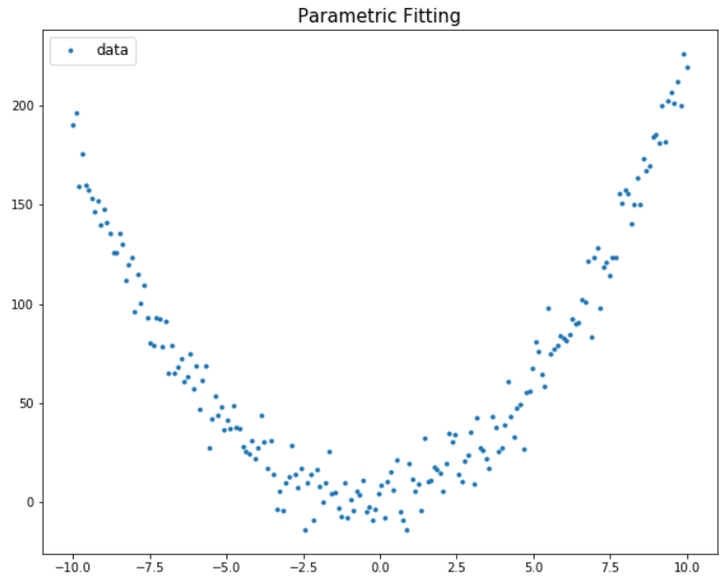
The What, Where, Why, and How of Gaussian Processes (GPs)

By: Ari Silburt

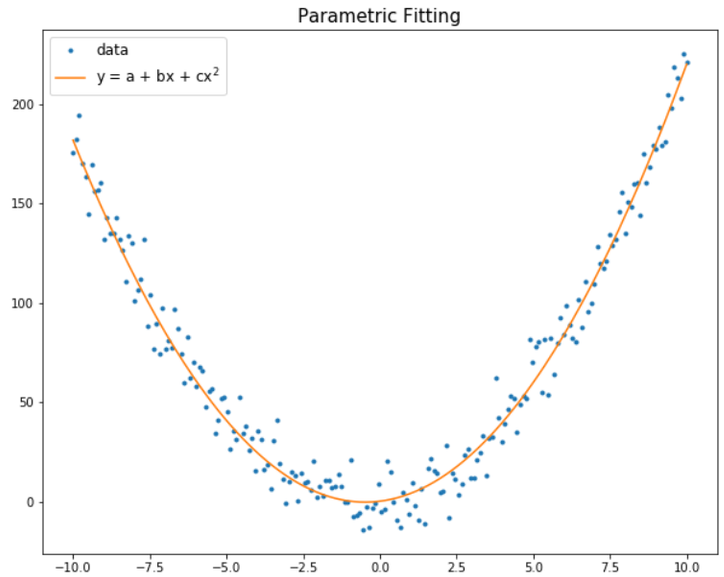
Why GPs?

Standard Fitting to Data: Parametric Modelling/Linear Regression



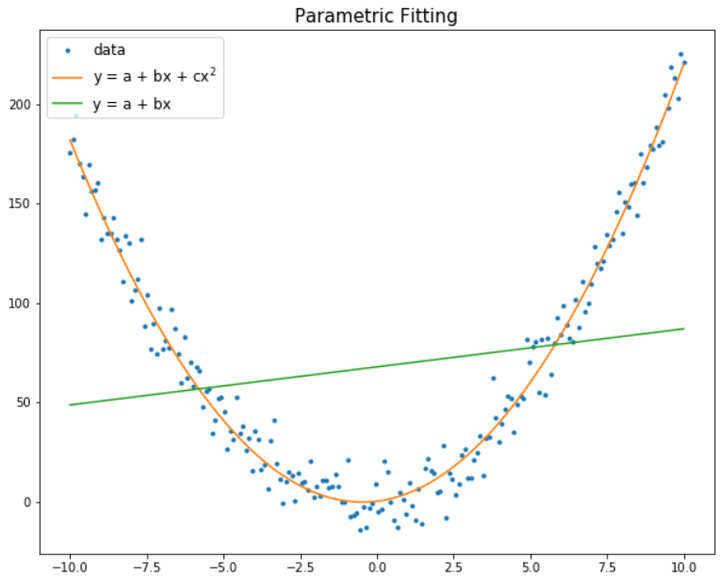
Why GPs?

