Science on Tap

How Random was That?

WIth James Risk, Assistant Professor of Mathematics and Statistics at Cal Poly Pomona

Come for an evening to talk about:

- What is the general framework used to predict uncertain events?
- What are the basics of machine learning and data science?
- What is a "random function"?
- How do you apply this concept to super-resolution (restoring high-frequency details of images) and mortality modeling?

Monday, October 4, 7-8 p.m.

Register: bit.ly/SciTap-Reg

What is a Function

- A **function** takes in an input and gives an output.

$$f(\texttt{input}) = \texttt{output}.$$

- Example:

$$f(\text{age}) = \text{age} + 1 \qquad \text{(birthday)}$$

$$f(x) = \sin(x) \qquad \text{(mathematical sin function)}$$

$$f(\texttt{messy hair}) = \texttt{clean head} \qquad \text{(haircut)}$$

Function Example

Example

A tree grows 20cm every year, so the height of the tree is related to its age using this function

$$f(\text{age}) = 20 \cdot \text{age}$$

- Is the above function realistic?

Statistical Modelling

- Statistical modelling[1] adds a **error term**.
- This could represent...
  - measurement error;
  - model inaccuracy;
  - etc.

$$f(\text{age}) = 20 \cdot \text{age} + \texttt{error}$$

- This is a **catch-all term**.
- A good model can *reduce* error using the data we have.
- Not all errors can be reduced.
  - Example: flip a coin a number of times, and consider a function that records the number of heads

$$f(\texttt{number of flips}) = ??$$

- A good statistical model will reduce predictable error and leave the irreducible error.

[1] or, machine learning model

Statistical Modelling

- Statistical modelling[1] adds a **error term**.
- This could represent...
  - measurement error;
  - model inaccuracy;
  - etc.

$$f(\text{age}) = 20 \cdot \text{age} + \texttt{error}$$

- This is a **catch-all term**.
- A good model can *reduce* error using the data we have.
- Not all errors can be reduced.
  - Example: flip a coin a number of times, and consider a function that records the number of heads

$$f(\texttt{number of flips}) = ??$$

- A good statistical model will reduce predictable error and leave the irreducible error.

[1] or, machine learning model

Example of Complex Data

- This dataset contains a subset of the fuel economy data that the EPA makes available on `https://fueleconomy.gov/`
- $n = 234$ cars
- $d = 11$ variables
  - `mpg` (miles per gallon)
  - `cylinders` (number of cylinders)
  - `horsepower` (engine horsepower)
  - `weight` (vehicle weight (lbs))
  - `year` (model year)
  - `origin` (origin of car (Amer, Euro, Japan)

$$\texttt{mpg} = f(\texttt{cylinders}, \texttt{horsepower}, \texttt{weight}, \texttt{year}, \texttt{origin})$$

Types of Statistical Models

- **Linear Regression**
  - Most common
  - Assumes a linear relationship

$$\texttt{mpg} = \alpha + \beta_1 \cdot \texttt{cylinders} + \beta_2 \cdot \texttt{horsepower} + \beta_3 \cdot \texttt{weight} + \beta_4 \cdot \texttt{year} + \beta_5 \cdot \texttt{origin} + \texttt{error}$$

- Coefficients $(\alpha, \beta_1, \ldots, \beta_5)$ are fitted from the data
- Produces a line[2] of best fit
- Assumptions of randomness are placed on error
- **Regression Spline**
  - Assumes some degree of smoothness on the relationship between mpg and its inputs
  - Most commonly, a collection of piecewise *third degree polynomials*
- Adds an error term to account for randomness

[2] a plane, in multiple dimensions

Types of Statistical Models

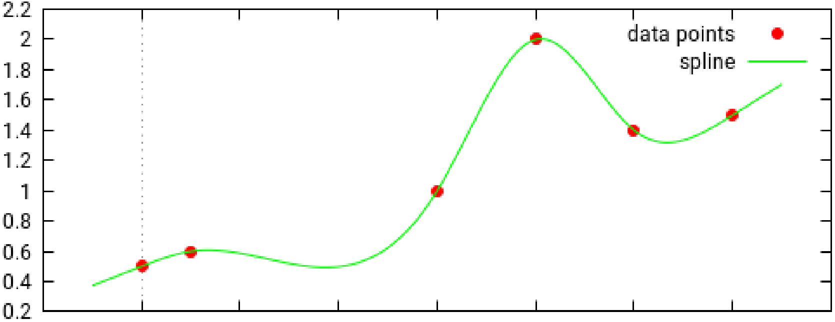
- **Linear Regression**
  - Most common
  - Assumes a linear relationship

$$\texttt{mpg} = \alpha + \beta_1 \cdot \texttt{cylinders} + \beta_2 \cdot \texttt{horsepower} + \beta_3 \cdot \texttt{weight} \\ + \beta_4 \cdot \texttt{year} + \beta_5 \cdot \texttt{origin} + \texttt{error}$$

