Institute for Health Metrics and Evaluation

Gaussian Process Regression

FRH Training

29 June 2016

Tara Templin

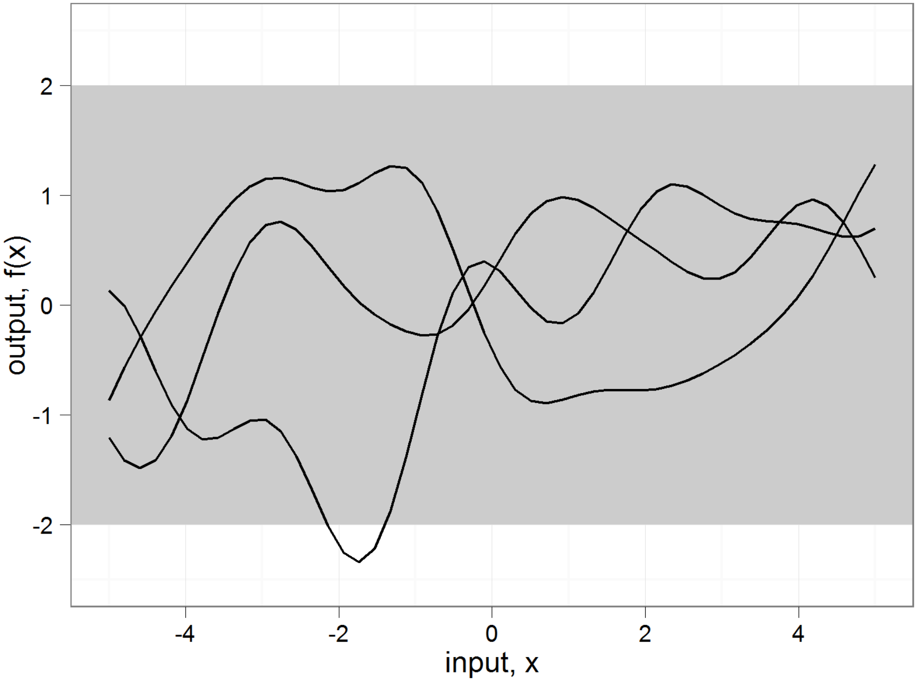
Researcher, FRH

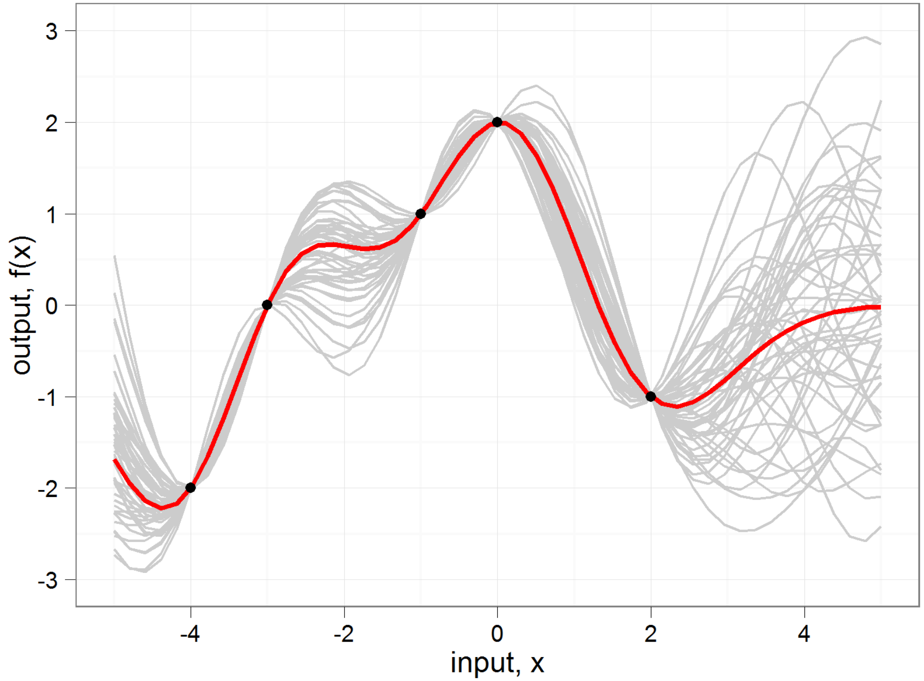
UNIVERSITY *of* WASHINGTON

Modeling approach

- At IHME you'll hear something like:

"We use a combination of space-time smoothing and Gaussian process regression (GPR) to estimate …"

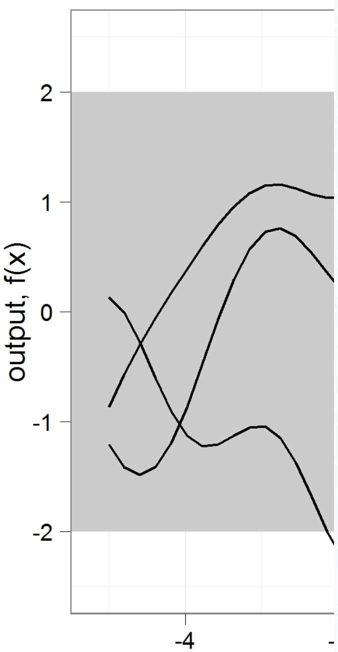


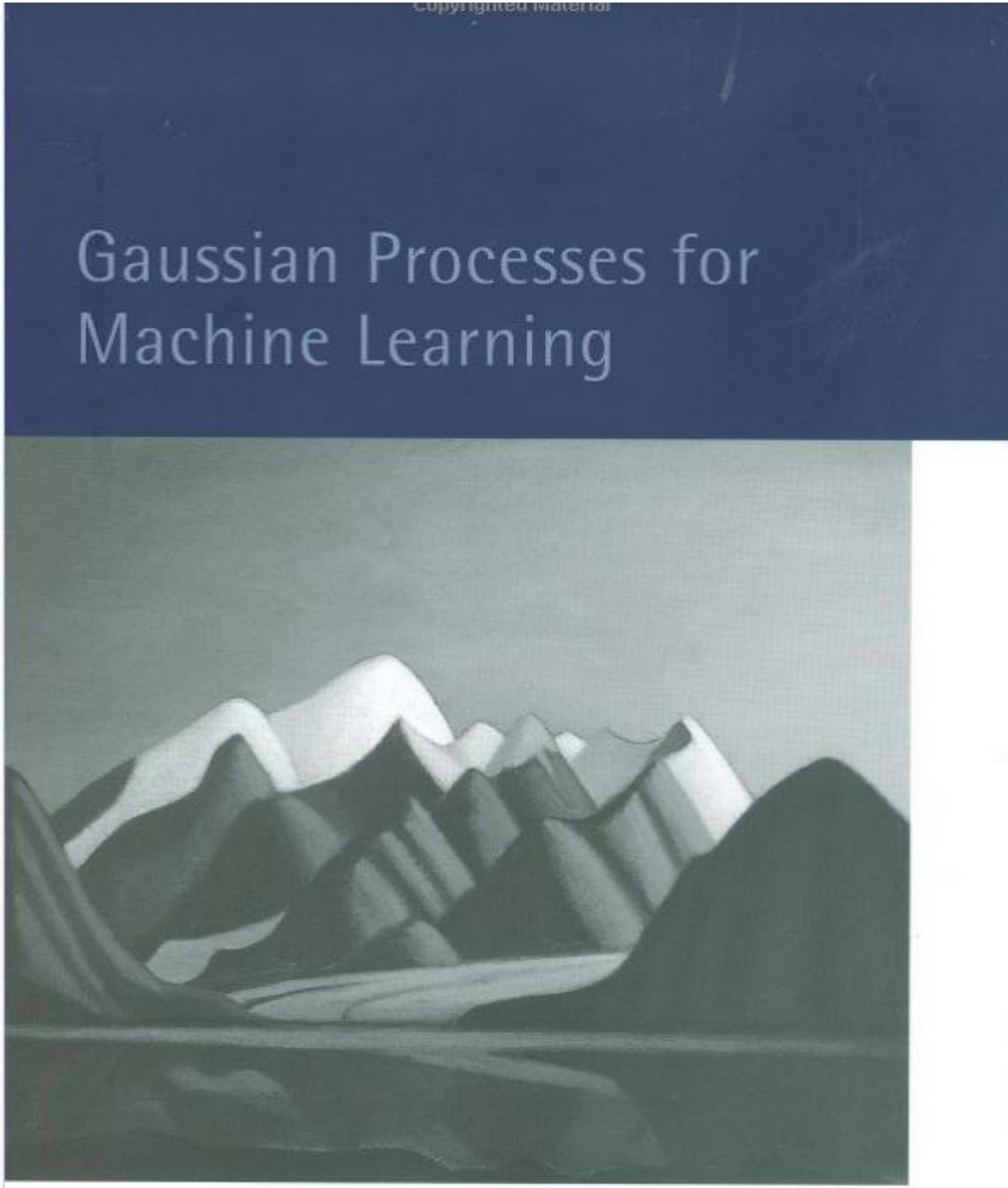


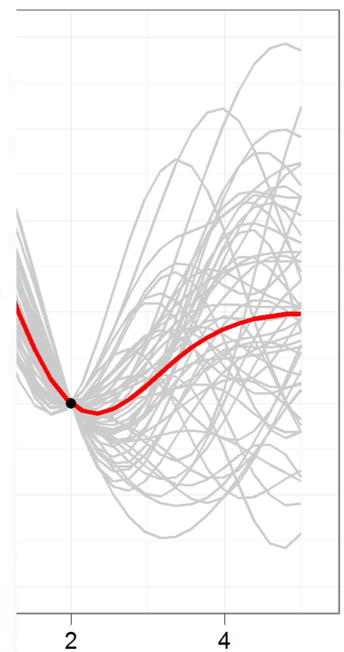