Standard Fitting to Data: Parametric Modelling/Linear Regression



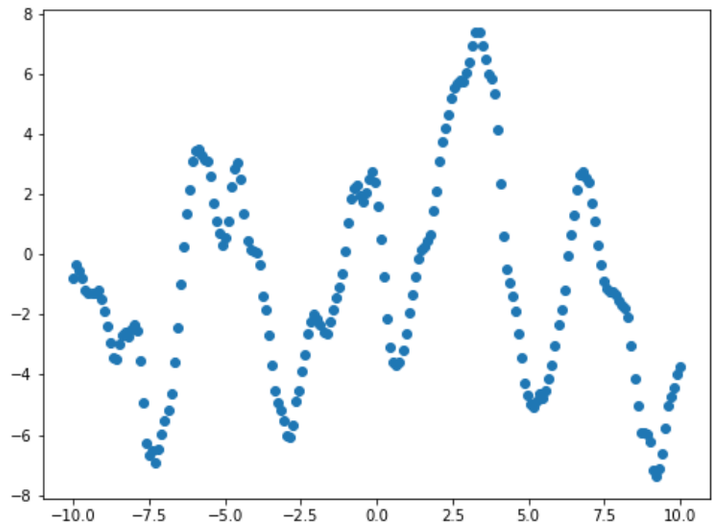
Why GPs?

Standard Fitting to Data: Parametric Modelling/Linear Regression



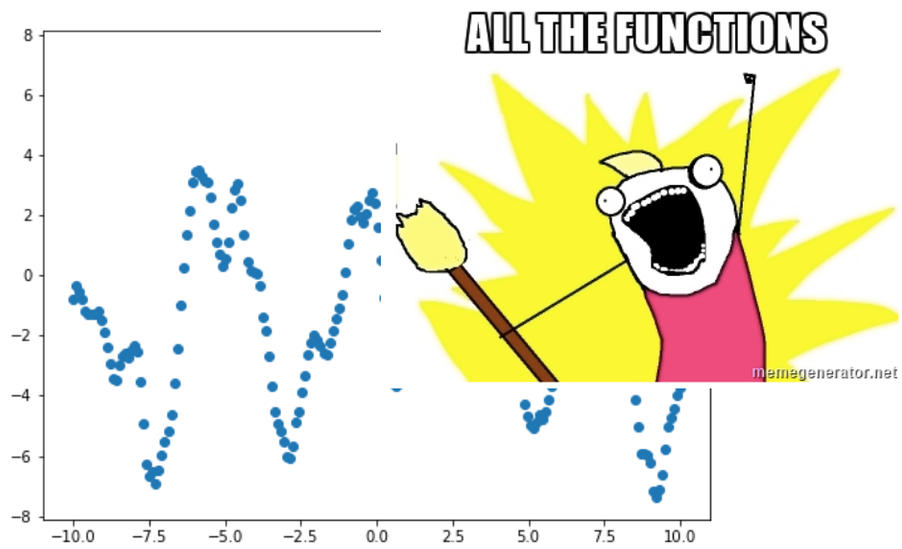
Why GPs?

So… what function do we want to fit this?



Why GPs?

