Introduction to kriging

Rodolphe Le Riche

Introduction to kriging

Rodolphe Le Riche[1]

[1] CNRS and Ecole des Mines de St-Etienne, FR

Class given as part of the “Modeling and Numerical Methods for Uncertainty Quantification” French-German summer school, Porquerolles, Sept. 2014

Course outline

1. Introduction to kriging (R. Le Riche)

   1.1. Gaussian Processes
   1.2. Covariance functions
   1.3. Conditional Gaussian Processes (kriging)
      1.3.1. No trend (simple kriging)
      1.3.2. With trend (universal kriging)
   1.4. Issues, links with other methods

Kriging : introduction

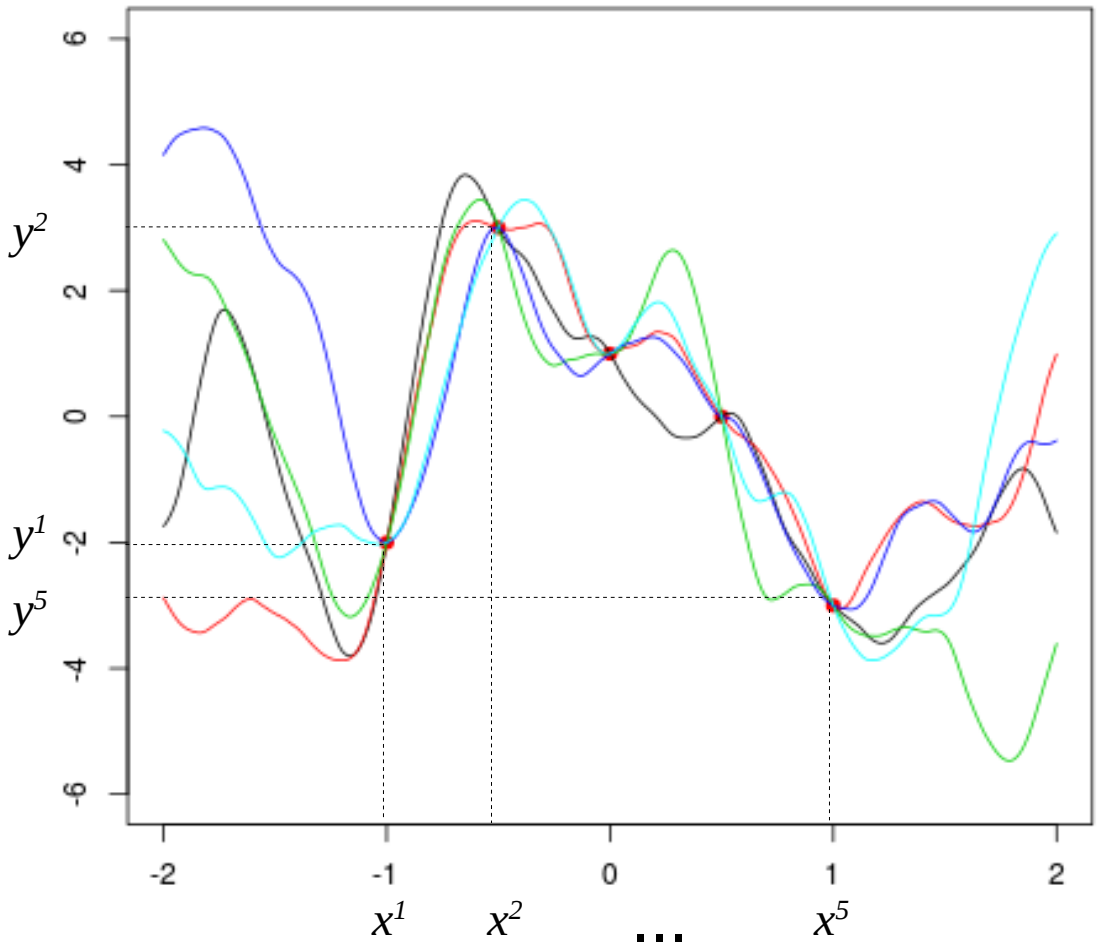
Context : scalar measurements $(y_{1,}\ldots,y_n)$

at n positions $(x^1,\ldots,x^n)$ in a d-dimensional space $\boldsymbol{X} \subset \mathbb{R}^d$

What can be said about possible measures at any x using probabilities ? (Krige, 1951; Matheron, 1963)

Here, kriging for regression.

Kriging = a family of metamodels (surrogate) with embedded uncertainty model.



Random processes

Random variable, Y

random event $\omega \in \Omega$
(e.g., throw dice)

get an instance y
Expl :
if dice ≤ 3 , y =1
3 < dice ≤ 5 , y = 2
dice = 6 , y = 3

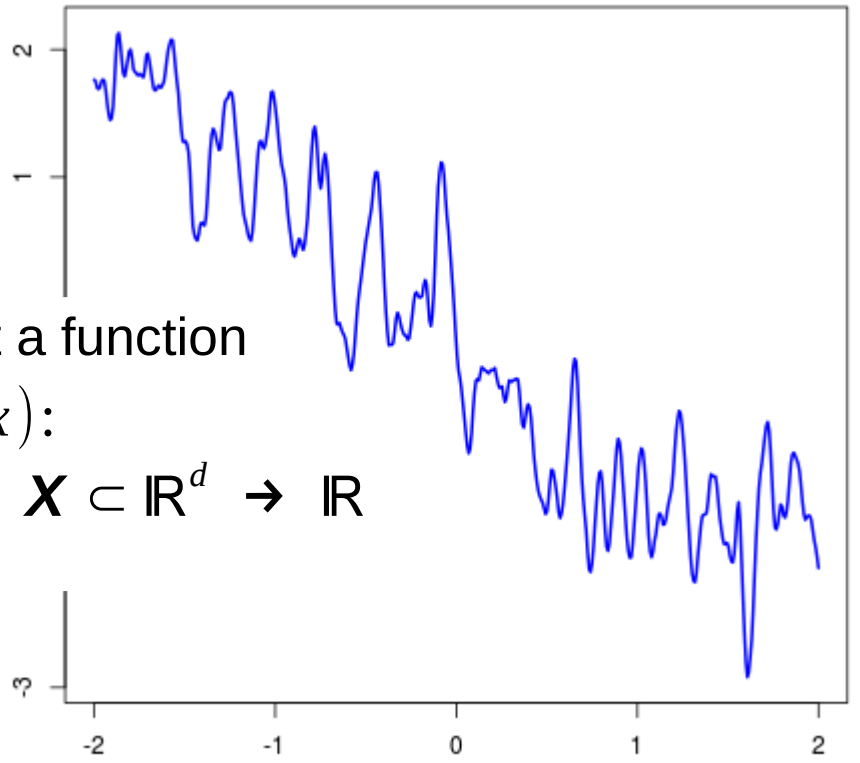
Random process, $Y(x)$

A set of RVs indexed by x

random event
$\omega \in \Omega$

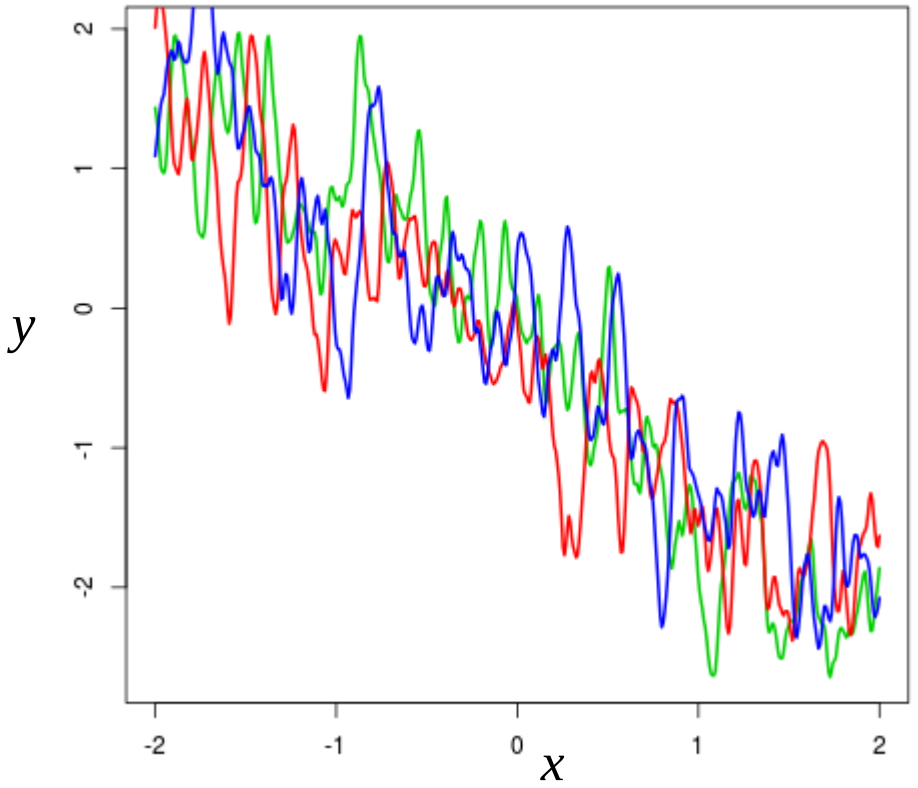
(e.g., wheather)

get a function
$y(x)$:
$x \in \boldsymbol{X} \subset \mathbb{R}^d \rightarrow \mathbb{R}$



Random processes

Repeat the random event



Ex : three events instances, three $y(x)$'s.
They are different, yet bear strong similarities.

