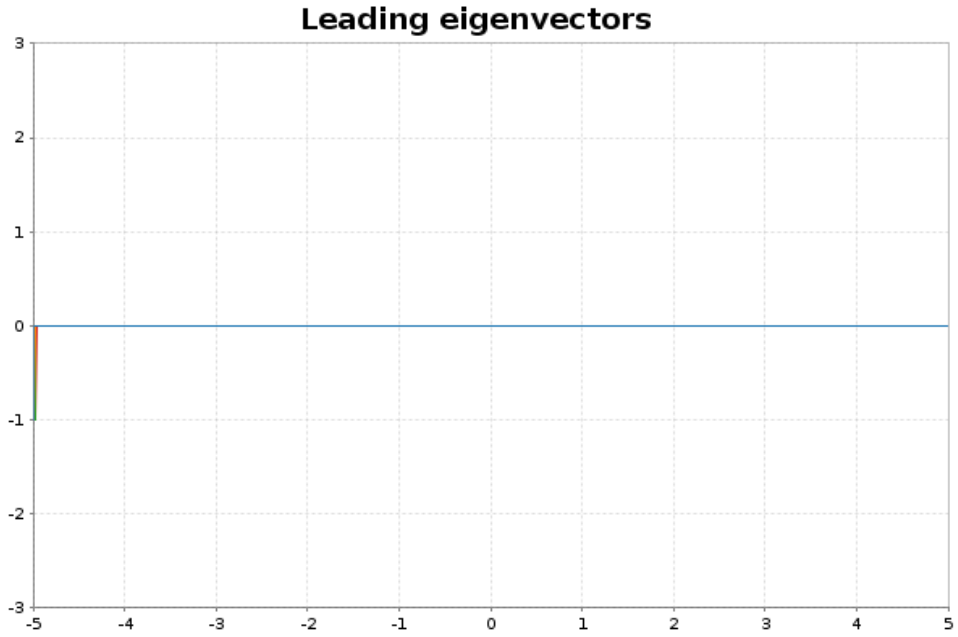




Example 4

$$k(x, x') = \delta(x, x')$$

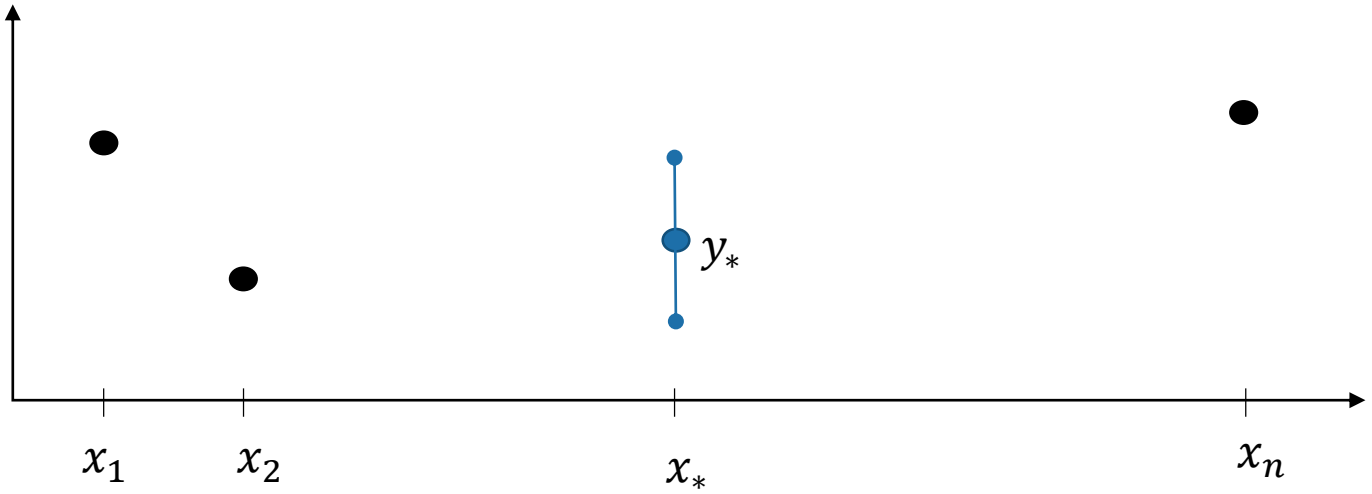




Gaussian process regression revisited

Gaussian process regression

- Given: observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$
- Model: $y_i = f(x_i) + \epsilon,\ f \sim GP(\mu, k)$
- Goal: compute $p(y_* | x_*, x_1, \dots, x_n, y_1, \dots, y_n)$



