Intro to Gaussian Processes

MA + AB (+help from MK)

La Serena School for Data Science 2009

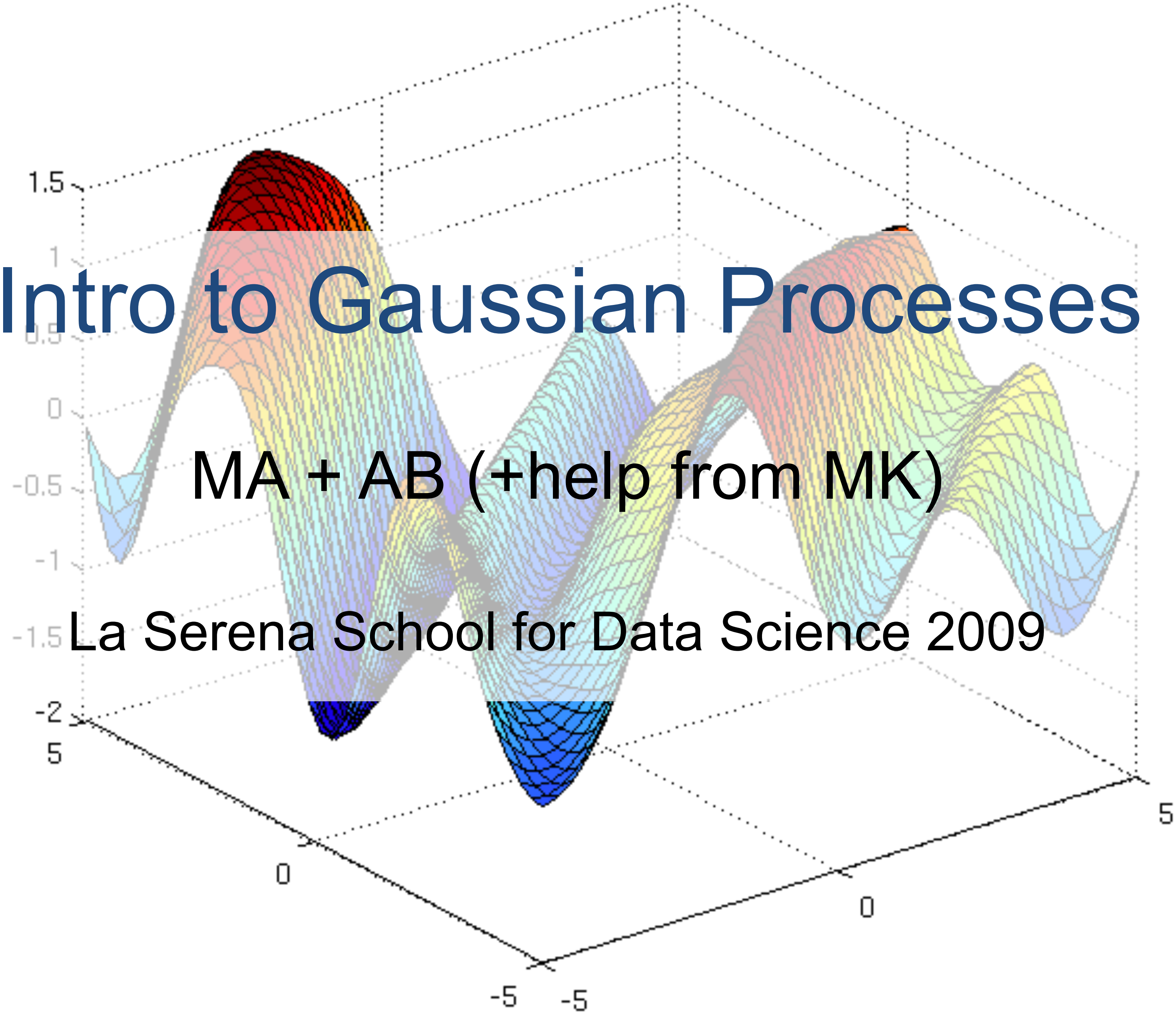
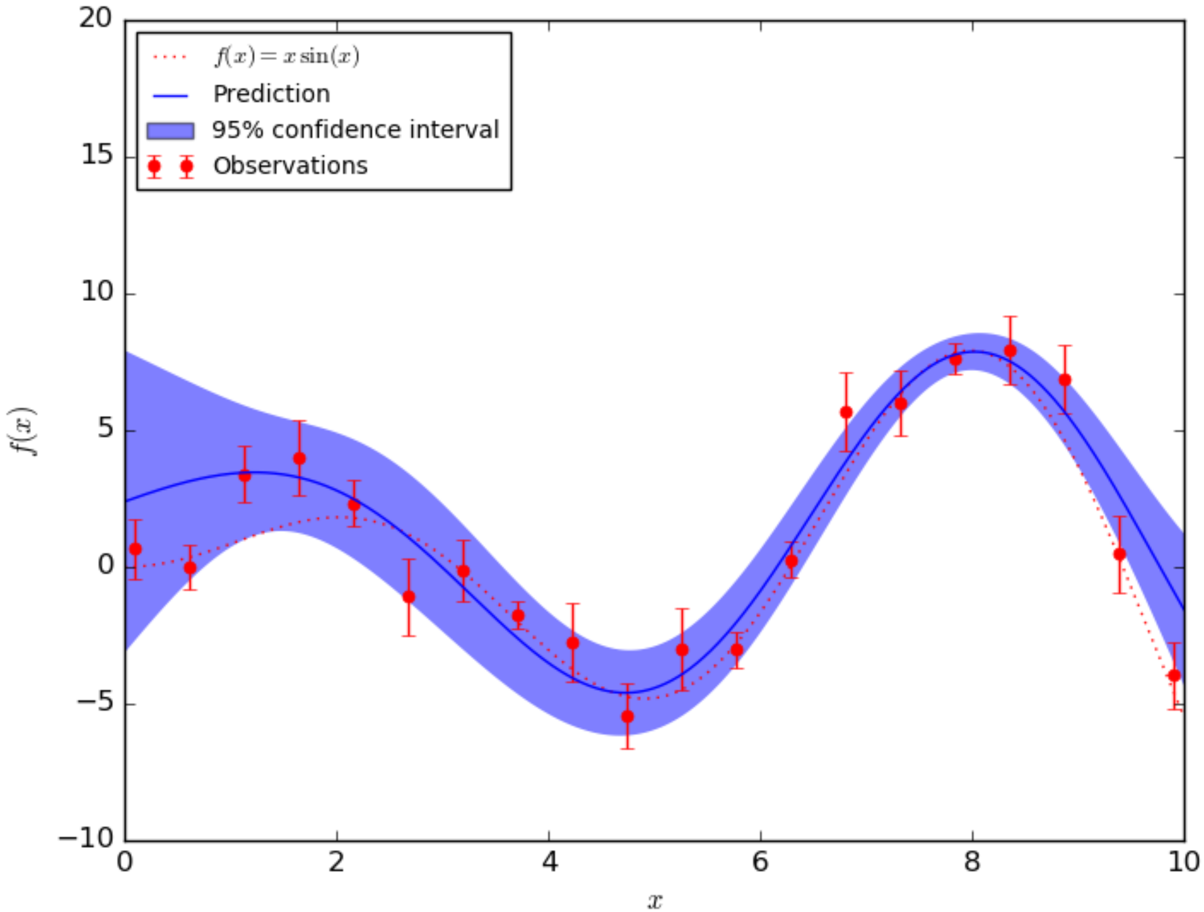