Model

- At IHME

“We use a aussian process r



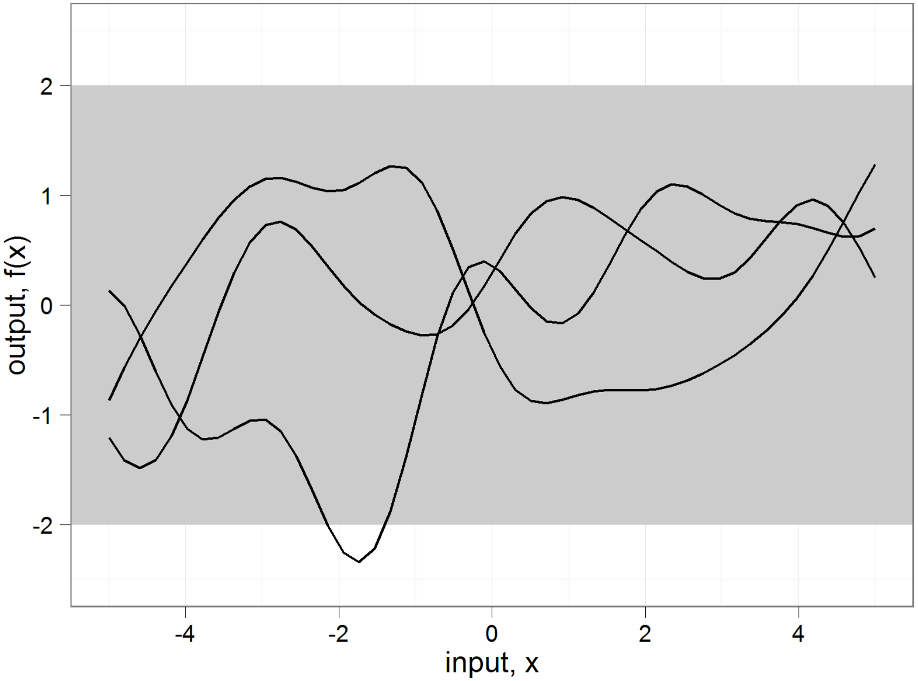


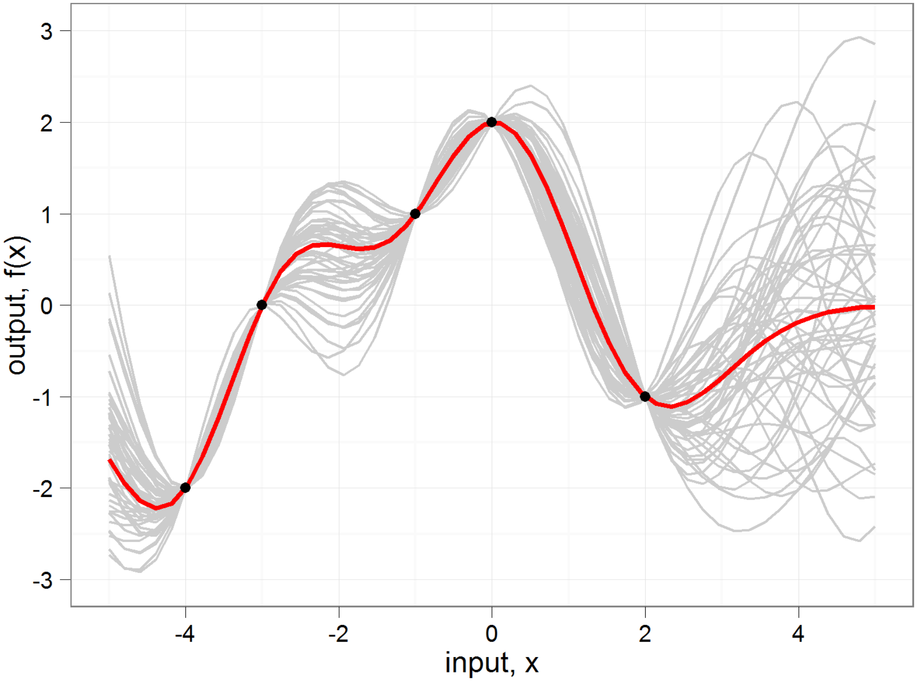


Modeling approach

- At IHME you'll hear something like:

