



## Introduction to kriging

Rodolphe Le Riche

### ► To cite this version:

| Rodolphe Le Riche. Introduction to kriging. Doctoral. France. 2014. <cel-01081304>

**HAL Id: cel-01081304**

<https://hal.archives-ouvertes.fr/cel-01081304>

Submitted on 7 Nov 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# **Introduction to kriging**

**Rodolphe Le Riche<sup>1</sup>**

**<sup>1</sup> 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

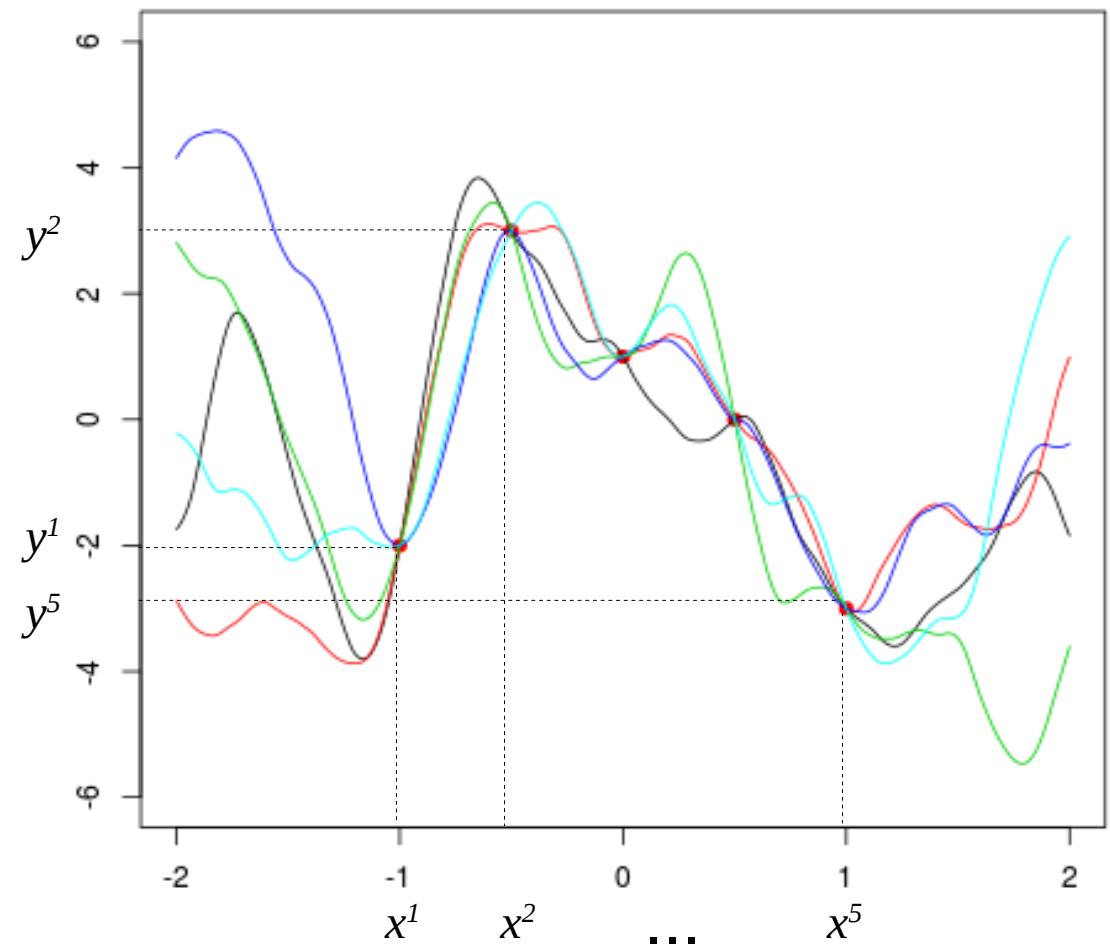
# Kriging : introduction

Context : scalar measurements  $(y_1, \dots, y_n)$   
at  $n$  positions  $(x^1, \dots, x^n)$  in a  $d$ -dimensional space  $X \subset \mathbb{R}^d$

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  $\leq 3$ ,  $y = 1$   
 $3 < \text{dice} \leq 5$ ,  $y = 2$   
 $\text{dice} = 6$ ,  $y = 3$

## Random process, $Y(x)$

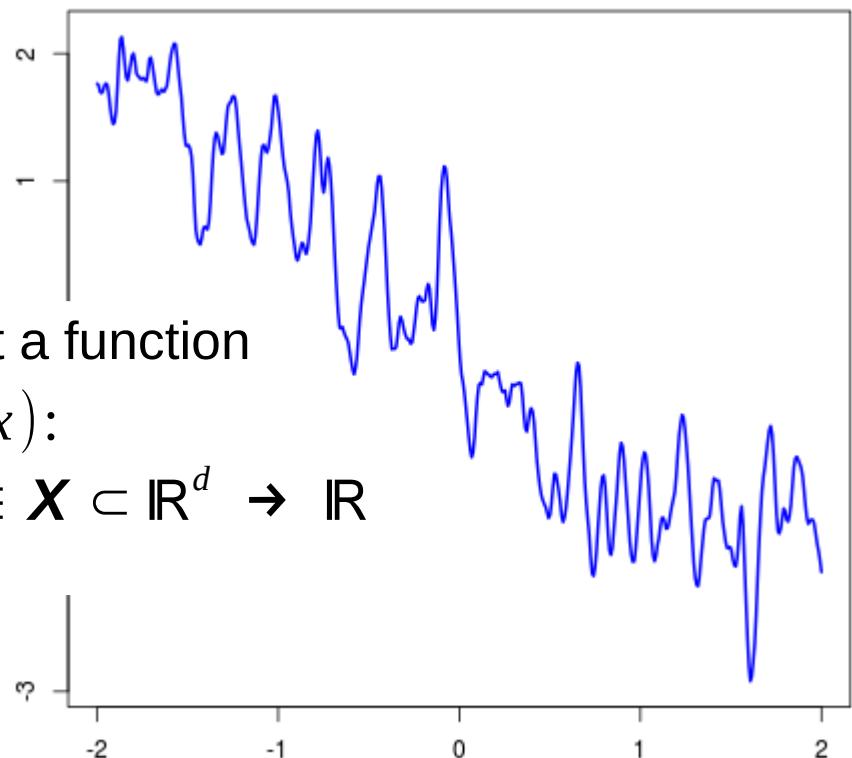
A set of RVs indexed by  $x$

random event  
 $\omega \in \Omega$

( e.g., wheather )



