

# Model-Based Interpolation, Prediction, and Approximation

— Statistical Computing for Uncertainty Analysis —

Antonio Possolo

Statistical Engineering Division  
Information Technology Laboratory

August 4, 2011



1/28

## Outline

### Statistical Computing

R — Statistical computing and graphics

### Interpolation

INFLUX experiment

Local regression

Kriging

### Prediction

Viral load in influenza A infection

### Approximation

Ensemble of solutions of ODEs

Projection pursuit regression

Ridge functions for viral load peak time

2/28



- ▶ Programming environment for statistical computing, data analysis, and graphics

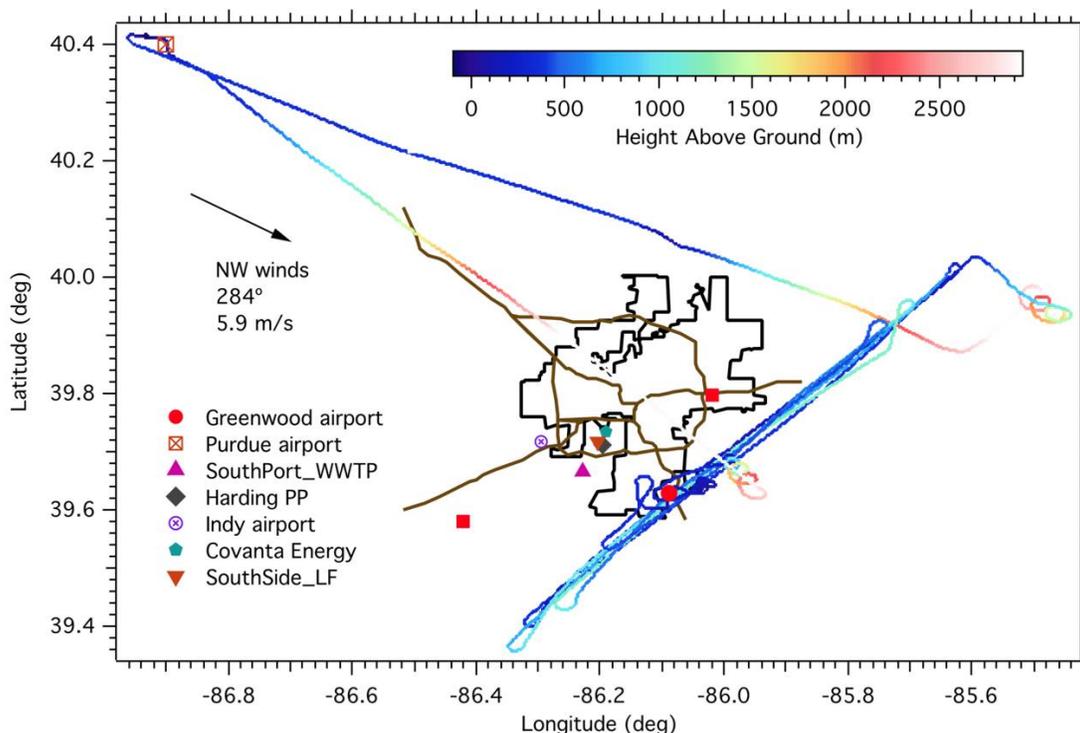
▶ [www.r-project.org](http://www.r-project.org)

- ▶ Free and open source
- ▶ *Lingua franca* of statistical computing: implementations of new statistical methods often first appear as R functions
- ▶ Ideal environment for uncertainty analysis, also well suited for prototyping general purpose scientific computing algorithms

