Gaussian Process Regression and Emulation

STAT8810, Fall 2017

M.T. Pratola

August 25, 2017

Today

Ball Drop Experiment

Ball Drop Experiment

- Recall our ball-drop experiment, which gave us a closed-form solution for drop time:

$$t_f = \beta\sqrt{h_0} = t_f(h_0)$$

  where $\beta = \sqrt{-2/g}$.
- So $t_f(h_0)$ describes the behavior of our ball, at least according to physics and under some assumptions.
- More generally, we might think of it as

$$t_f = \eta(h_0; g).$$

Ball Drop Experiment

- What if the initial condition $h'(0)$ is unknown and/or needs to be estimated?
- The solution becomes more complicated:

$$h(t) = \frac{1}{2}gt^2 + v_0 t + h_0$$

  where v_0 is the initial velocity at time 0.

- $t_f(g, v_0, h_0)$ is the solution to $\frac{1}{2}gt^2 + v_0 t + h_0 = 0$ since $h(t_f) = 0$:

$$t_f = \frac{-v_0 - \sqrt{v_0^2 - 2gh_0}}{g}$$

- Could also think of it as $t_f = \eta(h_0; g, v_0)$, which is some simulator model (e.g. Newton-Raphson) that “solves” the equation above.

Ball Drop Experiment

- Suppose we implement $t_f(h_0)$ as a computer code, let's call it $\eta(h_0)$.
- Suppose we have a *statistical emulator* $Z(\cdot)$.
- What is a statistical model for the drop time?
  - $t_{f\,i} = \eta(h_{0,i}) + \epsilon_1 = Z(h_{0,i}) + \epsilon_2$
  - $\epsilon_1 \sim N(0, \sigma_1^2)$
  - $\epsilon_2 \sim N(0, \sigma_2^2)$
- What does σ_1^2 represent? What about σ_2^2?

Modeling the Ball Drop Experiment

- What does σ_1^2 represent?
  - error from approximating square-root on computer
  - error of using the computers floating point representation of the real numbers.

https://en.wikipedia.org/wiki/Methods_of_computing_square_roots

Modeling the Ball Drop Experiment

- What does σ_2^2 represent?
    - all of the above + lack of fit from using our stat model $z(\cdot)$
    -

      ie. $\sigma_2^2 = \sigma_1^2 + \tilde{\sigma}^2 \equiv$ "computer code error" + "model error".

Modeling the Ball Drop Experiment

- What about the more complicated case of $t_f = \eta(h_0; g, v_0)$?
  - σ_1^2 represents the above plus other sources of error, such as error from the solver routine used to approximate the solution of $\frac{1}{2}gt^2 + v_0t + h_0 = 0$.
  - e.g. this may be related to the error tolerance of a Newton-Raphson scheme for solving this equation.
  - or, the accuracy of a Runge-Kutta differential equation solver.
  - etc. . .

Modeling the Ball Drop Experiment

- Simplifying assumptions: assume $\epsilon_1 = \epsilon_2 = 0$.
- This means our statistical model $Z(\cdot)$ should interpolate the output of $\eta(\cdot)$.
- This *seems* innocuous – in simple problems, this error is on the order of machine precision $\sim 1e - 16$.
- On the other hand, the mathematics literature is filled with decades of research on implementing numerical codes on computer for solving intractable math problems while controlling/minimizing the error of approximation.
- So how innocuous is it, really?

Emulating Outputs from a Simulator

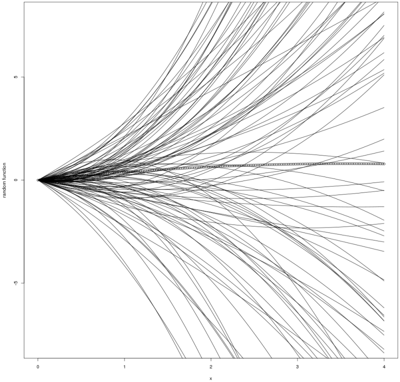
- Say we collect $\mathbf{y} = (y(x_1), \ldots, y(x_n))$ at $(x_1, \ldots, x_n)$ unique settings of the simulator inputs
- A statistical model for this data is to treat $\mathbf{y}$ as observations from an unknown realization y of a stochastic process $\{Y(x) : x \in [l, u] \subset \mathbb{R}\}$ where the realization was observed at $x_1, \ldots, x_n$. (I often will just write $Y(x)$).

Random Functions

- Recall: $Y \sim f_Y$ - Y is a random variable
- Recall: $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim f_{\mathbf{Y}}$ - Y is a random vector
- Can we have $\mathbf{Y} \sim f_{\mathbf{Y}}$ where $\mathbf{Y}$ is a random function? It turns out yes, we can.