Image: The Kernel Cookbook by David Duvenaud

Motivation GP: non-parametric regression method

Stochastic process: generalization of pdf to “functions of pdfs” (if the process is “Gaussianly behaved (in time)” things “become easy”



Bridging the gap between:

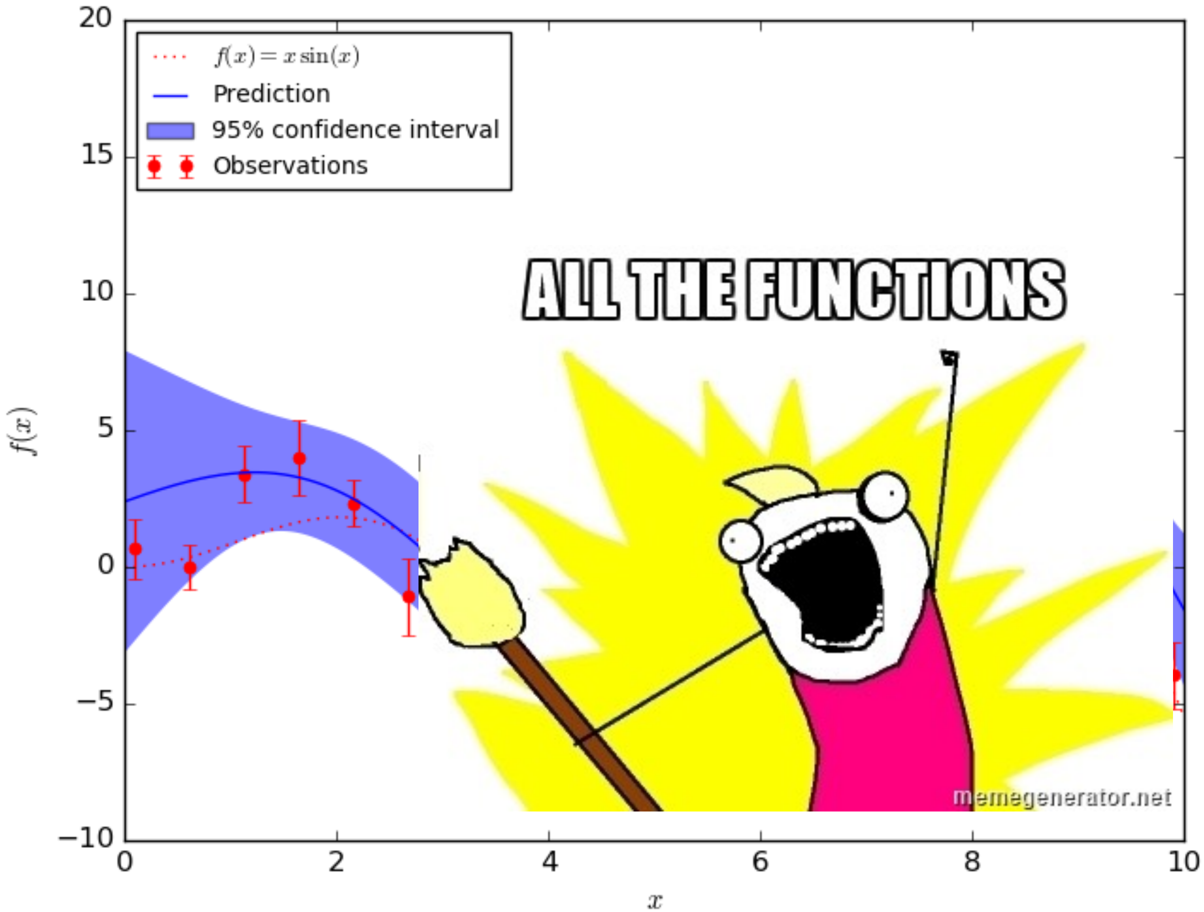
understanding data relationships

and

making predictions

Motivation GP: non-parametric regression method

Stochastic process: generalization of pdf to “functions of pdfs” (if the process is “Gaussianly behaved (in time)” things “become easy”



Bridging the gap between:

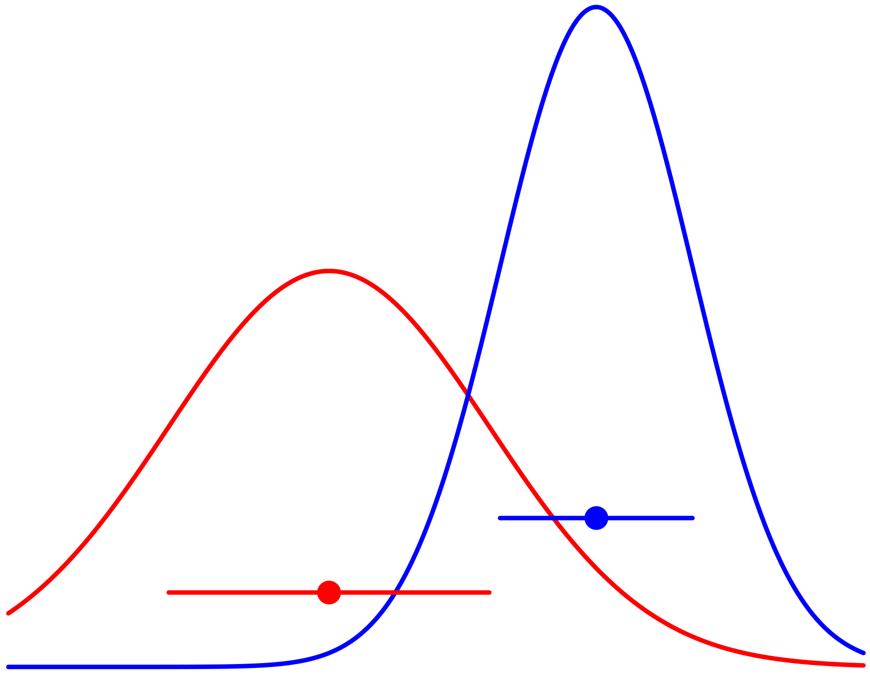
understanding data relationships

and

making predictions

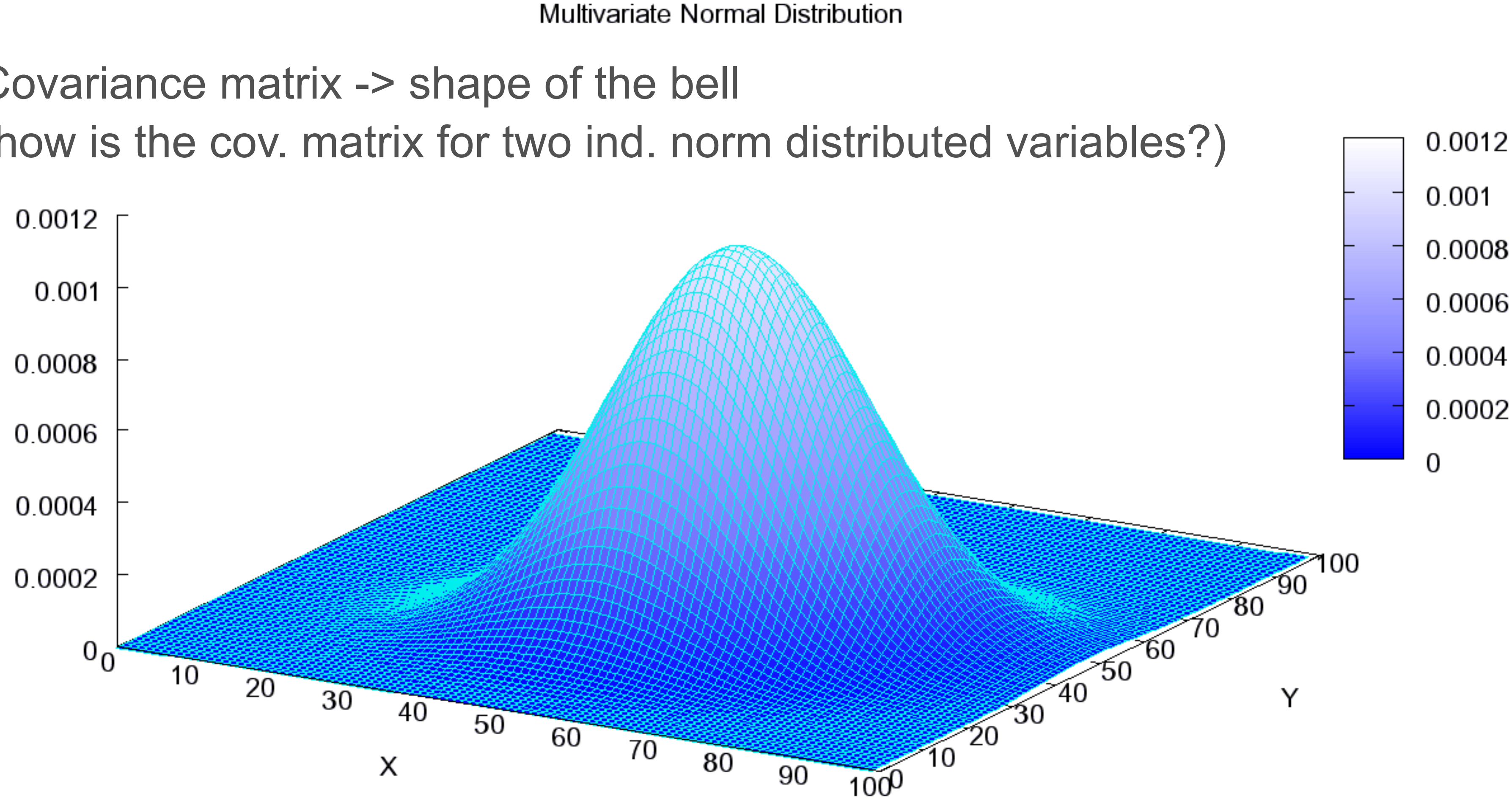
Motivation: the basics

A Gaussian distribution is described by a location μ and “a shape”: σ for one dimension and for several, a covariance matrix Σ



Motivation: the basics

Covariance matrix -> shape of the bell
(how is the cov. matrix for two ind. norm distributed variables?)



