More on kernels

Marcel Lüthi

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

Kernels everywhere

Integral and differential equations

- Aronszajn, Nachman. "Theory of reproducing kernels." *Transactions of the American mathematical society* (1950): 337-404.

Numerical analysis, Approximation and Interpolation theory

- Wahba, Grace. *Spline models for observational data*. Vol. 59. Siam, 1990.
- Schaback, Robert, and Holger Wendland. "Kernel techniques: From machine learning to meshless methods." *Acta Numerica* 15 (2006): 543-639.
- Hennig, Philipp, and Osborn, Michael: Probabilistic numerics

• Geostatistics (Gaussian processes)

- Stein, Michael L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

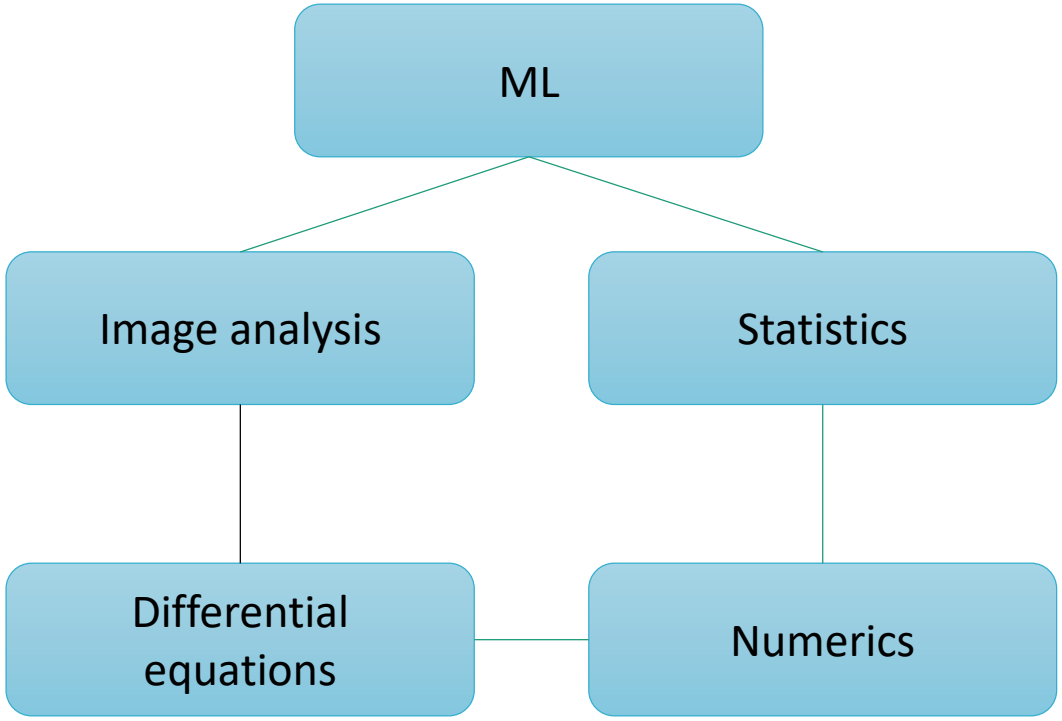
Kernels everywhere

- Learning Theory / Machine learning
  - Vapnik, Vladimir. *Statistical learning theory*. Vol. 1. New York: Wiley, 1998.
  - Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning." *The annals of statistics* (2008): 1171-1220.

- Shape modelling / Image analysis
  - Grenander, Ulf, and Michael I. Miller. "Computational anatomy: An emerging discipline." *Quarterly of applied mathematics* 56.4 (1998): 617-694.
  - Younes, Laurent: Shapes and diffeomorphisms, Springer 2010

What do they have in common?

- Solution space has a rich structure to be able to:
  - Predict unseen values
  - Deal with noisy or incomplete data
  - Capture a pattern
- Kernels ideally suited to define such structure
  - The resulting space of functions is mathematically “nice”.

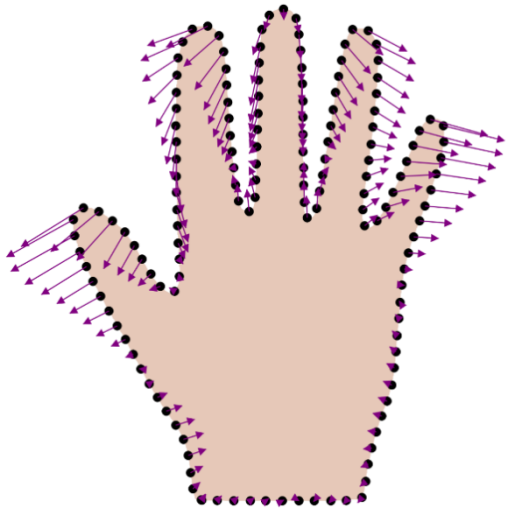


Back to basics: Scalar-valued GPs

Vector-valued (this course)

- Samples u are deformation fields:

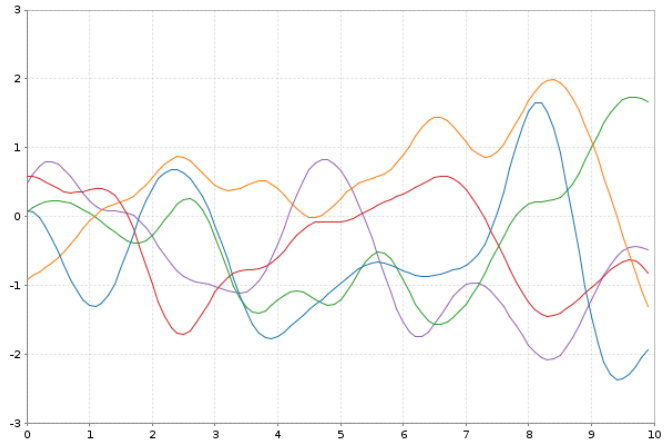
$$u: \mathcal{X} \to \mathbb{R}^d$$



Scalar-valued (more common)

- Samples f are real-valued functions

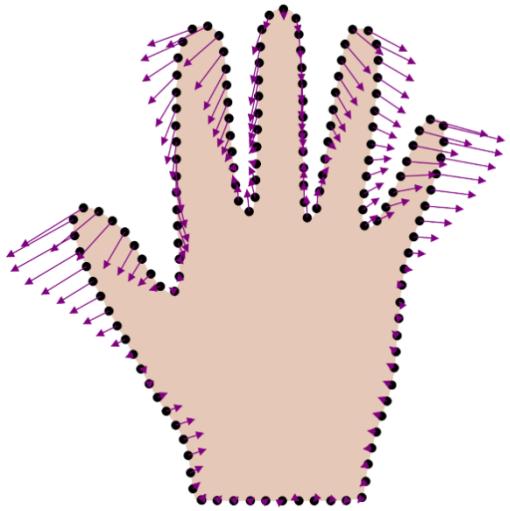
$$f : \mathcal{X} \to \mathbb{R}$$



Scalar-valued Gaussian processes

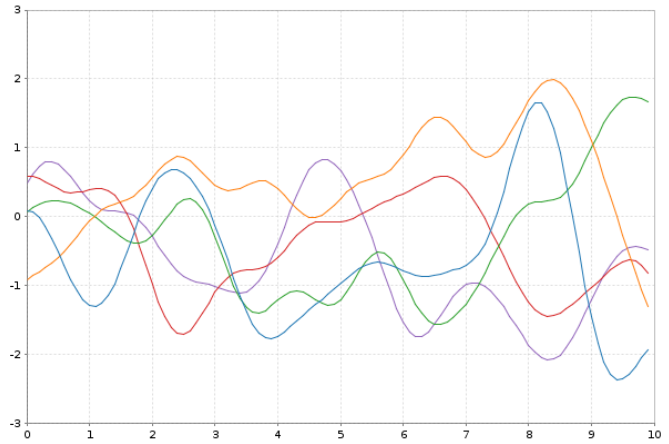
Vector-valued (this course)

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



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

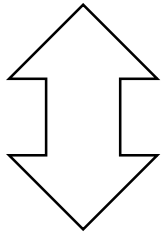
Scalar valued kernel: $k\colon \mathcal{X} \times (1..d) \times \mathcal{X} \times (1..d) \to \mathbb{R}$

Bijection: Define

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

Vector/scalar valued kernel matrices

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



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

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!