3/28

## INFLUX Experiment (Indianapolis, IN)

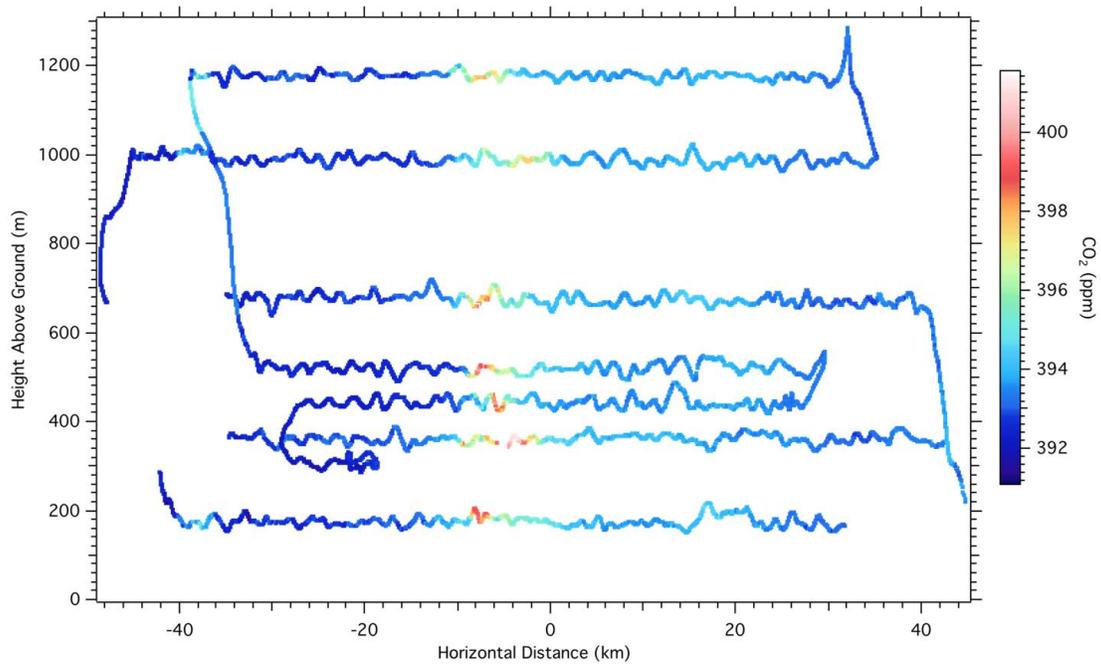
— FLIGHT PATH



4/28

# INFLUX Experiment

— CO<sub>2</sub> measurements on curtain flight



5/28

## Interpolation

— PROBLEM

- ▶ Given
  - Measured values  $y_1, \dots, y_m$  of real-valued function  $\theta$  at  $x_1, \dots, x_m$  in metric space  $\mathcal{X}$
- ▶ Estimate  $\theta(x)$  for any  $x$  “in the middle” of the  $\{x_i\}$
- ▶ Characterize uncertainty  $u(y)$  associated with estimate

6/28

# Model Based Interpolation

## — APPROACHES

- ▶ **Model observations probabilistically** — interpolation problem becomes statistical estimation problem
  - $y_i = \theta(x_i) + \epsilon_i$ 
    - ▶  $\{\epsilon_i\}$  realized values of non-observable random variables (**measurement errors**)
    - ▶ **Interpolate signal, not signal + noise**

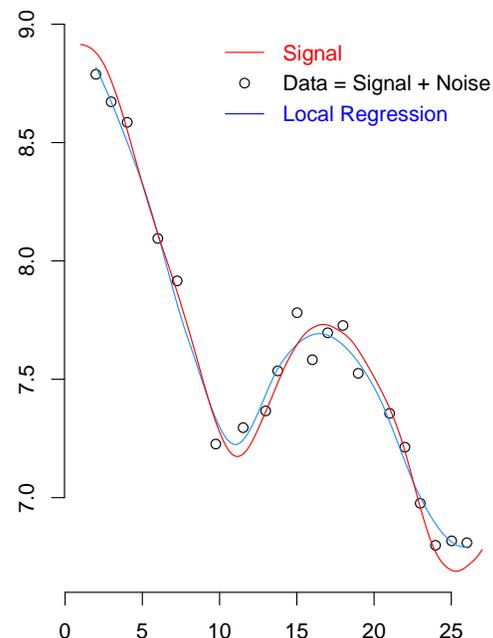
## Local regression vs. Kriging

- ▶  $\theta$  locally quadratic
- ▶  $\theta$  realized value of Gaussian random function  $\Theta$

7/28

## Local Regression

- ▶ Approximate  $\theta$  *locally* by parabola at each target location  $x$
- ▶ Fit each parabola by (robust) weighted least squares
  - Weights decrease to zero with increasing distance to target location