Motivation: the basics II

$$p(y|\theta) = \phi(y|\theta) = \phi(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Single variable

$$p(\mathbf{Y}|\boldsymbol{\mu}, K) = \phi(\mathbf{Y}|\boldsymbol{\mu}, K) = \frac{\exp\left(-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^{\mathrm{T}} K^{-1} (\mathbf{Y}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d \det K}}$$

Multi-variate

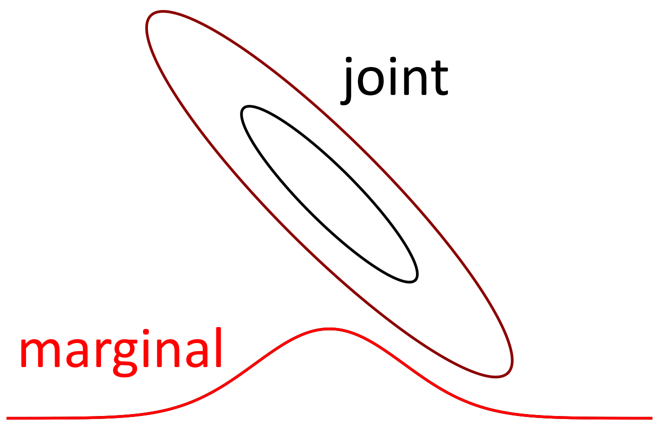
Cov matrix

$$K = \begin{bmatrix} \mathrm{E}[(Y_1-\mu_1)(Y_1-\mu_1)] & \cdots & \mathrm{E}[(Y_1-\mu_1)(Y_n-\mu_n)] \\ \mathrm{E}[(Y_2-\mu_2)(Y_1-\mu_1)] & \cdots & \mathrm{E}[(Y_2-\mu_2)(Y_n-\mu_n)] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[(Y_n-\mu_n)(Y_1-\mu_1)] & \cdots & \mathrm{E}[(Y_n-\mu_n)(Y_n-\mu_n)] \end{bmatrix}$$

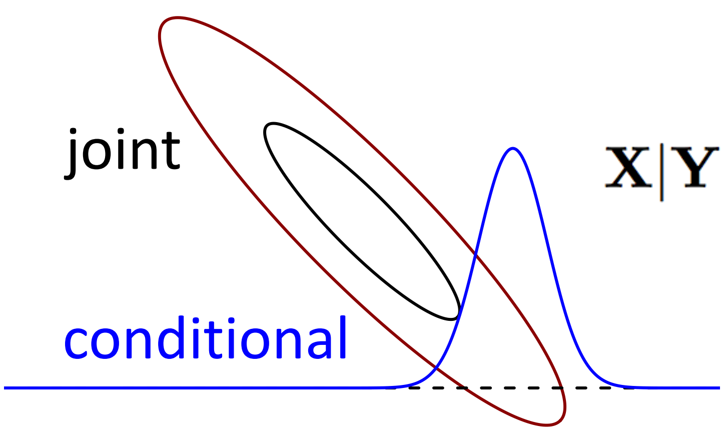
Motivation: the basics III

$$\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{X,Y}, \boldsymbol{\Sigma}_{X,Y})$$

$$\boldsymbol{\mu}_{X,Y} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{X,Y} = \begin{bmatrix} A & B \\ B^{\intercal} & C \end{bmatrix}$$



$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{a}, A) \qquad \mathbf{Y} \sim \mathcal{N}(\boldsymbol{b}, C)$$



$$\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + BC^{-1}(\mathbf{Y} - \mathbf{b}), A - BC^{-1}B^{\intercal})$$

Motivation: the problem?

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$$

Motivation: the problem?

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

Joint probability -> conditional probability -> posterior from the prior, but here we have the joint probability of the values of f(x) for all x

"observed", so: (observed -> f, non-observed -> f_*)

Motivation: the problem?

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

Joint probability -> conditional probability -> posterior from the prior, but here we have the joint probability of the values of f(x) for all x

“observed”, so: (observed -> f, non-observed -> f_*)

kernel function applied to x
“similarity” of each xi to all xi

“similarity” of each xi to non-observed xi that we want to predict

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{pmatrix}\right)$$

“similarity” of each non-observed xi to the other non-observed xi that we want to predict

Motivation: the problem?