The sampling space

The space of samples

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

Algorithm (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

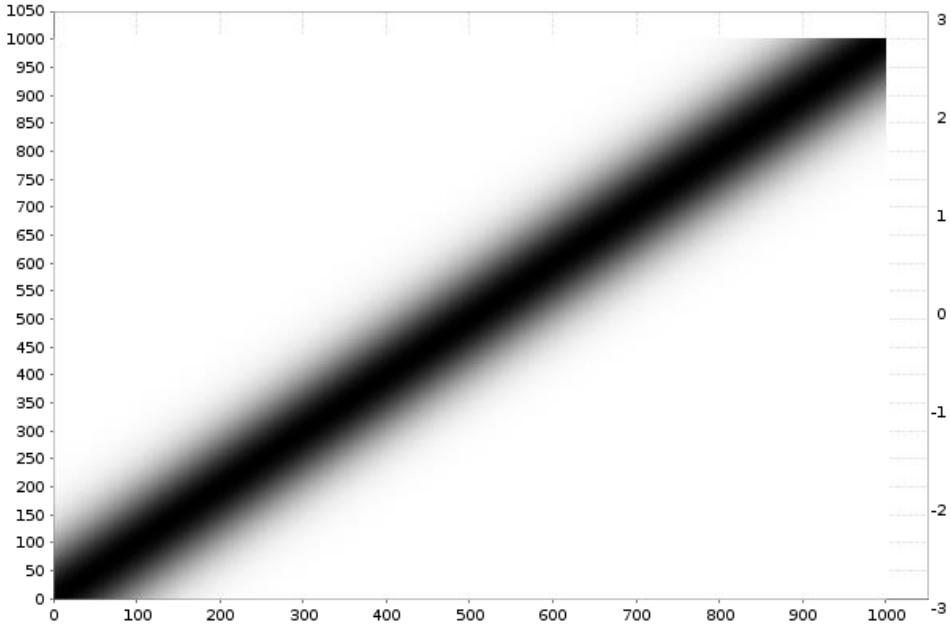
- From $\mathrm{K} = UD^2U^T$ (using that $U^TU = I$) we have that
$$\mathrm{K}UD^{-1} = UD$$
- A sample
$$s = \vec{\mu} + UD\alpha = \vec{\mu} + \mathrm{K}UD^{-1}\alpha$$
corresponds to linear combinations of the columns of K.

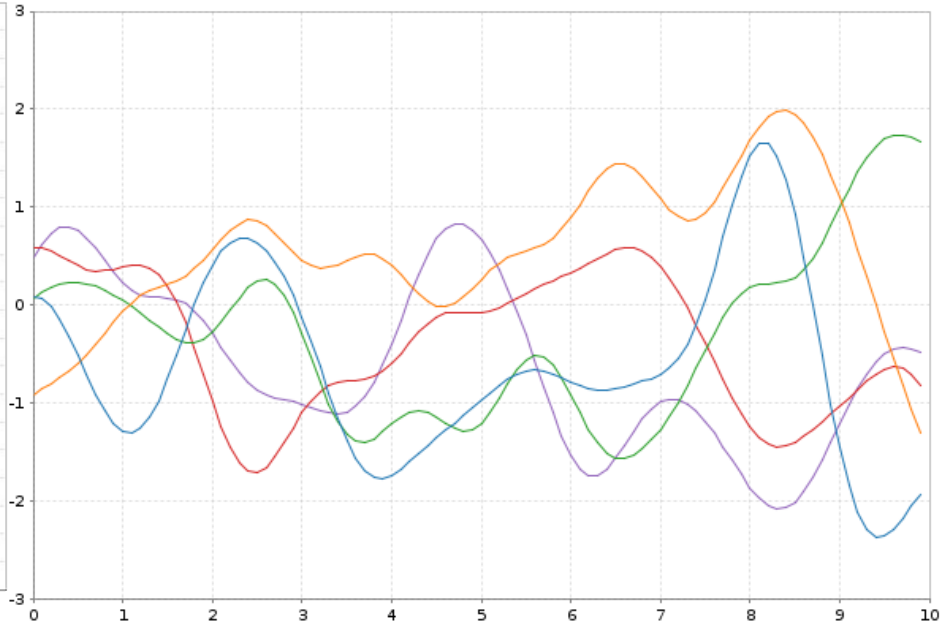
- K is symmetric → rows/columns can be used interchangeably