Gaussian process regression

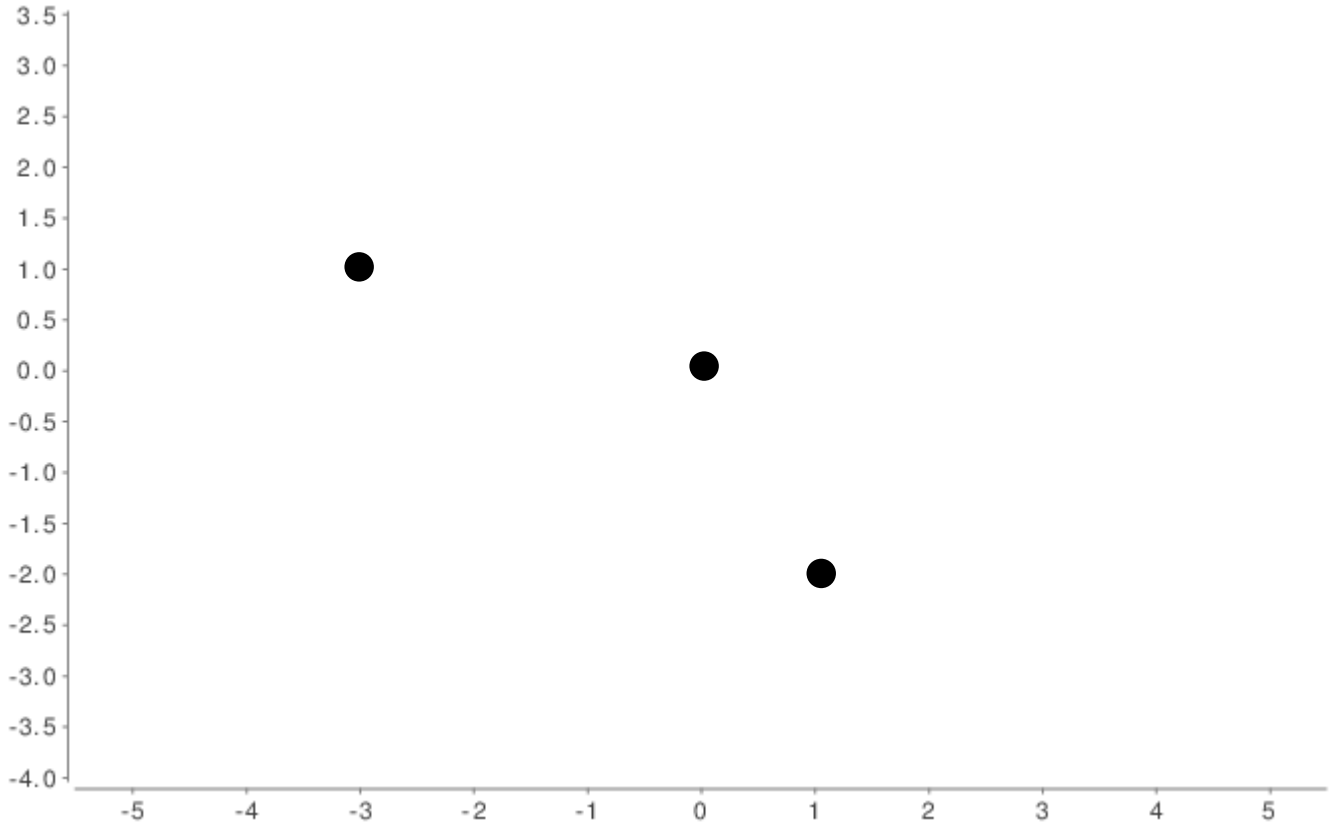
- Solution given by posterior process $GP(\mu_p, k_p)$ with

$$\mu_p(x_*) = K(x_*, X)[K(X, X) + \sigma^2 I]^{-1} y$$

$$k_p(x_*, x_*') = k(x_*, x_*') - K(x_*, X)[K(X, X) + \sigma^2 I]^{-1} K(X, x_*')$$

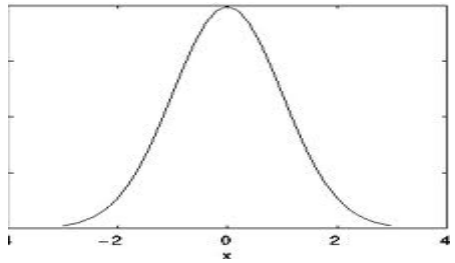
- The covariance is independent of the value at the training points
  - Structure of posterior GP determined solely by kernel.