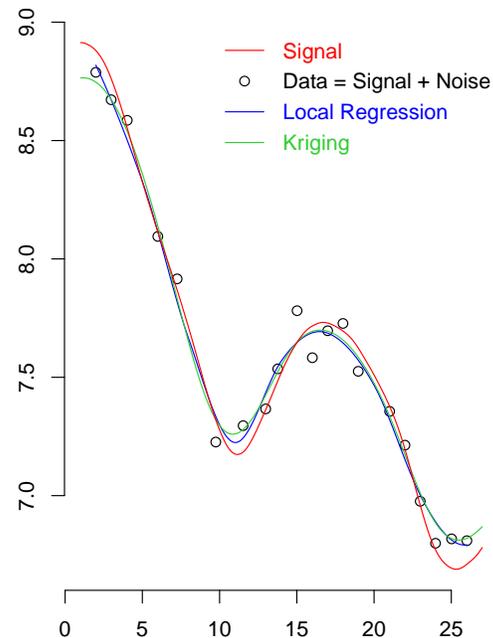
8/28

# Ordinary Kriging

- ▶  $\{\Theta(x)\}$  Gaussian RVs with mean  $\mu$  and covariance function

$$\gamma(h) = \text{Cov}(\Theta(x), \Theta(x+h))$$

- ▶  $\hat{\theta}(x)$  is weighted average of data  $\{y_i\}$ 
  - Weights depend on  $\gamma(h)$  and on variances of  $\{\epsilon_i\}$



9/28

# Kriging Assessment of Uncertainty

- ▶ Kriging often heralded as providing assessments of uncertainty of interpolations automatically
- ▶ In many instances of application, kriging's built-in assessments *underestimate* uncertainty because one pretends that  $\hat{\gamma} = \gamma$ 
  - Bayesian kriging provides means to account for this often neglected component of uncertainty

10/28

# Interpolation Uncertainty

— COMPONENTS AND ASSESSMENT

## COMPONENTS

- ▶ Measurement error —  $\{\epsilon_i\}$  in  $y_i = \theta(x_i) + \epsilon_i$
- ▶ Model selection and calibration — different results corresponding to different choices of functional form for  $\theta$ , and parameter estimation