Example: Squared exponential

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

$$\sigma = 1$$

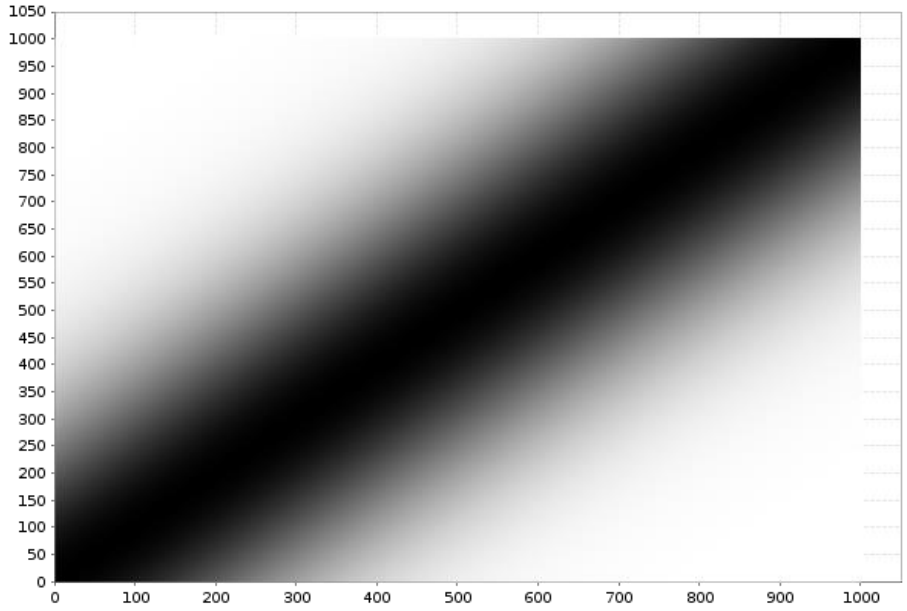


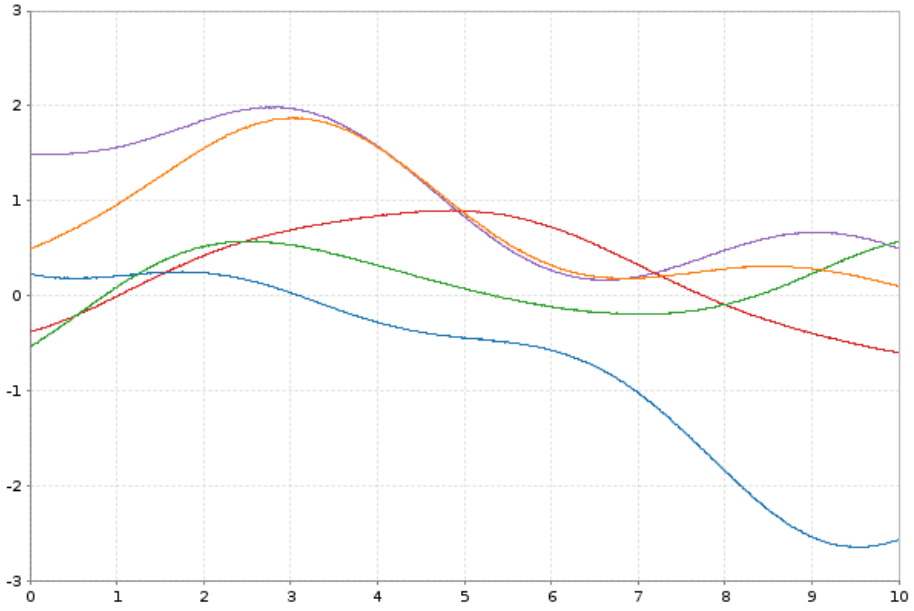


Example: Squared exponential

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

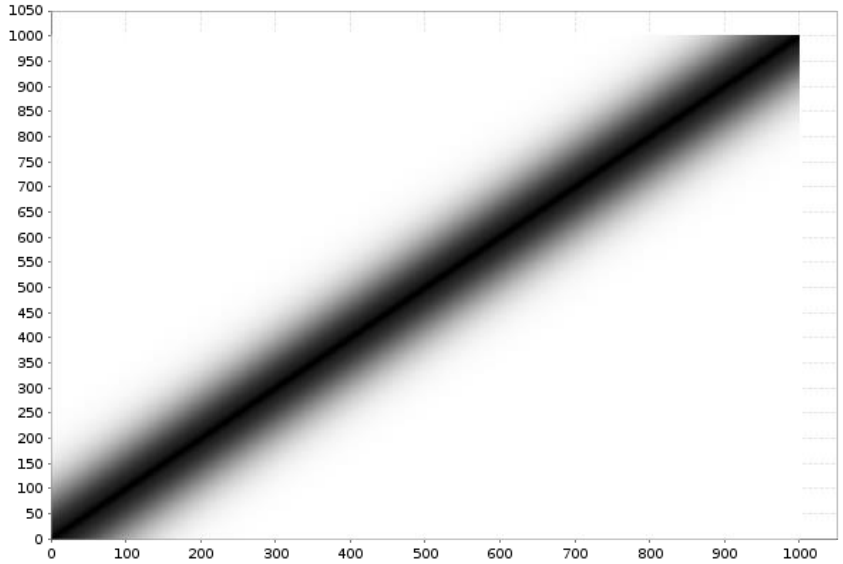
$$\sigma = 3$$

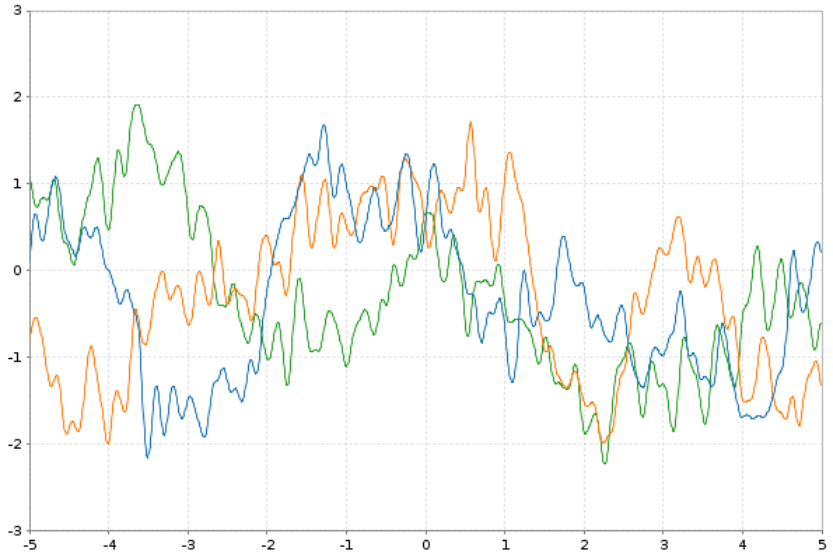




Multi-scale signals

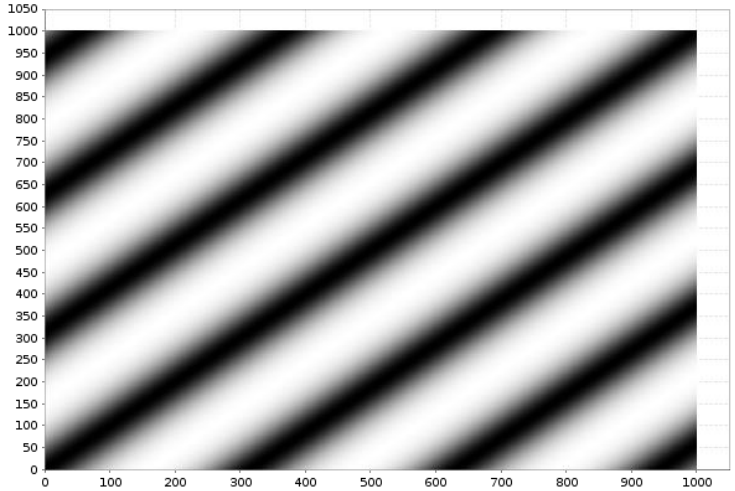
- $\mathrm{k}(\mathrm{x}, \mathrm{x}') = \exp\left(-\left\|x - \frac{x'}{1}\right\|^2\right) + 0.1 \exp\left(-\left\|x - \frac{x'}{0.1}\right\|^2\right)$

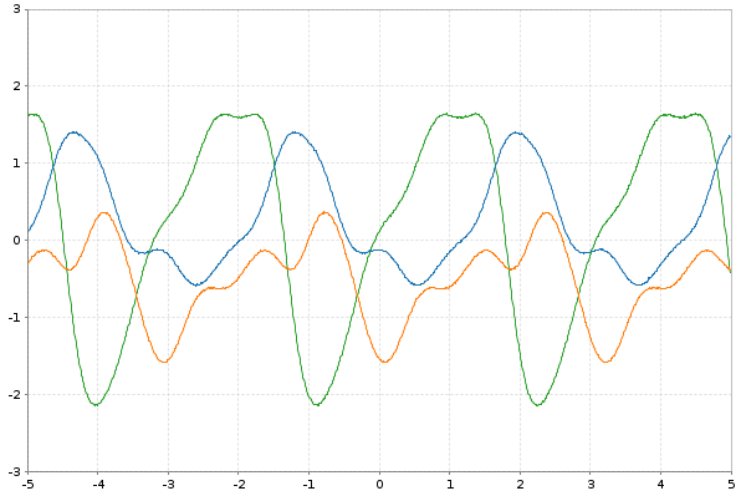




Periodic kernels

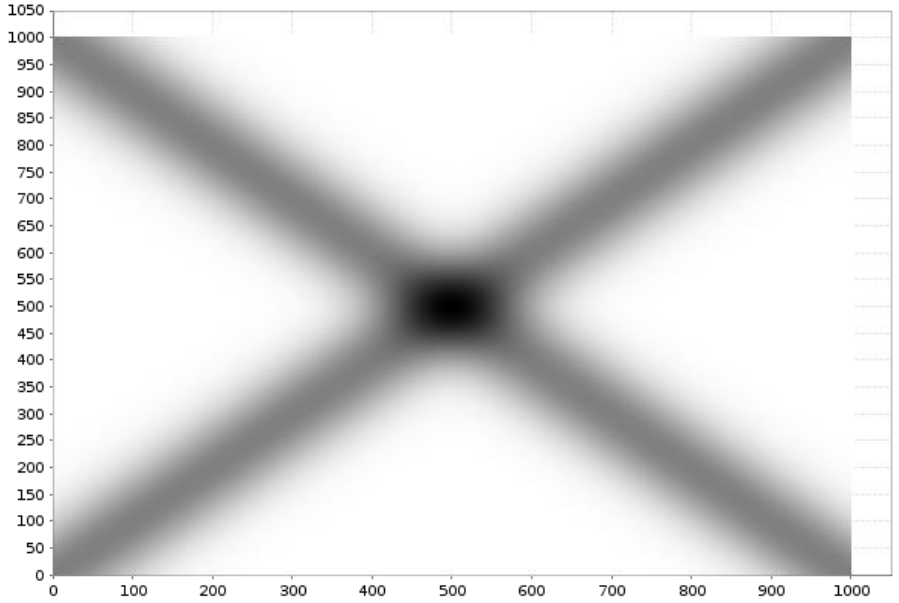
- Define $u(x) = \begin{pmatrix} \cos(x) \\ \sin(x) \end{pmatrix}$
- $k(x, x') = \exp(-\|(u(x) - u(x')\|^2 = \exp(-4\sin^2\left(\frac{\|x - x'\|}{\sigma^2}\right))$

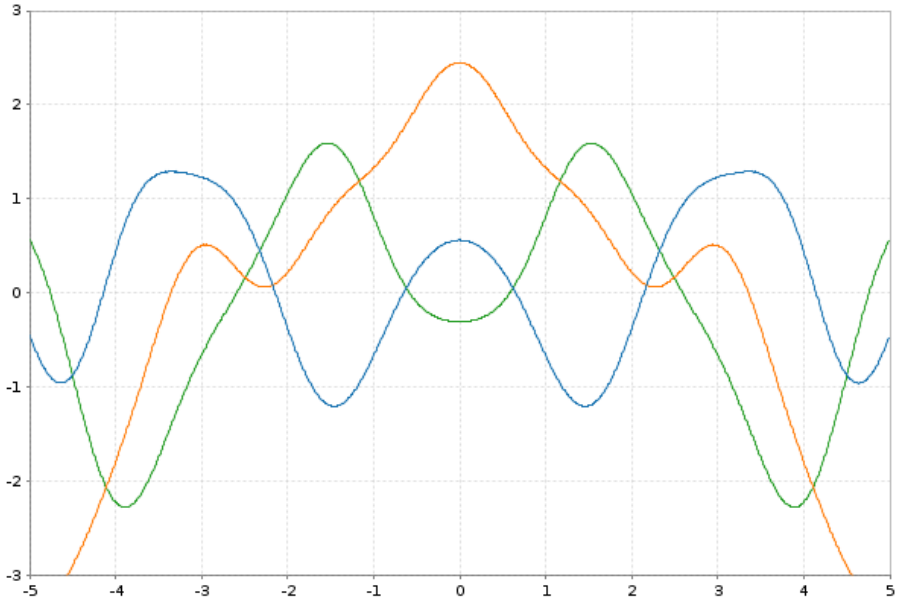




Symmetric kernels

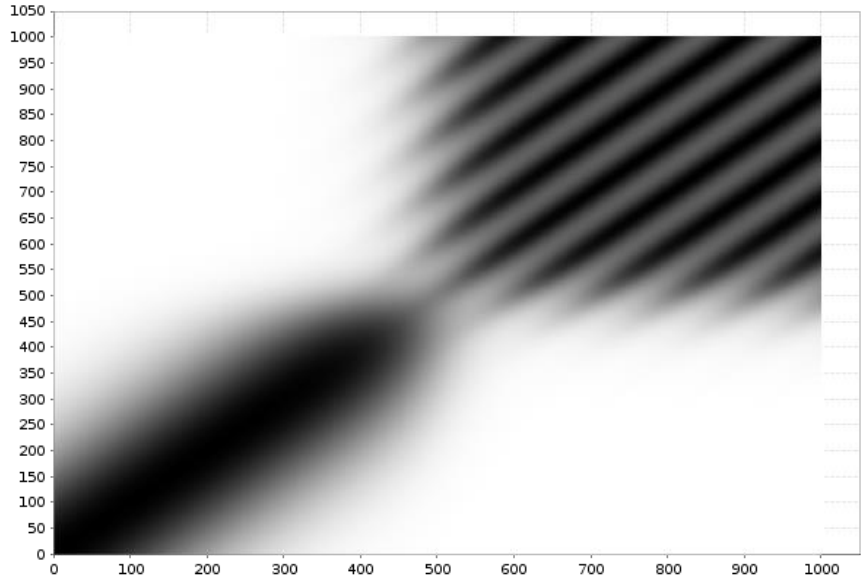
- *Enforce that f(x) = f(-x)*
- $k(x, x') = k(-x, x') + k(x, x')$