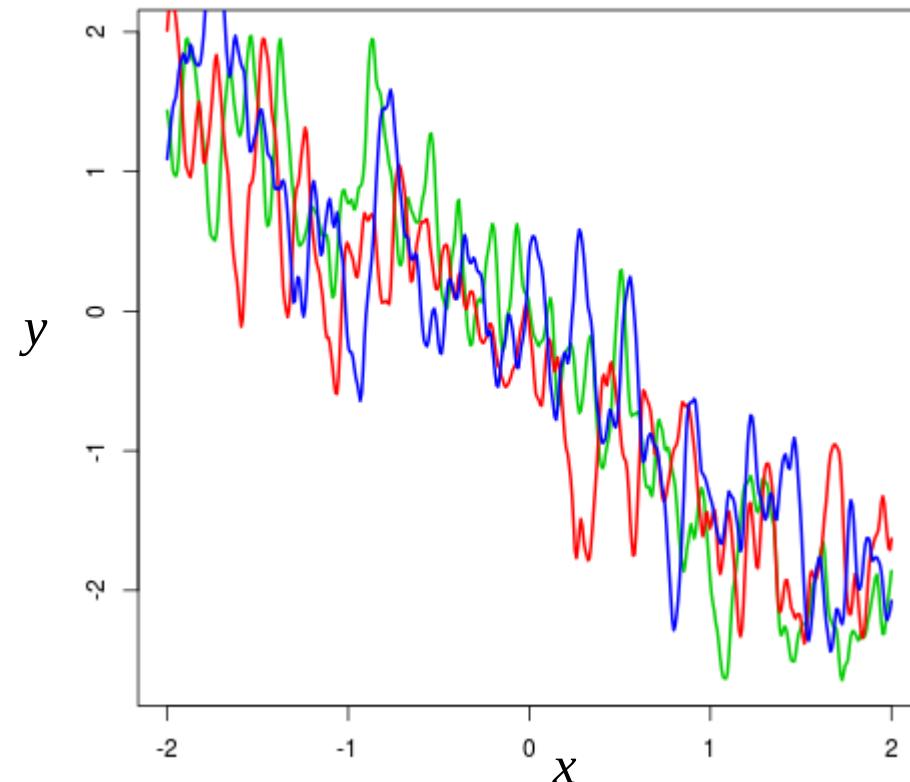
get a function  
 $y(x)$ :  
 $x \in X \subset \mathbb{R}^d \rightarrow \mathbb{R}$



# 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**

$$Y(x) \sim N(\mu(x), C(x, x))$$

and,

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

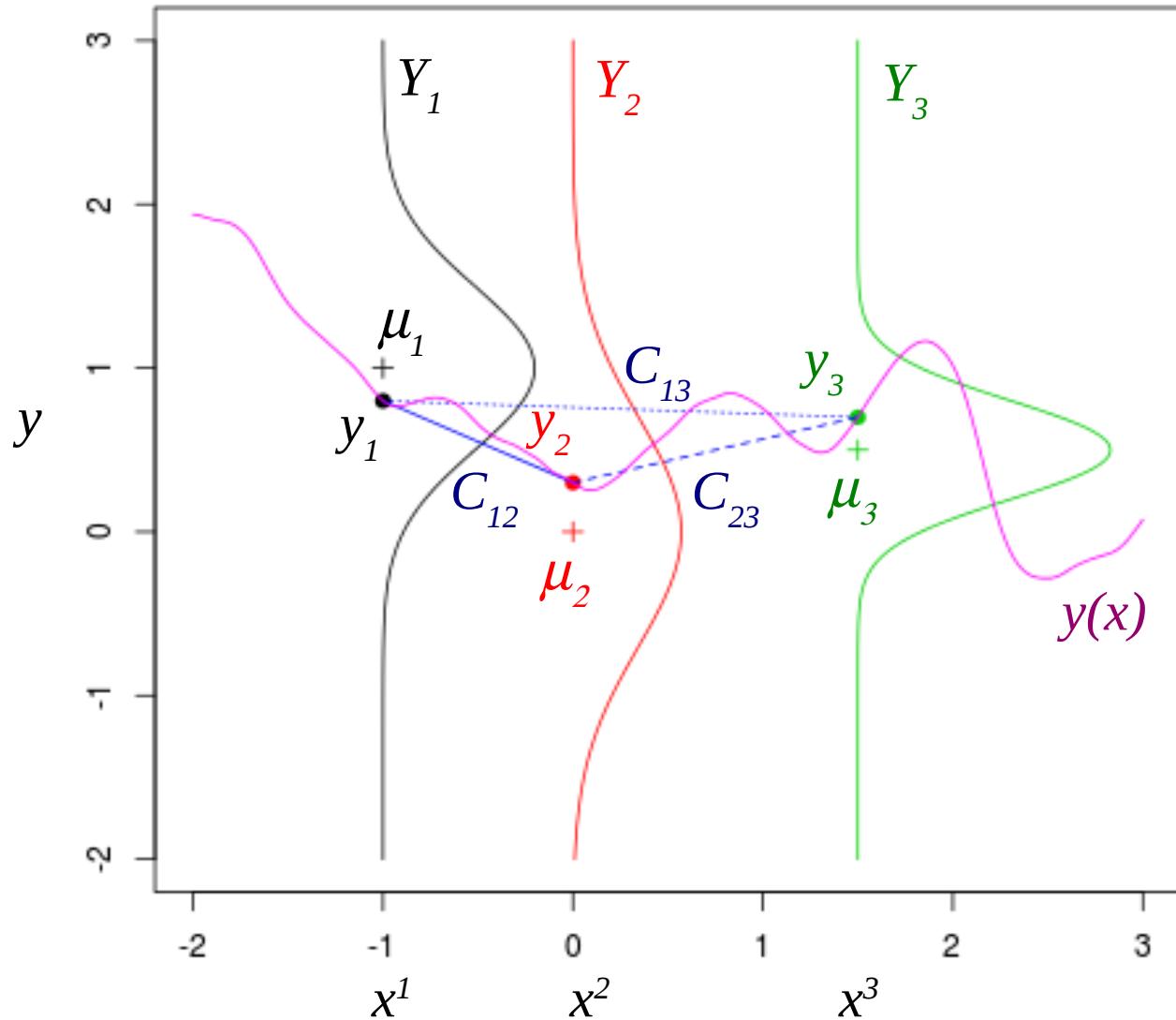
with probability density function (multi-Normal law),