Gaussian processes

Each $Y(x)$ follows a Gaussian law

$$Y(x) \sim \boldsymbol{N}(\mu(x), C(x,x))$$

and,

$$\forall \boldsymbol{X} = \begin{pmatrix} x^1 \\ \cdots \\ x^n \end{pmatrix} \in \boldsymbol{X^n} \subset \mathbb{R}^{d \times n} \quad , \quad \boldsymbol{Y} = \begin{pmatrix} Y(x^1) \\ \cdots \\ Y(x^n) \end{pmatrix} = \begin{pmatrix} Y_1 \\ \cdots \\ Y_n \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{C})$$

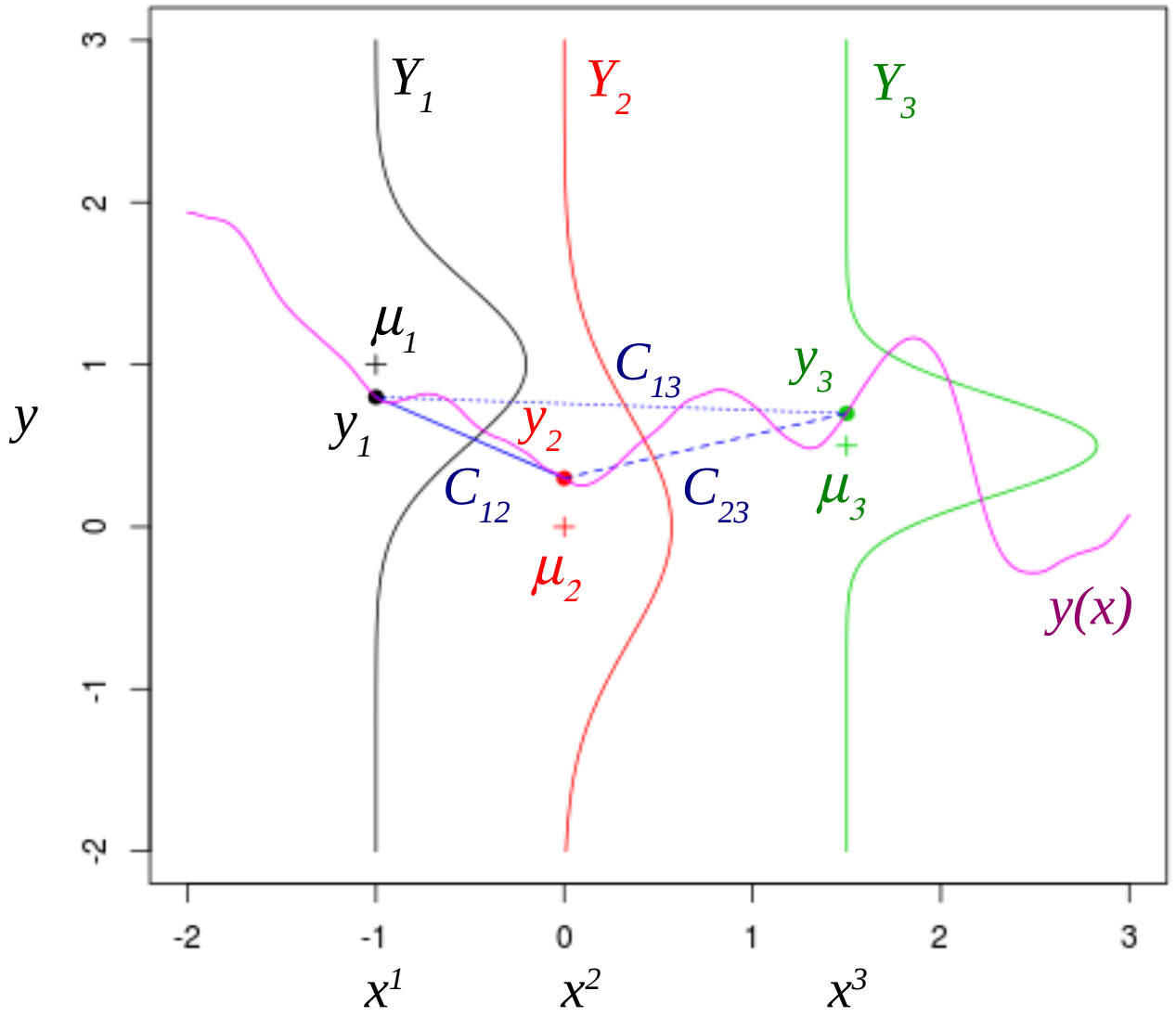
$$\boldsymbol{C_{ij}} = C(x^i, x^j)$$

with probability density function (multi-Normal law),

$$p(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2} \det^{1/2}(\boldsymbol{C})} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

Note : $\boldsymbol{C}$ is called Gram matrix in the SVM class (J.-M. Bourinet)

Gaussian processes (illustration)



$n = 3$

C_{ij}'s covariances between Y_i and Y_j (linear couplings)

Other possible illustration : contour lines of $p(\mathbf{y})$ as ellipses with principal axes as eigen-vectors/values of $\mathbf{C}^{-1}$

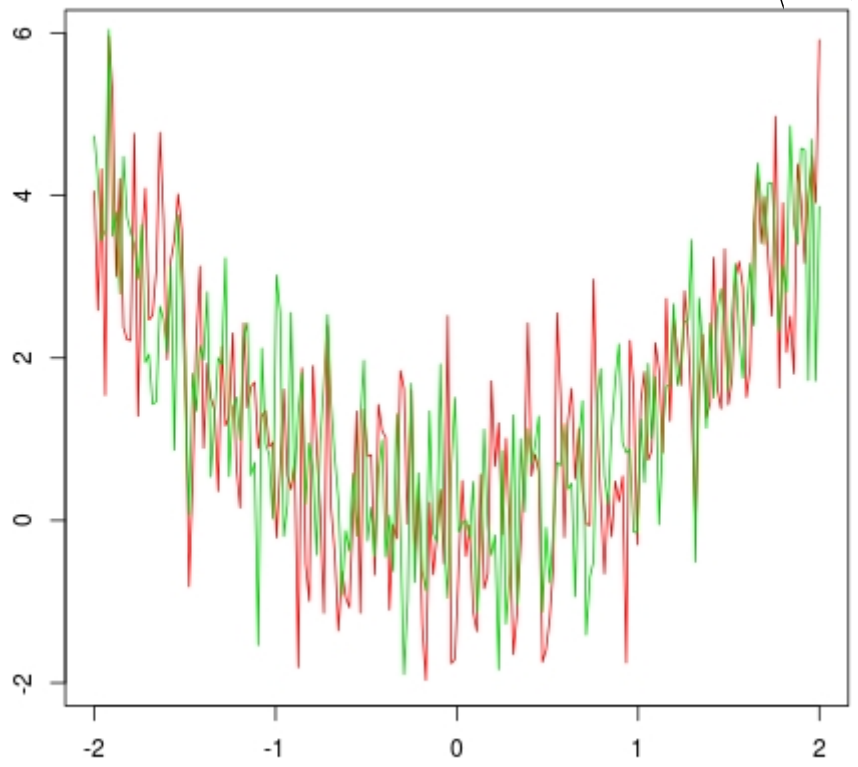
Special case : no spatial covariance

$Y(x) \sim N(\mu(x), \sigma(x)^2)$ i.e., GP generalize a trend with white noise

" " regression

$Y(x) = \mu(x) + E(x)$ where $E(x) \sim N(0, \sigma(x)^2)$

At n observation points, $\boldsymbol{Y} = \begin{pmatrix} Y(x^1) \\ \dots \\ Y(x^n) \end{pmatrix} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \sim N\left(\boldsymbol{\mu}, \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix}\right)$



Example :

$\mu(x) = x^2 \quad , \quad \sigma(x) = 1$

Numerical sampling of a GP

To plot GP trajectories (in 1 and 2Ds), perform the eigen analysis

$\boldsymbol{C} = U D^2 U^T$ and then notice $\boldsymbol{Y} = \boldsymbol{\mu} + U D \mathrm{E}$, $\mathrm{E} \sim N(0, \boldsymbol{I})$

(prove that $E Y = \boldsymbol{\mu}$ and $\mathrm{Cov}(Y) = \boldsymbol{C}$,
+ a Gaussian vector is fully determined by its 2 first moments)

In R,

```
Ceig <- eigen(C)
y <- mu + Ceig$vectors %*% diag(sqrt(Ceig$values))
                       %*% matrix(rnorm(n))
```

(Cf. previous illustrations, functions plotted with large n's)

More efficient implementation : cf. Carsten Proppe class, sl. “Discretization of random processes”

Definition of covariance functions

How do we build $\boldsymbol{C}$ and $\boldsymbol{\mu}$ and account for the data points (x^i, y_i) ?
→ Start with $\boldsymbol{C}$.