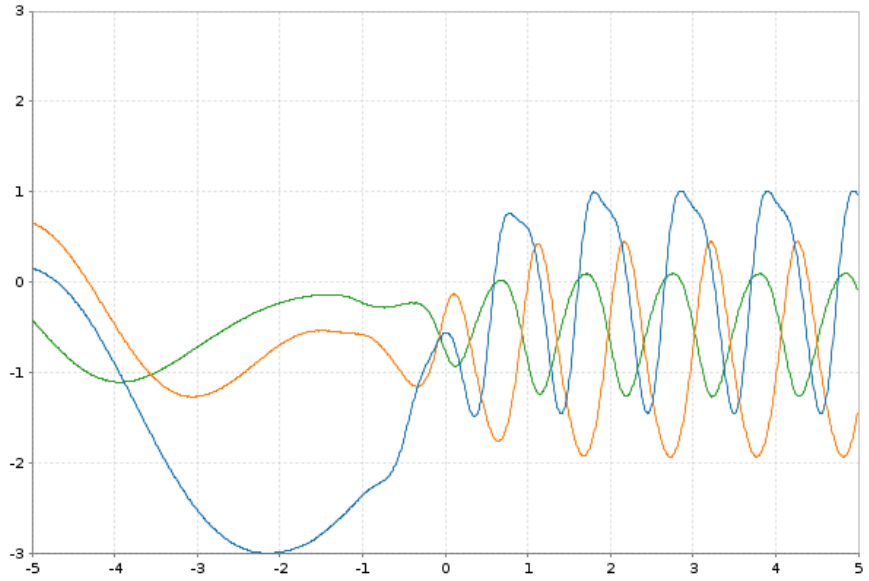




Changepoint kernels

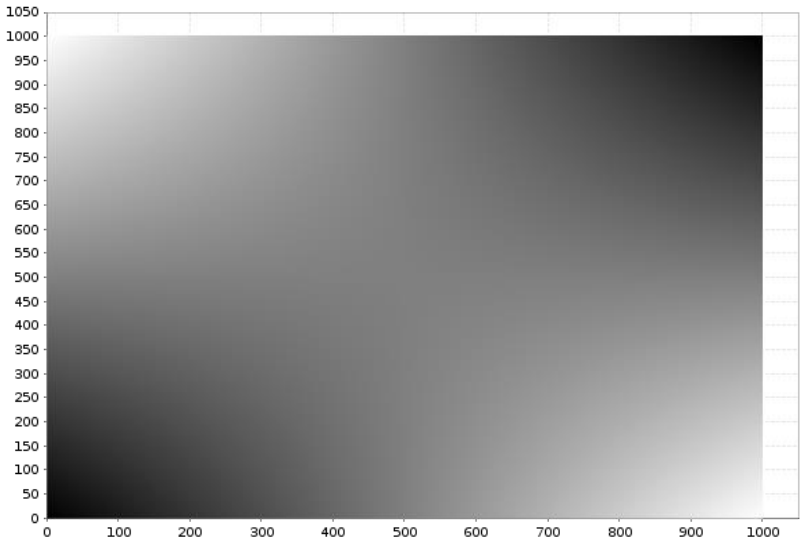
- $k(x, x') = s(x)k_1(x, x')s(x') + (1 - s(x))k_2(x, x')(1 - s(x'))$
- $\mathrm{s}(x) = \frac{1}{1+\exp(-x)}$