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

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

---

# Gaussian processes (illustration)



$$n = 3$$

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

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

## Special case : no spatial covariance

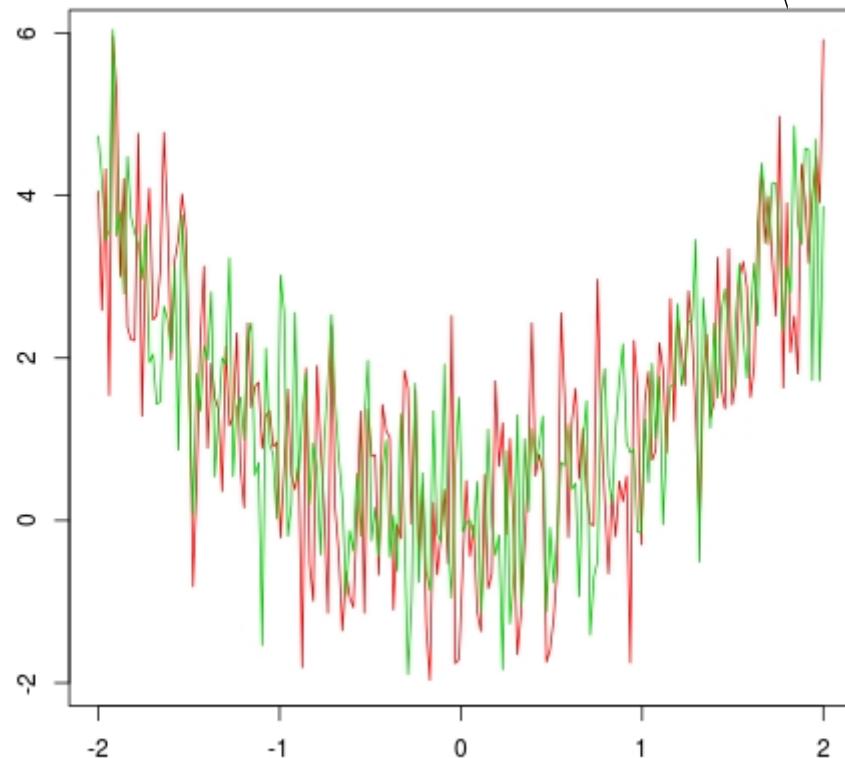
---

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

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

At  $n$  observation points,

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



Example :

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

## Numerical sampling of a GP

---

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

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

( prove that  $E \mathbf{Y} = \boldsymbol{\mu}$  and  $\text{Cov}(\mathbf{Y}) = \mathbf{C}$  ,  
+ 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  $C$  and  $\mu$  and account for the data points ( $x^i, y_i$ ) ?

→ Start with  $C$  .

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

$$\text{Cov}(Y(x), Y(x')) = C(x, x')$$

is only a function of  $x$  and  $x'$ ,

$$C : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

The kernel defines the covariance matrix through

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

# Valid covariance functions

---

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

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

$$\forall u \in \mathbb{R}^n, \quad u^T \mathbf{C} u \geq 0$$

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

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

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

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

# Stationary covariance functions

---

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

$$\begin{aligned}\text{Cov}(Y(x), Y(x')) &= C(x, x') = C(x - x') = C(\tau) \\ C(0) &= \sigma^2, \quad C(\tau) = \sigma^2 R(\tau) \quad (R \text{ the correlation})\end{aligned}$$

Example : the squared exponential covariance function (Gaussian)

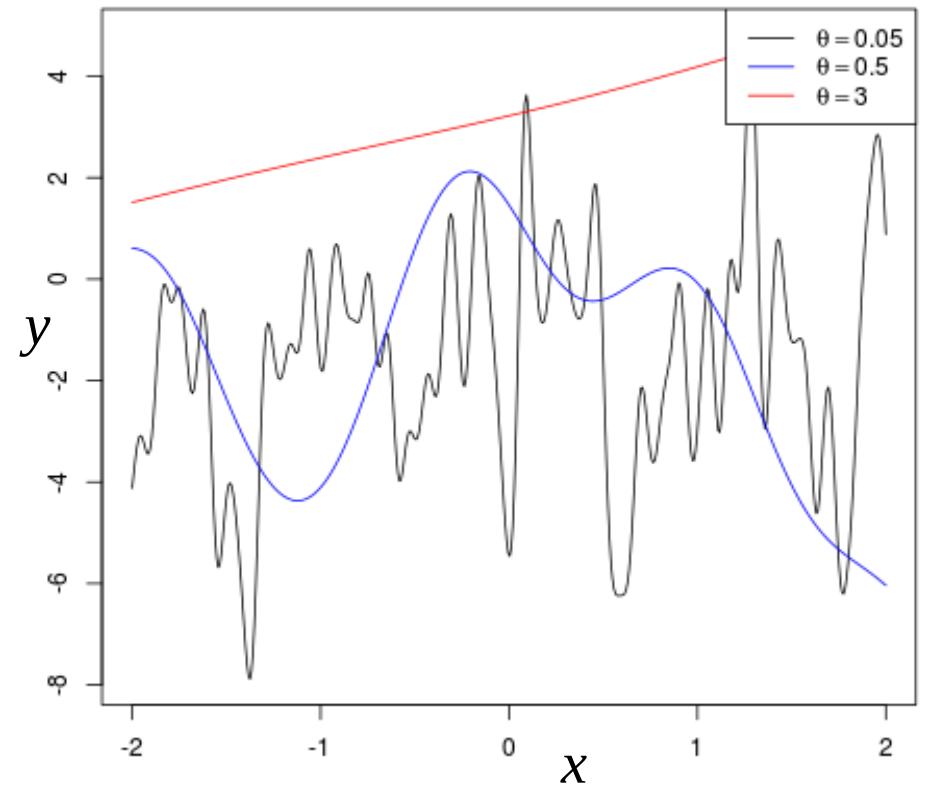
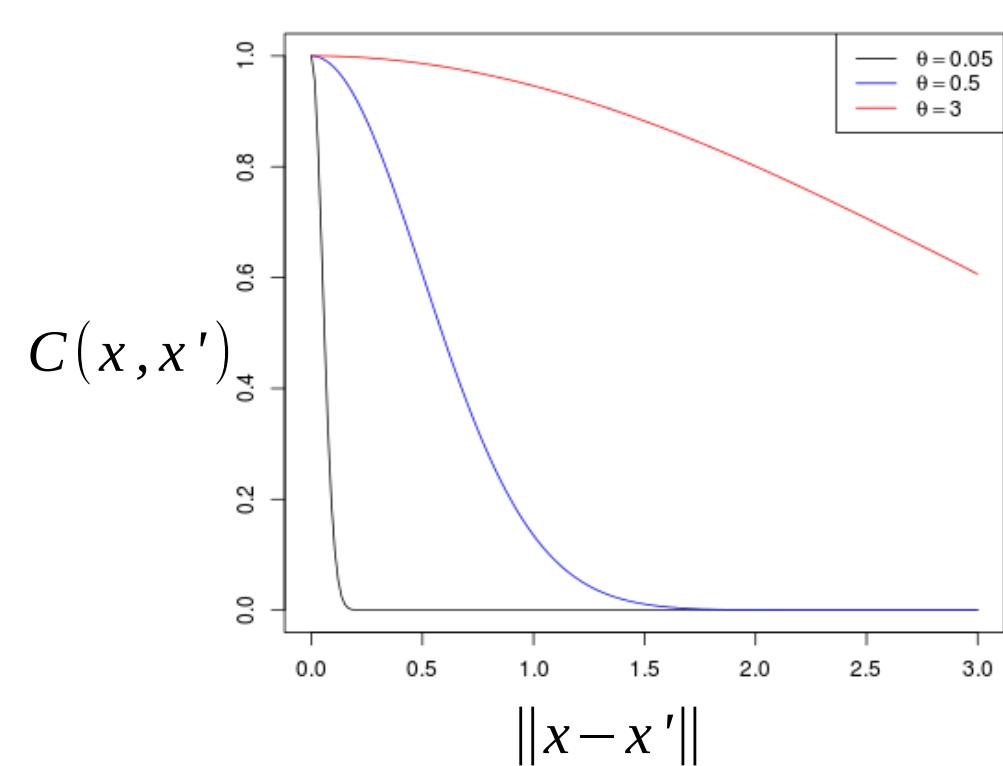
$$\begin{aligned}\text{Cov}(Y(x), Y(x')) &= C(x - x') = \sigma^2 \exp\left(-\sum_{i=1}^d \frac{|x_i - x'_i|^2}{2\theta_i^2}\right) \\ &= \sigma^2 \prod_{i=1}^d \exp\left(-\frac{|\tau_i|^2}{2\theta_i^2}\right)\end{aligned}$$