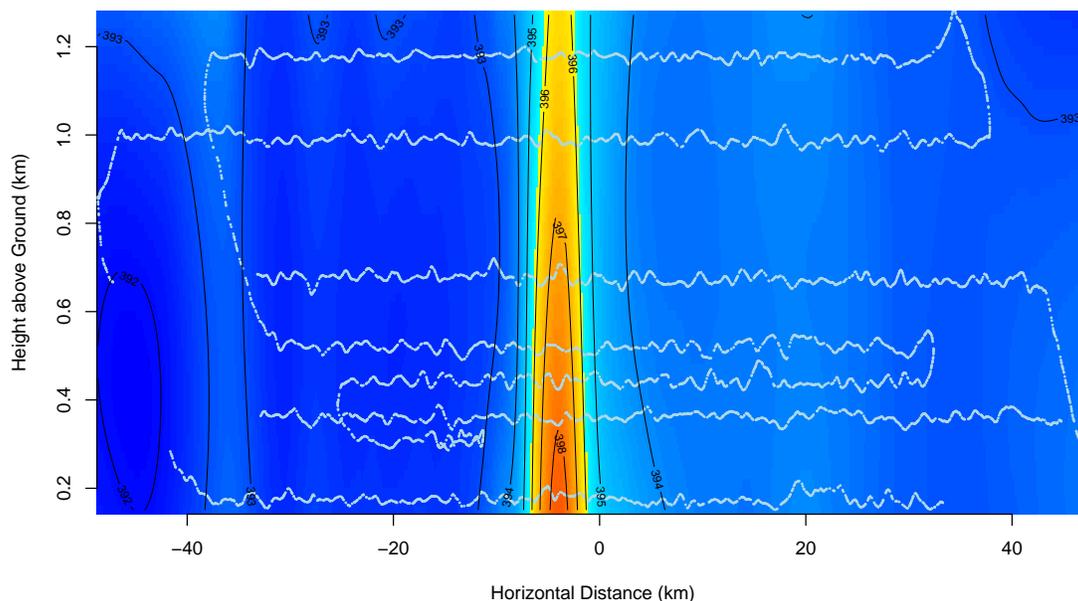
## ASSESSMENT

- ▶ Cross-validation (*leave-some-out*)
- ▶ Model inter-comparisons

11/28

# Local Regression Interpolation

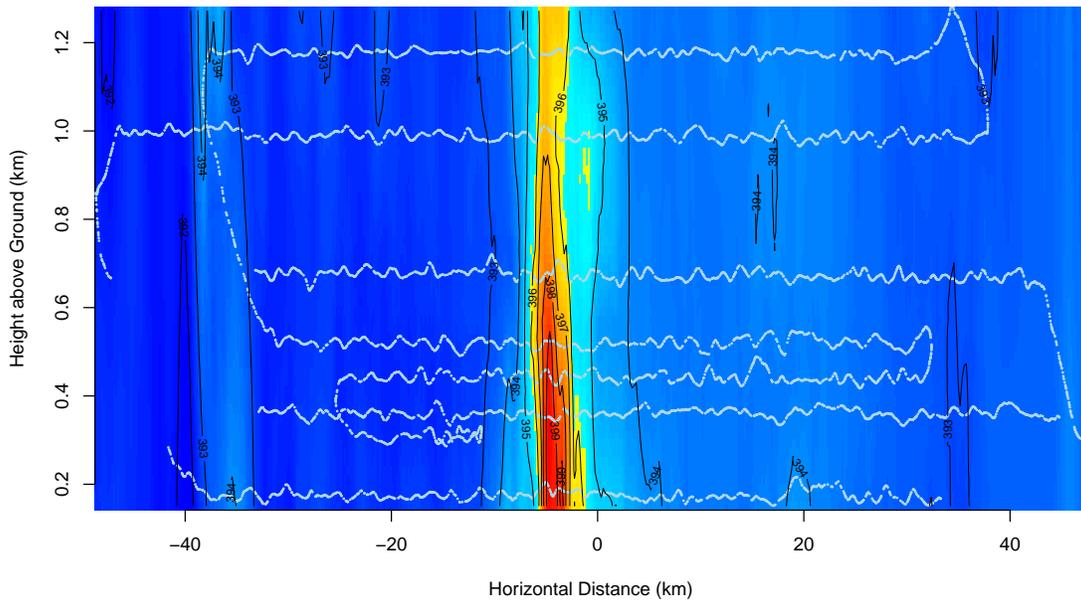
— INFLUX EXPERIMENT: CO<sub>2</sub>



12/28

# Kriging Interpolation

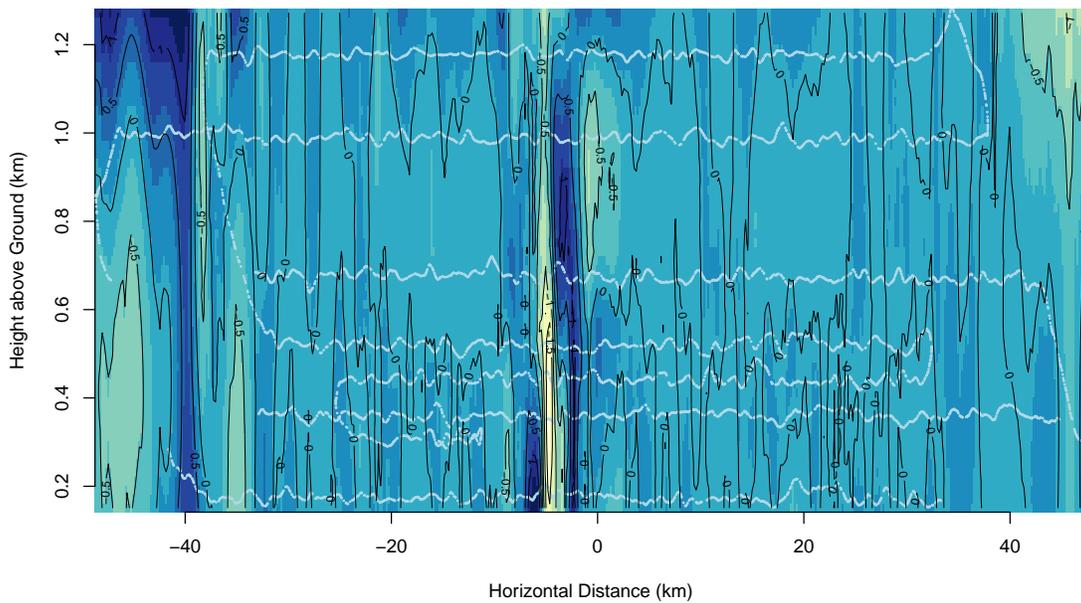
— INFLUX EXPERIMENT: CO<sub>2</sub>



13/28

# Local Regression vs. Kriging

— INFLUX EXPERIMENT: CO<sub>2</sub>



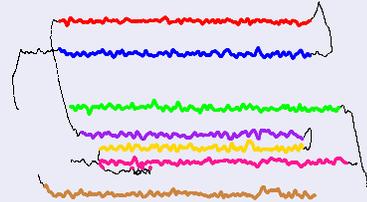
14/28

# Cross-Validation & Model Uncertainty

— INFLUX EXPERIMENT: CO<sub>2</sub>

## CROSS-VALIDATION

- ▶ Partition data into **training** and **testing** subsets: fit models using former, assess performance on latter
- ▶ Partition may be random, or may include consideration for particulars of situation



## MODEL UNCERTAINTY

- ▶ Compare predictions made by different models

15/28

# Uncertainty Budget

— INFLUX EXPERIMENT: CO<sub>2</sub>

SOURCE	EVALUATION	STD. UNCERT.
Model selection	CV	0.36
Interpolation	CV	0.91
Instr. calibration	LAB+CERT	0.034
Instr. repeatability	MANUF*	0.2
Instr. drift	MANUF*	0.2
Atmospheric temperature	MANUF*	0.0075
Atmospheric pressure	MANUF*	0.7
<b>Expanded Uncertainty</b>	<b><math>U_{95\%} = 2.5</math> ppmv</b>	

\* Picarro G2301-m Flight

$$2.5 = 2\sqrt{0.36^2 + \dots + 0.7^2}$$

16/28

# Influenza A Virus Infection in Humans

Baccam *et al.* (Aug, 2006) Journal of Virology

## PROGRESSION

- ▶ Initial exponential growth of viral load
- ▶ Peaking 2–3 days post-infection