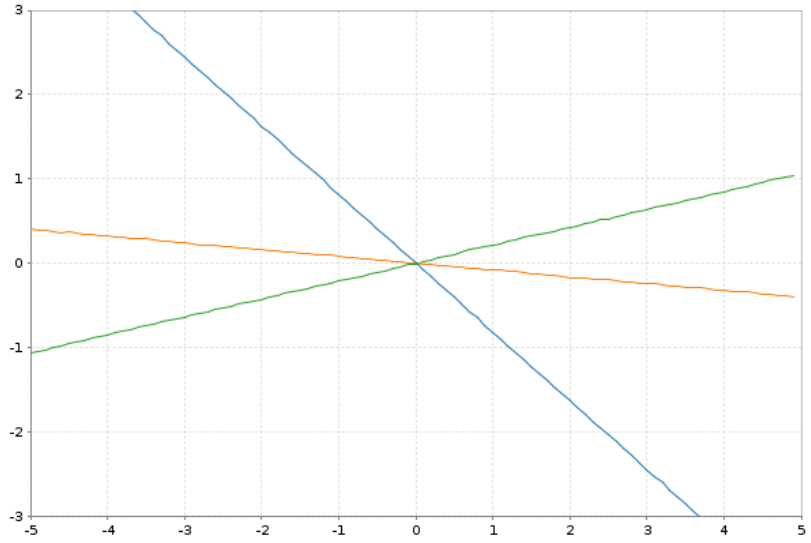




Combining existing functions

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

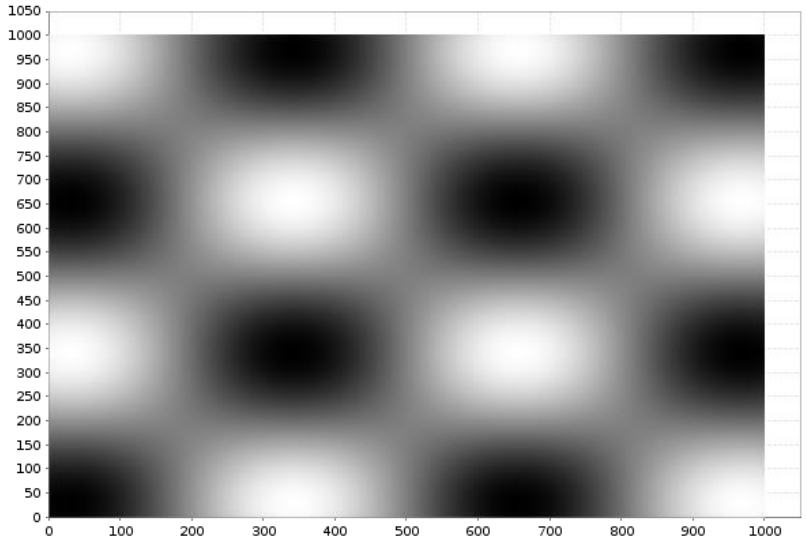


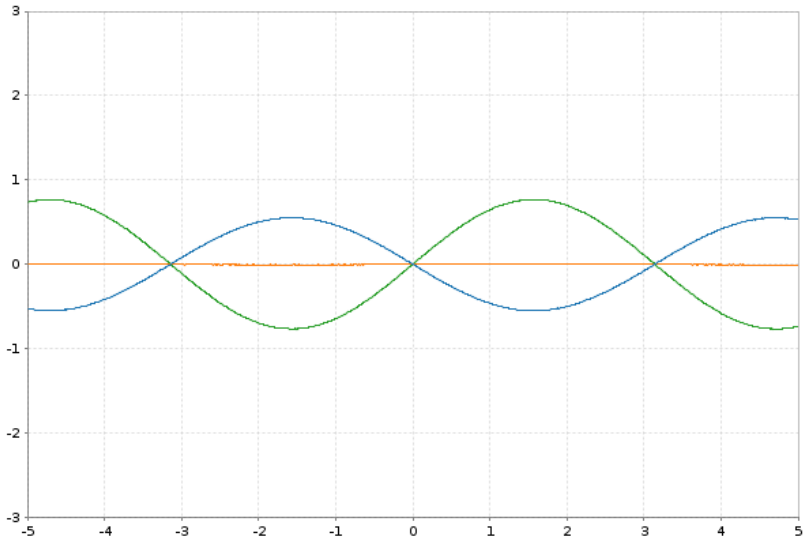


$$f(x) = x$$

Combining existing functions

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

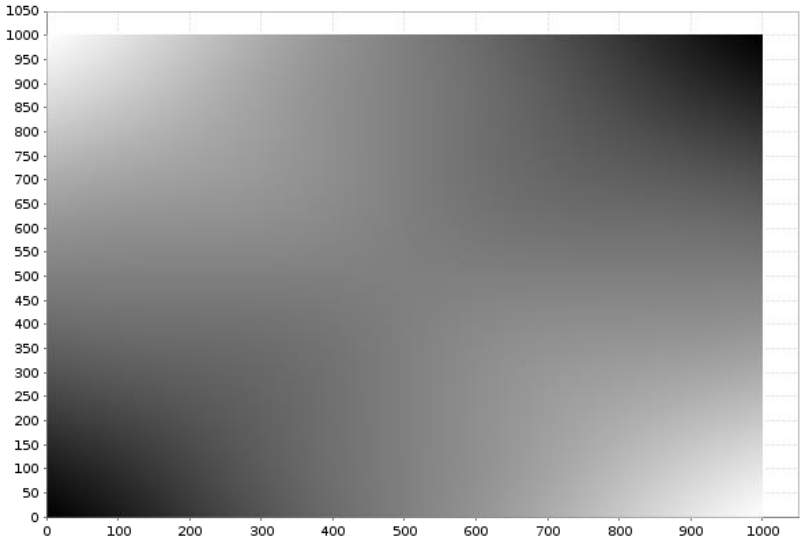


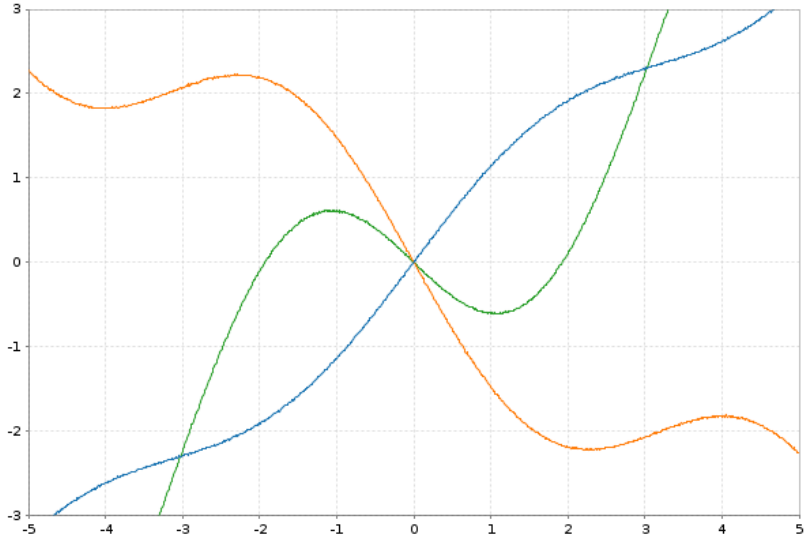


$$f(x) = \sin(x)$$

Combining existing functions

$$k(x, x') = \sum_i f_i(x) f_i(x')$$





$$\{f_1(x) = x, f_2(x) = \sin(x)\}$$

Reproducing Kernel Hilbert Space

- Define the space of functions

$$H = \{f | f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i), \qquad n \in \mathbb{N}, x_i \in X, \alpha_i \in \mathbb{R}\}$$

For $f(x) = \sum_i \alpha_i k(x_i, x)$ and $g(x) = \sum_j \alpha'_j k(x_j, x)$ we define the inner product

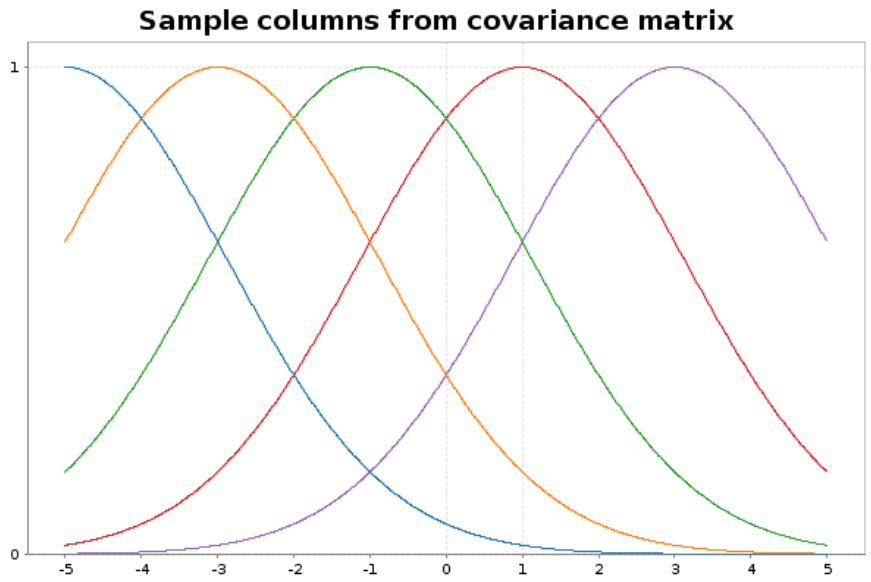
$$(f, g)_k = \sum_{i,j} \alpha_i \alpha'_j k(x_i, x_j)$$

The space H called a **Reproducing Kernel Hilbert Space (RKHS).**

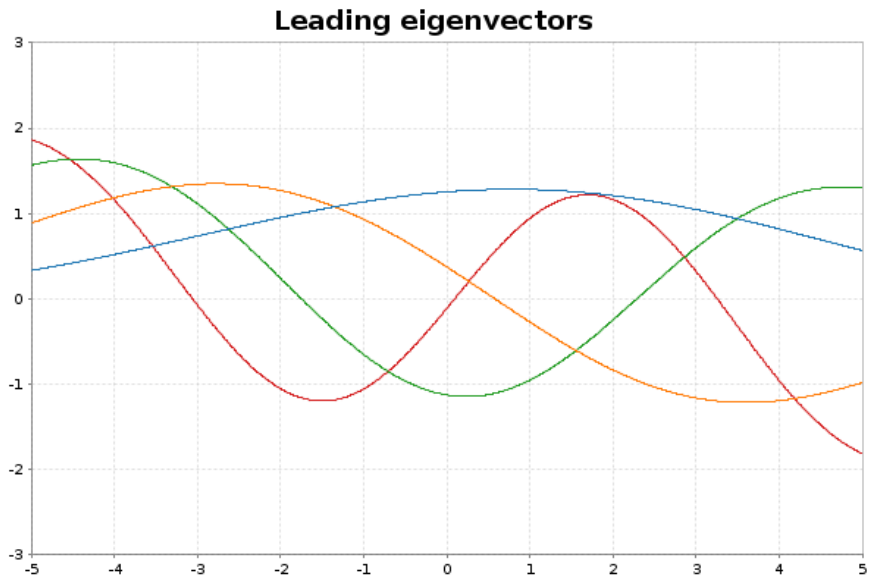
Two differnet basis for the RKHS

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

- Kernel basis



- Eigenbasis (KL-Basis)