The **covariance function**, a.k.a. **kernel**,

$$\mathrm{Cov}\left(Y(x), Y(x')\right) = C(x, x')$$

is only a function of x and x' ,

$$C \;:\; \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}$$

The kernel defines the covariance matrix through

$$\boldsymbol{C_{ij}} = C\left(x^i, x^j\right)$$

Valid covariance functions

All functions $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are **not** valid kernels.

Kernels must yield **positive semidefinite** covariance matrices, $\boldsymbol{C}$:

$$\forall u \in \mathbb{R}^n \ , \ u^T \boldsymbol{C} u \geqslant 0$$

(but $C(x,x') < 0$ may happen)

Functional view (cf. Mercer's theorem) : let ϕ_i be square integrable eigenfunctions of x and $\lambda_i \geq 0$ the associated eigenvalues,

$$C(x, x') = \sum_{i=1}^{N} \lambda_i \phi_i(x) \phi_i(x') \ , \ N = \infty$$ but finite for degenerate kernels

Interpretation : kernels actually work in an N - (possibly infinite) dimensional feature space where a point is $(\phi_1(x), \ldots, \phi_N(x))^T$

Stationary covariance functions

The covariance function depends only on $\tau = x\text{-}x'$, but not on the position in space

$$\mathrm{Cov}\left(Y(x), Y(x')\right) = C(x, x') = C(x - x') = C(\tau)$$
$$C(0) = \sigma^2 \quad , \quad C(\tau) = \sigma^2 R(\tau) \quad (R \text{ the correlation})$$

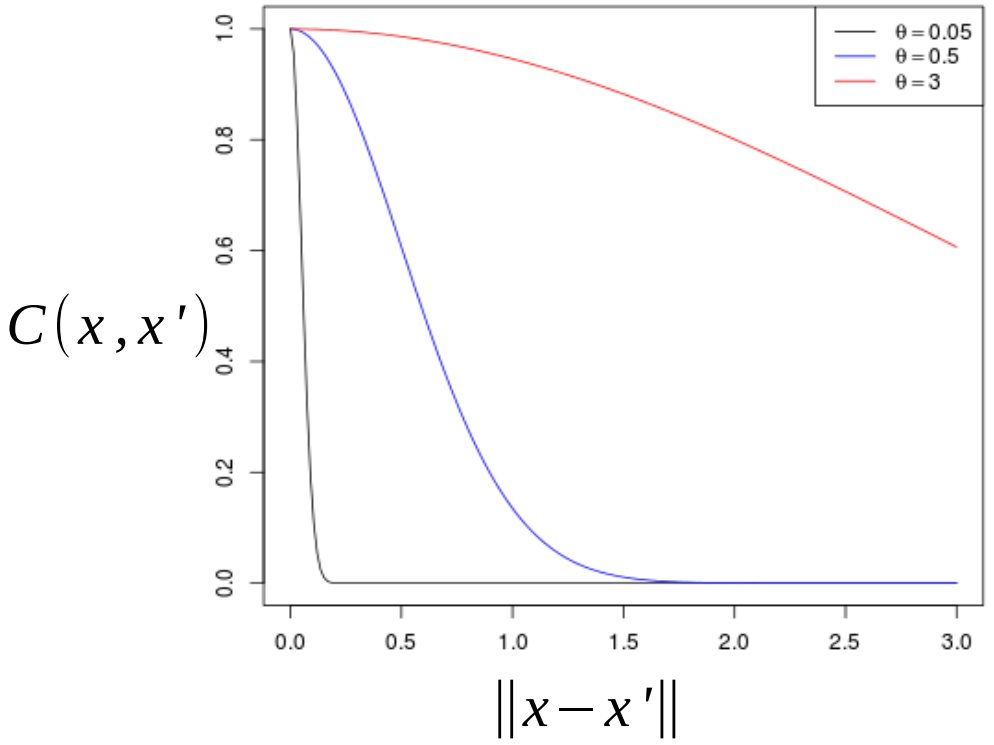
Example : the squared exponential covariance function (Gaussian)

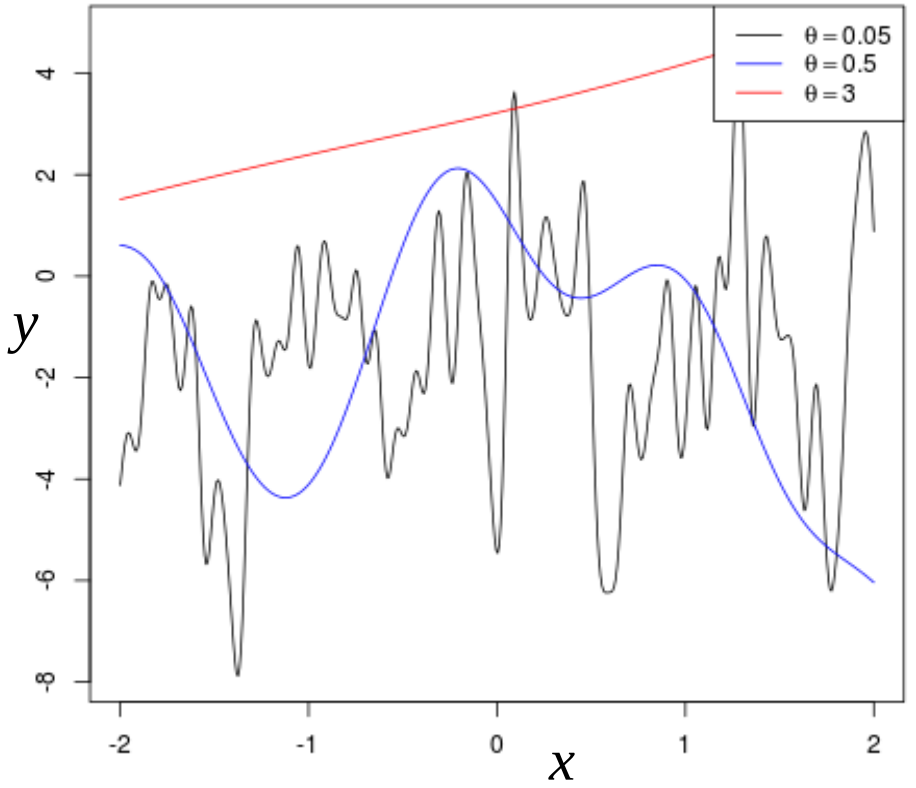
$$\mathrm{Cov}\left(Y(x), Y(x')\right) = C(x - x') = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{|x_i - x_i'|^2}{2\theta_i^2} \right)$$
$$= \sigma^2 \prod_{i=1}^{d} \exp\left(-\frac{|\tau_i|^2}{2\theta_i^2} \right)$$

Note : θ_i is a length scale $\equiv$ bandwidth (SVM class) $\equiv 1/\gamma$

Gaussian kernel, illustration

$$\mathrm{Cov}\left(Y(x), Y(x')\right) = C(x - x') = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{|x_i - x_i'|^2}{2\theta_i^2}\right)$$





The regularity and frequency content of the trajectories is controlled by the covariance functions. The θ_i's act as length scales.

Regularity of covariance functions

For stationary processes,
the trajectories $y(x)$ are p times differentiable (in the mean square sense) if $C(\tau)$ is $2p$ times differentiable at $\tau=0$.

→ The properties of $C(\tau)$ at $\tau=0$ define the regularity of the process.