"We use a combination of space-time regression and Gaussian process regression (GPR) to estimate …"

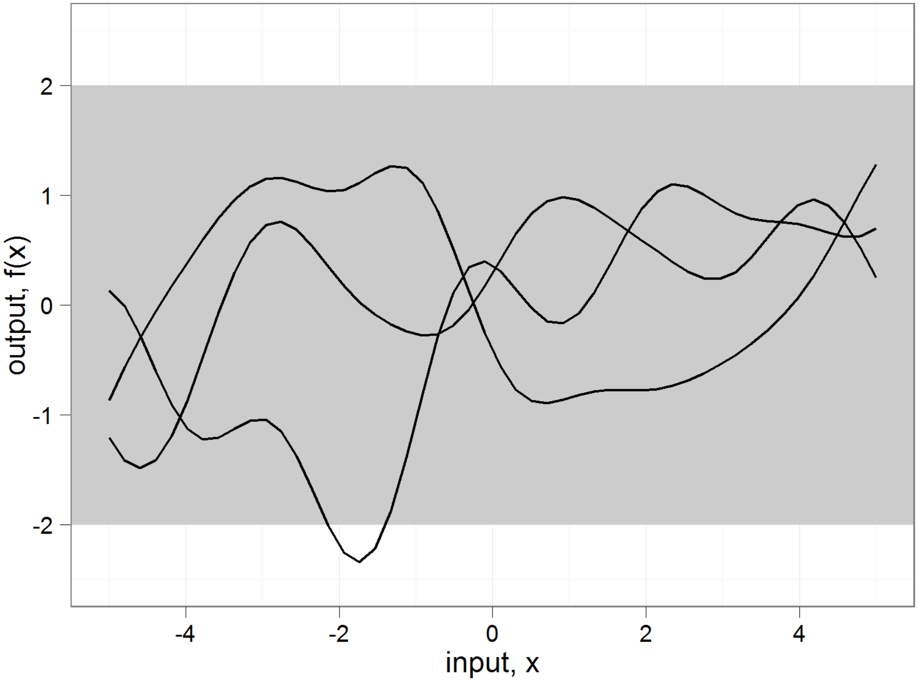




Intro to Gaussian Processes

- Gaussian processes (GPs) are probability distributions for functions.