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

---

# Gaussian kernel, illustration

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



The regularity and frequency content of the trajectories is controlled by the covariance functions. The  $\theta_i$ '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,

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

# 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 :

$\sigma$  and the  $\theta_i$ 's to learn from data.

# Tuning of hyperparameters

---

Our first use of the observations  $(X, Y)$

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

Three paths to selecting hyperparameters ( $\sigma$  and the  $\theta_i$ '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

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

where  $\mathbf{C}_{ij} = C(x^i, x^j ; \sigma, \theta) = \sigma^2 R(x^i, x^j ; \theta)$   
and  $\mu_i = \mu(x^i)$

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

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

---

## Maximum of likelihood estimate (2/2)

---

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

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

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

So, numerically, minimize the “concentrated” likelihood,

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

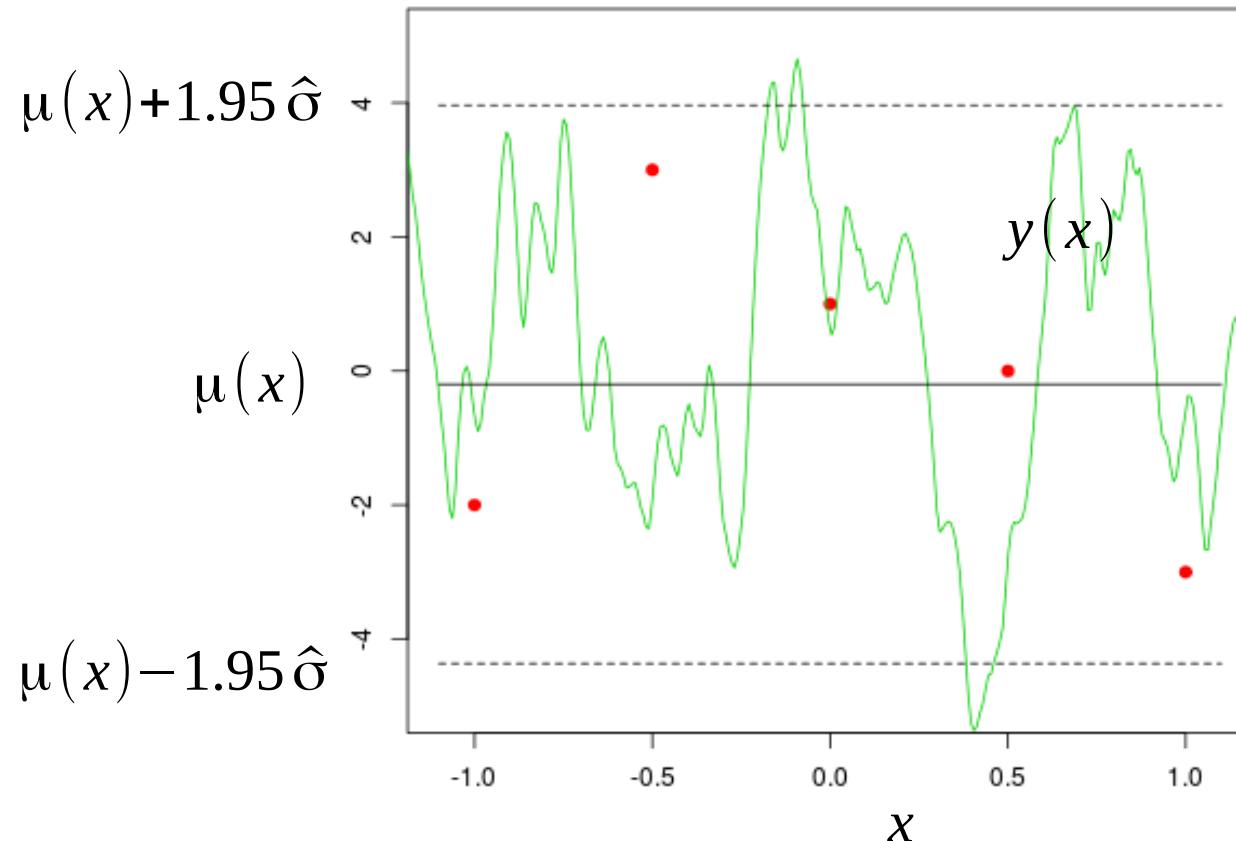
where  $\theta^{\min} > 0$

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

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

- ✓ 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

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

which directly yields the simple kriging\* formula

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

$$m_{SK}(x^*) = \mu(x^*) + C(x^*, X)C^{-1}(y - \mu) \quad (\text{identical to LS-SVR formula?})$$