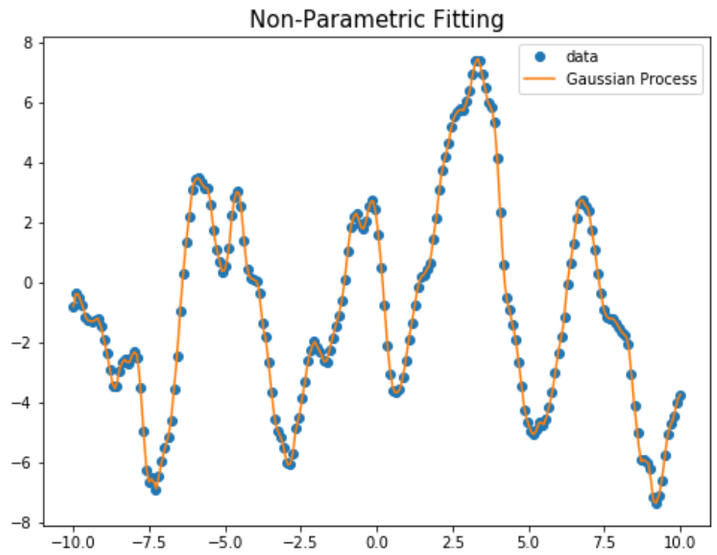
So… what function do we want to fit this?



We’d like to consider every possible function that matches our data, with however many parameters are involved. That’s what non-parametric means: it’s not that there aren’t parameters, it’s that there are infinitely many parameters.

Why GPs?

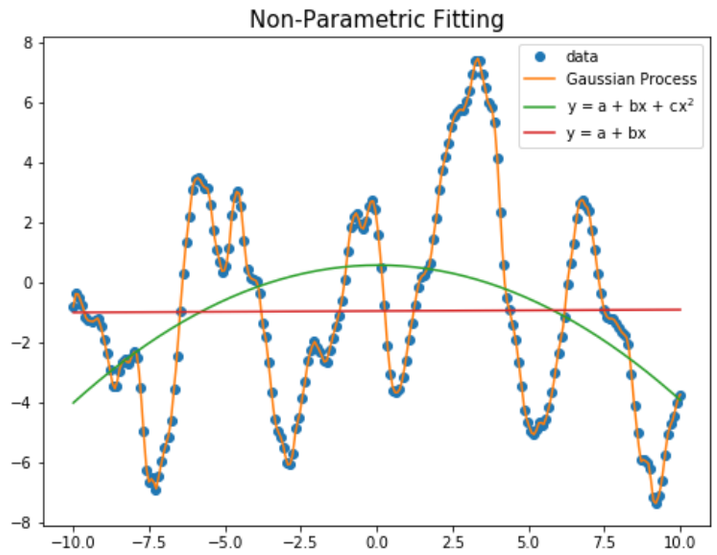
Nonparametric = infinitely many parameters (not zero parameters)



GPs encode general properties: smooth, continuous and variations in the function take place over characteristic time scales (not too slowly yet not so fast) and have typical amplitude.

Why GPs?

Nonparametric = infinitely many parameters (not zero parameters)



-We can see in this case that the wrong parametric form gives terribly incorrect answers. You could spend ages trying to hand-tune the right parametric form.
-In addition, we may only know that our observations come from an underlying process that is smooth, continuous, have variations over characteristic time scales (not too slowly yet not so fast) and have a typical amplitude. Surprisingly, we may work mathematically with the infinite space of all functions that have these properties.
-So, why not use GPs?

What (are) GPs?

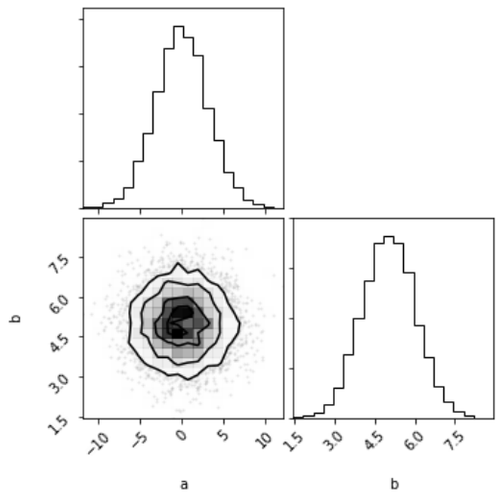
Carl Rasmussen (i.e. GP God):

> *A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.*

What (are) GPs?