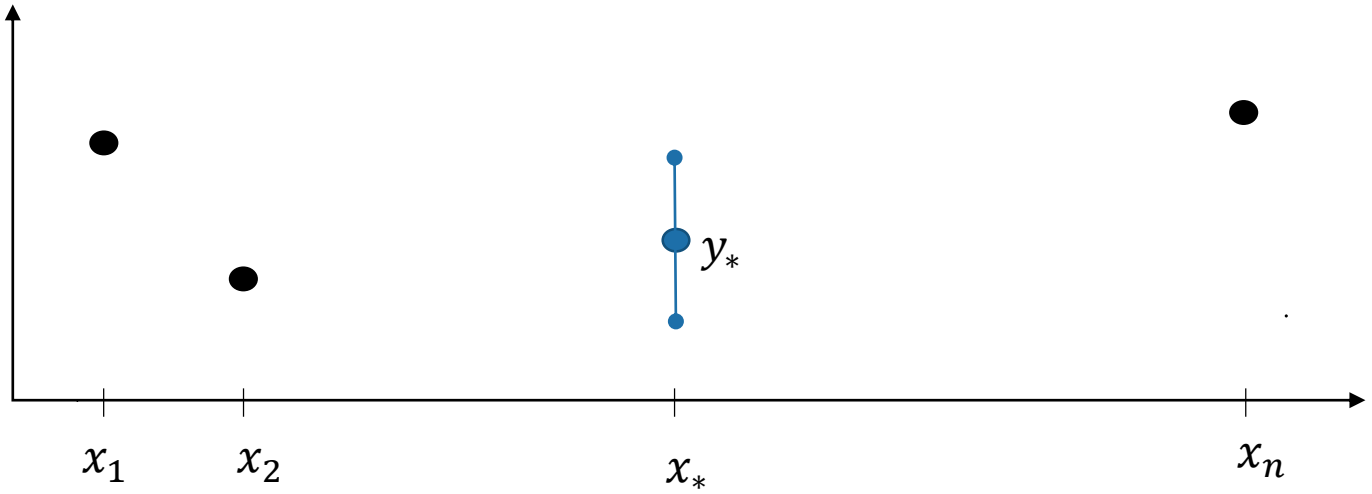


Gaussian process regression

Gaussian process regression

- Given :
  Observations: $\{(x_1, y_1), \dots, (x_n, y_n)\}$

- Goal:
  compute $p(y_* | x_*, x_1, \dots, x_n, y_1, \dots, y_n)$



Gaussian process regression

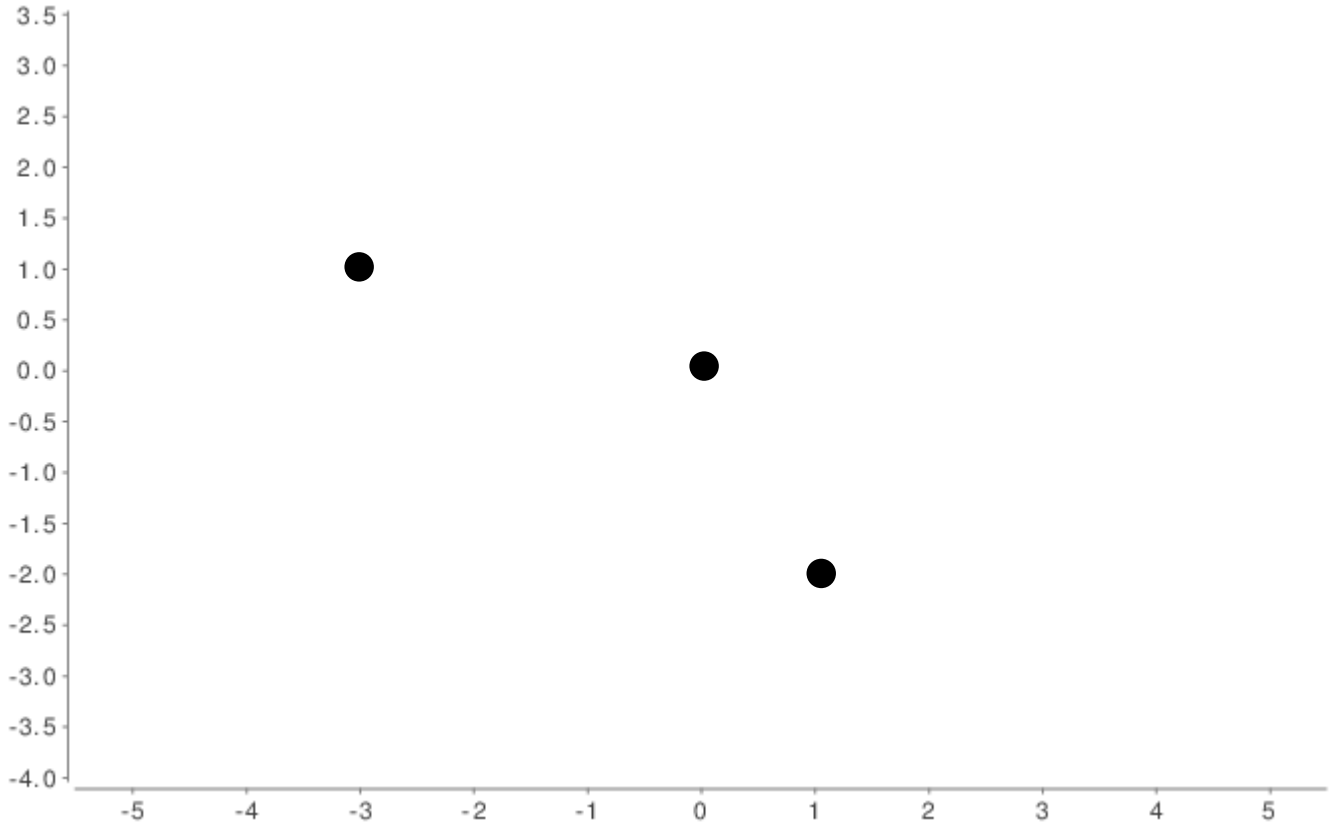
- Solution given by posterior process $GP(\mu_p, k_p)$ with

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

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

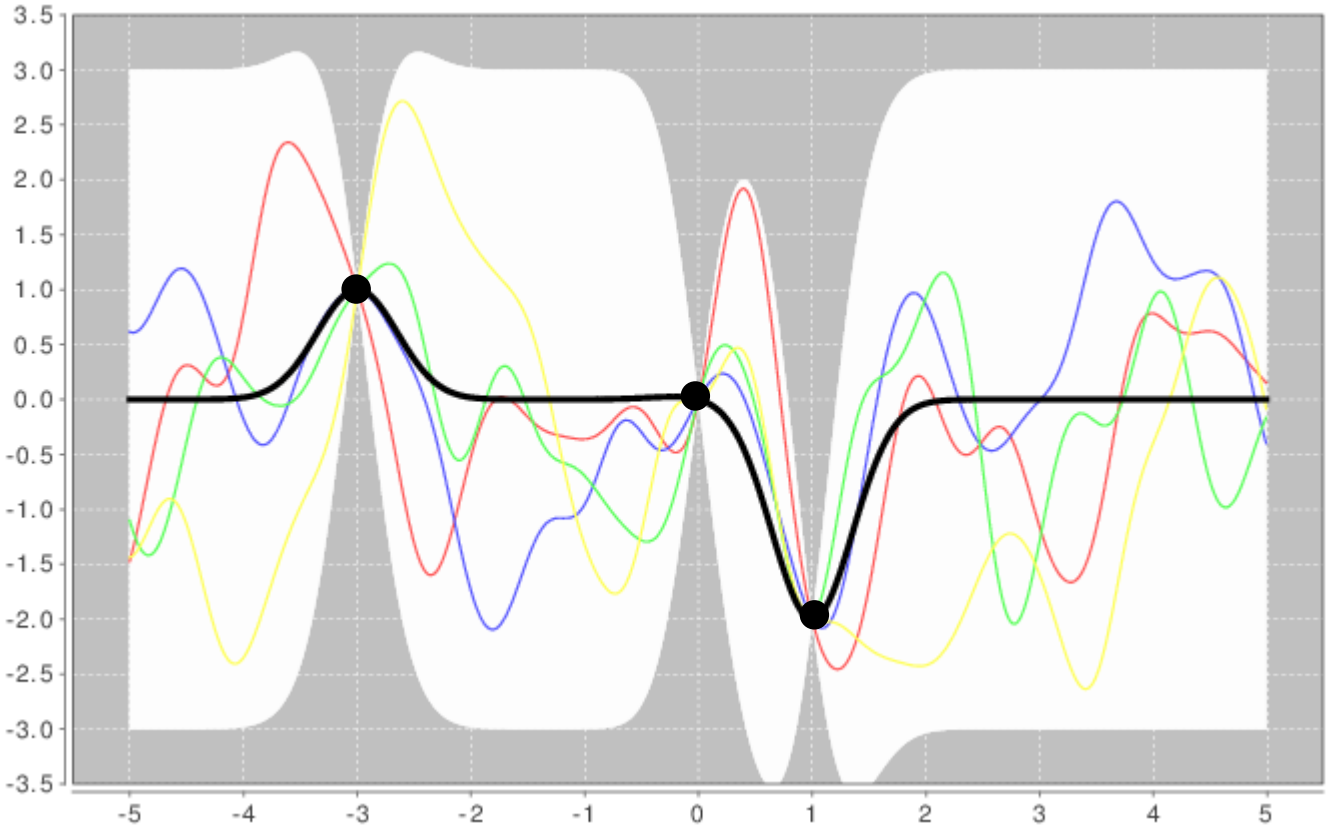
- We can sample from the posterior.

Examples



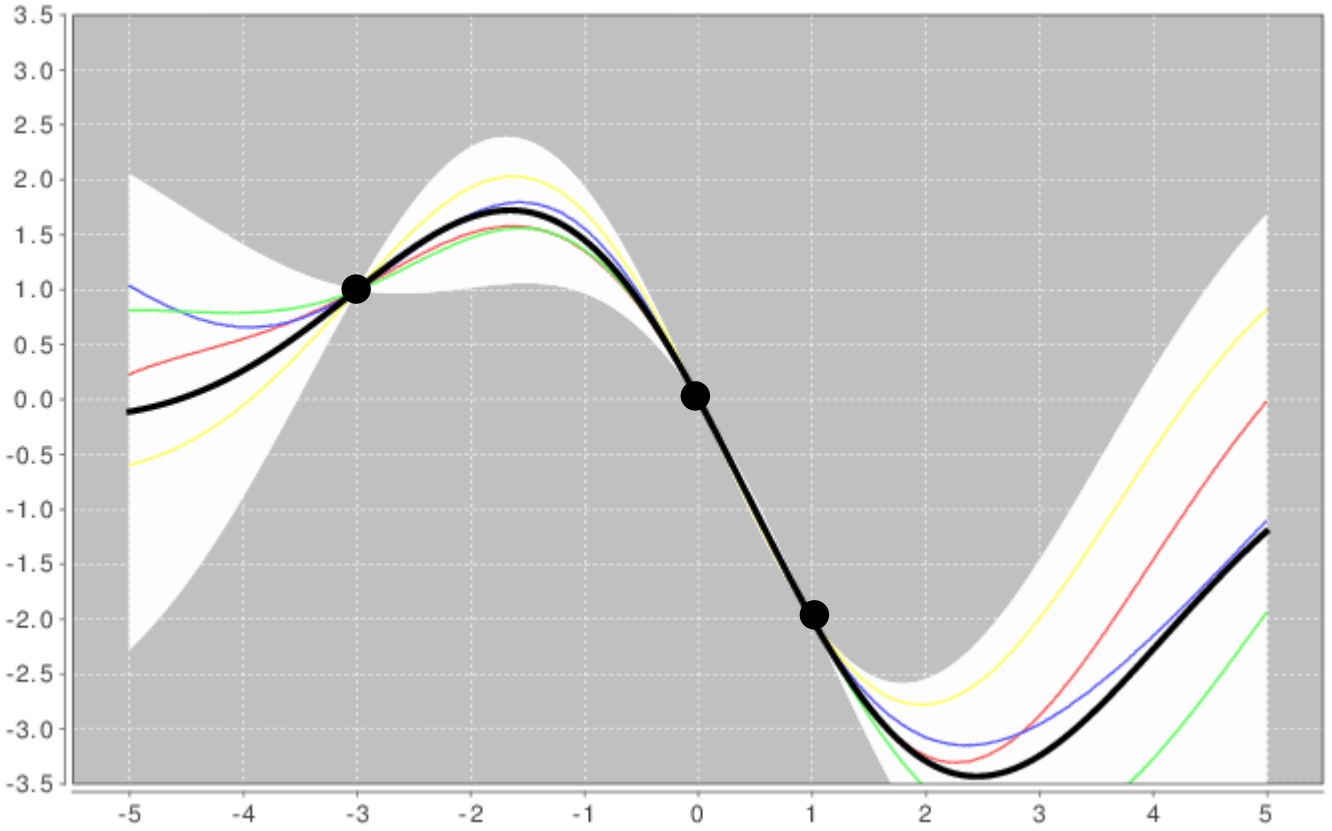
Examples

Gaussian kernel ($\sigma = 1$)