- Coefficients $(\alpha, \beta_1, \ldots, \beta_5)$ are fitted from the data
- Produces a line[2] of best fit
- Assumptions of randomness are placed on `error`
- **Regression Spline**
  - Assumes some degree of smoothness on the relationship between `mpg` and its inputs
  - Most commonly, a collection of piecewise *third degree polynomials*
- Adds an `error` term to account for randomness

[2] a plane, in multiple dimensions

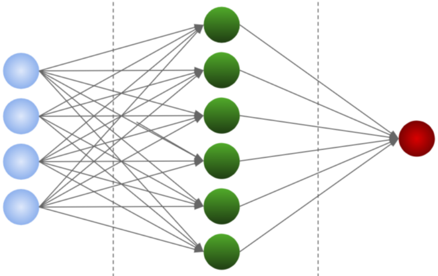
Spline Example



(Taken from `https://github.com/ttk592/spline`)

More Complicated Machinery (Neural Network)

- **Neural Network**
  - Designed to mimic how the brain handles information
  - Compromised of many parameters, including
    - the number of hidden layers (*1, in the example below*)
    - the number of neurons per layer (*6, in the example below*)
  - Very powerful model
  - $\texttt{output} = f(\texttt{input})$ is compared to a "black-box:"
    1. Plug in `input`
    2. Magic happens *(black-box)*
    3. Get `output`
  - Difficult to understand and analyze



More Complicated Machinery (Gaussian Process)

- **Gaussian Process**
  - Gaussian Processes originated as a probabilistic concept in the early 1920's.
  - Although mathematically difficult, they have rich theoretical properties and are interpretable methods.
- Assumes the function f itself is **random**

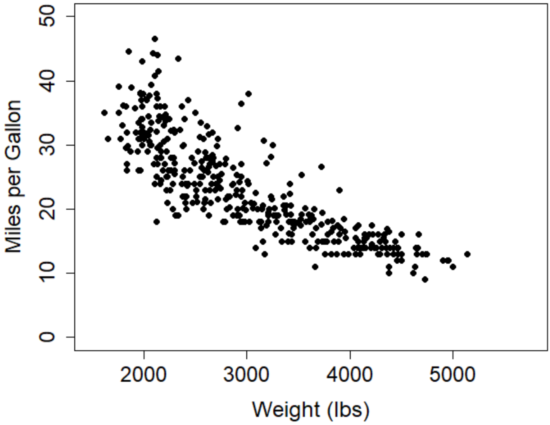
$$\underbrace{f}_{\text{random}}(\texttt{input}) = \texttt{output} \qquad \text{(Gaussian process)}$$

- Compare to linear regression:

$$f(\texttt{input}) = \underbrace{\alpha + \beta \cdot \texttt{input}}_{\text{deterministic}} + \underbrace{\texttt{error}}_{\text{random}} \qquad \text{(Linear Regression)}$$

Motivation: Car MPG Data

- Car MPG data from 1970–1982
- Produce a new car with weight 5500.
  - Best guess for MPG?



Linear Regression

$$\texttt{mpg} = f(\texttt{lbs})$$
$$\texttt{mpg} = \alpha + \beta \cdot \texttt{lbs} + \texttt{error}$$

- Line of best fit[3]

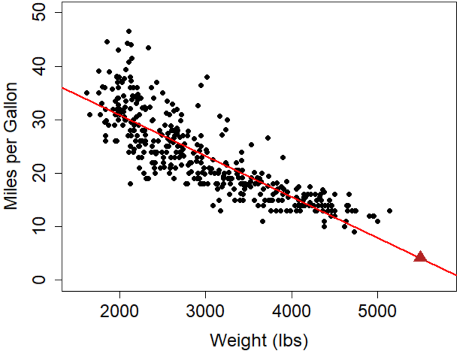
$$\texttt{mpg} = 46.22 - 0.0076 \cdot \texttt{lbs} + \texttt{error}$$

[3] minimizes squared distance to data points

Prediction (MPG)

- Prediction at `lbs` $= 5500$:

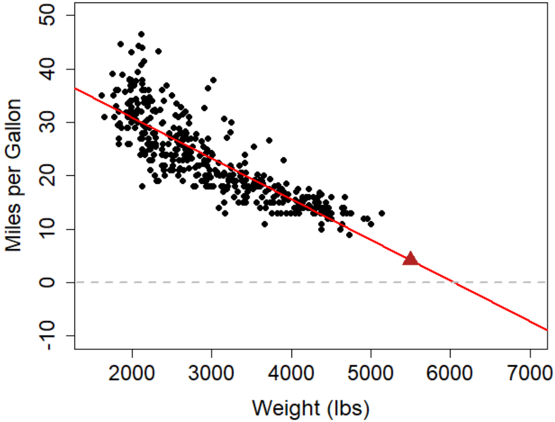
$$\widehat{\texttt{mpg}} = 46.22 - 0.0076 \cdot \texttt{lbs}$$
$$= 46.22 - 0.0076 \cdot 5500$$
$$= 4.42$$



Oops...

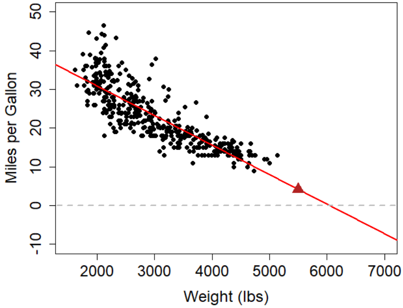
- Prediction:

$$\widehat{\texttt{mpg}} = 46.22 - 0.0076 \cdot \texttt{lbs}$$



What Went Wrong (Part 1)

- **extrapolate**[4]: *extend the application of (a method or conclusion, especially one based on statistics) to an unknown situation by assuming that existing trends will continue or similar methods will be applicable.*
  "the results cannot be extrapolated to other patient groups"
- In general, extrapolating can lead to trouble.



What Went Wrong (Part 1)

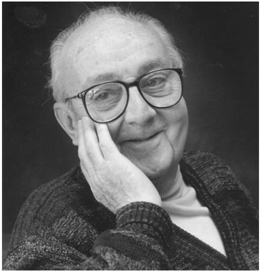
- Linear regression **assumes** the mathematical relationship for f

$$\texttt{mpg} = f(\texttt{lbs}, \texttt{error})$$
$$\texttt{mpg} = \alpha + \beta \cdot \texttt{lbs} + \texttt{error}$$

- incorrect assumption $\Rightarrow$ incorrect predictions

All Models are Wrong

"*All models are wrong, but some are useful*" - George Box



Functional Approach (Take 2)

- Assume a different mathematical relationship for f

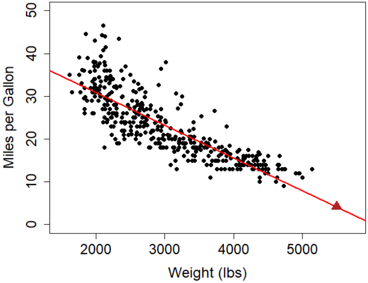
$$\text{mpg} = f(\text{lbs}, \text{error})$$
$$\text{mpg} = \alpha + \beta_1 \cdot \text{lbs} + \beta_2 \cdot \text{lbs}^2 + \text{error}$$

- "Curve" of best fit:[5]

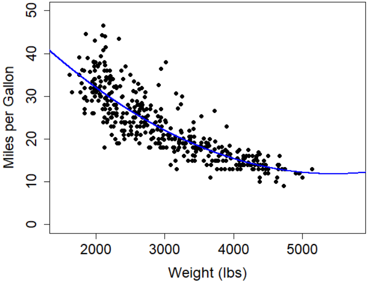
$$\text{mpg} = 62.26 - 0.0185 \cdot \text{lbs} + 0.0000017 \cdot \text{lbs}^2 + \text{error}$$

[5] minimizes squared distance to data points

Better...

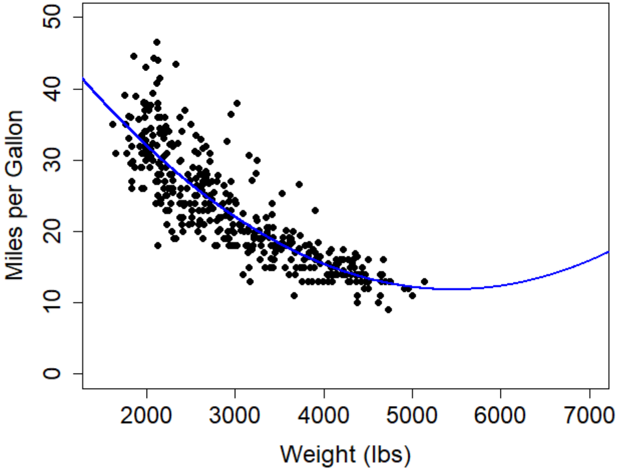


Miles per Gallon
0
10
20
30
40
50
2000
3000
4000
5000
Weight (lbs)