- ▶ Exponential decrease to undetectable levels at 6–8 days

## PREDICTION

- ▶ Predict time when **viral load** peaks
- ▶ Estimate **basic reproductive number** of infection

17/28

## Influenza A — Kinetics

- $T$  No. of uninfected target cells
- $I$  No. of productively infected cells
- $V$  Viral load

$$\frac{dT}{dt} = -\beta TV \quad \frac{dI}{dt} = \beta TV - \delta I \quad \frac{dV}{dt} = \rho I - \gamma V$$

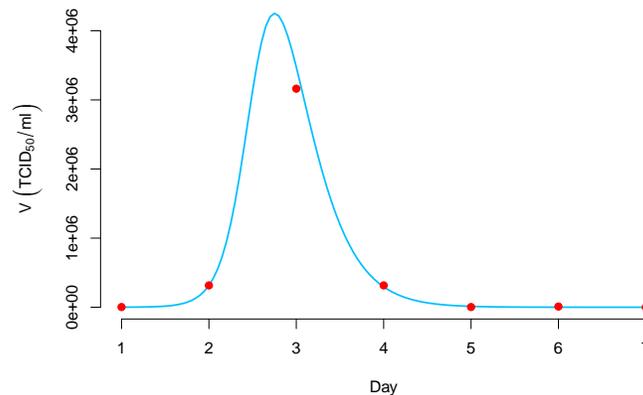
- $\beta$  Infection rate
- $1/\delta$  Lifespan of infected cell
- $\rho$  Increment to viral load per infected cell
- $\gamma$  Viral clearance rate

**SOLUTION:** ODEPACK (Livermore Solver for Ordinary Differential Equations, LSODA) — R package **deSolve**

18/28

# Influenza A — Data & Statistical Model

- ▶ Patient 4 (Table 1, Baccam *et al.*, 2006)



- ▶ Generalized non-linear model for viral load  $V$ 
  - $\log_{10} V \sim \text{GAU}(\nu, \tau^2)$
  - $\nu = \nu(\beta, \delta, \rho, \gamma)$  — solution of kinetic model

19/28

# Influenza A — Prediction & Estimation

- ▶ Predict time  $\text{argmax}_t V_t$  when viral load peaks  
*TCID<sub>50</sub> — 50 % Tissue Culture Infective Dose per milliliter of nasal wash*
- ▶ Estimate Basic Reproductive Number

$$R_0 = \frac{\rho\beta T_0}{\gamma\delta}$$

- Average number of second-generation infections produced by single infected cell placed among susceptible cells
  - ▶ If  $R_0 > 1$  infection progresses full course
  - ▶ If  $R_0 < 1$  infection dies out prematurely