Carl Rasmussen (i.e. GP God):

A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.



$$\text{mean} = [0, 5]$$

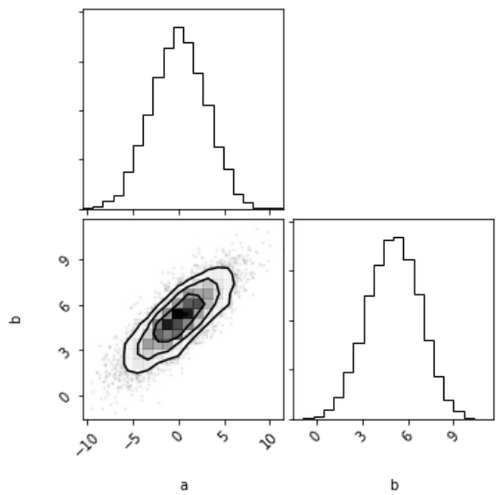
$$\text{cov} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$

$$D = N(\text{mean}, \text{cov})$$

What (are) GPs?

Carl Rasmussen (i.e. GP God):

A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.



$$\text{mean} = [0, 5]$$

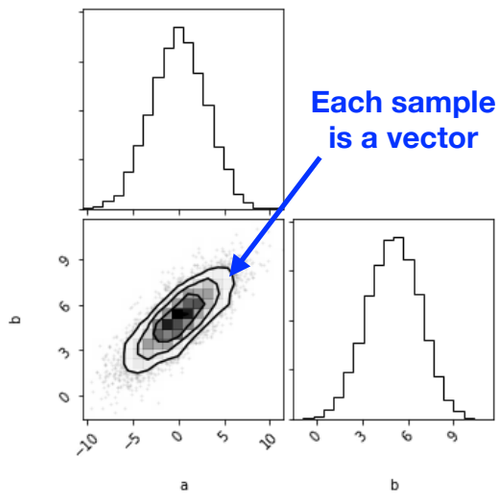
$$\text{cov} = \begin{bmatrix} 10 & 5 \\ 0 & 1 \end{bmatrix}$$

$$D = N(\text{mean}, \text{cov})$$

What (are) GPs?

Carl Rasmussen (i.e. GP God):

A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.



$$\text{mean} = [0, 5]$$

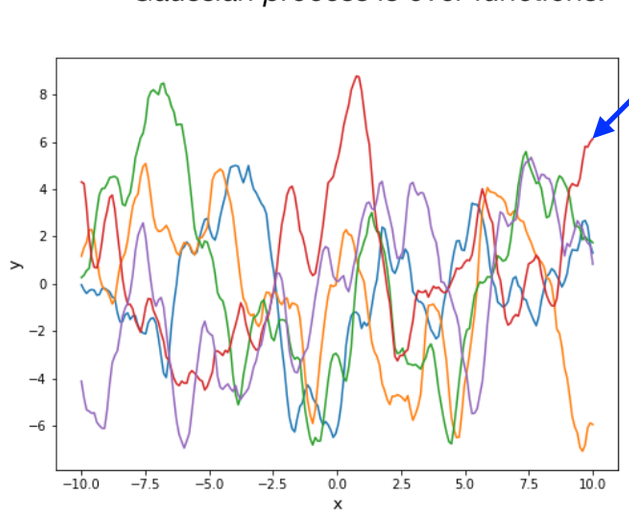
$$\text{cov} = \begin{bmatrix} 10 & 5 \\ 0 & 1 \end{bmatrix}$$

$$D = N(\text{mean}, \text{cov})$$

What (are) GPs?