Expl : trajectories with Gaussian kernels are infinitely differentiable (very – unrealistically ? – smooth)

Recycling covariance functions

The product of kernels is a kernel,

$$C(x, x') = \prod_{i=1}^{N} C_i(x, x')$$

(expl, ***d>1*** **kernels** like the Gaussian kernel)

The sum of kernels is a kernel,

$$C(x, x') = \sum_{i=1}^{N} C_i(x, x')$$

Let $W(x) = a(x)\, Y(x)$, where $a(x)$ is a deterministic function. Then,

$$\mathrm{Cov}(W(x), W(x')) = a(x) C(x, x') a(x')$$

Examples of stationary kernels

(the ones implemented in the DiceKriging R package)

General form , $C(x,x') = \sigma^2 \prod_{i=1}^{d} R(|x_i - x'_i|)$

Gaussian , $R(\tau) = \exp\left(-\frac{\tau^2}{2\theta^2}\right)$ (infinitely differentiable trajectories)

Matérn $\nu = 5/2$, $R(\tau) = \left(1 + \frac{\sqrt{5}|\tau|}{\theta} + \frac{5\tau^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}|\tau|}{\theta}\right)$ (twice diff. tr.)

Matérn $\nu = 3/2$, $R(\tau) = \left(1 + \frac{\sqrt{3}|\tau|}{\theta}\right)\exp\left(-\frac{\sqrt{3}|\tau|}{\theta}\right)$ (once diff. tr.)

Power-exponential , $R(\tau) = \exp\left(-\frac{|\tau|^p}{\theta}\right)$, $0 < p \leq 2$

(tr. not diff. except for $p = 2$)

Matérn $\nu = 5/2$ is the default choice.

They are functions of $d+1$ hyperparameters :

σ and the θ_i's to learn from data.

Tuning of hyperparameters

Our first use of the observations $(\boldsymbol{X}, \boldsymbol{Y})$

where $\boldsymbol{X} = \begin{pmatrix} x^1 \\ \cdots \\ x^n \end{pmatrix}$ and $\boldsymbol{Y} = \begin{pmatrix} Y(x^1) \\ \cdots \\ Y(x^n) \end{pmatrix}$

Three paths to selecting hyperparameters (σ and the θ_i's) :

maximum likelihood, cross-validation, Bayesian.

(discussed here)

Current statistical model : $Y(x) \sim N(\mu(x), C(x,x))$

equivalently , $Y(x) = \mu(x) + Z(x)$

where $\mu(x)$ known (deterministic) and $Z(x) \sim N(0, C(x,x))$

Maximum of likelihood estimate (1/2)

Likelihood : the probability of observing the observations as a function of the hyperparameters

$$L(\sigma,\boldsymbol{\theta}) = p(y|\sigma,\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\det^{1/2}(\boldsymbol{C})}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{C_{ij}} = C(x^i, x^j\ ;\ \sigma,\boldsymbol{\theta}) = \sigma^2 R(x^i, x^j\ ;\ \boldsymbol{\theta})$

and $\boldsymbol{\mu_i} = \mu(x^i)$

$$\max_{\sigma,\boldsymbol{\theta}} L(\sigma,\boldsymbol{\theta}) \ \Leftrightarrow\ \min_{\sigma,\boldsymbol{\theta}} \overbrace{-\log L(\sigma,\boldsymbol{\theta})}^{\text{mLL}}$$

Note : compare the likelihood $L(\sigma,\boldsymbol{\theta})$ to the regularized loss function of the SVM class, $L(\sigma,\boldsymbol{\theta}) = \text{reg\_terms}(\sigma,\boldsymbol{\theta}) \times \exp(-\text{loss\_function}(\sigma,\boldsymbol{\theta}))$

Maximum of likelihood estimate (2/2)

$$\mathrm{mLL}(\sigma,\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi) + n\log(\sigma) + \frac{1}{2}\log\left(\det(\boldsymbol{R})\right) + \frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

$$\frac{\partial \mathrm{mLL}}{\partial \sigma} = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{R}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

However, the calculations cannot be carried out analytically for $\boldsymbol{\theta}$.

So, numerically, minimize the “concentrated” likelihood,

$$\min_{\boldsymbol{\theta}\in[\boldsymbol{\theta}^{\min},\boldsymbol{\theta}^{\max}]} \mathrm{mLL}(\hat{\sigma}(\boldsymbol{\theta}),\boldsymbol{\theta})$$

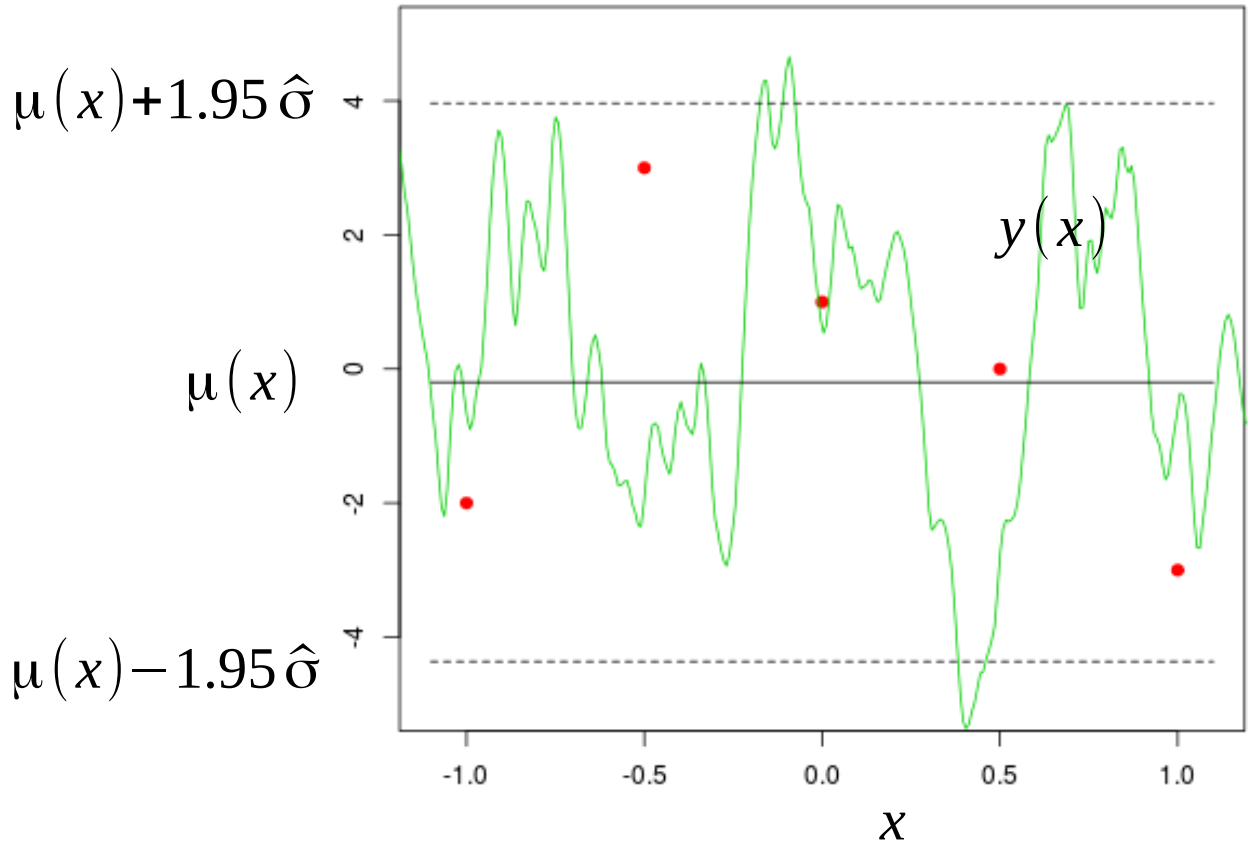
where $\boldsymbol{\theta}^{\min} > 0$

A nonlinear, multimodal optimization problem. In R / DiceKriging , solved by a mix of evolutionary – global – and BFGS – local – algorithms.

Where do we stand ?

$Y(x)$ normal with known average $\mu(x)$

and we have learned the covariance $C(x-x')$



Further use the observations *(**X** , **Y**=**y**)*. Make such a model **interpolate** them → last step to kriging.

Conditioning of a Gaussian random vector

Let U and V be jointly Gaussian random vectors,

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix}, \begin{bmatrix} C_U & C_{UV} \\ C_{UV}^T & C_V \end{bmatrix} \right)$$

then the conditional distribution of U knowing $V = v$ is

$$U \mid V = v \sim N\left(\underbrace{\mu_U + C_{UV} C_V^{-1} (v - \mu_V)}_{\text{cond. mean}} \;,\; \underbrace{C_U - C_{UV} C_V^{-1} C_{UV}^T}_{\text{cond. covar.}} \right)$$

- the conditional distribution is still Gaussian
- the conditional covariance does not depend on the observations v

and this is all we need ...