Miles per Gallon
0
10
20
30
40
50
2000
3000
4000
5000
Weight (lbs)

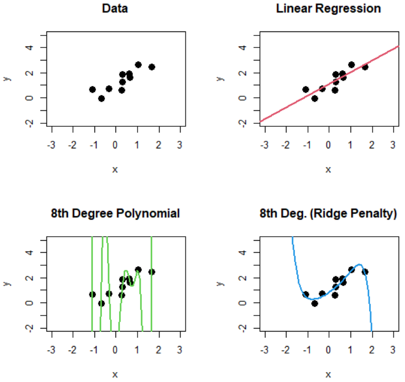
Oops



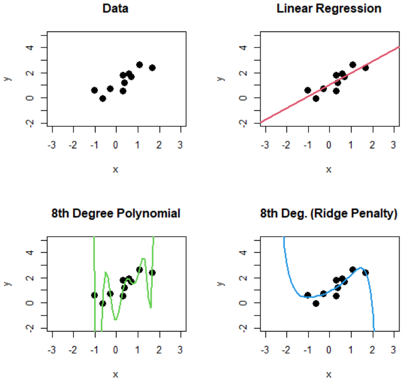
Miles per Gallon
0
10
20
30
40
50
2000
3000
4000
5000
6000
7000
Weight (lbs)

Think Smarter, Not Harder

Data



Slightly Perturbed Data



Adding More Data

- In a world of big data, we often have more than one variable (e.g. `cylinder`, `horsepower`, etc.)
  - More Power
  - More Complex Models
  - More Possibility of Bad Assumptions
- More difficult to visualize
  - Compare with 2-d scatterplot
  - How to visualize in 5-d?
- Unnecessary variables can complicate things
  - Example: what if `car color` was included in the data set?
  - The blind modeller would include it, but it would hinder (not help) the model
  - This is an extreme example, but it happens more than you would think in our world of big data.

Adding More Data

- In a world of big data, we often have more than one variable (e.g. `cylinder`, `horsepower`, etc.)
  - More Power
  - More Complex Models
  - More Possibility of Bad Assumptions
- More difficult to visualize
  - Compare with 2-d scatterplot
  - How to visualize in 5-d?
- Unnecessary variables can complicate things
  - Example: what if `car color` was included in the data set?
  - The blind modeller would include it, but it would hinder (not help) the model
  - This is an extreme example, but it happens more than you would think in our world of big data.

Adding More Data

- In a world of big data, we often have more than one variable (e.g. `cylinder`, `horsepower`, etc.)
  - More Power
  - More Complex Models
  - More Possibility of Bad Assumptions
- More difficult to visualize
  - Compare with 2-d scatterplot
  - How to visualize in 5-d?
- Unnecessary variables can complicate things
  - Example: what if `car color` was included in the data set?
  - The blind modeller would include it, but it would hinder (not help) the model
  - This is an extreme example, but it happens more than you would think in our world of big data.

Gaussian Processes

- A Gaussian process (GP) is a type of model for a function f that uses "nearby data" to produce a prediction.
- Determining if a data point is "close" depends on a kernel function
  - Determines the properties of the underlying f
- For example, choosing the kernel function lets you decide if the data is...
  - linear?
  - continuous (*no jumps*)?
  - periodic (*repeating patterns*)?
  - smooth vs jagged?
  - combination of above?
- A GP provides full probabilistic properties
  - How variable are future predictions?
  - What is the probability that my prediction will be above 150? Below 25? *(etc.)*

Gaussian Processes

- A Gaussian process (GP) is a type of model for a function f that uses "nearby data" to produce a prediction.
- Determining if a data point is "close" depends on a kernel function
  - Determines the properties of the underlying f
- For example, choosing the kernel function lets you decide if the data is...
  - linear?
  - continuous (*no jumps*)?
  - periodic (*repeating patterns*)?
  - smooth vs jagged?
  - combination of above?
- A GP provides full probabilistic properties
  - How variable are future predictions?
  - What is the probability that my prediction will be above 150? Below 25? *(etc.)*

Gaussian Process Details

Definition (Gaussian Process)

Let $f : \mathbb{R}^d \to \mathbb{R}$. Then f is a **Gaussian process** if for all $n = 1, 2, 3, \ldots$, the vector $[f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)]^\top$ is multivariate normal.