- They are often used as priors for functions whose forms are unknown.



What are Gaussian processes good for?

- In some cases several candidate forms exist, but it is not possible to rule all of them out or even to ascertain that they are the only possibilities.

- Encode many types of knowledge about functions, yet remain much less restrictive than priors based on particular functional forms.

Gaussian processes and the multivariate normal distribution

- Gaussian processes generalize the multivariate normal distribution from vectors to functions, like the multivariate normal distribution generalizes the univariate normal distribution from scalars to vectors.

$y \sim N(\mu, V)$: y, μ, V are scalars

$\mathbf{y} \sim N(\boldsymbol{\mu}, C)$: $\mathbf{y}$ and $\boldsymbol{\mu}$ are vectors, C is a matrix

$f \sim GP(M, C)$: f and M are functions of one variable, C is a function of two variables

Creating a Gaussian process

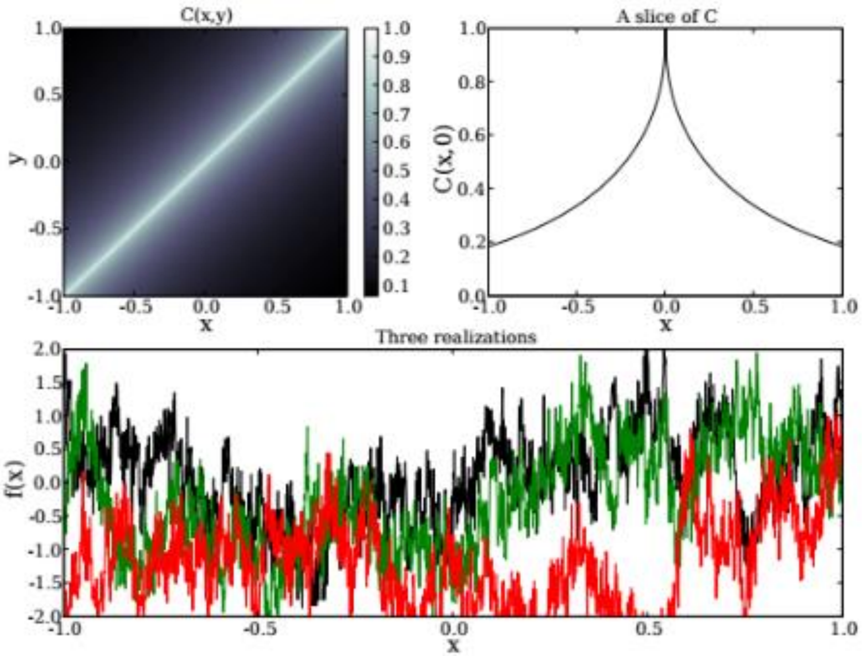
- f ~ GP(M, C)

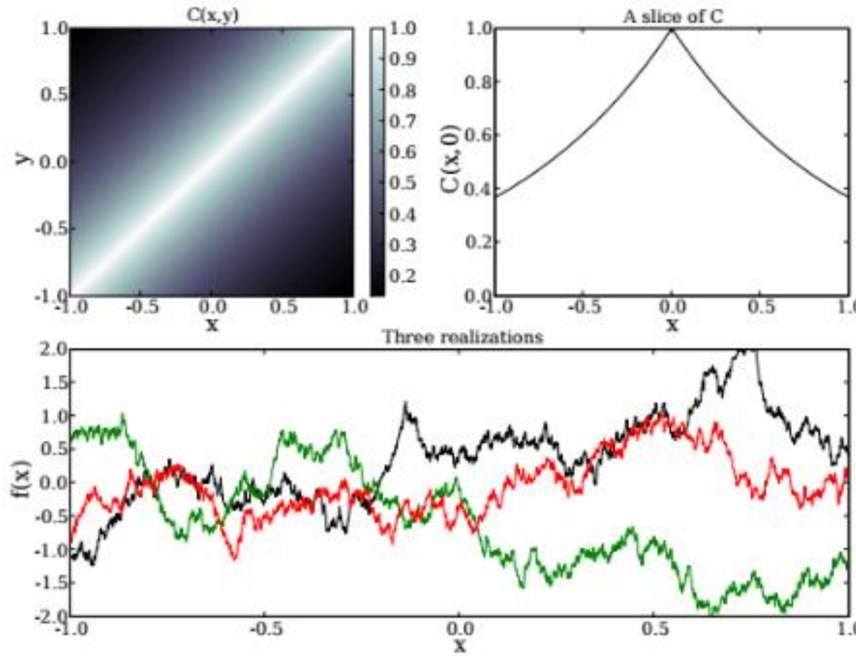
- The mean function can be interpreted as a prior guess for the GP