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

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

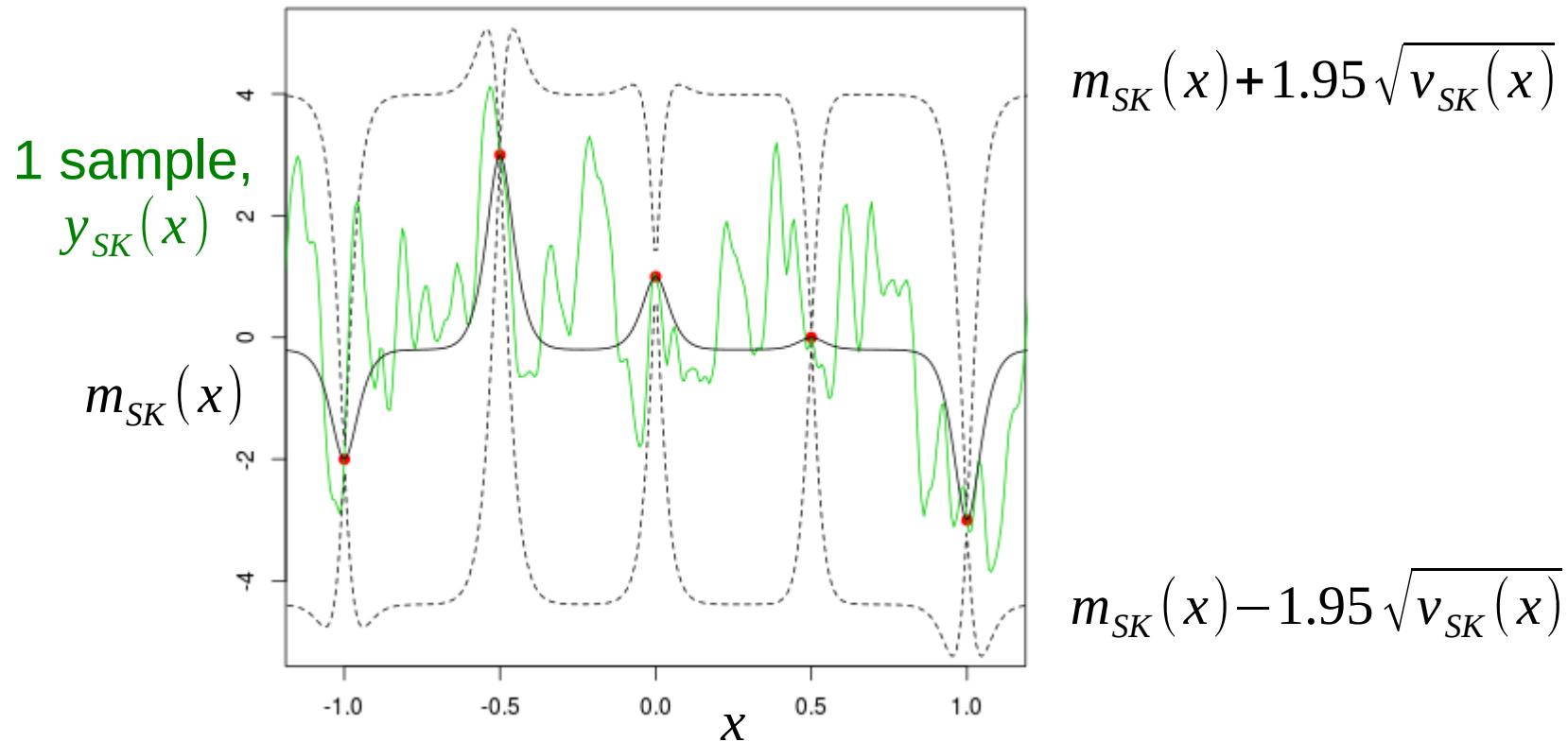
For prediction at 1 point,  $C(x^*, x^*) = \sigma^2 R(x^* - x^*) = \sigma^2$