- Specified by
  - mean function μ: $\mathbb{E}[f(x)] = \mu(x)$
  - covariance kernel k: $\mathrm{cov}(f(x), f(x')) = k(x, x')$
- Generalization of a multivariate normal distribution to infinite dimensional indices
- Yes, this is very technical. See the next few slides for a simplification.

Clarifying Randomness

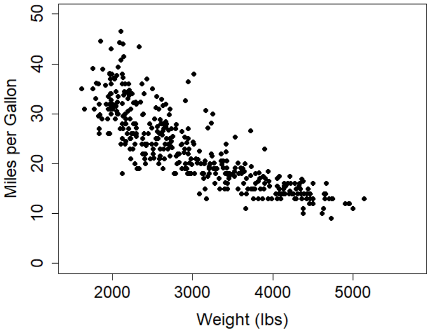
- Think of **randomness** as like **flipping a coin**.
- Prior to the experiment, the outcome is unknown (modelled as random)
- When we flip the coin, we get H or T
  - This is called a realization *(of the coin flipping experiment)*

Experiment	Realization
 justflipacoin.com	

Randomness in Statistical Models

Phenomenon	Model	Realization
Coin Flipping	$\mathbb{P}(H) = 0.5, \mathbb{P}(T) = 0.5$	H or T
Linear Regression	$f(\texttt{input}) = \underbrace{\alpha + \beta \cdot \texttt{input}}_{\text{deterministic}} + \underbrace{\texttt{error}}_{\text{random}}$	`output` (for given `input`)
Gaussian Process	Determined by Kernel	The entire function f

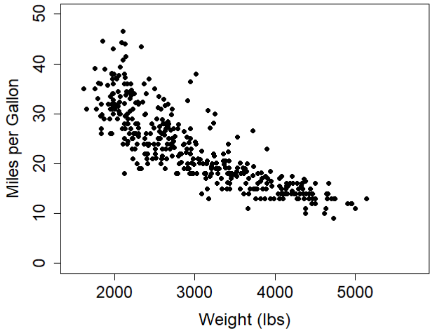
Which one is correct?



Linear Regression vs Gaussian Process

- Neither is "correct"!
- There isn't a right answer.
- Remember the quote: *"All models are wrong, but some are useful"*

Which one is correct?



Linear Regression vs Gaussian Process

- Neither is "correct"!
- There isn't a right answer.
- Remember the quote: "*All models are wrong, but some are useful*"

Understanding Models with Randomness

$$f(\texttt{input}) = \underbrace{\alpha + \beta \cdot \texttt{input}}_{\text{deterministic}} + \underbrace{\texttt{error}}_{\text{random}} \qquad \text{(Linear Regression)}$$

- Data is assumed to be a realization of this process
- Just like flipping a coin multiple times produces a sequence

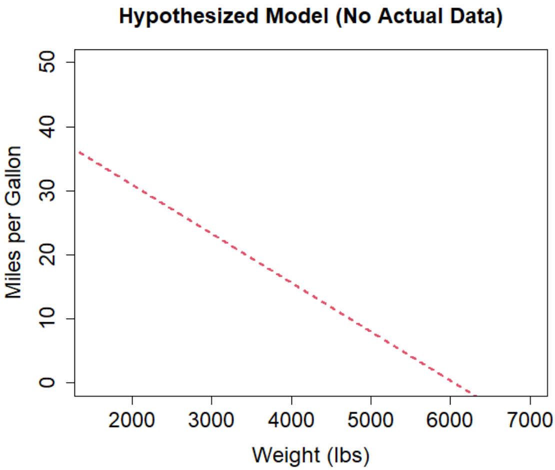
$$H, T, T, H, H, H, T, H, H, T, H, T, T, \ldots$$

Phenomenon	Model	Realization
Coin Flipping	$\mathbb{P}(H) = 0.5, \mathbb{P}(T) = 0.5$	H or T
Linear Regression	$f(\texttt{input}) = \underbrace{\alpha + \beta \cdot \texttt{input}}_{\text{deterministic}} + \underbrace{\texttt{error}}_{\text{random}}$	`output` (for given `input`)
Gaussian Process	Determined by Kernel	The entire function f

Visualization

Suppose our hypothesized model follows

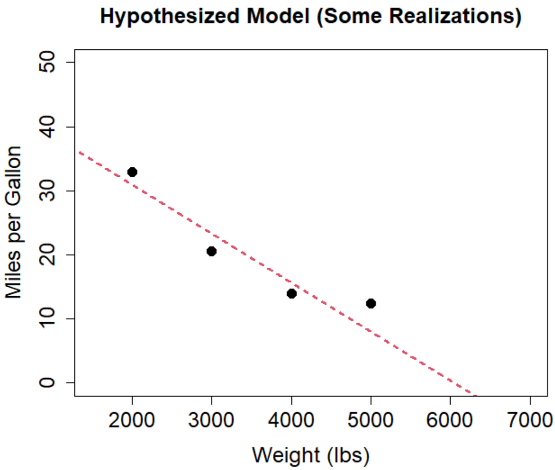
$$\texttt{mpg} = 46.22 - 0.0076 \cdot \texttt{lbs} + \texttt{error}$$