Simple kriging

Apply the conditioning result to the vector

$$\begin{bmatrix} Y(x^*) \\ Y(x^1) \\ \cdots \\ Y(x^n) \end{bmatrix} = \begin{bmatrix} Y(x^*) \\ \boldsymbol{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu(x^*) \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \sigma^2 & C(x^*, \boldsymbol{X}) \\ C(x^*, \boldsymbol{X})^T & \boldsymbol{C} \end{bmatrix} \right)$$

which directly yields the simple kriging* formula

$$Y(x^*) | \boldsymbol{Y} = \boldsymbol{y} \sim N\left(m_{SK}(x^*) , v_{SK}(x^*) \right)$$

$$m_{SK}(x^*) = \mu(x^*) + C(x^*, \boldsymbol{X}) \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$$

(identical to LS-SVR formula ?)

$$v_{SK}(x^*) = \sigma^2 - C(x^*, \boldsymbol{X}) \boldsymbol{C}^{-1} C(x^*, \boldsymbol{X})^T$$

* simple kriging : the trend ($\mu(x^*)$, $\boldsymbol{\mu}$) is assumed to be known

For prediction at 1 point, $C(x^*, x^*) = \sigma^2 R(x^* - x^*) = \sigma^2$

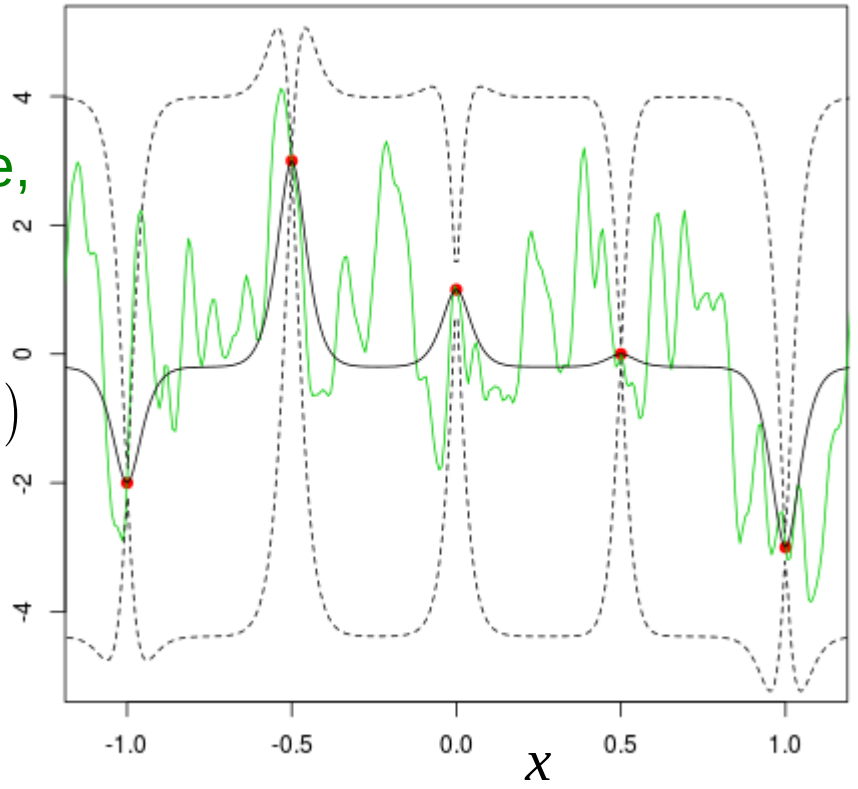
For joined prediction at many points, same formula but

$$x^* \rightarrow \boldsymbol{x}^* \; , \; \sigma^2 \rightarrow C(\boldsymbol{x}^*, \boldsymbol{x}^*) \; , \; v_{SK}(x^*) \rightarrow C_{SK}(\boldsymbol{x}^*, \boldsymbol{x}^*) = \boldsymbol{C}^*_{SK}$$

Simple kriging (illustration)

1 sample, $y_{SK}(x)$

$m_{SK}(x)$



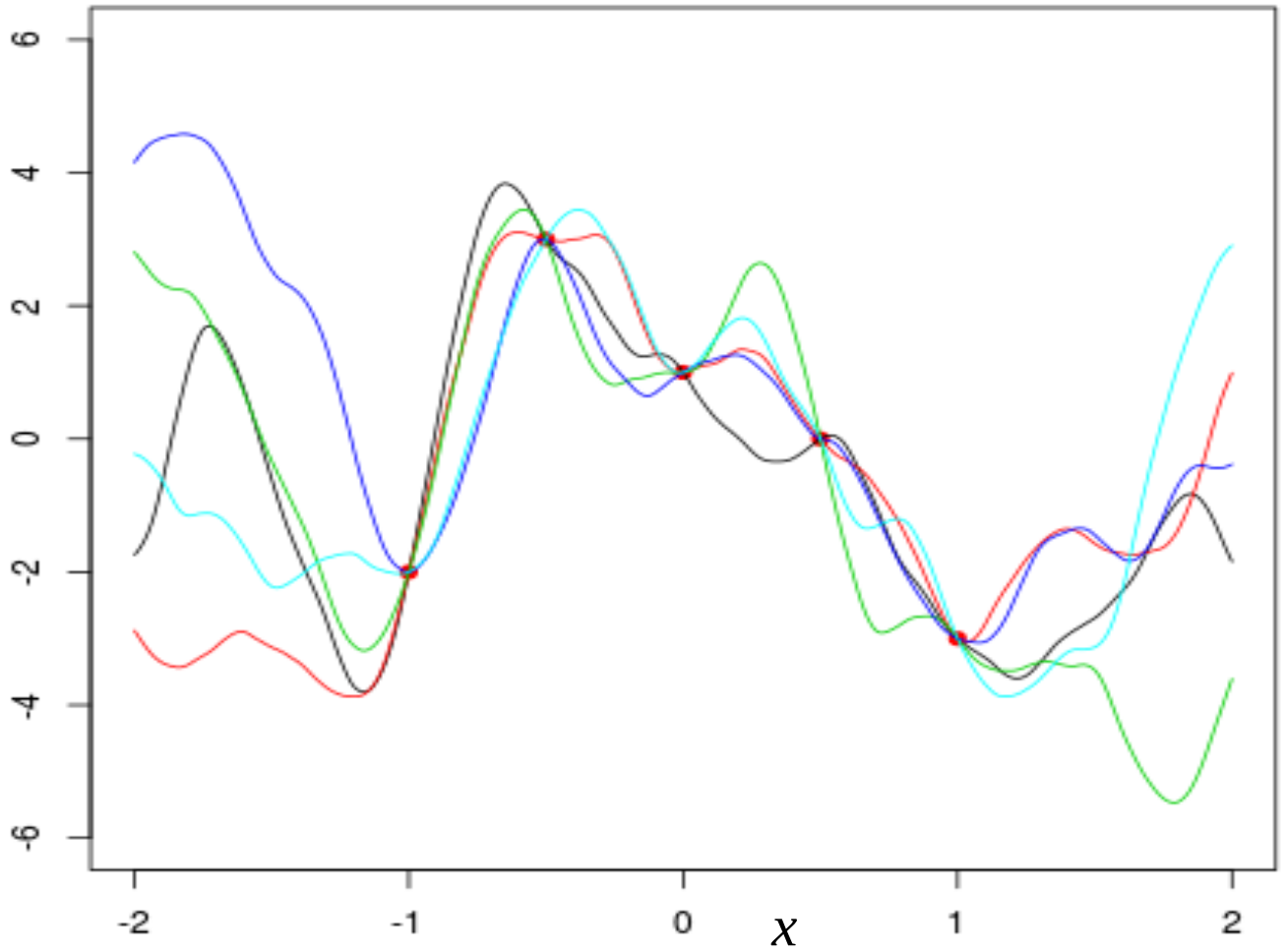
$m_{SK}(x)+1.95\sqrt{v_{SK}(x)}$

$m_{SK}(x)-1.95\sqrt{v_{SK}(x)}$

Matérn 5/2 kernel, constant trend – $\mu(x) = -0.2$ – $\theta \approx 0.02$ by MLE.
Not a good statistical model for these data.

Simple kriging (other illustration)

Since $\boldsymbol{Y}_{KS} \sim N\left(\boldsymbol{m}_{KS}, \boldsymbol{C}_{KS}\right)$, kriging trajectories can be sampled



Matérn 5/2 kernel, constant trend – $\mu(x) = -0.2$ – θ=0.5 fixed a priori.