Examples

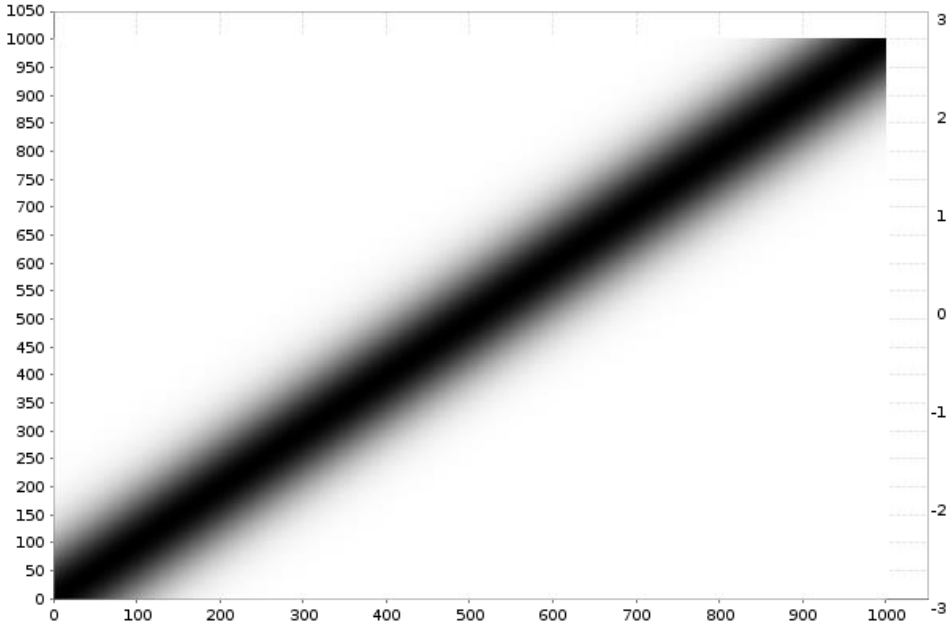


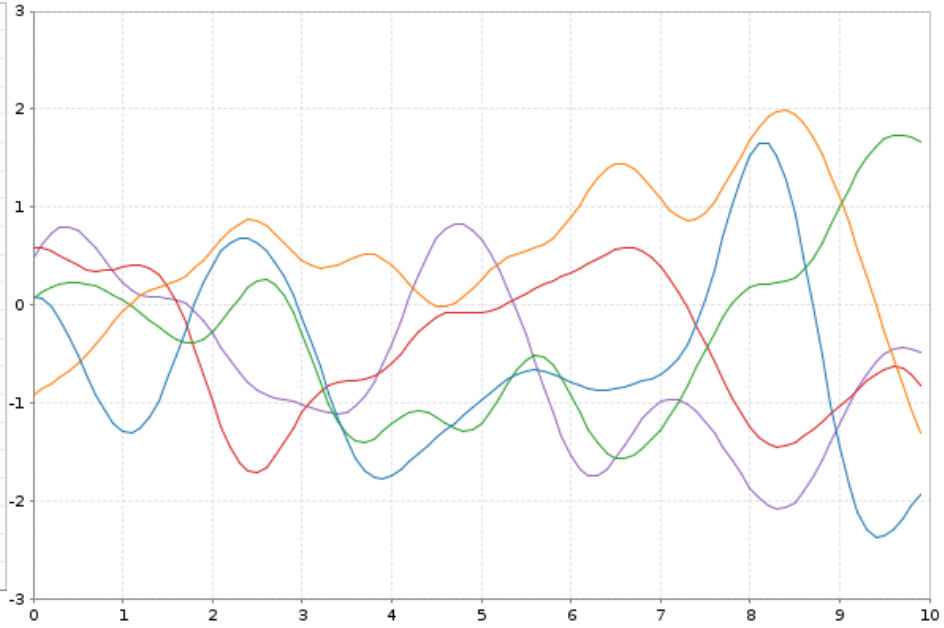
Example: Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right)$$



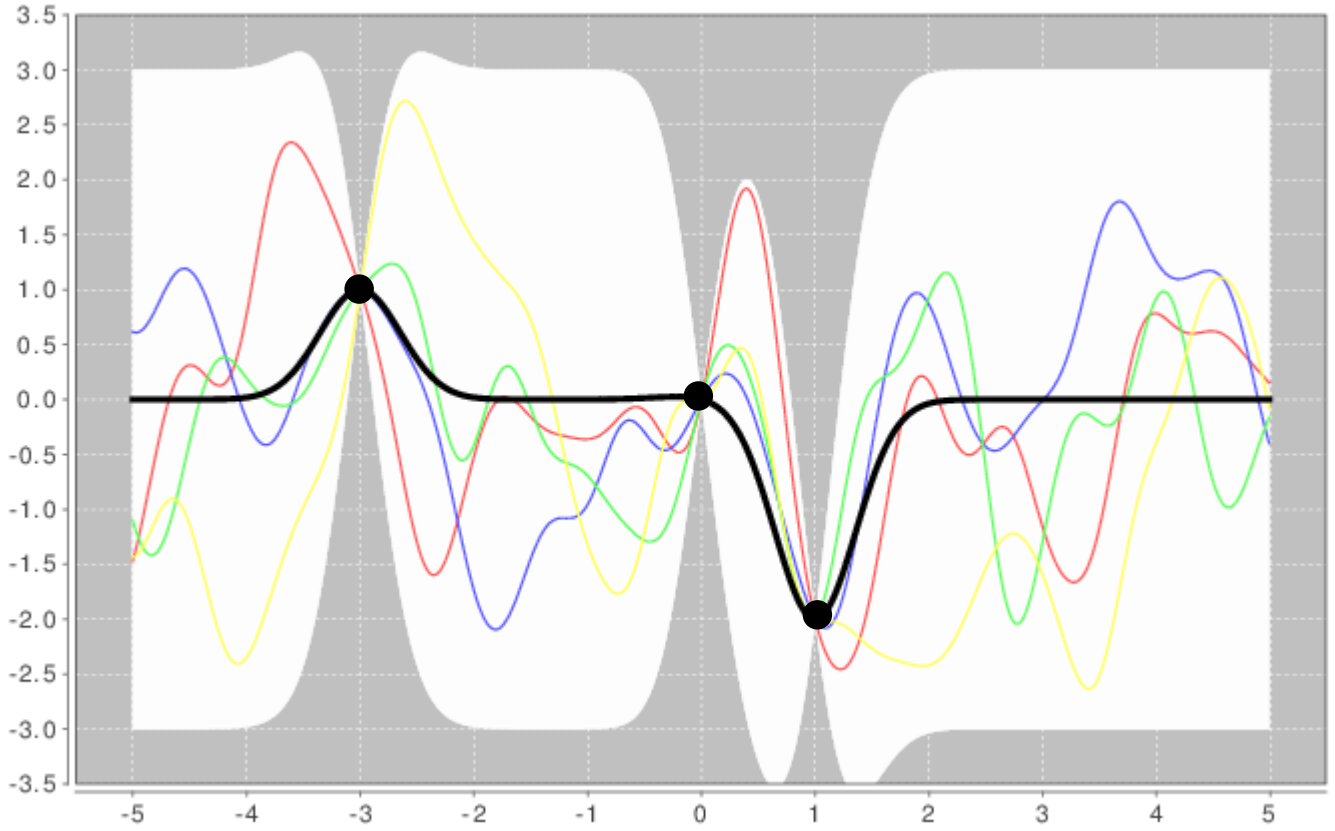
$$\sigma = 1$$





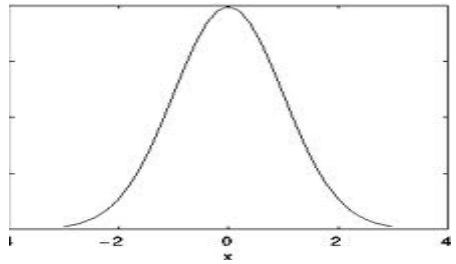
Examples

- Gaussian kernel ($\sigma = 1$)