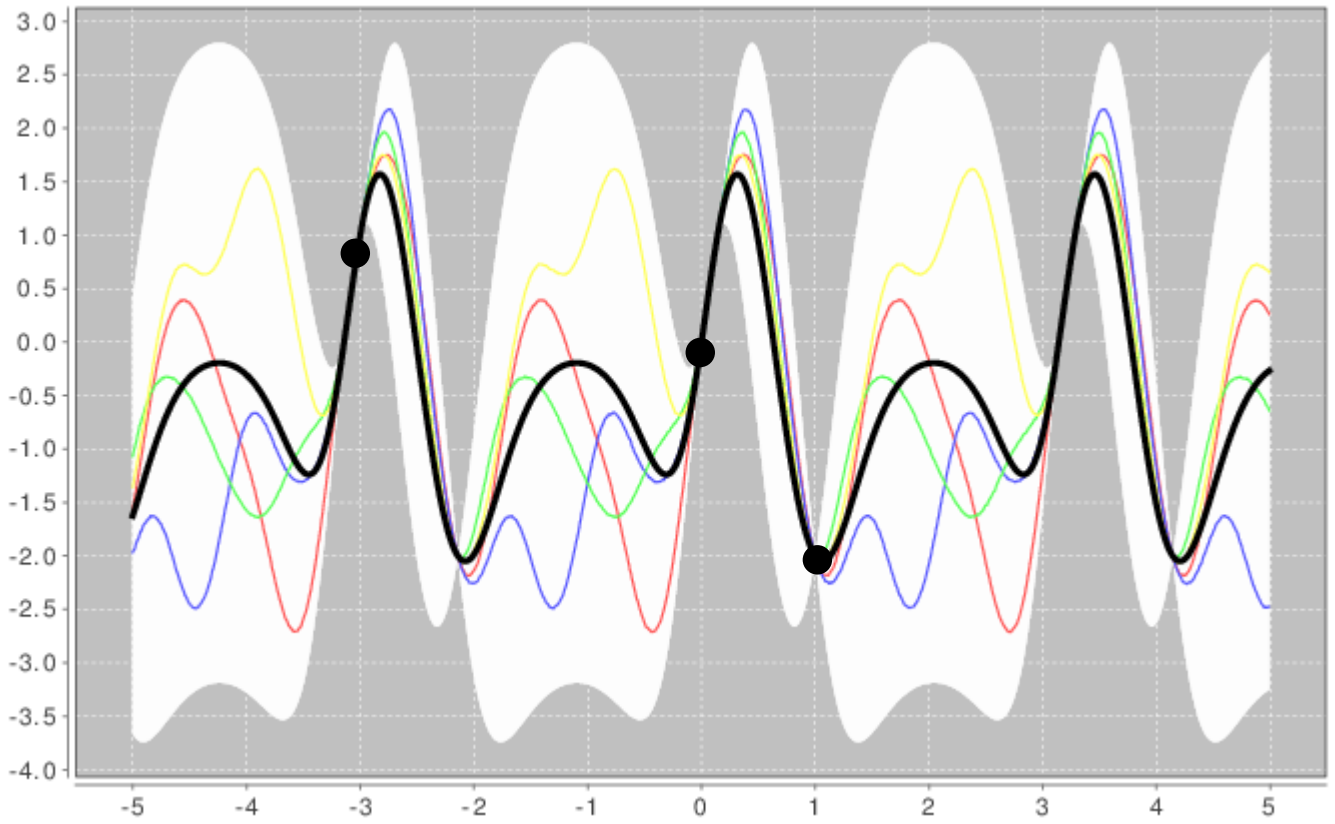
Examples

Gaussian kernel ($\sigma = 5$)