For joined prediction at many points, same formula but

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

---

# Simple kriging (illustration)

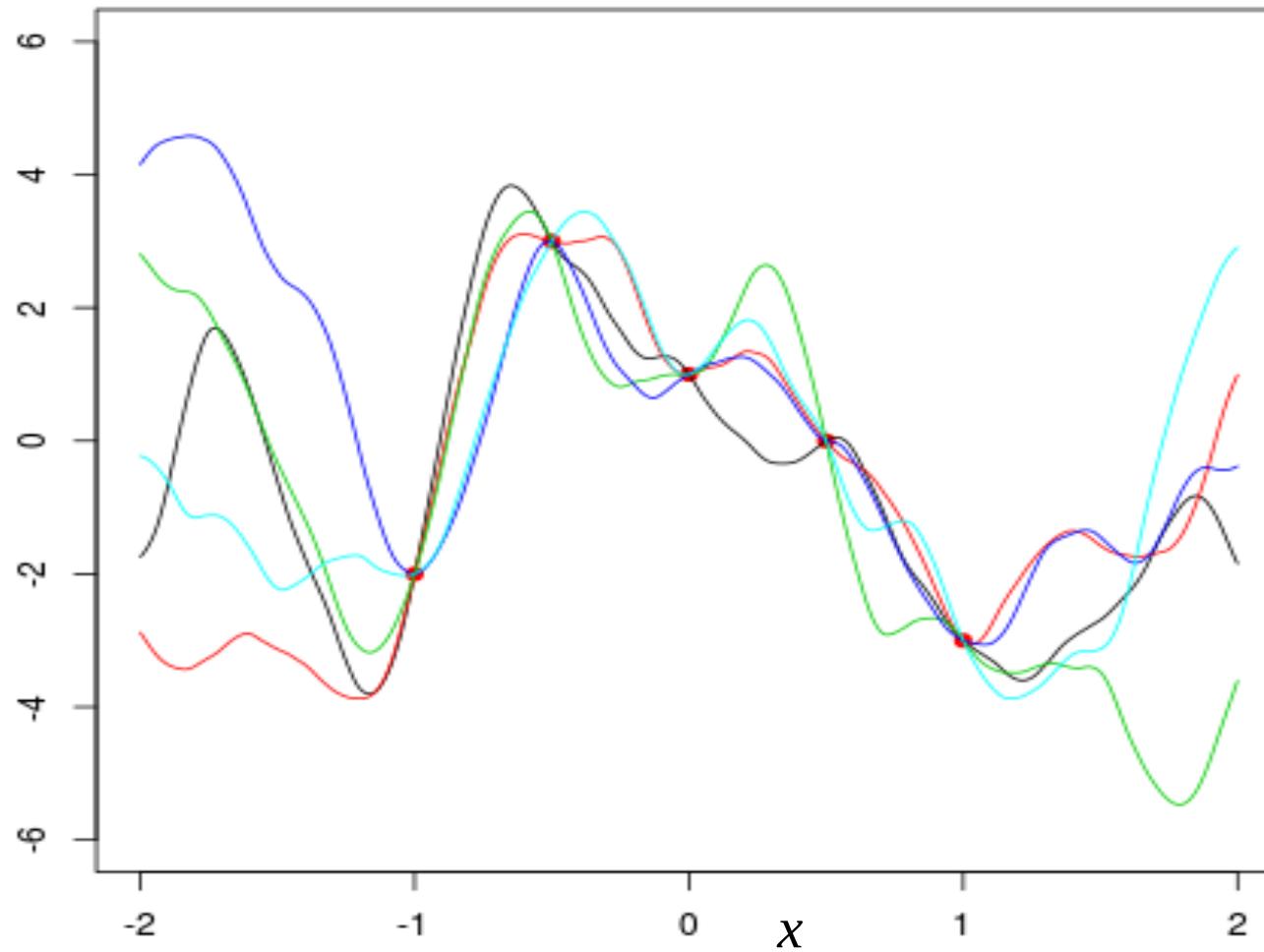


Matérn 5/2 kernel, constant trend –  $\mu(x) = -0.2$  –  $\theta \approx 0.02$  by MLE.  
Not a good statistical model for these data.

## Simple kriging (other illustration)

---

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



Matérn 5/2 kernel, constant trend –  $\mu(x) = -0.2$  –  $\theta=0.5$  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  $v_{SK}$  :  $C(x^i, X)^T = C^i$ ,  $i$ -th column of  $C$

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

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

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

$Y(x^i) | Y=y \sim N(y_i, 0)$ , 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** :

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

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

$$\text{best , } \boldsymbol{\lambda}(x) = \arg \min_{\boldsymbol{\lambda} \in \mathbb{R}^n} E(\|\hat{Y}(x) - Y(x)\|^2)$$

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

$$m_K(x) = E(\hat{Y}(x)|\mathbf{Y}=\mathbf{y}) \text{ and } v_K(x) = E((\hat{Y}(x)-Y(x))^2)$$

But the link with GP interpretation is typically not discussed

$$E(Y(x)|\mathbf{Y}=\mathbf{y}) \stackrel{?}{=} E(\hat{Y}(x)|\mathbf{Y}=\mathbf{y}), E((Y(x)-\mu(x))^2|\mathbf{Y}=\mathbf{y}) \stackrel{?}{=} E((\hat{Y}(x)-Y(x))^2)$$

**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,

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

where  $Z(x) \sim N(0, C(x, x))$  and  $\boldsymbol{\beta} \sim N(\mathbf{b}, \mathbf{B})$   
i.e., Gaussian prior on the trend weights

Consider the Gaussian vector,

new points  $\begin{pmatrix} Y(x^{*1}) \\ \dots \\ Y(x^{*m}) \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^* \end{pmatrix}$  and note  $\begin{bmatrix} \mathbf{a}(x^1)^T \\ \dots \\ \mathbf{a}(x^n)^T \end{bmatrix} = \mathbf{A}$

observation points  $\begin{pmatrix} Y(x^1) \\ \dots \\ Y(x^n) \end{pmatrix}$   $\begin{bmatrix} \mathbf{a}(x^{*1})^T \\ \dots \\ \mathbf{a}(x^{*n})^T \end{bmatrix} = \mathbf{A}^*$

## Universal kriging (2/4)

---

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

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

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

and the kriging covariance is the conditional covariance,

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

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

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

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

# Universal kriging (3/4)

---

Proof for  $m_{UK}(.)$

Two matrix inversion lemma are used :

$$(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A \quad (1)$$

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

$$\begin{aligned} (\mathbf{C} + \mathbf{ABA}^T)^{-1} &\stackrel{(2)}{=} \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{A}(\mathbf{B}^{-1} + \mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}^{-1} \\ &\stackrel{(1)}{=} \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{A}[(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1} - (\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}((\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1} + \mathbf{B})^{-1}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}]\mathbf{A}^T\mathbf{C}^{-1} \\ \Rightarrow \mathbf{A}^* \mathbf{B} \mathbf{A}^T (\mathbf{C} + \mathbf{ABA}^T)^{-1} \mathbf{y} &= \mathbf{A}^* \mathbf{B} ((\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1} + \mathbf{B})^{-1} (\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1} \mathbf{A}^T\mathbf{C}^{-1} \mathbf{y} \\ &\xrightarrow[\lambda \rightarrow \infty]{} \mathbf{A}^* \hat{\beta} \end{aligned} \quad (3)$$

$$\begin{aligned} C(\mathbf{X}^*, \mathbf{X})(\mathbf{C} + \mathbf{ABA}^T)^{-1} \mathbf{y} &= C(\mathbf{X}^*, \mathbf{X})[\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{A}(\mathbf{B}^{-1} + \mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}^{-1}] \mathbf{y} \\ &\xrightarrow[\lambda \rightarrow \infty]{} C(\mathbf{X}^*, \mathbf{X})\mathbf{C}^{-1}(\mathbf{y} - \mathbf{A}\hat{\beta}) \end{aligned} \quad (4)$$