- The covariance kernel function takes in any two points, and outputs the covariance between them.
  - Determines how strongly linked (correlated) these two points are.
  - A common choice is to have the covariance decrease as the distance between the points grows

Covariance functions

- Matern covariance functions with varying degrees of differentiability





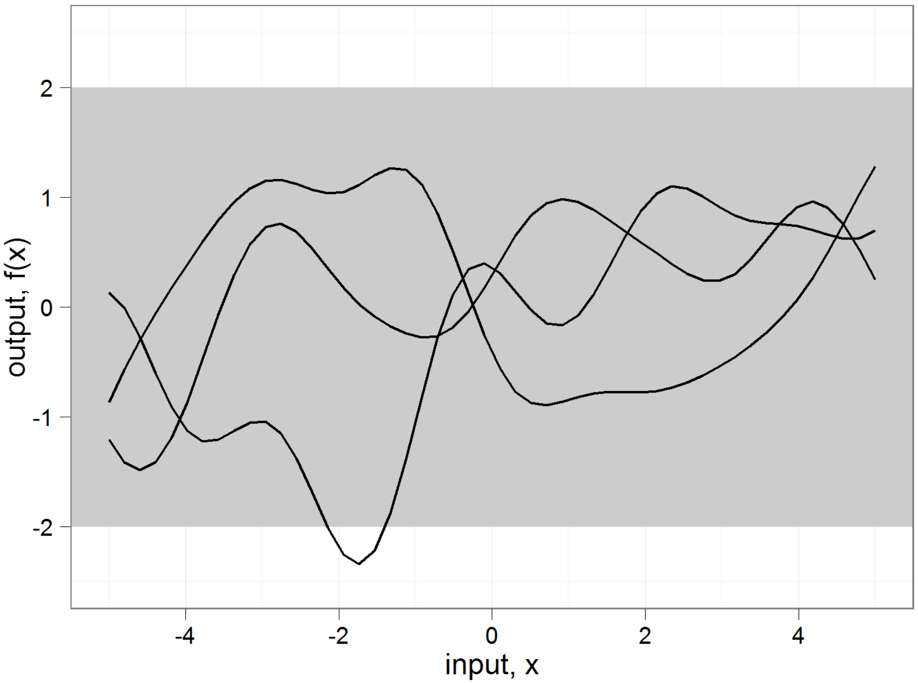
Covariance functions

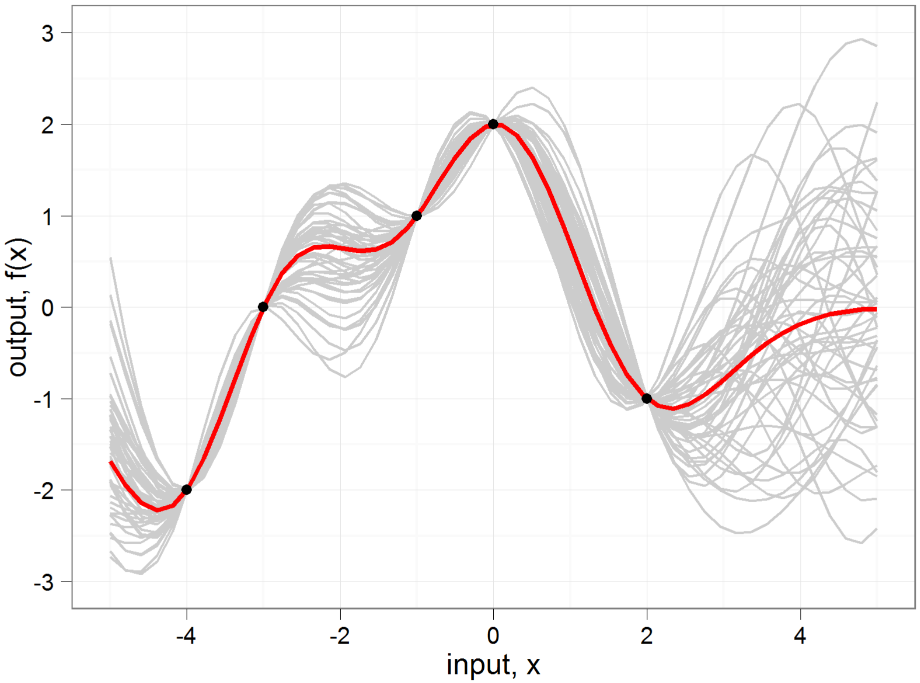
- Many types of covariance

covariance function	expression
constant	σ_0^2
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}}{\ell} r\right)$
exponential	$\exp(-\frac{r}{\ell})$
γ-exponential	$\exp\left(-\left(\frac{r}{\ell}\right)^{\gamma}\right)$
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^{\top}\Sigma\tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^{\top}\Sigma\tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^{\top}\Sigma\tilde{\mathbf{x}}')}}\right)$

Observing Gaussian processes

$$\left.\begin{array}{l} \text{data}_i \overset{\text{ind}}{\sim} \mathrm{N}(f(o_i), V_i) \\ f \sim \mathrm{GP}(M, C) \end{array}\right\} \Rightarrow f|\text{data} \sim \mathrm{GP}(M_o, C_o).$$