Interpolation properties

- The kriging mean and trajectories are interpolating the data points
- The kriging variance is null at data points

Proof for v_{SK} : $C(x^i, \boldsymbol{X})^T = \boldsymbol{C^i}$, i-th column of $\boldsymbol{C}$

$\boldsymbol{C^{-1}C} = \boldsymbol{I} \Rightarrow \boldsymbol{C^{-1}C^i} = \boldsymbol{e^i}$, i-th basis vector

$$v_{SK}(x^i) = \sigma^2 - \boldsymbol{C^{iT}C^{-1}C^i} = \sigma^2 - \boldsymbol{C^{iT}}e^i = \sigma^2 - C(x^i, x^i) = 0$$

Same idea with $m_{SK}(x^i) = y_i$

$Y(x^i) | \boldsymbol{Y} = \boldsymbol{y} \sim N(y_i, 0)$, thus the trajectories are interpolating

Other paths towards kriging

We just used a Bayesian approach (GP conditioning) to justify kriging.

Often, kriging is introduced as a **best linear estimator** :

linear , $\widehat{Y}(x) = \sum_{i=1}^{n} \lambda_i(x) Y_i = \boldsymbol{\lambda}(x)^T \boldsymbol{Y}$

unbiased , $E\widehat{Y}(x) = \boldsymbol{\lambda}(x)^T E\boldsymbol{Y} = EY(x) = \mu(x)$

best , $\boldsymbol{\lambda}(x) = \underset{\boldsymbol{\lambda}\in\mathbb{R}^n}{arg\,min}\, E\left(\|\widehat{Y}(x) - Y(x)\|^2\right)$

This constrained optimization problem is solved in $\boldsymbol{\lambda}(x)$ and the kriging equations are recovered through

$m_K(x) = E(\widehat{Y}(x)|\boldsymbol{Y}=\boldsymbol{y})$ and $v_K(x) = E((\widehat{Y}(x) - Y(x))^2)$

But the link with GP interpretation is typically not discussed

$E(Y(x)|\boldsymbol{Y}=\boldsymbol{y}) \overset{?}{=} E(\widehat{Y}(x)|\boldsymbol{Y}=\boldsymbol{y})$, $E((Y(x)-\mu(x))^2|\boldsymbol{Y}=\boldsymbol{y}) \overset{?}{=} E((\widehat{Y}(x) - Y(x))^2)$

RKHS view : kriging as minimum norm interpolator in the Reproducing Kernel Hilbert Space generated by $C(.,.)$.
Cf. Nicolas Durrande's PhD thesis.

Universal kriging (1/4)

Finally, let us learn the **trend** within the same framework.

The UK statistical model is,

$$Y(x) = \sum_{i=1}^{p} a_i(x)\beta_i + Z(x) = \boldsymbol{a}(x)^T \boldsymbol{\beta} + Z(x)$$

where $Z(x) \sim N(0, C(x,x))$ and $\boldsymbol{\beta} \sim N(\boldsymbol{b}, \boldsymbol{B})$

i.e., Gaussian prior on the trend weights

Consider the Gaussian vector,

new points $\begin{Bmatrix} Y(x^{*1}) \\ \cdots \\ Y(x^{*m}) \end{Bmatrix}$

observation points $\begin{Bmatrix} Y(x^{1}) \\ \cdots \\ Y(x^{n}) \end{Bmatrix}$

$= \begin{Bmatrix} \boldsymbol{Y}^* \\ \boldsymbol{Y} \end{Bmatrix}$ and note $\begin{bmatrix} \boldsymbol{a}(x^1)^T \\ \cdots \\ \boldsymbol{a}(x^n)^T \end{bmatrix} = \boldsymbol{A}$

$\begin{bmatrix} \boldsymbol{a}(x^{*1})^T \\ \cdots \\ \boldsymbol{a}(x^{*n})^T \end{bmatrix} = \boldsymbol{A}^*$

Universal kriging (2/4)

$$\begin{Bmatrix} \boldsymbol{Y}^* \\ \boldsymbol{Y} \end{Bmatrix} \sim N\left(\begin{matrix} \boldsymbol{A}^* \boldsymbol{b} \\ \boldsymbol{A}\boldsymbol{b} \end{matrix} , \begin{bmatrix} \boldsymbol{C}^* + \boldsymbol{A}^* \boldsymbol{B} \boldsymbol{A}^{*T} & C(\boldsymbol{X}^*, \boldsymbol{X}) + \boldsymbol{A}^* \boldsymbol{B} \boldsymbol{A}^T \\ C(\boldsymbol{X}, \boldsymbol{X}^*) + \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{*T} & \boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T \end{bmatrix} \right)$$

and apply the Gaussian vector conditioning formula (see earlier slide). The kriging mean is the conditional average,

$$m(\boldsymbol{X}^*) = \boldsymbol{A}^* \boldsymbol{b} + \left(C(\boldsymbol{X}^*, \boldsymbol{X}) + \boldsymbol{A}^* \boldsymbol{B} \boldsymbol{A}^T \right) \left(\boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T \right)^{-1} \left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{b} \right)$$

and the kriging covariance is the conditional covariance,

$$v(\boldsymbol{X}^*) = \boldsymbol{C}^* + \boldsymbol{A}^* \boldsymbol{B} \boldsymbol{A}^{*T} - \left(C(\boldsymbol{X}^*, \boldsymbol{X}) + \boldsymbol{A}^* \boldsymbol{B} \boldsymbol{A}^T \right) \left(\boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T \right)^{-1} \left(C(\boldsymbol{X}, \boldsymbol{X}^*) + \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{*T} \right)$$

The universal kriging formula are obtained by taking the limit of $m(\boldsymbol{X}^*)$ and $v(\boldsymbol{X}^*)$ when the prior on the weights becomes non informative

$$\lambda \in \mathbb{R} \ , \ \boldsymbol{B} = \lambda \bar{\boldsymbol{B}} \ ,$$

$$\lim_{\substack{\lambda \to \infty \\ \boldsymbol{b} = 0}} m(\boldsymbol{X}^*) = m_{UK}(\boldsymbol{X}^*) \quad , \quad \lim_{\substack{\lambda \to \infty \\ \boldsymbol{b} = 0}} v(\boldsymbol{X}^*) = v_{UK}(\boldsymbol{X}^*)$$

Universal kriging (3/4)

Proof for $m_{UK}(.)$

Two matrix inversion lemma are used :

$$\left(A^{-1}+B^{-1}\right)^{-1} = A - A\left(A+B\right)^{-1}A \qquad (1)$$

$$\left(Z+UWV^{T}\right)^{-1} = Z^{-1} - Z^{-1}U\left(W^{-1}+V^{T}Z^{-1}U\right)^{-1}V^{T}Z^{-1} \qquad (2)$$

$$\left(\boldsymbol{C}+\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{T}\right)^{-1} \overset{(2)}{=} \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{A}\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}$$

$$\overset{(1)}{=} \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{A}\left[\left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1} - \left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\left(\left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\right]\boldsymbol{A}^{T}\boldsymbol{C}^{-1}$$

$$\Rightarrow \boldsymbol{A}^{*}\boldsymbol{B}\boldsymbol{A}^{T}\left(\boldsymbol{C}+\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{T}\right)^{-1}\boldsymbol{y} = \boldsymbol{A}^{*}\boldsymbol{B}\left(\left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{y}$$

$$\underset{\lambda\to\infty}{\to} \boldsymbol{A}^{*}\hat{\boldsymbol{\beta}} \qquad (3)$$

$$C\left(\boldsymbol{X}^{*},\boldsymbol{X}\right)\left(\boldsymbol{C}+\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{T}\right)^{-1}\boldsymbol{y} = C\left(\boldsymbol{X}^{*},\boldsymbol{X}\right)\left[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{A}\left(B^{-1}+\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}\right]\boldsymbol{y}$$

$$\underset{\lambda\to\infty}{\to} C\left(\boldsymbol{X}^{*},\boldsymbol{X}\right)\boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}\right) \qquad (4)$$

(3) + (4) yield $m_{UK}(.)$ □

Universal kriging (4/4)

$$m_{UK}(\boldsymbol{X}^*) = \underbrace{\boldsymbol{A}(\boldsymbol{X}^*)\hat{\boldsymbol{\beta}}}_{\text{linear part}} + \underbrace{C(\boldsymbol{X}^*, \boldsymbol{X})\boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})}_{\text{local correcting part}}$$