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$$

Joint probability -> conditional probability -> posterior from the prior, but here we have the joint probability of the values of f(x) for all x

“observed”, so: (observed -> f, non-observed -> f_*)

The magic? (algebra)

From the joint distribution to the conditional one

$$p(f, f_*) \longrightarrow p(f_* | f)$$

kernel function applied to x

“similarity” of each xi to all xi

“similarity” of each xi to non-observed xi that we want to predict

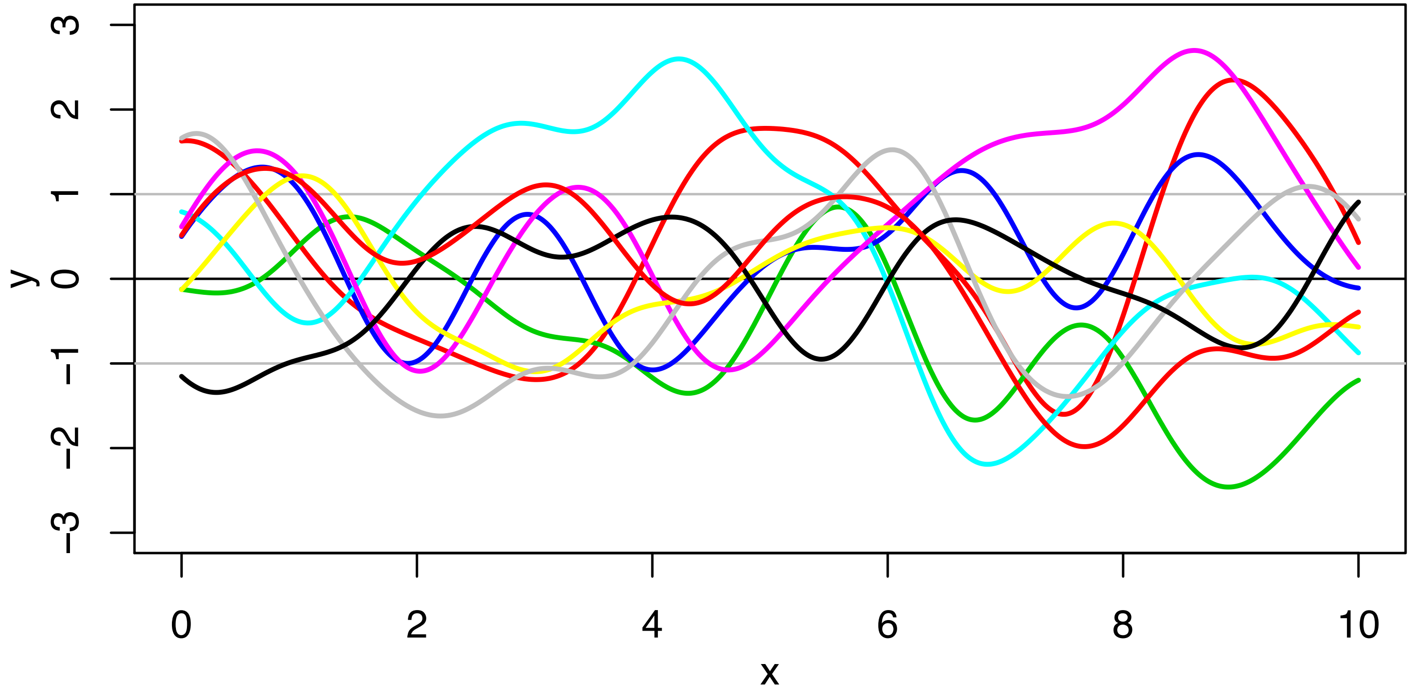
$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{pmatrix} \right)$$

“similarity” of each non-observed xi to the other non-observed xi that we want to predict