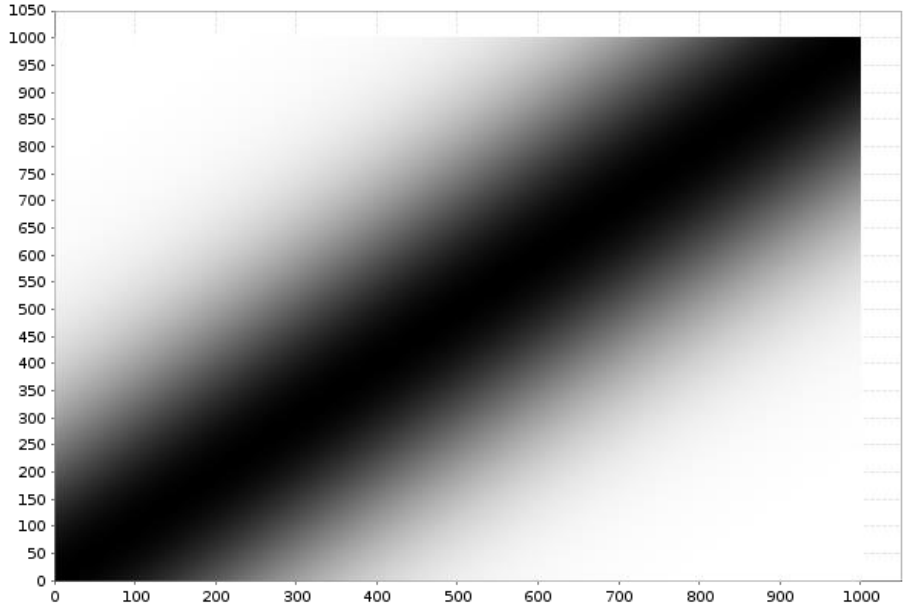


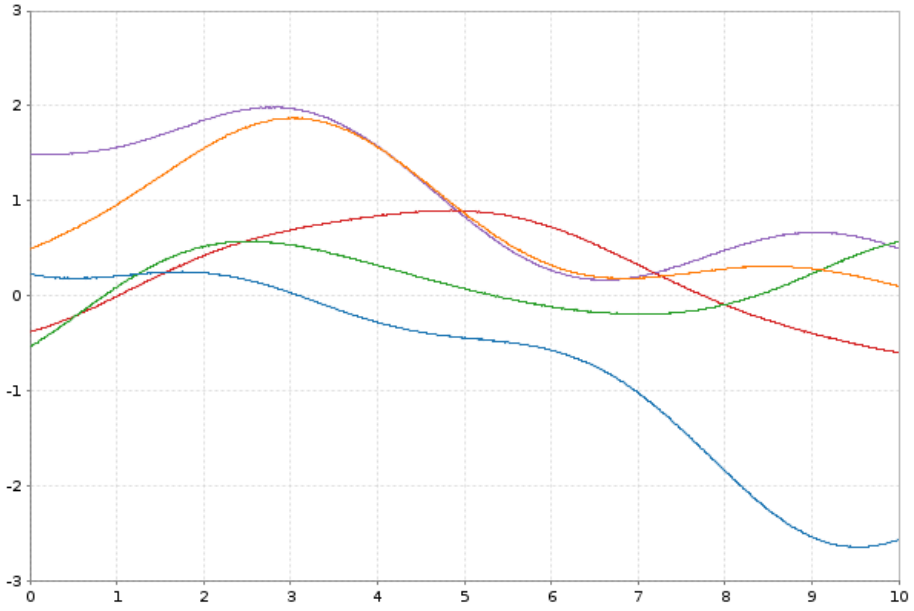
Example: Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right)$$