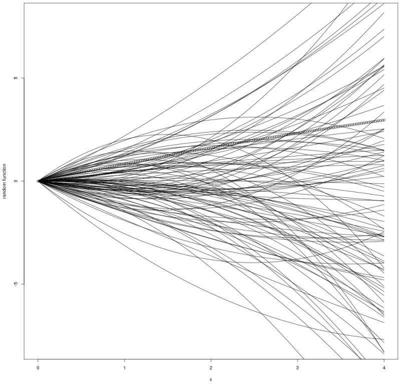
Random Polynomial Function Example

- Consider $y = \beta_1 x + \beta_2 x^2 = X^T\beta$ with $\beta = (\beta_1, \beta_2)^T \sim N(0, \Sigma)$. Here, $\Sigma = I$.



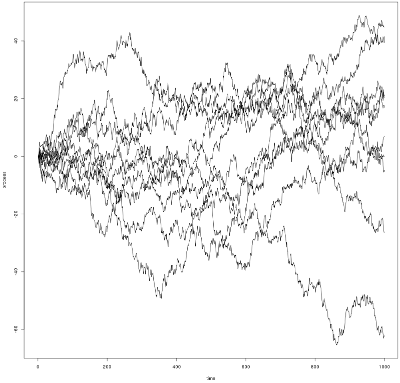
Random Polynomial Function Example

- Consider $y = \beta_1 x + \beta_2 x^2 = X^T \beta$ with $\beta = (\beta_1, \beta_2)^T \sim N(0, \Sigma)$. Here, $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = -0.9$.



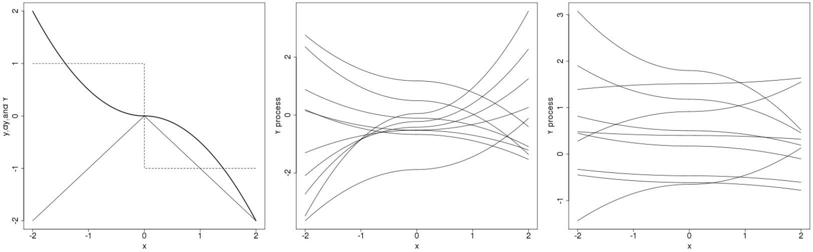
Wiener Process

- Say $\{Z_i\}_{i=1}^{n}$ are iid standard normal random variables and $t \in [0,1]$ and let $[nt]$ be the integer part of nt. Define $w_{n,t} = \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} z_i$. Then $w_{n,t}$ converges in distribution to a Wiener process on $[0,1]$ as $n \to \infty$. Wiener process is continuous but nowhere differentiable.



Example of a Once-Differentiable Random Functions

- Recall that $y = abs(x)$ is continuous but not differentiable as the derivative has a discontinuity at $x = 0$. It follows that the integral is a once-differentiable function: $y(x) = \frac{1}{2}\beta_1 x^2 + \beta_0$ when $x \geq 0$ and $y(x) = -\frac{1}{2}\beta_1 x^2 + \beta_0$ otherwise. And take $\beta \sim N(0, \Sigma)$



Gaussian Random Function Models

- Since Gaussian distributions are fully specified by their mean & covariance, our function space will be defined through specifying these two parameters.
- Want flexibility so we can model a wide range of “function data”
- Want predictability - data should inform about function values at unobserved x’s, and the closer the data to such an x the more predictive the data is expected to be.

Gaussian Process

- A stochastic process $Y(x), x \in \chi \subset \mathbb{R}^d$ is a collection of random variables with underlying probability space $(\Omega, \mathcal{F}, P)$.

$$\begin{tikzcd} & Y(\cdot):\mathbb{R}^d \to \mathbb{R} \text{ (a function)} \\ Y(x,\omega), x\in\chi, \omega\in\Omega \arrow[ur, "\text{fix } \omega"] \arrow[dr, "\text{fix } x"'] & \\ & Y(\cdot):\Omega\to\mathbb{R} \text{ (a random variable)} \end{tikzcd}$$

- A stochastic process $Y(x)$ is an infinite-dimensional *Gaussian Process* (GP) if for any $x_1, \ldots, x_n \in \chi$ and any n finite, the joint distribution of $y(x_1), \ldots, y(x_n) \equiv \mathbf{y}$ is

$$\mathbf{y} \sim MVN\left(\mu(\mathbf{x}), \Sigma(\mathbf{x})\right)$$

- We will assume $Y(x)$ is a stationary process.

Strict Stationarity

- $\{Y(x)\}_{x \in \chi}$ is strictly stationary if for any $k \geq 1$ and any $x_1, \ldots, x_k \in \chi$ and any h s.t. $x_1 + h, \ldots, x_k + h \in \chi$ then

$$P(Y(x_1), \ldots, Y(x_k)) = P(Y(x_1 + h), \ldots, Y(x_k + h))$$

- Properties:
  - (i.) $P(Y(x))$ is the same for all x. E.g. for a GP, this means the variance of the marginal distribution of $Y(x)$ is the same $\forall x \in \chi$.
  - (ii.) Suppose $\{Y(x)\}_{x \in \chi}$ is strictly stationary and $Y(x)$ has finite mean and variance. Then $Cov(Y(x_i), Y(x_j)) = c(x_j - x_i) = c(h)$ where $h = x_j - x_i$. $c(\cdot)$ is called the *covariance function*.