Carl Rasmussen (i.e. GP God):

A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.



$$\text{mean} = 0$$

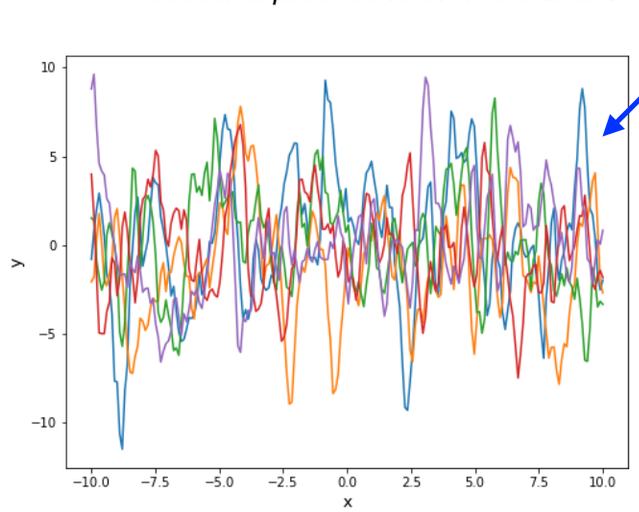
$$\text{cov} = \exp(-(\text{x} - \text{x}')^2)$$

$$f = GP(\text{mean}, \text{cov})$$

What (are) GPs?

Carl Rasmussen (i.e. GP God):

A Gaussian process is fully specified by its mean function m(x) and covariance function k(x, x'). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.



$$\text{mean} = 0$$

$$\text{cov} = \exp(-0.1(\text{x} - \text{x}')^2)$$

$$f = GP(\text{mean}, \text{cov})$$

How (to use) GPs

<u>GPs in a Bayesian Framework:</u>

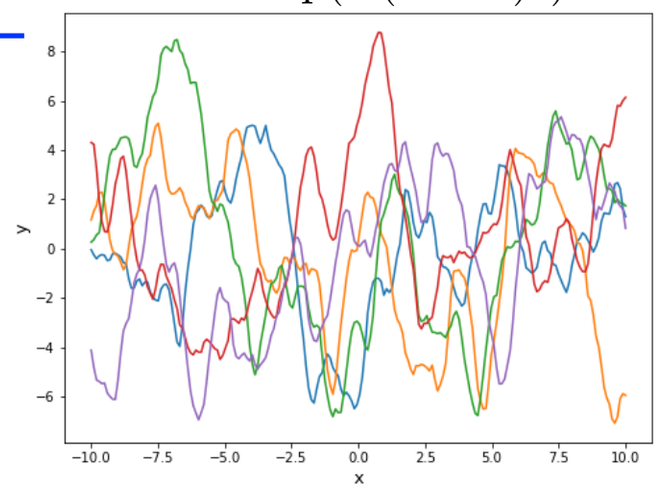
- Prior information about model.
- See some data.
- Update model by conditioning on data, arriving at a posterior distribution.

How (to use) GPs

GPs in a Bayesian Framework:

- Prior information about model.

$$\mathrm{cov} = \exp(-(\mathrm{x} - \mathrm{x}')^2)$$



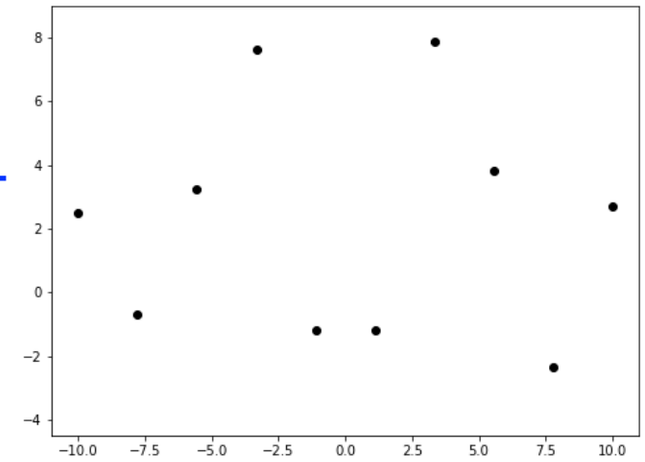
- See some data.