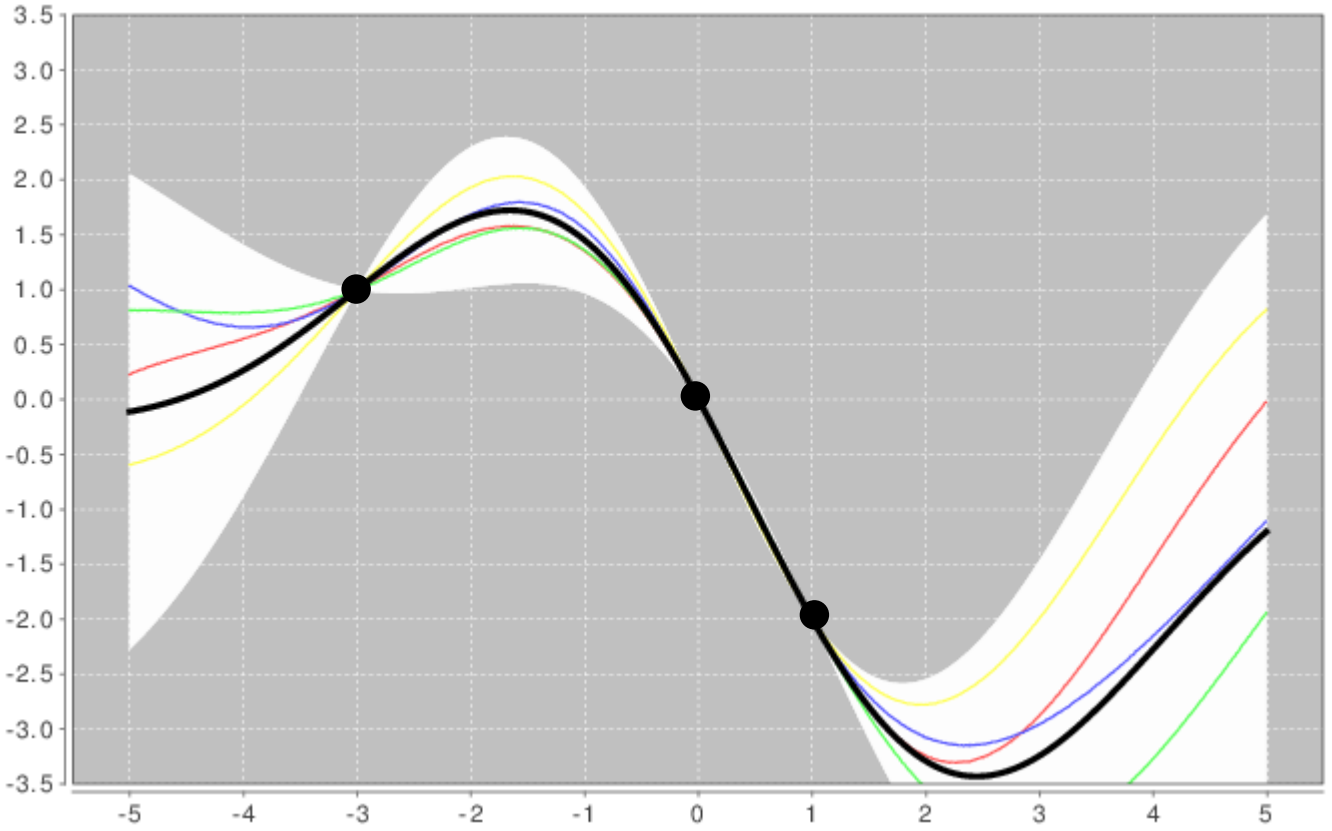
$$\sigma = 3$$





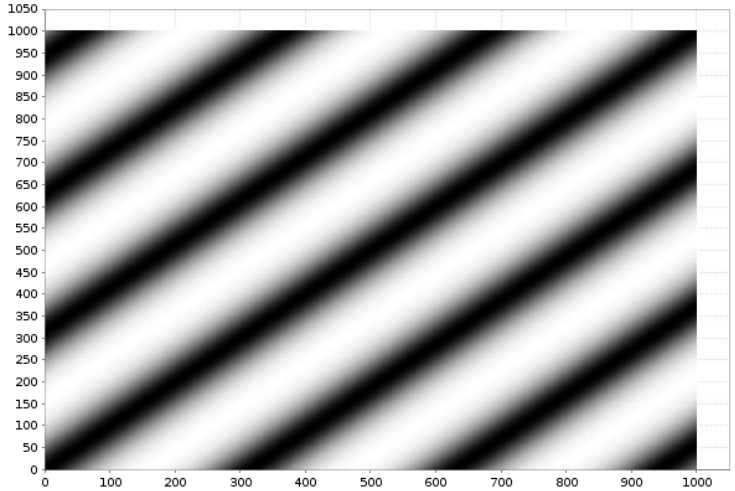
Examples

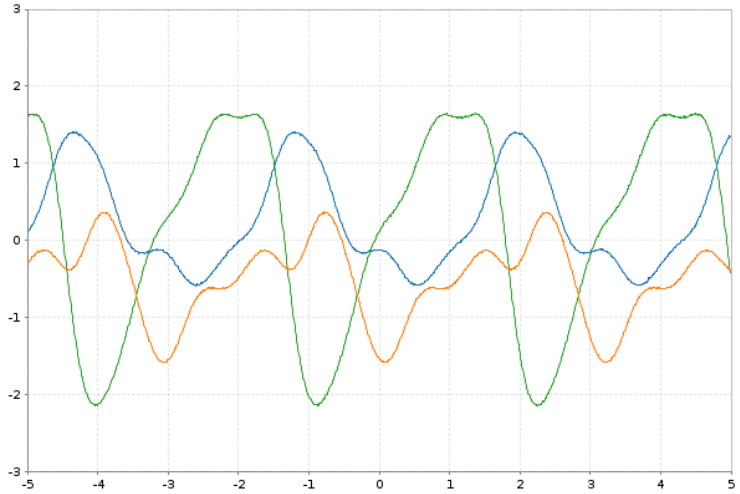
- Gaussian kernel ($\sigma = 5$)