IHME: Gaussian process regression

- GPR refers to inference of continuous values using a Gaussian process prior; aka kriging
- Instead of specifying one function, specify a distribution of functions

$$f \sim GP(M, C)$$

- *M* is a function of time capturing the average, underlying trend in the country.
- *C* encodes smoothness in the trend and correlation over time
- GPR updates the function M using information from the observed data (including uncertainty around the data) according to several parameters C

Space-time for smoking prevalence

- Step 1 – Linear model ($X\beta$)

$$logit(p_{c,a,s,t}) = \beta_0 + \beta_1 \mathrm{CPC}_{\mathrm{c,t}} + \sum_{k=2}^{21} \beta_k I_{R[c]} + \sum_{k=22}^{35} \beta_k I_{A[a]} + \varepsilon_{c,a,s,t}$$

  - $\mathrm{CPC}_{\mathrm{crt}}$: consumption per capita
  - $I_{R[c]}$: GBD region dummy
  - $I_{A[a]}$: Age group dummy

- Step 2 – Space-time-age smoothing of residuals ($h(r_{c,a,s,t})$)
  - Space weight: 0.95
  - Time weight: 1
  - Age weight: 2

GPR for Prevalence

- Let $p_{c,a,s,t}$ be the predicted time series of smoking prevalence for country t , age a, sex s in period t