Weak Stationarity a.k.a. Covariance Stationarity

- Suppose $\{Y(x)\}_{x \in \chi}$ has finite second moments $\forall x$.

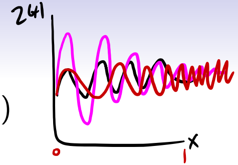
Definition: a process $\{Y(x)\}_{x \in \chi}$ with second moments is covariance stationary if

- (i.) $E[Y(x)]$ is the same $\forall x$.
- (ii.) $Cov(Y(x_i), Y(x_j)) = c(x_j - x_i)$.

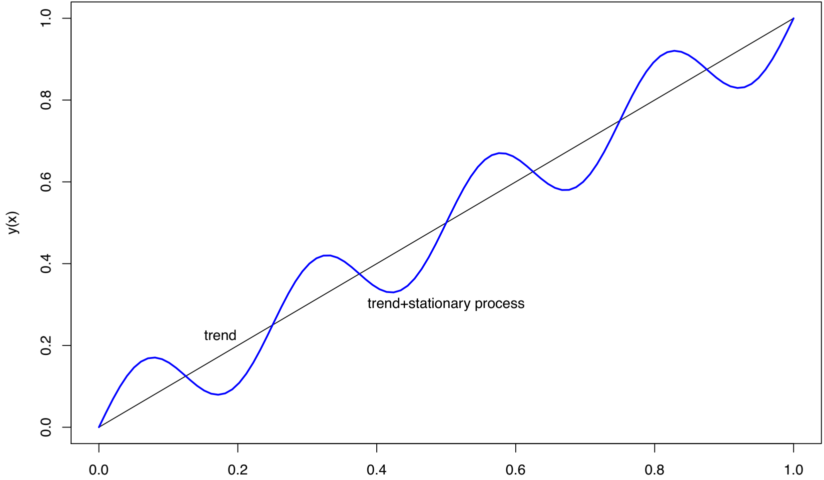
Note: if $\{Y(x)\}$ is strictly stationary and has finite second moments, then $\{Y(x)\}$ is covariance stationary.

- Fact: the GP $\{Y(x)\}_{x \in \chi}$ is strictly stationary $\Leftrightarrow$ $\{Y(x)\}_{x \in \chi}$ is covariance stationary (the MVN distribution is fully characterized by it's mean and covariance).

Gaussian Process



- $Y(x) = \sum_i \beta_i f_i(x) + Z(x), Z(x) \sim GP(0, c(\cdot))$
  - $\sum_i \beta_i f_i(x)$ captures large-scale trends
  - $Z(x)$ captures smaller-scale variability



Covariance Functions

- Suppose $\{Z(x)\}_{x \in \chi}$ is a stationary GP, so $\text{Cov}(Z(x_i), Z(x_j)) = c(x_j - x_i), x_i \in \chi \subset \mathbb{R}^d$ and where $c(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a covariance function.

Property 1:

every valid covariance function must satisfy $c(h) = c(-h)$, that is, covariance functions are *even*.

Property 2:

every valid covariance function must be non-negative definite, that is:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j c(x_j - x_i) \geq 0$$

for any k and any $\alpha_i, \alpha_j \in \mathbb{R}$ and any $x_i, x_j \in \chi$.

Covariance Functions

- One way to think of this is as

$$Var\left(\sum_i w_i Y(x_i)\right) \geq 0$$

- Another way results from the p.s.d. requirement of covariance matrices, since it must hold that the quadratic form

$$\mathbf{y}^T \Sigma^{-1} \mathbf{y} \geq 0.$$

Correlation Functions

- It is actually much more popular to work with correlation functions:

$$V(Z(x)) = \sigma^2 \forall x \in \chi$$

$$Cor(Z(x_i), Z(x_j)) = \frac{Cov(Z(x_i), Z(x_j))}{\sqrt{\sigma^2}\sqrt{\sigma^2}} = \frac{c(x_j - x_i)}{\sigma^2} = R(x_j - x_i).$$

- If $x_i = x_j$ then $Cor(x_i, x_j) = \sigma^2/\sigma^2 = 1$, i.e. $R(x_j - x_i) = R(0) = 1$.
- Correlation functions must also satisfy the non-negative definite property.

Isotropy

- A more restrictive correlation function is

$$Cor(x_i, x_j) = R(||x_j - x_i||)$$

  where $|| \cdot ||$ denotes Euclidean distance.

  - this model implies *rotational invariance*.

Anisotropy

- A correlation function is anisotropic if

$$Cor(x_i, x_j) = R(||x_j - x_i||_K)$$

where

$$||x_j - x_i||_K^2 = (x_j - x_i)^T K(x_j - x_i).$$

  - this model implies stretching/scaling along axes (when K is diagonal) and possibly axis rotation (much like PCA regression).
  - the most popular anisotropic models take K to be a diagonal matrix.