Periodic kernels

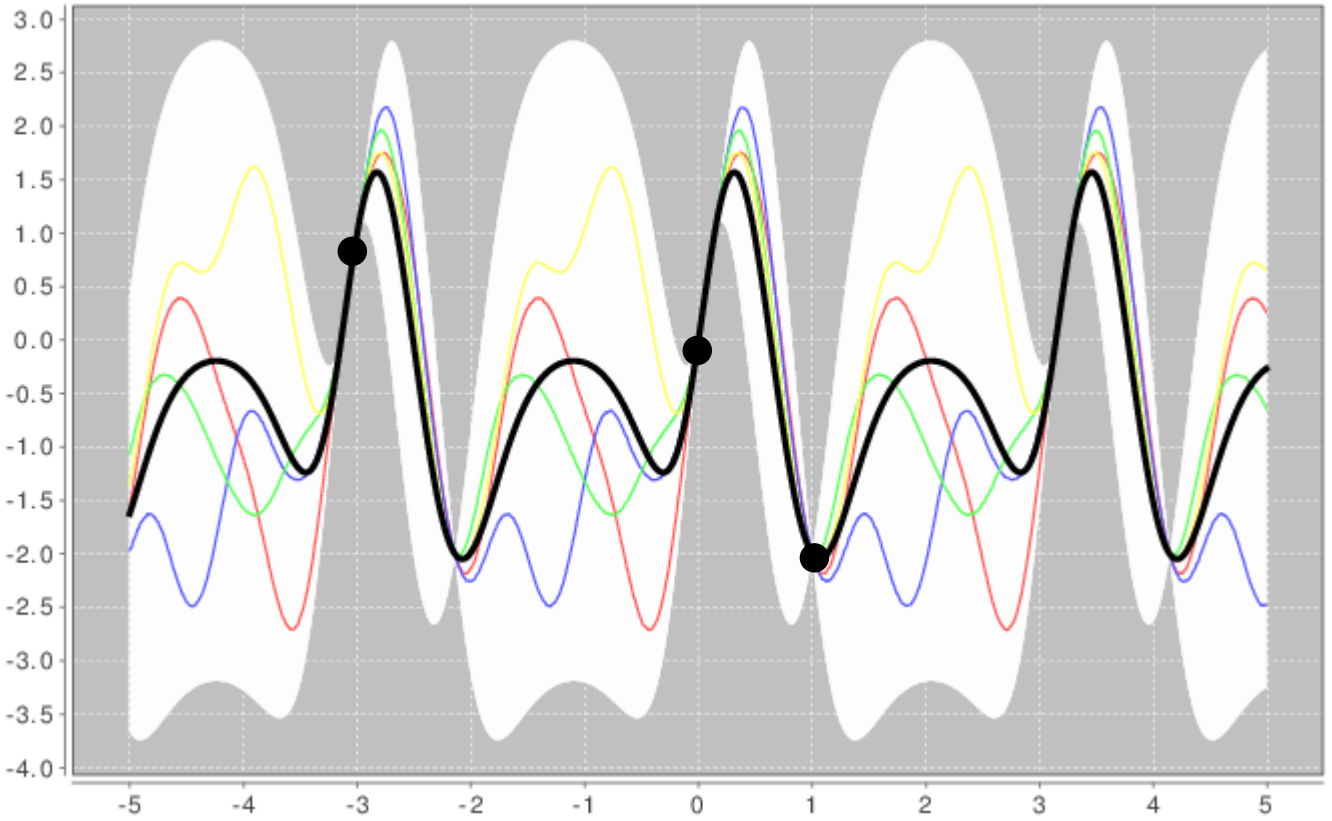
- Define $u(x) = \begin{pmatrix} \cos(x) \\ \sin(x) \end{pmatrix}$
- $k(x, x') = \exp(-\|(u(x) - u(x')\|^2 = \exp(-4 \sin^2\left(\frac{\|x - x'\|}{\sigma^2}\right))$