- Update model by conditioning on data, arriving at a posterior distribution.

How (to use) GPs

GPs in a Bayesian Framework:

- Prior information about model.
- See some data.
- Update model by conditioning on data, arriving at a posterior distribution.

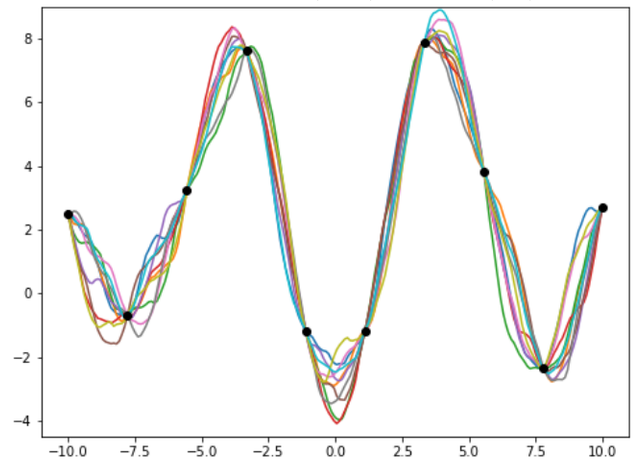


How (to use) GPs

GPs in a Bayesian Framework:

- Prior information about model.
- See some data.
- Update model by conditioning on data, arriving at a posterior distribution.

$$\mathrm{cov} = \exp(-(\mathrm{x} - \mathrm{x}')^2)$$



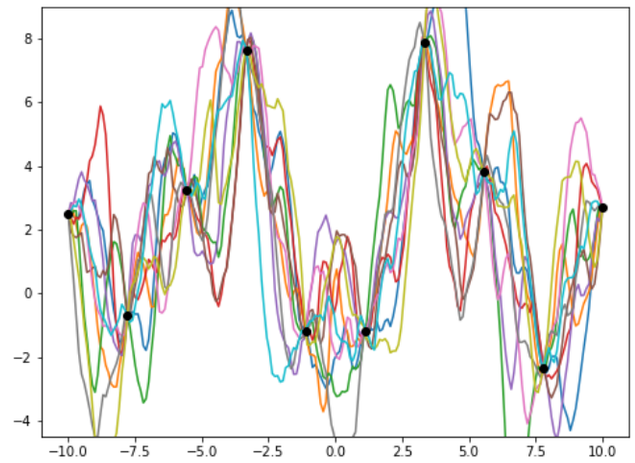
Conditioning means, “Given my data, what subset of those infinite number of functions are still valid”?

How (to use) GPs

GPs in a Bayesian Framework:

- Prior information about model.
- See some data.
- Update model by conditioning on data, arriving at a posterior distribution.

$$\mathrm{cov} = \exp(-0.1(x - x')^2)$$



What are the optimal parameters?

Conditioning means, “Given my data, what subset of those infinite number of functions are still valid”?

How (to use) GPs

Maximize the Marginal Likelihood:

$$L = \log p(y|x,\theta) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu) - \frac{n}{2}\log(2\pi)$$

How (to use) GPs

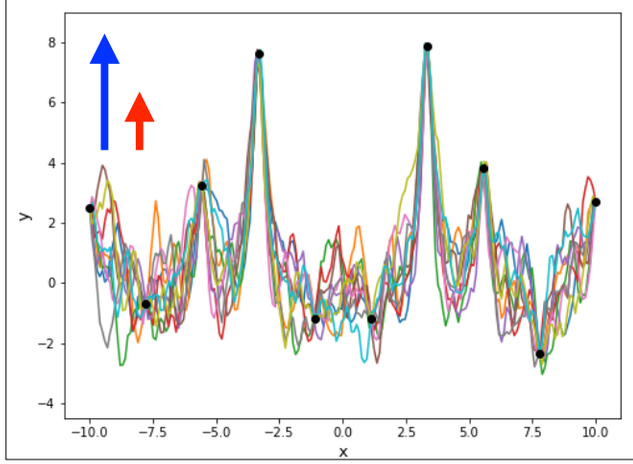
Maximize the Marginal Likelihood:

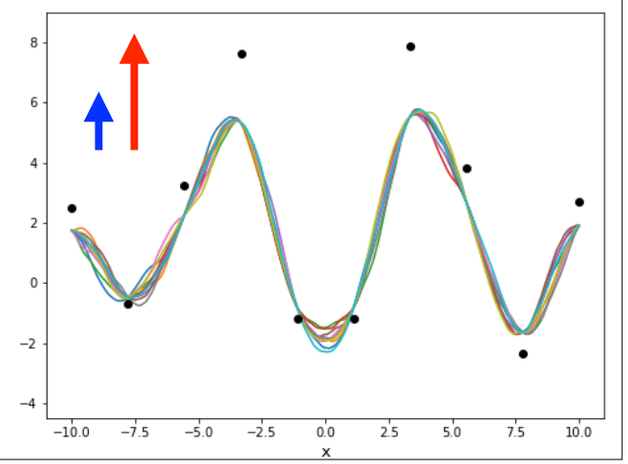
$$L = \log p(y|x,\theta) = \underbrace{-\frac{1}{2}\log|\Sigma|}_{\text{Complexity penalty term}} \underbrace{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)}_{\text{Goodness of fit term}} \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{Normalization term (useless)}}$$

How (to use) GPs

Maximize the Marginal Likelihood:

$$L = \log p(y|x,\theta) = \underbrace{-\frac{1}{2}\log|\Sigma|}_{\text{Complexity penalty term}} \underbrace{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)}_{\text{Goodness of fit term}} \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{Normalization term (useless)}}$$





Large arrow = bad. Since the likelihood is a sum of 3 negative terms, maximizing it actually means minimizing the absolute value of each term. I.e. the best fit has the smallest sum of red/blue arrows.