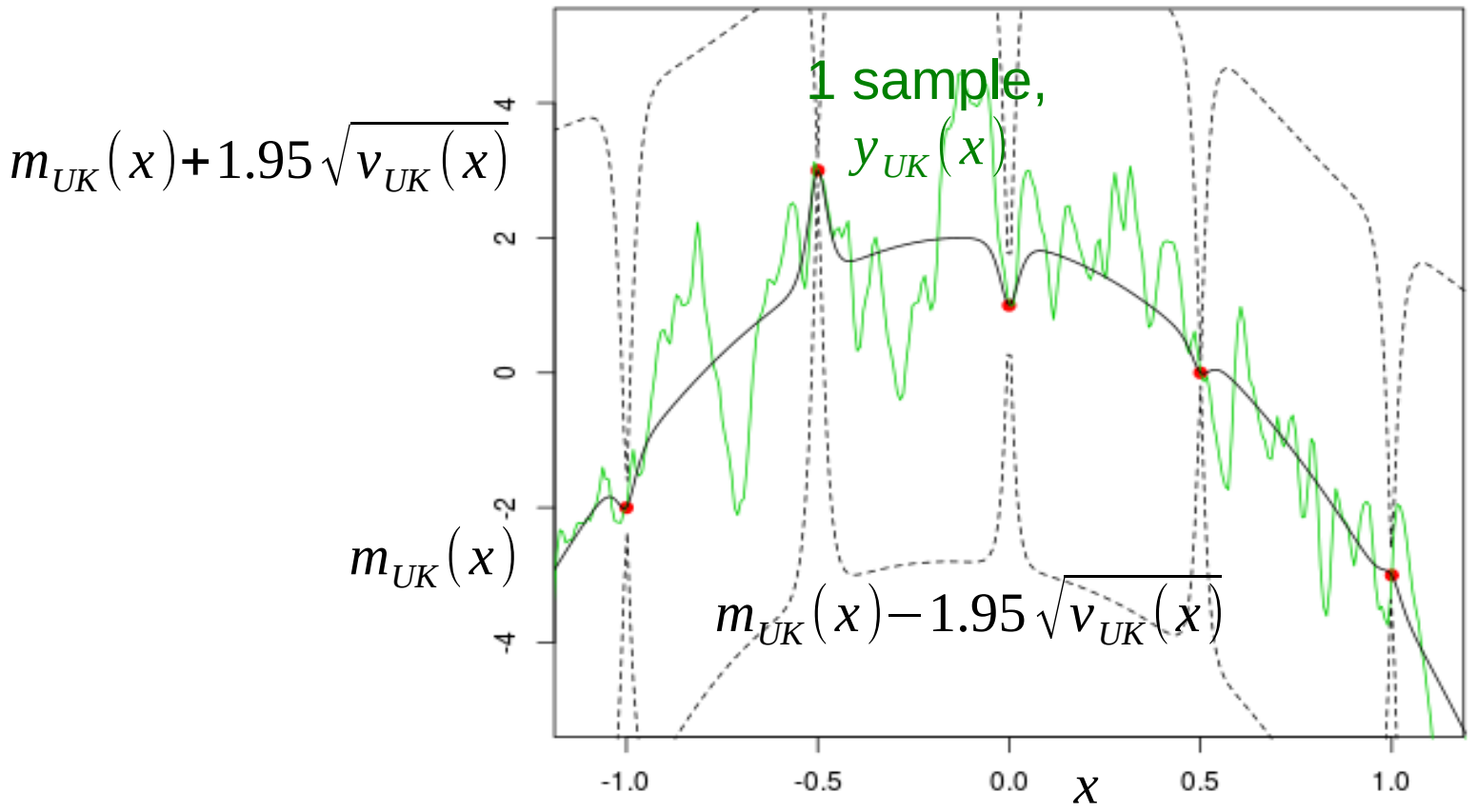
where $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{y}$

$$v_{UK}(\boldsymbol{X}^*) = \underbrace{\boldsymbol{C}^* - C(\boldsymbol{X}^*, \boldsymbol{X})\boldsymbol{C}^{-1} C(\boldsymbol{X}, \boldsymbol{X}^*)}_{= v_{SK}} + \underbrace{\boldsymbol{U}^T \left(\boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{U}}_{\text{addit. part due to } \beta \text{ estimation}}$$

where $\boldsymbol{U} = \boldsymbol{A}^T \boldsymbol{C}^{-1} C(\boldsymbol{X}, \boldsymbol{X}^*) - \boldsymbol{A}(\boldsymbol{X}^*)^T$

Note : **Ordinary kriging** = universal kriging with constant trend

Universal kriging (illustration)

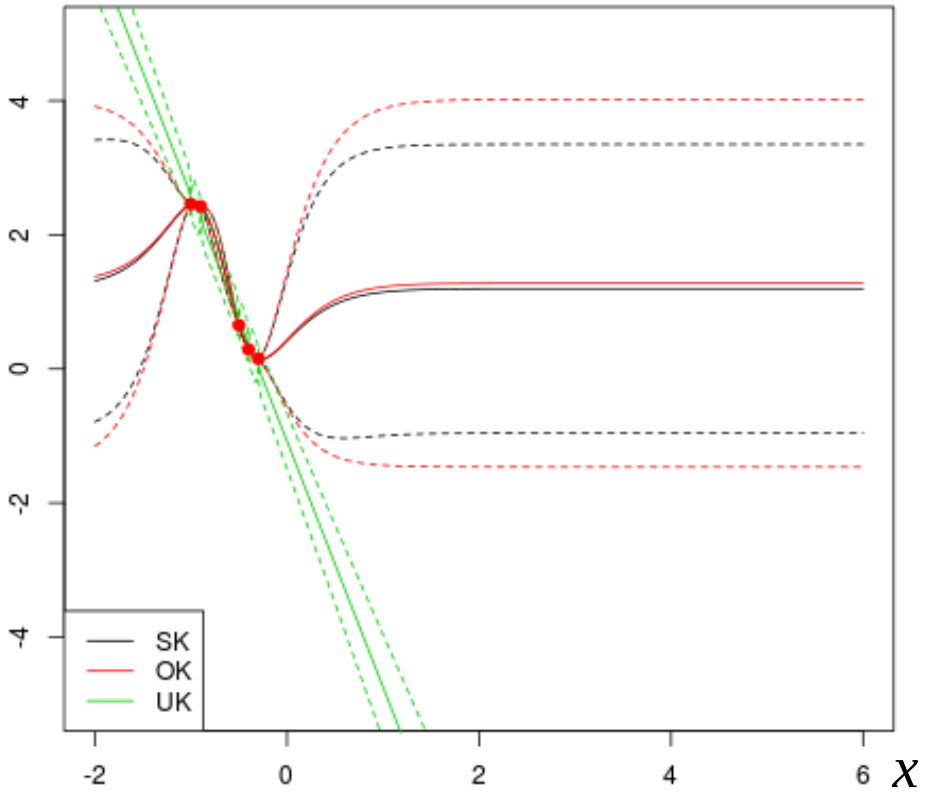


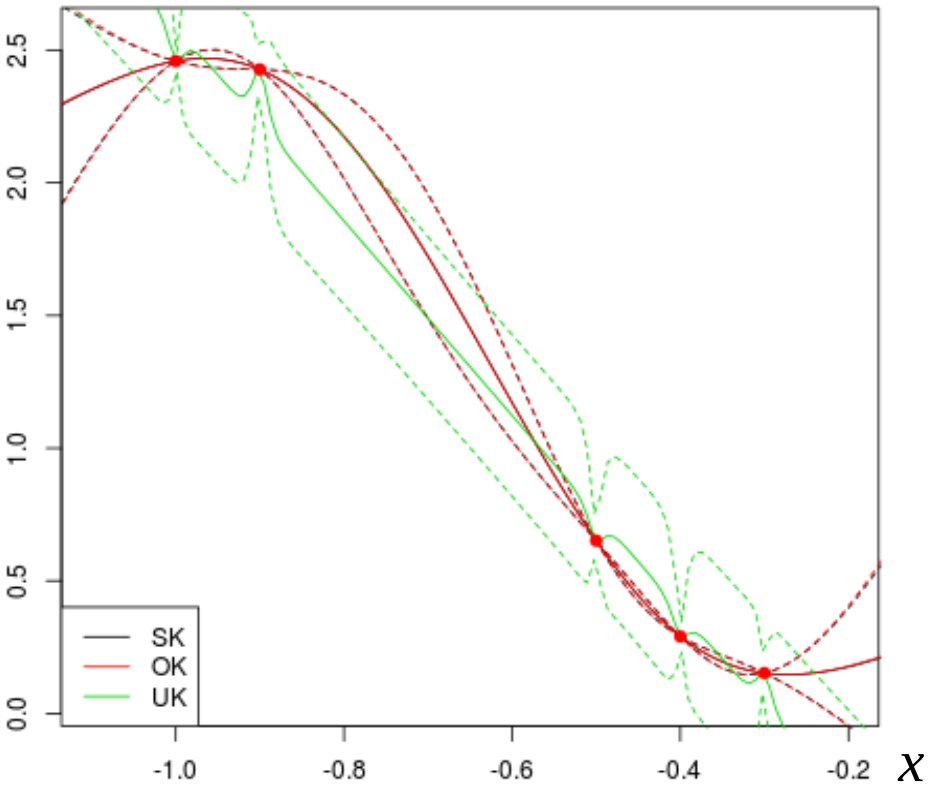
Matérn 5/2 kernel, quadratic trend – $\boldsymbol{a}(x)^T = (\,1\,,\,x\,,\,x^2\,)$ – $\theta \approx 0.02$ by MLE.