(3) + (4) yield  $m_{UK}(.)$

□

## Universal kriging (4/4)

---

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

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

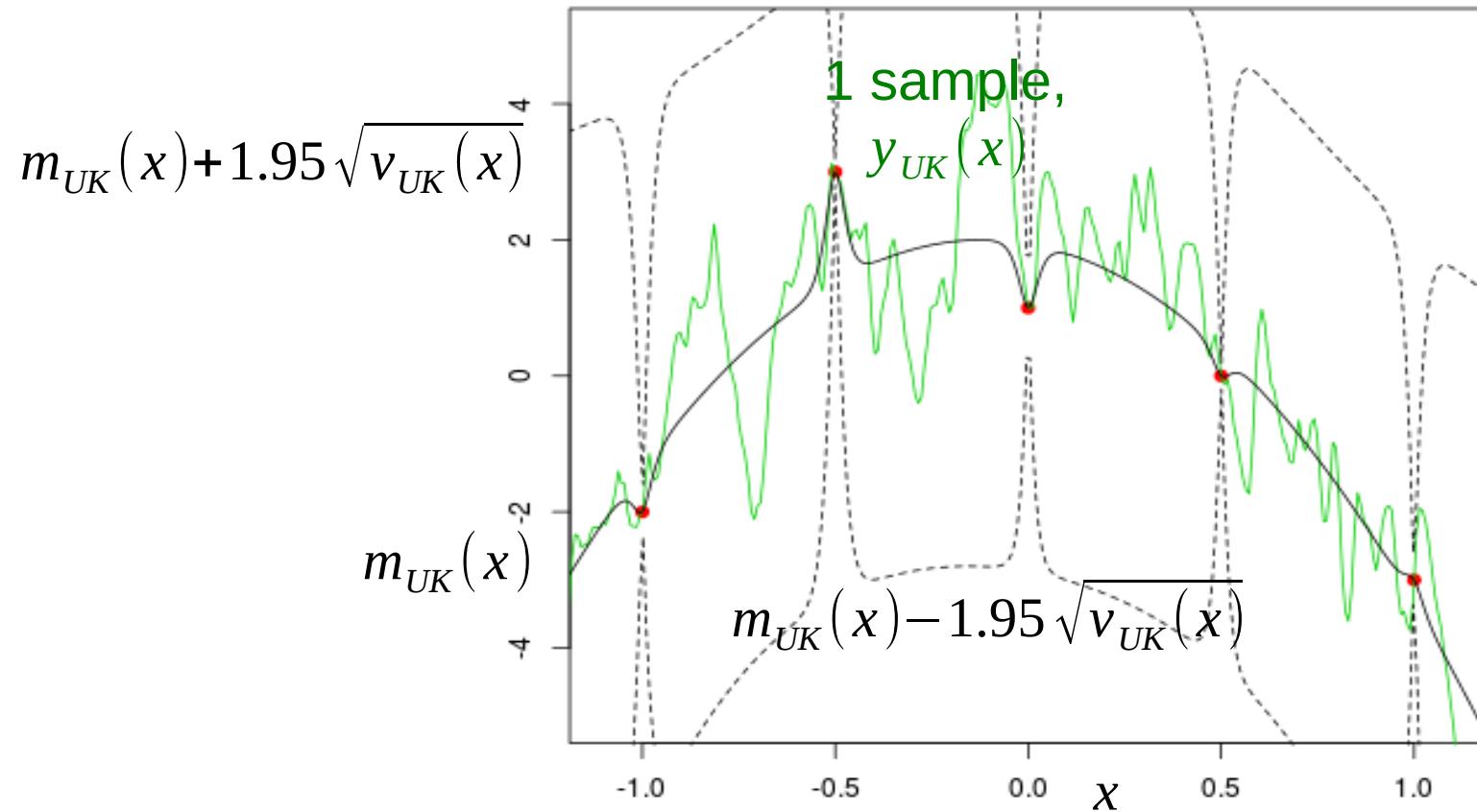
$$\begin{aligned} v_{UK}(X^*) &= \underbrace{C^* - C(X^*, X)C^{-1}C(X, X^*)}_{= v_{SK}} + \underbrace{U^T (A^T C^{-1} A)^{-1} U}_{\text{addit. part due to } \beta \text{ estimation}} \\ &= v_{SK} \end{aligned}$$

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

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

---

# Universal kriging (illustration)

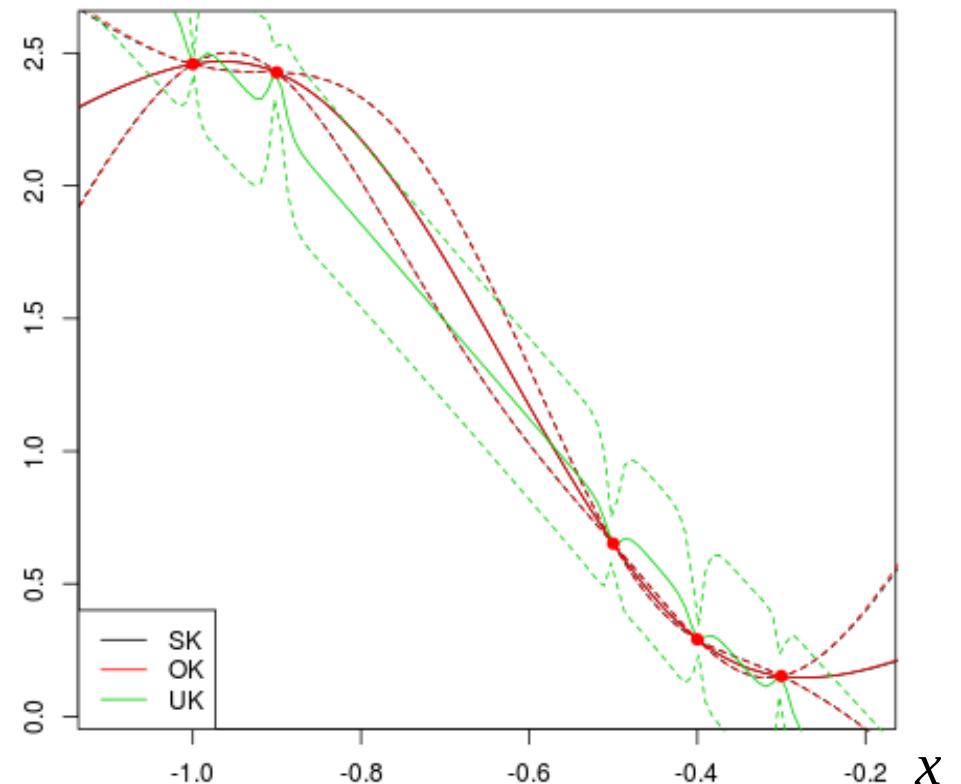
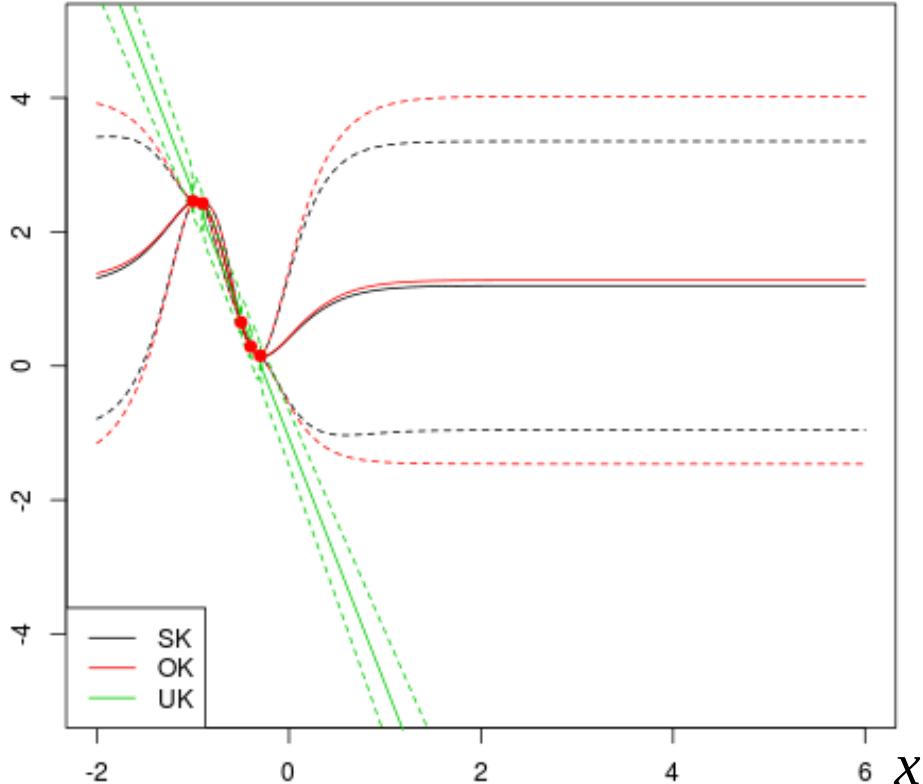