$$logit(p_{c,a,s,t}) = g_{c,a,s}(t) + \epsilon_{c,a,s,t}$$

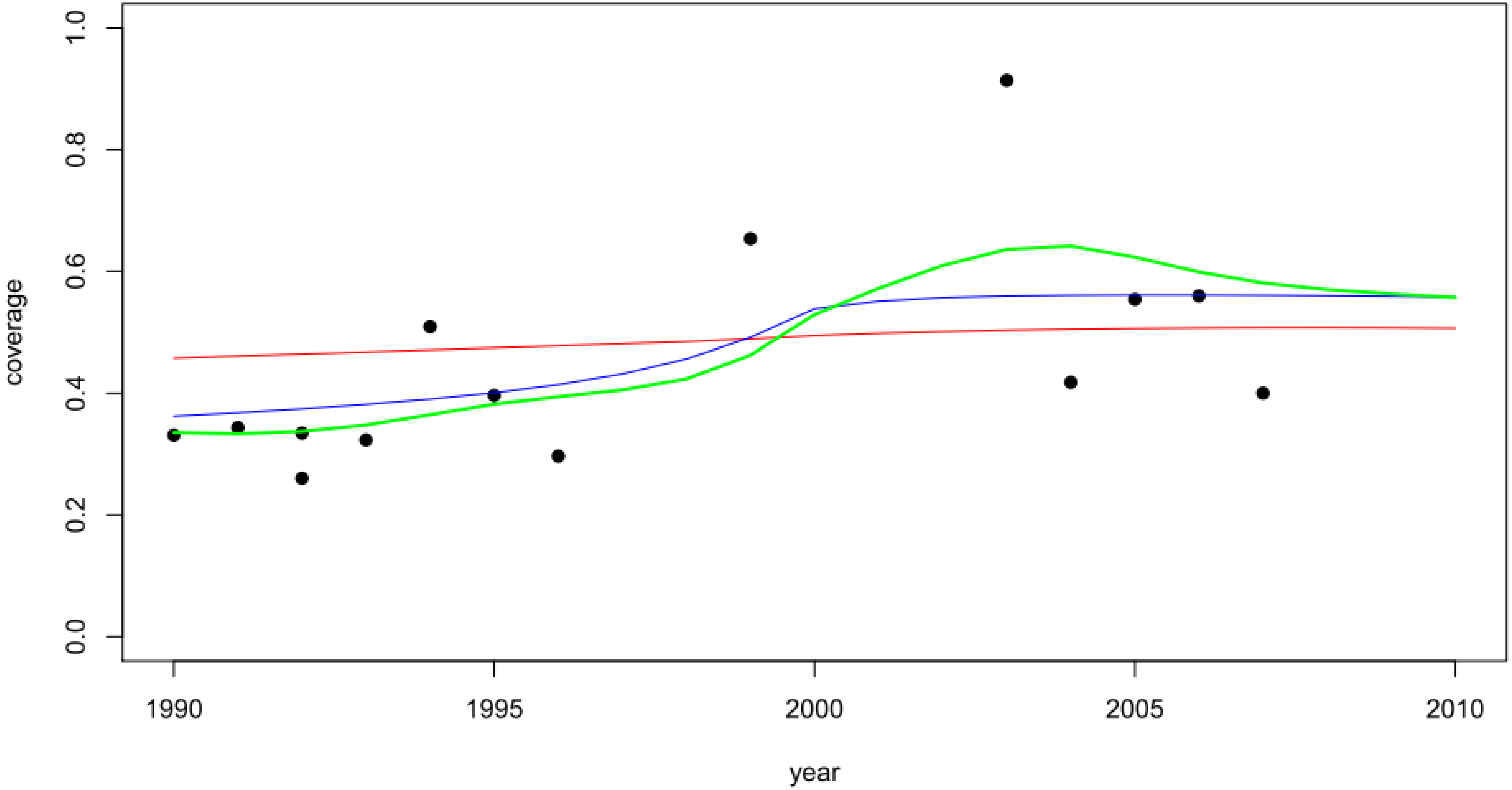
where

$$\epsilon_{c,a,s,t} \sim Normal(0, \sigma_p^2),$$

$$g_{c,a,s}(t) \sim GP\left(m_{c,a,s}(t), Cov\left(g_{c,a,s}(t)\right)\right).$$

- σ_p^2: error variance (squared standard error of the observed data point as well as the prediction errors from the cross-walk models)
- $m_{c,a,s}(t)$: Mean function
- $Cov\left(g_{c,a,s}(t)\right)$: Covariance function

Gaussian process regression - example



GPR advantages and disadvantages

- Advantages
  - Can be combined with other predictive models
  - Incorporates uncertainty
  - Very flexible
- Disadvantages
  - Parameter choice can be arbitrary (Cross validate!)
  - Very flexible