20/28

# Influenza A — Uncertainty Assessment

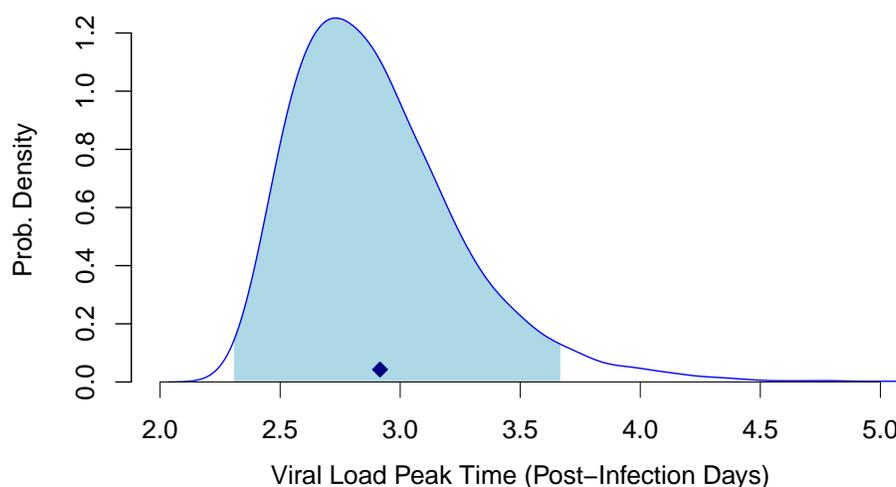
## PARAMETRIC BOOTSTRAP

- ▶ Compute numerical approximation to Hessian  $H(\beta, \delta, \rho, \gamma)$  of negative log-likelihood used to fit kinetic model to data for Patient 4
- ▶ For  $k = 1, \dots, K$ 
  - Draw sample  $(\beta_k, \delta_k, \rho_k, \gamma_k)$  from multivariate Gaussian distribution with mean  $(\hat{\beta}, \hat{\delta}, \hat{\rho}, \hat{\gamma})$  and covariance matrix  $H^{-1}(\hat{\beta}, \hat{\delta}, \hat{\rho}, \hat{\gamma})$
  - Draw one sample from uniform distribution for each initial condition  $T_0 \pm 0.1T_0, I_0 \pm 0.1I_0, V_0 \pm 0.1V_0$
  - Solve kinetic model with *perturbed* parameters and compute  $\psi(\beta_k, \delta_k, \rho_k, \gamma_k)$

21/28

# Influenza A — Uncertainty Assessment

## RESULTS $K = 10\,000$ — VIRAL LOAD PEAK

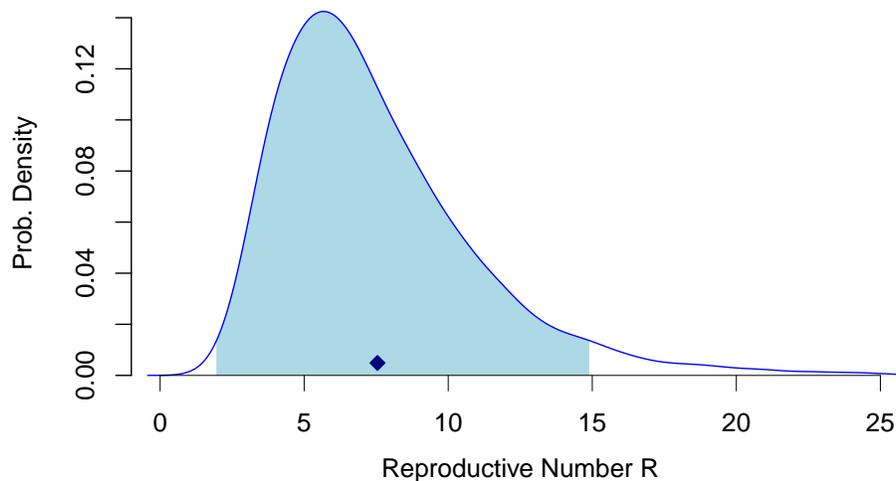


- ▶  $\operatorname{argmax}_t V_t = 2.9 \text{ PID}, u(\operatorname{argmax}_t V_t) = 0.4 \text{ PID}$
- ▶ Shortest 95 % probability interval (2.3 PID, 3.7 PID)