Illustrative comparison of krigings

Comparison of kriging mean ± 1.95 std. dev. for simple, ordinary and universal kriging (SK, OK, UK, resp.)

Matérn 5/2 kernel, linear trend – $\boldsymbol{a}(x)^T = (\,1\,,\,x\,)$ – for UK, MLE estimation





$\hat{\theta}_{SK} = 0.5$, $\hat{v}_{SK} = 1.2$; $\hat{\theta}_{OK} = 0.5$, $\hat{v}_{OK} = 1.2$; $\hat{\theta}_{UK} = 0.01$, $\hat{v}_{UK} = 0.02$

Note how SK and OK go back to average, how OK uncertainty slightly > that of SK , uncertainty of UK small because of small estimated σ

Kriging in more than 1D

Sample code from DiceKriging R package :

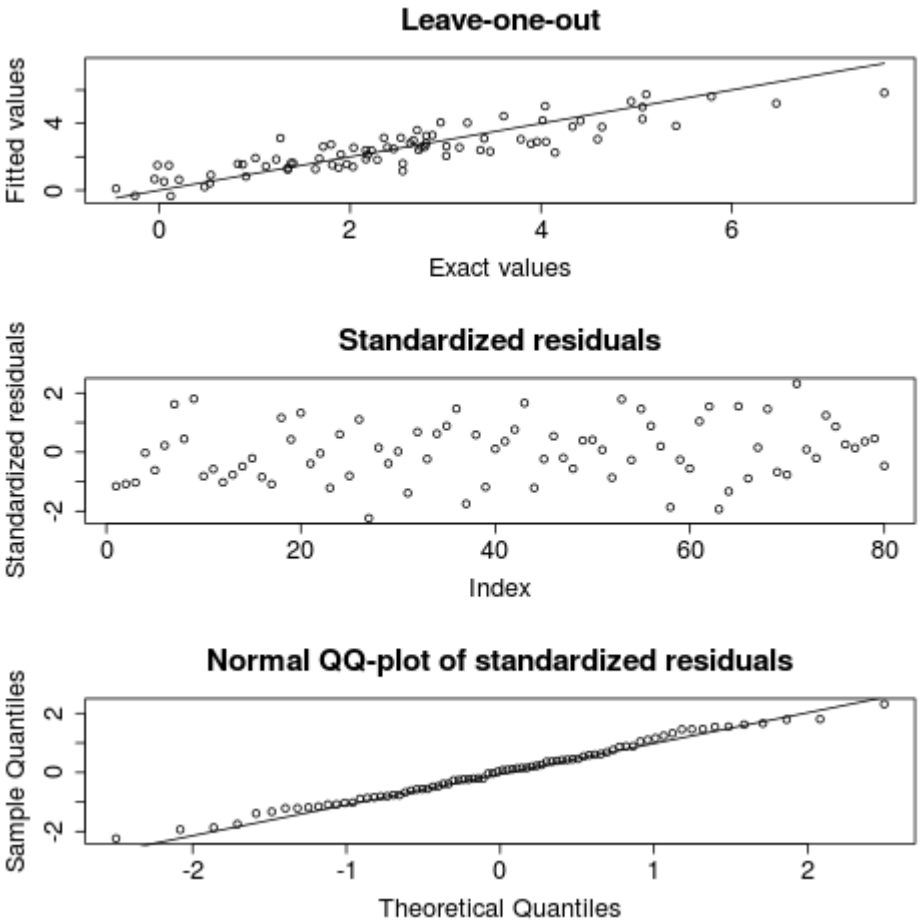
```
library(DiceKriging)
library(lhs)

n <- 80
d <- 6

X <- optimumLHS(n, d)
X <- data.frame(X)
y <- apply(X, 1, hartman6)

mlog <- km(design = X,
           response = -log(-y))

plot(mlog)
```



Rough sensitivity analysis from kriging

Estimated length scales (the θ's) can serve for rapid sensitivity analysis. The larger θ, the least function variation, the least sensitivity.

Expl. in 4D.

True function : $f(x) = 2 + x_1^2 - 3x_3 + x_1^3 x_3^2 + \sin(3x_1 x_4)$

```
n <- 80
d <- 4
X <- optimumLHS(n, d)
X <- data.frame(X)
y <- apply(X, 1, truefunc)
model <- km(design = X, response = y, upper=c(5,5,5,5))
```

$\Rightarrow$ $\hat{\theta}_1 = 1.2$, $\hat{\theta}_2 = 5$, $\hat{\theta}_3 = 2.9$, $\hat{\theta}_4 = 1.7$

one checks that the function is the least sensitive to x_2, the most to x_1 .

Links with other methods (1/2)

SVR

LS-SVR ~ m_K, cf. previous comments, but regularization control (C SVR parameter) embedded in the likelihood.

Bayesian linear regression

$$Y(x) = \sum_{i=1}^{N} W_i \phi_i(x) + \mathrm{E} = \boldsymbol{\phi}(x)^T \boldsymbol{W} + \mathrm{E}$$

where $\boldsymbol{\phi}(x) = \begin{bmatrix} \phi_1(x) \\ \cdots \\ \phi_N(x) \end{bmatrix}$, $\boldsymbol{W} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_w)$, $\mathrm{E} \sim N(0, \sigma_n^2)$

The posterior distribution of the model follows the SK equations,

$$(\boldsymbol{\phi}(x)^T \boldsymbol{W} \mid \boldsymbol{Y} = \boldsymbol{y}) \sim N(m_{SK}(x), v_{SK}(x))$$

provided $C(x, x') = \boldsymbol{\phi}(x)^T \boldsymbol{\Sigma}_w \boldsymbol{\phi}(x)$ (GPML, p.12)

$\Rightarrow$ kriging is equivalent to Bayesian linear regression in the N (possibly infinite) dimensional feature space $(\phi_1(x), \ldots, \phi_N(x))$

Links with other methods (2/2)

Bayesian linear regression (cont)

Case of the Gaussian kernel

The feature space associated to the Gaussian kernel is an infinite series of gaussians with varying centers

$$C(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\theta^2}\right) = \lim_{N \to \infty} \sum_{c=1}^{N} \frac{\sigma^2}{N} \phi_c(x) \phi_c(x')$$

where $\phi_c(x) = \exp\left(-\frac{\|x - c\|^2}{\theta^2}\right)$ (see GPML, p.84)

Kriging issues

C not invertible

If there are linear dependencies between the covariances of subset of data points. This happens in particular if data points are close to each other according to the cov function ← frequent with the very smooth Gaussian kernel.
Solution : **regularization**, either by adding a small >0 quantities on the diagonal or by replacing C^{-1} by the pseudo-inverse $C^{\dagger}$.
(cf. Mohammadi et al., 2013)

Too many data points

It is not standard to invert a matrix beyond 1000 data points.
Solution : regional kriging, still a research issue.

Choosing the covariance function

Max. likelihood is a multimodal optimization problem. More generally, how to choose the trend and covariance model ?

Bibliography

O'Hagan, A., *Bayesian analysis of computer code outputs: a tutorial*, Reliability Engineering and System Safety, 91, pp.1290-1300, 2006.

Mohammadi, H., Le Riche, R., Touboul, E and Bay, X., An analytic comparison of regularization methods for Gaussian processes, submitted to J. of multivariate statistics, 2014.

Rasmussen, C.E., and Williams, C.K.I., *Gaussian Processes for Machine Learning*, (GPML), the MIT Press, 2006.

Roustant, O., Ginsbourger, D. and Deville, Y., *DiceKriging, DiceOptim: Two R Packages for the analysis of Computer Experiments by Kriging-based Metamodeling and Optimization*, J. of Statistical Software, 51(1), pp.1-55, 2012.

Compatibility note

From now on, Bruno Sudret will write

$$m_K(x) \rightarrow \mu_{\hat{Y}}(\boldsymbol{x})$$
$$v_K(x) \rightarrow \sigma^2_{\hat{Y}}(\boldsymbol{x})$$