Visualization

Suppose our hypothesized model follows

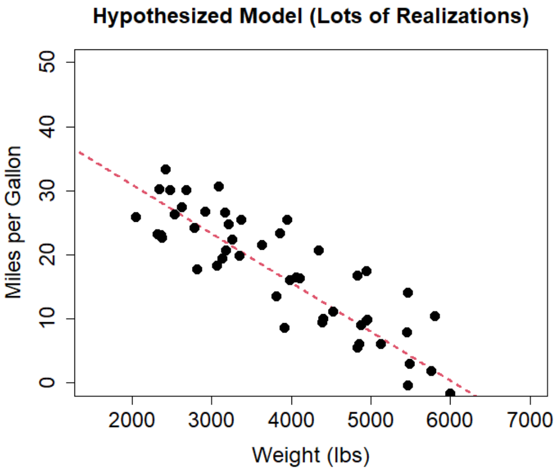
$$\texttt{mpg} = 46.22 - 0.0076 \cdot \texttt{lbs} + \texttt{error}$$



Visualization

Suppose our hypothesized model follows

$$\texttt{mpg} = 46.22 - 0.0076 \cdot \texttt{lbs} + \texttt{error}$$

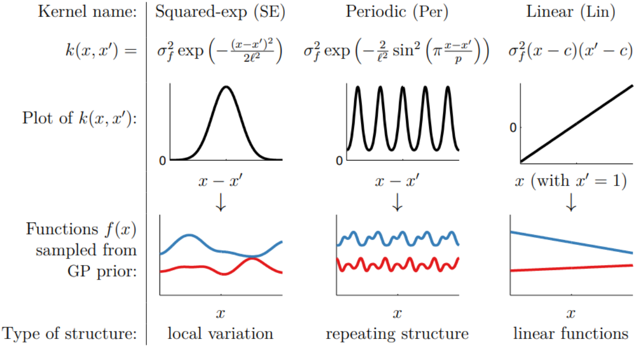


Gaussian Process Visualization

- A Gaussian process assumes the entire function is random

$$\underbrace{f}_{\text{random}}(\texttt{input}) = \texttt{output} \qquad \text{(Gaussian process)}$$

- The function properties are determined by its covariance kernel



Automatic Model Construction with Gaussian Processes by Duvenaud

Evidence As A Competitive Method (Gaussian Processes)

Individual Life Experience Committee Mortality Prediction and Presentation Contest

SOA Individual Life Experience Committee 2021 Mortality Forecasting Contest Results

We are happy to announce that we have 3 winners for the SOA Individual Life Experience Committee 2021 Mortality Forecasting Contest. The competition provided for one first place entry and two second place entries. The winners are as follows:

First Place

Nhan Huynh
Mike Ludkovski
James Risk

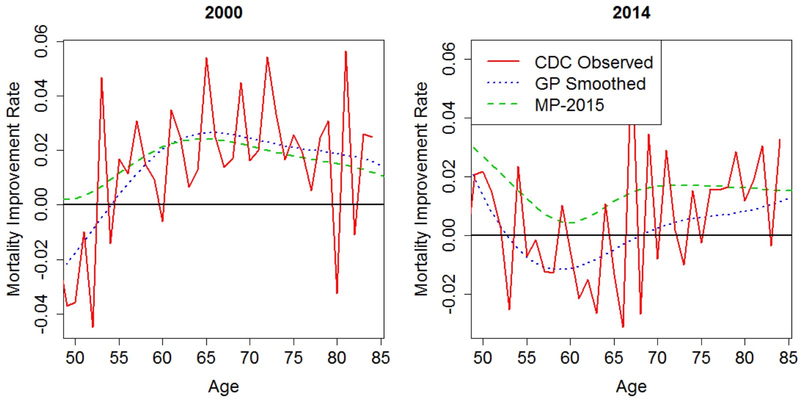
Second Place

1. Zach Stenberg, ASA, MAAA
2. Shuxian Ning
   Shuyu Zhu

We will follow up shortly with a write-up from the judges providing more details regarding the submissions received.