The definition

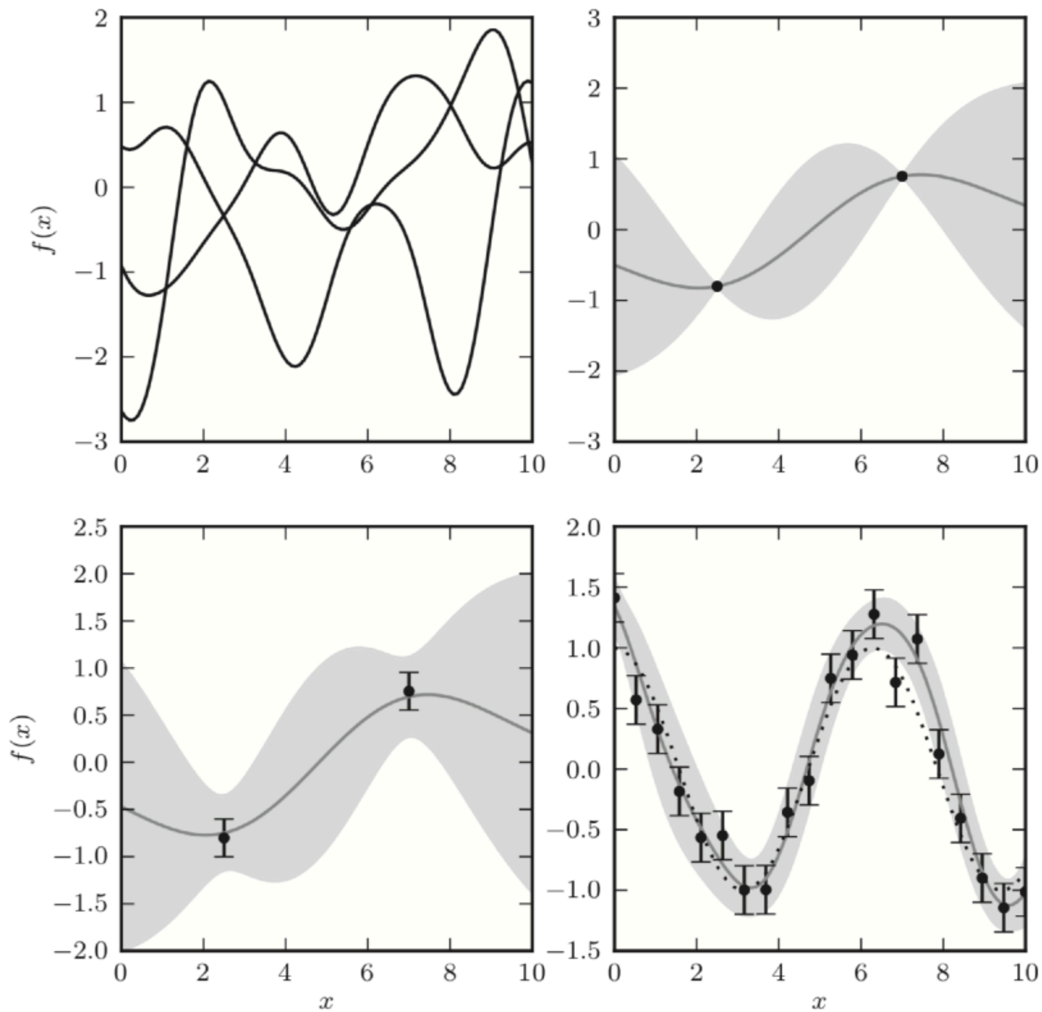
"A Gaussian process is a collection of random variables with the property that the joint distribution of any finite subset is a Gaussian. "

$$f(x) \sim \mathcal{GP}(\boldsymbol{\mu}, k(x, x'))$$



The Interpretation

The Gaussian kernel can be interpreted as a prior on the functions (instead of on the parameters when doing parametric statistics)