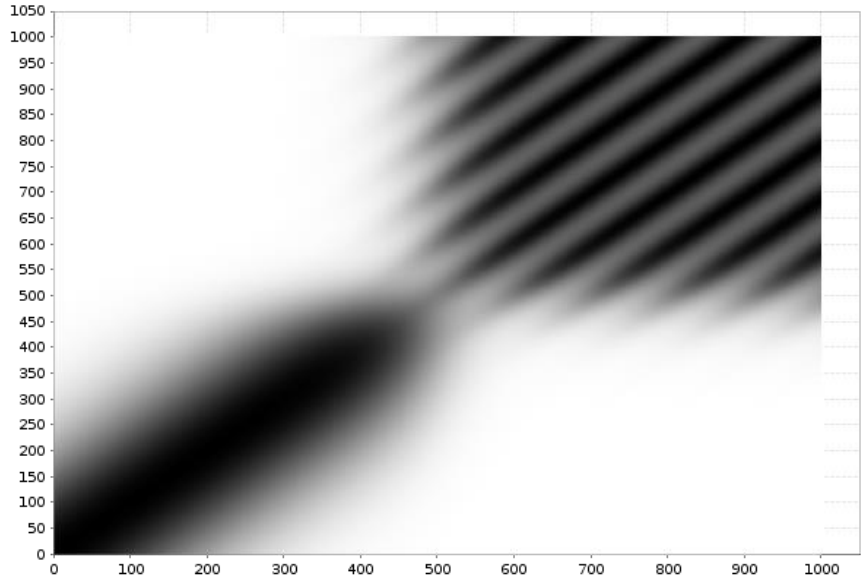
Examples

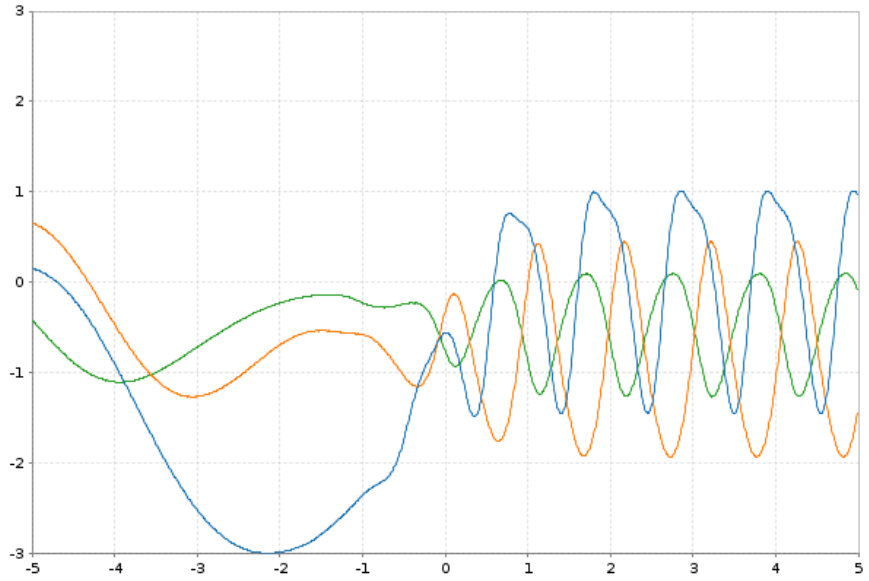
- Periodic kernel



Changepoint kernels

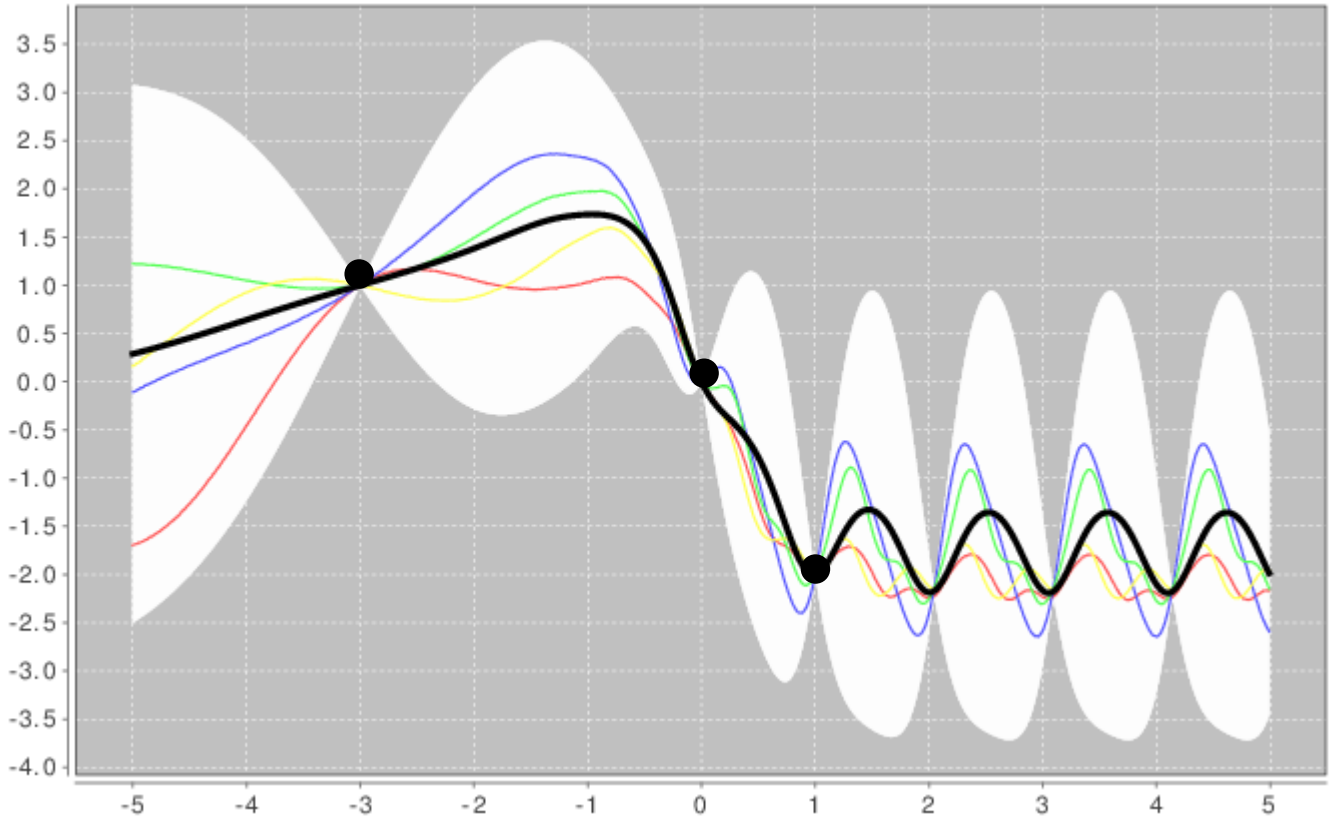
- $k(x, x') = s(x)k_1(x, x')s(x') + (1 - s(x))k_2(x, x')(1 - s(x'))$
- $\mathrm{s}(x) = \frac{1}{1+\exp(-x)}$