Source: https://www.soa.org/research/opportunities/2021-individual-life-experience-contest/

Mortality Modelling Example

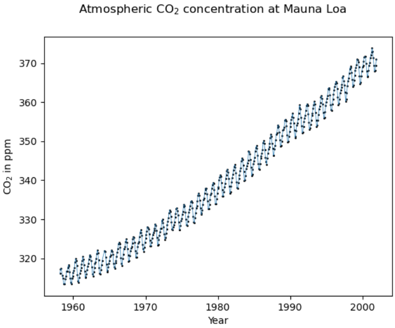


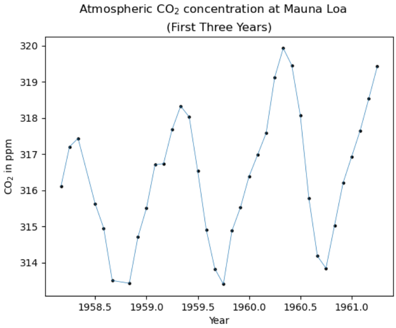
GAUSSIAN PROCESS MODELS FOR MORTALITY RATES AND IMPROVEMENT FACTORS by Jimmy Risk, Mike Ludkovski, and Howard Zail (ASTIN Bulletin 2018)

- **CDC Observed**: Actual mortality improvement data
- **GP Smoothed**: Gaussian process smoothed mortality improvement ($f(\texttt{age}, \texttt{calendar year})$)
- **MP-2015**: Society of Actuaries Gold Standard of Mortality Improvement *(at the time)*

Example: Mauna Loa Data Set

- y: **monthly** average atmospheric CO_2 concentrations (in ppm by volume) derived from air samples at the Mauna Loa Observatory, Hawaii, between 1958 and 2003, with some missing values
- x: month





Example: Mauna Loa Data Set (Kernel Choice)

Model the apparent features[6]:

- Long term rising trend

$$k_1(x,x') = \theta_1^2 \exp\left(-\frac{(x-x')^2}{2\theta_2^2}\right)$$

where θ_1 is the amplitude, and θ_2 is the characteristic length-scale

- Yearly decaying periodicity

$$k_2(x,x') = \theta_3^2 \exp\left(-\frac{(x-x')^2}{2\theta_4^2}\right) \exp\left(-\frac{2\sin^2(\pi(x-x'))}{2\theta_5^2}\right)$$

where θ_3 is the magnitude, θ_4 is the decay-time, and θ_5 is the smoothness of the periodic component.

[6] This particular construction is taken from Gaussian Processes for Machine Learning by Rasmussen and Williams

Example: Mauna Loa Data Set (Kernel Choice)

Model the apparent features[6]:

- Long term rising trend

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right)$$

where θ_1 is the amplitude, and θ_2 is the characteristic length-scale

- Yearly decaying periodicity

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2}\right) \exp\left(-\frac{2\sin^2(\pi(x - x'))}{2\theta_5^2}\right)$$

where θ_3 is the magnitude, θ_4 is the decay-time, and θ_5 is the smoothness of the periodic component.

[6]This particular construction is taken from Gaussian Processes for Machine Learning by Rasmussen and Williams

Example: Mauna Loa Data Set (Kernel Choice, Continued)

- (Small) medium term irregularities

$$k_3(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8 \theta_7^2}\right)^{-\theta_8}$$

where θ_6 is the magnitude, θ_7 is the typical length-scale, and θ_8 is the shape parameter

- Noise term

$$k_4(x, x') = \theta_9^2 \exp\left(-\frac{(x - x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{x=x'},$$

where θ_9 is the magnitude of the correlated noise component, θ_{10} is its length-scale, and θ_{11} is the magnitude of the independent noise component.

Final covariance function:

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x')$$

Example: Mauna Loa Data Set (Kernel Choice, Continued)

- (Small) medium term irregularities

$$k_3(x,x') = \theta_6^2 \left(1 + \frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

where θ_6 is the magnitude, θ_7 is the typical length-scale, and θ_8 is the shape parameter

- Noise term

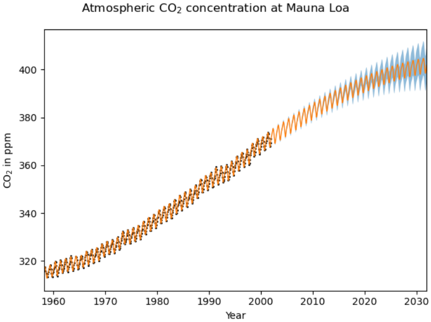
$$k_4(x,x') = \theta_9^2 \exp\left(-\frac{(x-x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{x=x'},$$

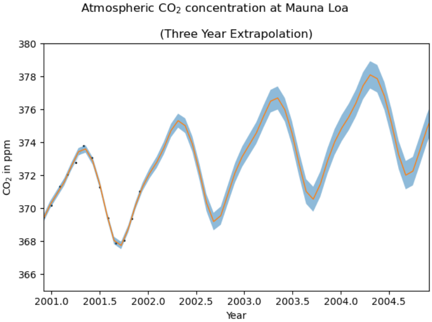
where θ_9 is the magnitude of the correlated noise component, θ_{10} is its length-scale, and θ_{11} is the magnitude of the independent noise component.