22/28

# Influenza A — Uncertainty Assessment

RESULTS  $K = 10\,000$  — REPRODUCTIVE NUMBER



- ▶  $\hat{R} = 7.5$ ,  $u(R) = 3.5$
- ▶ Shortest 95 % probability interval  $(2, 15)$
- ▶  $R > 5$  with probability 76 %

23/28

## Approximation

- ▶ For unknown function  $\psi : \mathcal{X} \mapsto \mathbb{R}$  that is “expensive” to evaluate, observe

$$(x_1, \psi(x_1) + \epsilon_1), \dots, (x_m, \psi(x_m) + \epsilon_m)$$

- Non-observable measurement errors  $\epsilon_1, \dots, \epsilon_m$
- ▶ Develop approximant  $\varphi$  and assess its quality

- EXAMPLE

$$\begin{aligned} \operatorname{argmax}_t V_t &= \psi(\beta, \delta, \rho, \gamma) \\ &\approx \varphi(\beta, \delta, \rho, \gamma) \end{aligned}$$

24/28

# Projection Pursuit Magic

— AT A PRICE

- ▶ Finds **interesting** low-dimensional projections of a high-dimensional point cloud
  - Builds predictors out of these projections
- ▶ Automatically sets aside variables with little predictive power
- ▶ Bypasses **curse of dimensionality** by focussing on functions of linear combinations of the original variables

*PRICE: Compute-intensive technique*

25 / 28

# Universal Approximant

PROJECTION PURSUIT

- ▶ Friedman & Tukey (1974)

*The algorithm seeks to find one- and two-dimensional linear projections of multivariate data that are relatively highly revealing*

- ▶ Projection Pursuit Regression  
— Friedman & Stuetzle (1981)

$$\psi(x) \approx \varphi(x) = \alpha_0 + \sum_{k=1}^K \varphi_k(\alpha_k^T x_i)$$

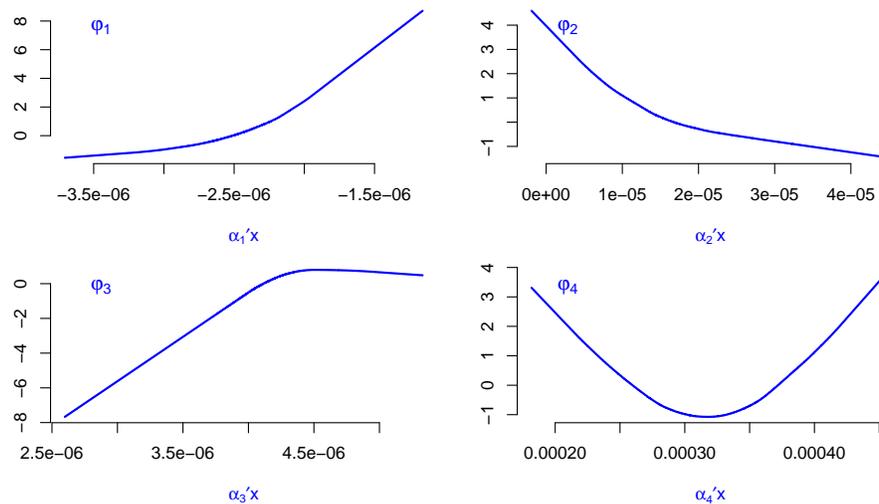
- **IMPLEMENTATION:** R function `ppr`
- ▶ Diaconis & Shahshahani (1984)

26 / 28

# Influenza A — Projection Pursuit

RIDGE FUNCTIONS FOR VIRAL LOAD PEAK TIME

$$\operatorname{argmax}_t V_t = \psi(\beta, \delta, \rho, \gamma)$$



Cross-validated rel. approxim. error: 3%

27/28

## Summation

- ▶ **Non-linear, computationally expensive models** — in medicine, atmospheric science, oceanography, etc. — challenge traditional uncertainty analysis toolkit
- ▶ **R** is state-of-the-art platform for statistical modeling and uncertainty analysis, also offering ample capabilities for general scientific computing
- ▶ **Model sampling, cross-validation** and the **statistical bootstrap** are general-purpose tools for realistic uncertainty assessment

28/28