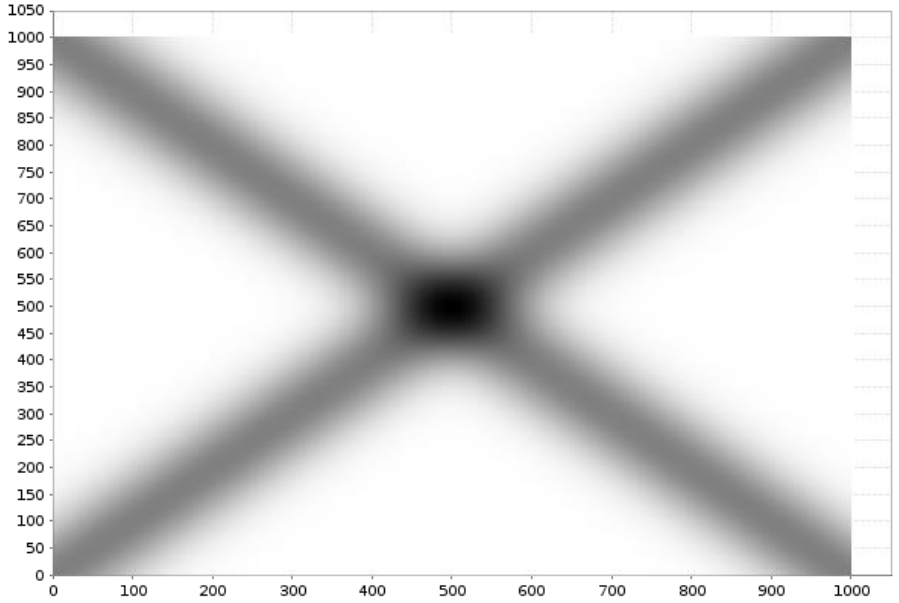
Examples

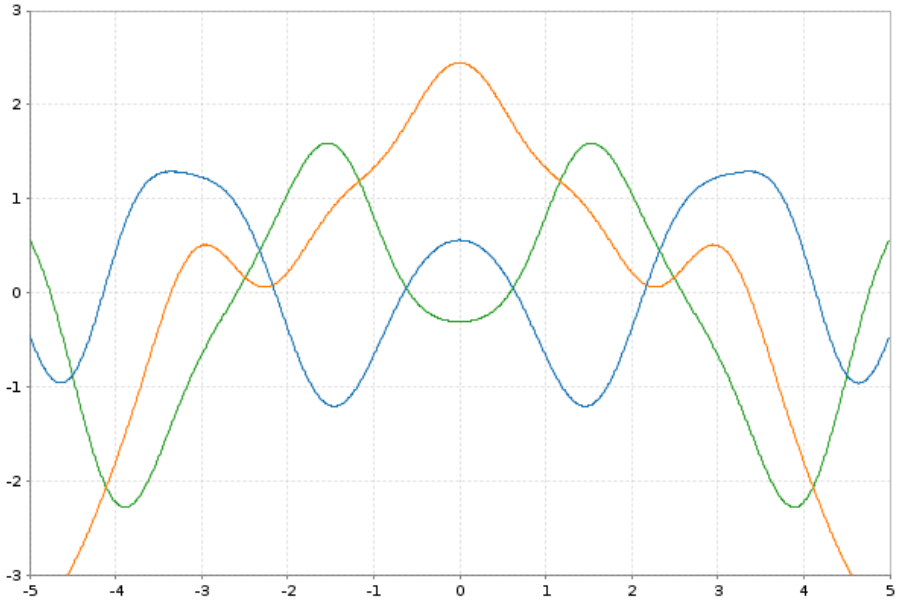
- Changepoint kernel



Symmetric kernels

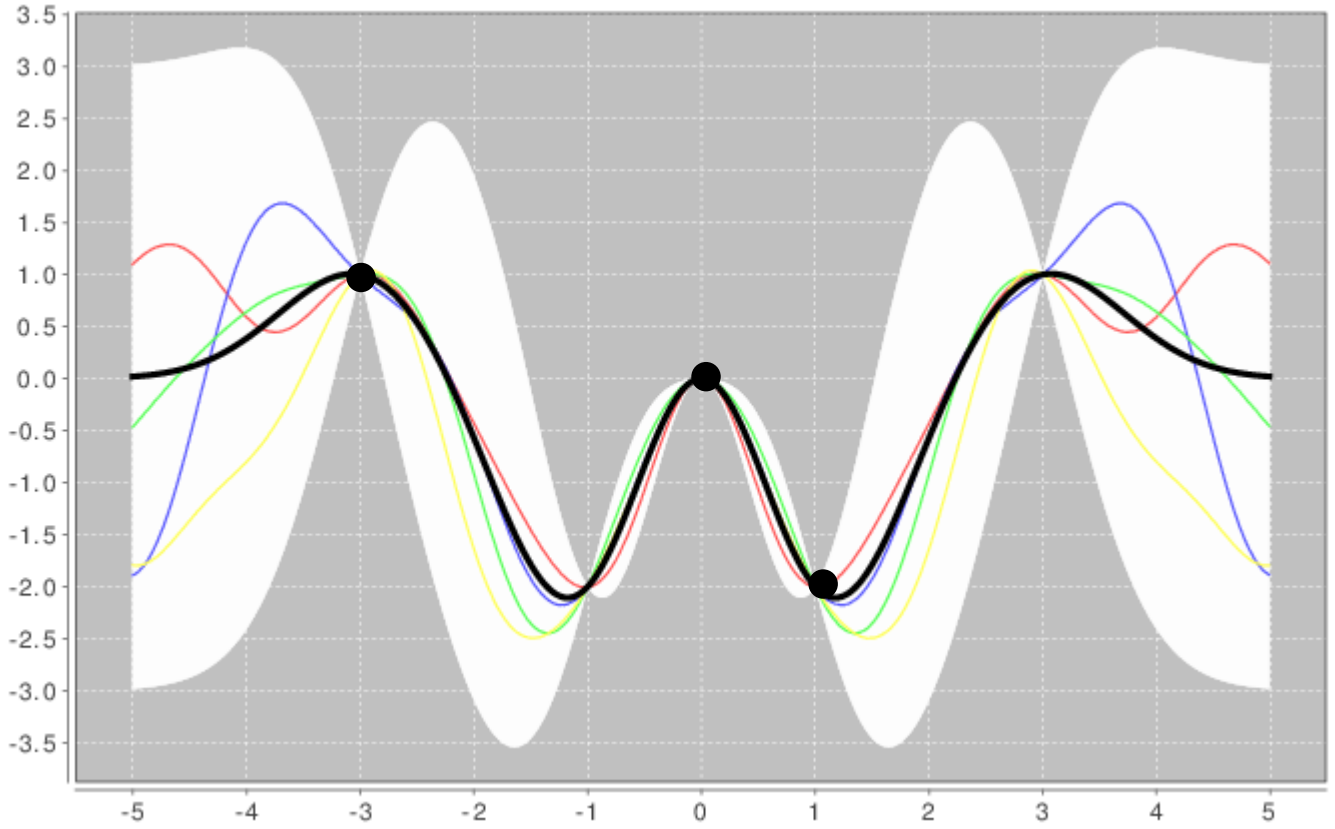
- *Enforce that f(x) = f(-x)*
- $k(x, x') = k(-x, x') + k(x, x')$





Examples

- Symmetric kernel



Summary

- Gaussian processes are an extremely rich toolbox for modelling functions / deformation fields

- Possible functions are linear combinations of the kernels $k(\cdot, x)$, fixed at one point x
  - Kernels $k(\cdot, x)$ form the basis of the space of possible functions
  - Regularity/smoothness of kernels is transferred to samples

- In inference tasks, the structure of the kernel determines the prediction
  - => Extremely important to model it well