Final covariance function:

$$k(x,x') = k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x')$$

Example: Mauna Loa Data Set (Posterior Prediction)






```
Learned kernel:
2.63**2 * RBF(length_scale=51.6) +
0.155**2 * RBF(length_scale=91.5) * ExpSineSquared(length_scale=1.48,
                                                   periodicity=1) +
0.0314**2 * RationalQuadratic(alpha=2.89, length_scale=0.968) +
0.011**2 * RBF(length_scale=0.122) + WhiteKernel(noise_level=0.000126)
```

Gaussian Process Superresolution

- **Super-resolution** is the task of reconstructing high-resolution (HR) images from one or more observed low-resolution (LR) image
- Different from smoothing out noise in images (*does not restore high resolution details*)
- Seminole work *Single Image Super-Resolution Using Gaussian Process Regression* by He, et. al. uses only the **squared-exponential kernel**
  - A popular and flexible kernel
  - Has its limits
- Our idea:
  - Explore using other kernels
  - Construct an algorithm to search over kernels based on image
  - Identify what kernels are useful for determining certain features in images

Gaussian Process Superresolution

- **Super-resolution** is the task of reconstructing high-resolution (HR) images from one or more observed low-resolution (LR) image
- Different from smoothing out noise in images (*does not restore high resolution details*)
- Seminole work *Single Image Super-Resolution Using Gaussian Process Regression* by He, et. al. uses only the **squared-exponential kernel**
  - A popular and flexible kernel
  - Has its limits
- Our idea:
  - Explore using other kernels
  - Construct an algorithm to search over kernels based on image
  - Identify what kernels are useful for determining certain features in images

Staircase (Test Image)

(a) Ground Truth (b) Low Resolution (c) Bicubic Interpolation

Figure: Effects of varying kernels on image reconstruction. Image was downscaled from Ground truth 192×192 to 96×96 before applying Bicubic Interpolation.

Kernel Effects on Gaussian Process Staircase SR

(a) Linear Kernel

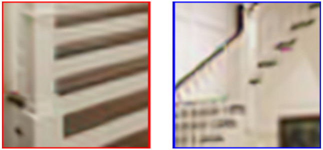
(b) RBF (Smooth) Kernel

(c) Non-Smooth Kernel

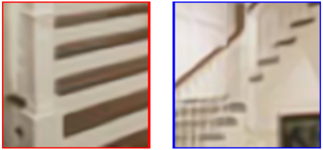
(d) Periodic Kernel

Kernel Effects on Gaussian Process Staircase SR

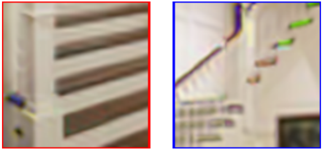
(a) Linear Kernel



(b) RBF (Smooth) Kernel



(c) Non-Smooth Kernel



(d) Periodic Kernel

(e) Bicubic Interpolation

Automatic Kernel Searching

- We applied a automatic kernel search algorithm from *Automatic Model Construction with Gaussian Processes* (Duvenaud)
- The kernel it came up with was

$$\text{MAT}_{\frac{3}{2}} + \text{Linear} + \text{Periodic}$$

  - $\text{MAT}_{\frac{3}{2}}$: Measures similarity according to spatial closeness
  - Linear: Produces a linear trend effect
  - Periodic: Adds a periodic component

Staircase (Final Comparison)

(a) Ground Truth

(b) Low Resolution

(c) Bicubic Interpolation

(d) GP (Best Kernel)

Final Comparison (Highlights)

(a) GP (Best Kernel)

(b) Bicubic Interpolation

Thank You!

Our work:

- Ludkovski, Mike, Jimmy Risk, and Howard Zail. "Gaussian process models for mortality rates and improvement factors." *ASTIN Bulletin: The Journal of the IAA 48.3 (2018): 1307-1347.*
- Amelin, Charles P. *GAUSSIAN PROCESS SUPER-RESOLUTION.* Diss. California State Polytechnic University, Pomona, 2021.