Matérn 5/2 kernel, quadratic trend –  $a(x)^T = (1, x, x^2)$  –  $\theta \approx 0.02$  by MLE.

# Illustrative comparison of krigings

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

Matérn 5/2 kernel, linear trend –  $a(x)^T = (1, x)$  – 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  $\sigma$

# Kriging in more than 1D

Sample code from DiceKriging R package :

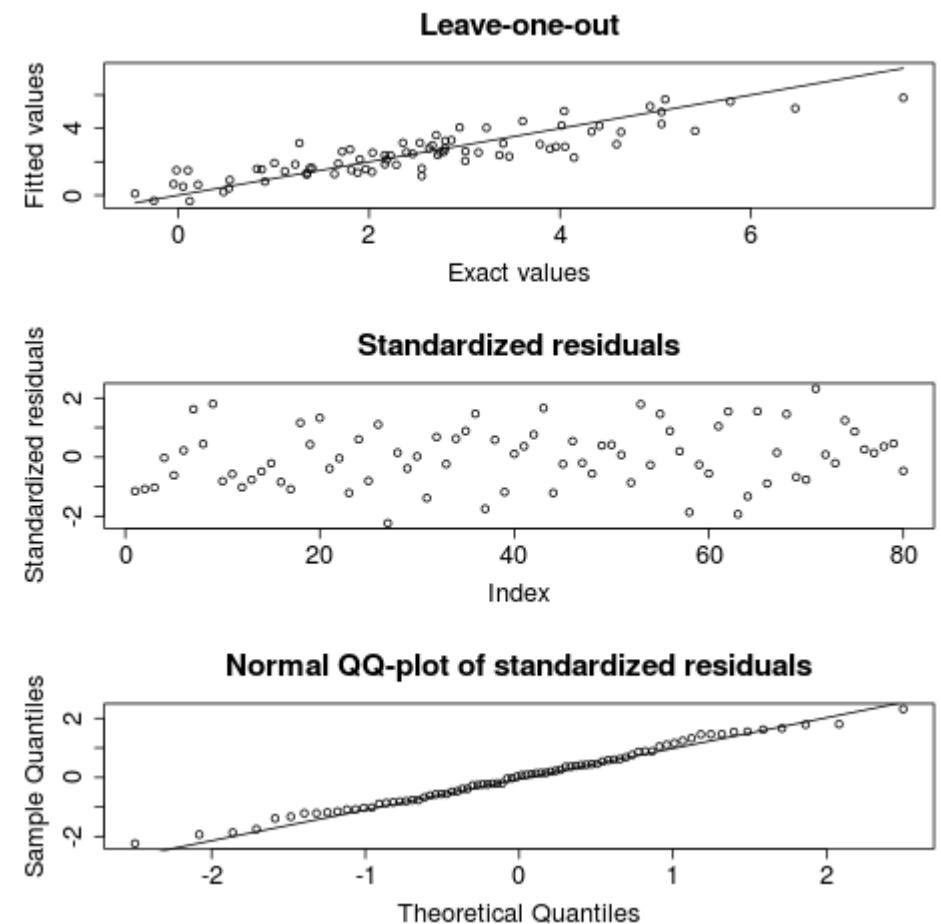
```
library(DiceKriging)
library(lhs)

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  $\theta$ 's) can serve for rapid sensitivity analysis. The larger  $\theta$ , 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))
```

$$\begin{aligned}\hat{\theta}_1 &= 1.2 \\ \Rightarrow \hat{\theta}_2 &= 5 \\ \hat{\theta}_3 &= 2.9 \\ \hat{\theta}_4 &= 1.7\end{aligned}$$

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

## SVR

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

## Bayesian linear regression

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

$$\text{where } \phi(x) = \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_N(x) \end{bmatrix}, \quad W \sim N(\mathbf{0}, \Sigma_w), \quad E \sim N(0, \sigma_n^2)$$

The posterior distribution of the model follows the SK equations,

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

provided  $C(x, x') = \phi(x)^T \Sigma_w \phi(x')$  ( GPML, p.12 )

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

### 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

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

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  $\leftarrow$  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

$$\begin{aligned} m_K(x) &\rightarrow \mu_{\hat{Y}}(\mathbf{x}) \\ v_K(x) &\rightarrow \sigma_{\hat{Y}}^2(\mathbf{x}) \end{aligned}$$