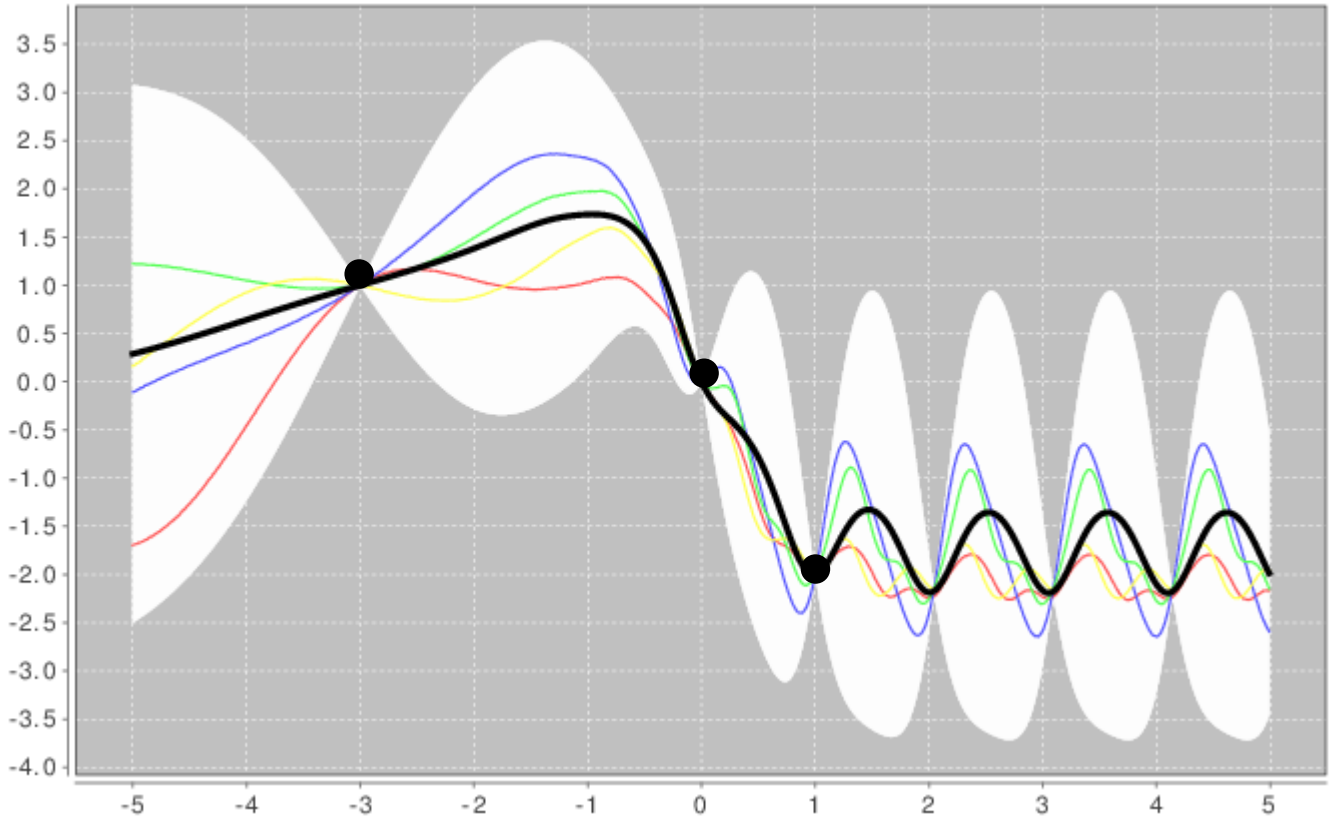
Examples

Periodic kernel



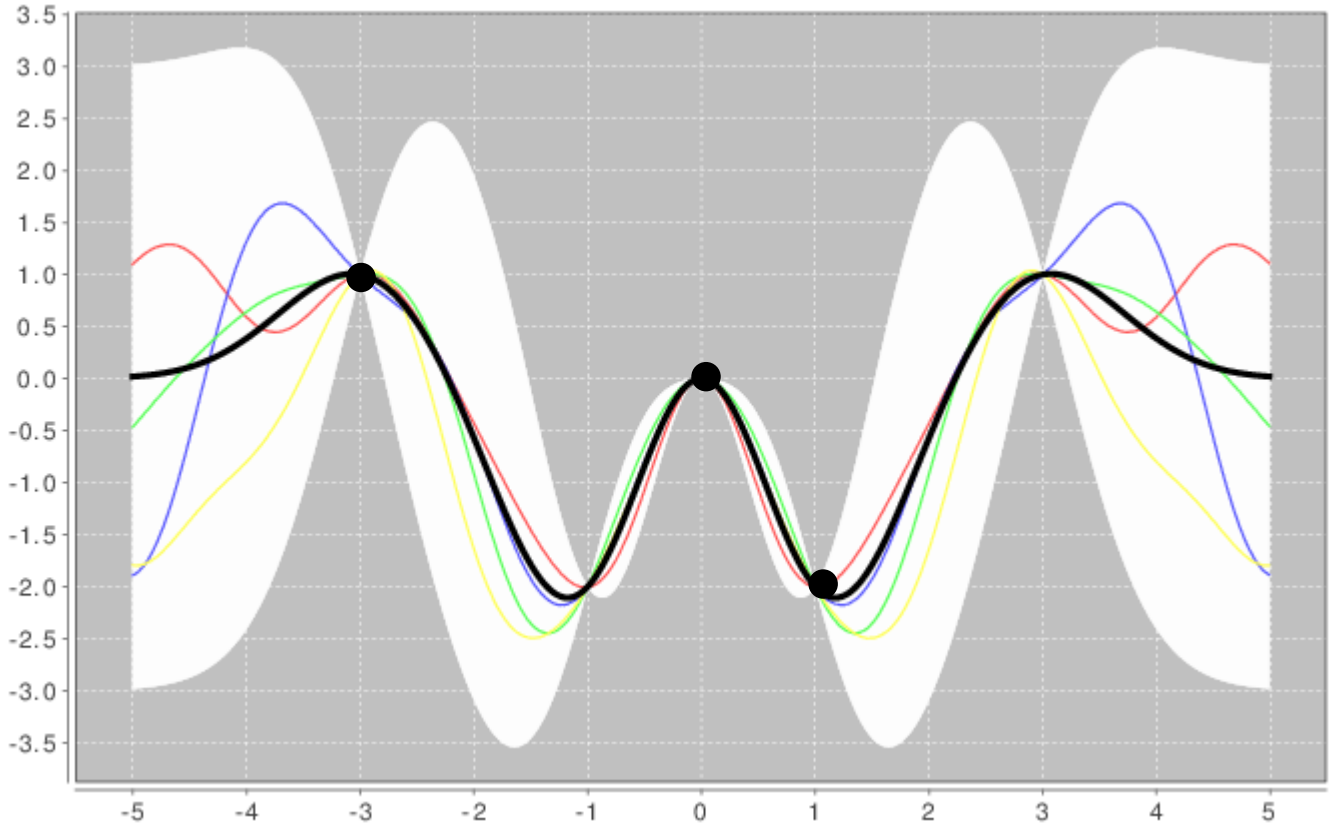
Examples

Changepoint kernel



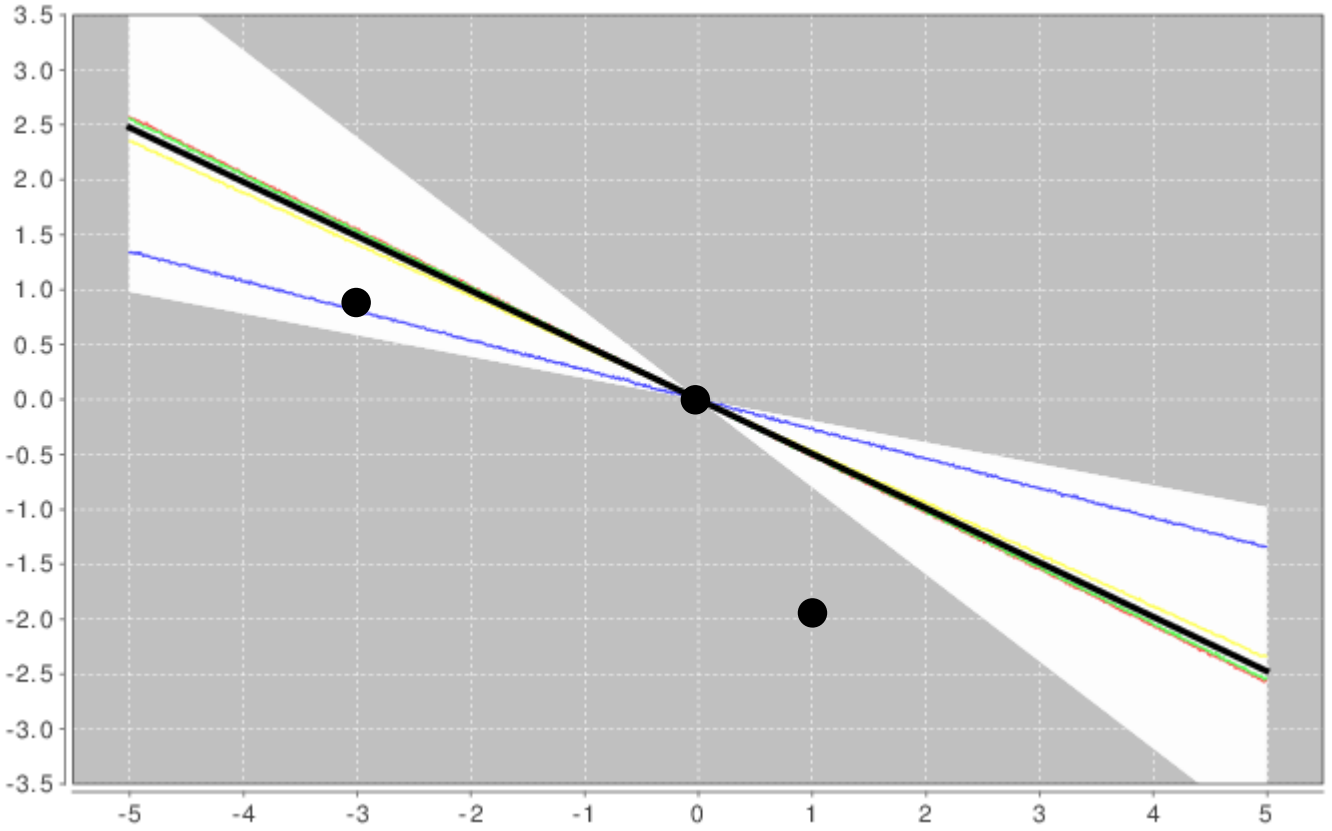
Examples

Symmetric kernel



Examples

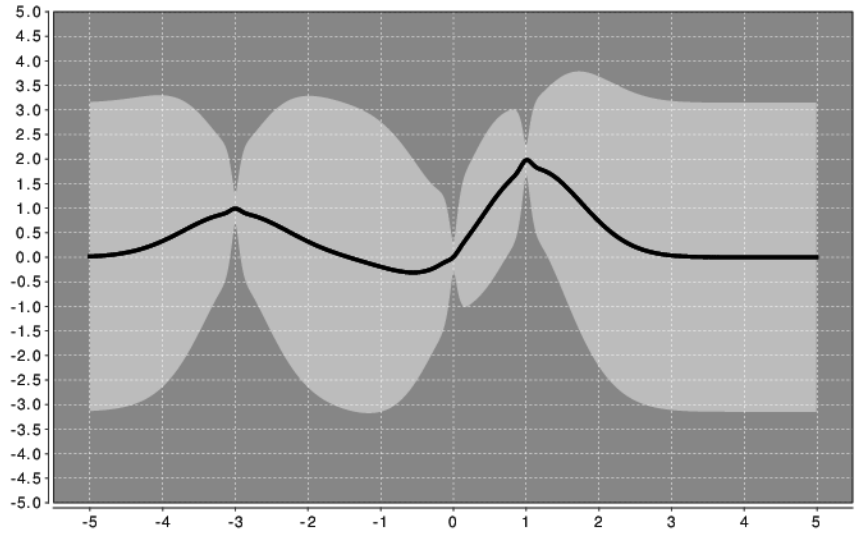
Linear kernel

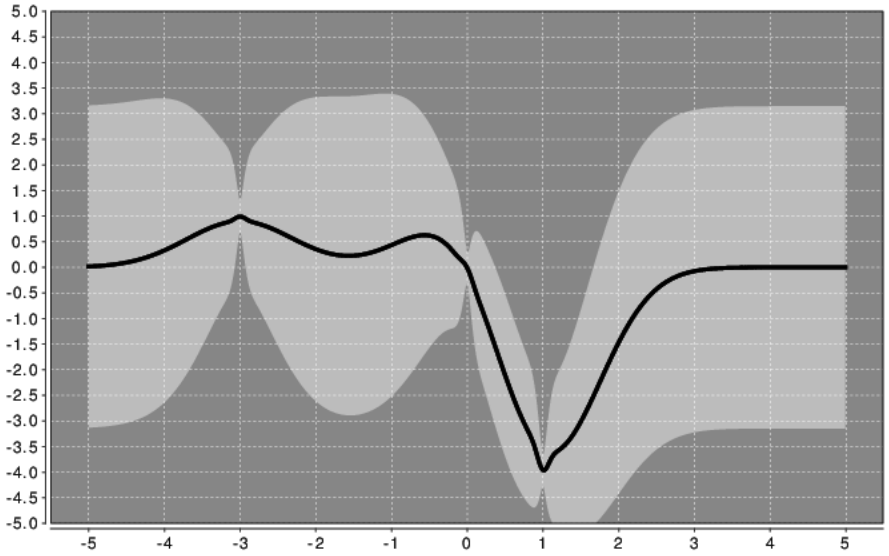


Observations about the solution

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

- The covariance is independent of the value at the training points





Kernels and associated structures

An enlightening paper

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Kernels, regularization and differential equations

Florian Steinke*, Bernhard Schölkopf

Max-Planck-Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 18 March 2008
Received in revised form 3 June 2008
Accepted 5 June 2008

Keywords:
Positive definite kernel
Differential equation
Gaussian process
Reproducing kernel Hilbert space

ABSTRACT

Many common machine learning methods such as support vector machines or Gaussian process inference make use of positive definite kernels, reproducing kernel Hilbert spaces, Gaussian processes, and regularization operators. In this work these objects are presented in a general, unifying framework and interrelations are highlighted.

With this in mind we then show how linear stochastic differential equation models can be incorporated naturally into the kernel framework. And vice versa, many kernel machines can be interpreted in terms of differential equations. We focus especially on ordinary differential equations, also known as dynamical systems, and it is shown that standard kernel inference algorithms are equivalent to Kalman filter methods based on such models.

In order not to cloud qualitative insights with heavy mathematical machinery, we restrict ourselves to finite domains, implying that differential equations are treated via their corresponding finite difference equations.