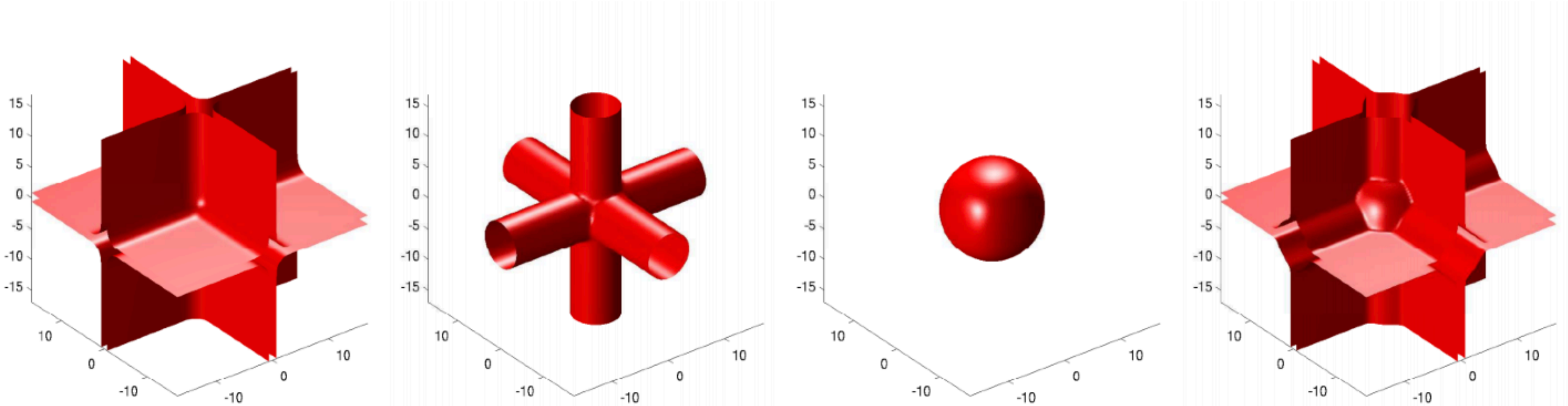


covariance matrix ensures that values that are close together in input space will produce output values that are close together.

Is it all about the kernel?

different kernels?!
additive kernels?!

Kind of defining the "radius of influence" differently at different spatial scales



1st order interactions	2nd order interactions	3rd order interactions	All interactions
$k_1 + k_2 + k_3$	$k_1 k_2 + k_2 k_3 + k_1 k_3$	$k_1 k_2 k_3$	
		(Squared-exp kernel)	(Additive kernel)

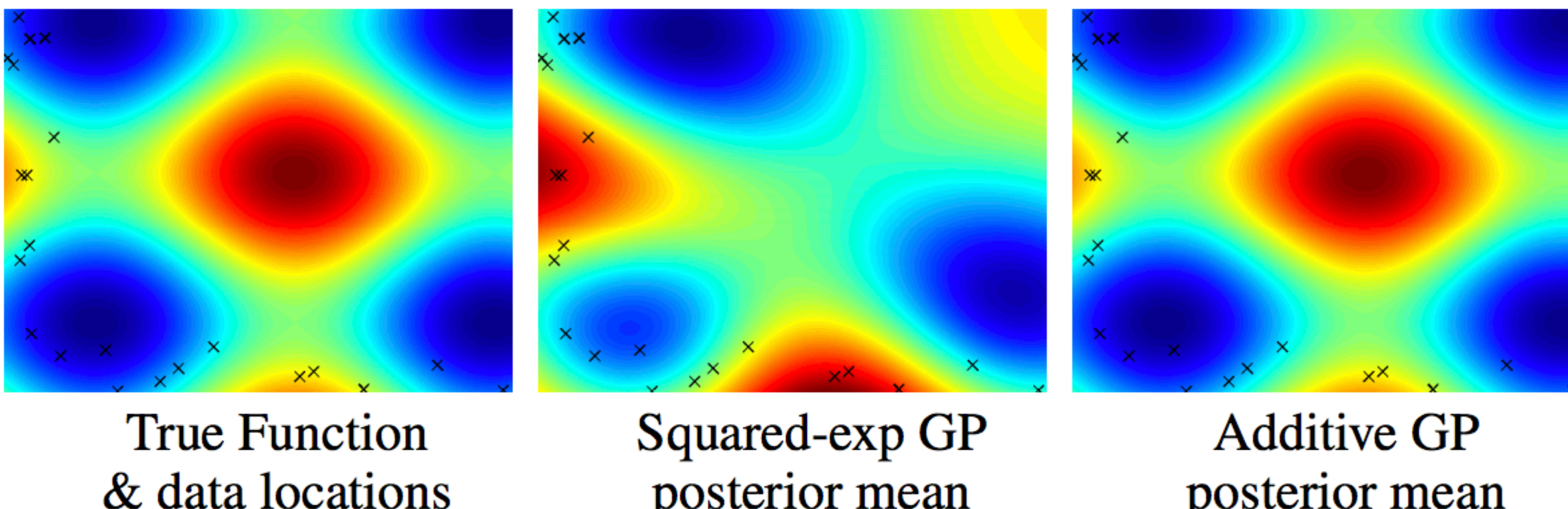
Duvenaud+11

Is it all about the kernel?

different kernels?!
additive kernels?!

Kind of defining the "radius of influence" differently at different spatial scales

Good at discovering non-local structure!!



Duvenaud+11

The implementation?