How (to use) GPs

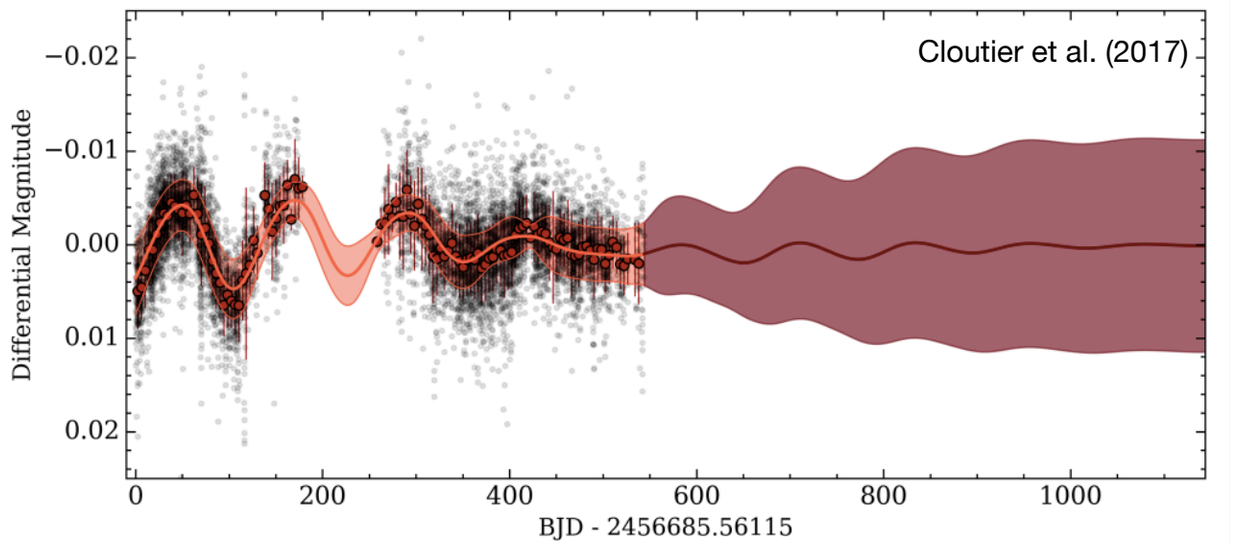
Daniel Foreman-Mackey's live demo

http://dfm.io/gp.js/

-first slider is amplitude (a), second slider is length scale (lambda, which controls complexity).

Where (are) GPs (used)

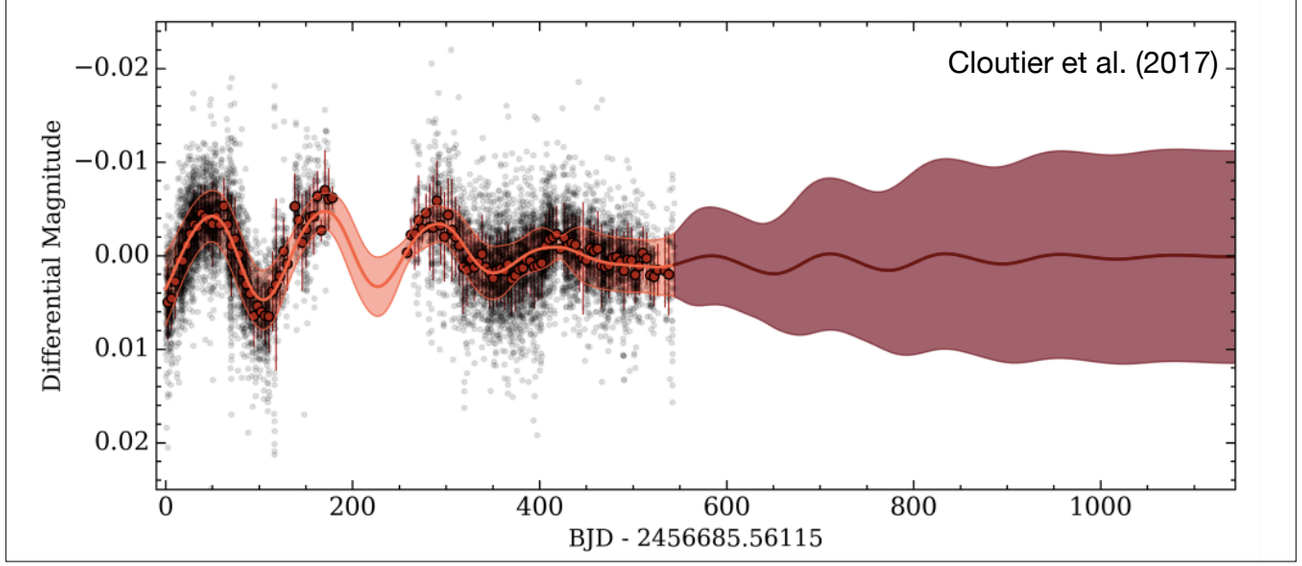
"Parametric models of stellar variability due to active regions feature degenerate model parameters including the sizes and spatial distribution of active regions thus making it difficult to accurately constrain model parameters of active regions."



-Lightcurve, with transiting planets removed, measuring stellar noise only. This is unlike RVs, which could still have hidden signals of undetected planets.

Where (are) GPs (used)

$$\mathrm{cov} = a^2 \exp\left[-\frac{(x-x')^2}{2\lambda^2} - \Gamma^2 \sin^2\left(\frac{\pi|x-x'|}{P_{rot}}\right)\right]$$

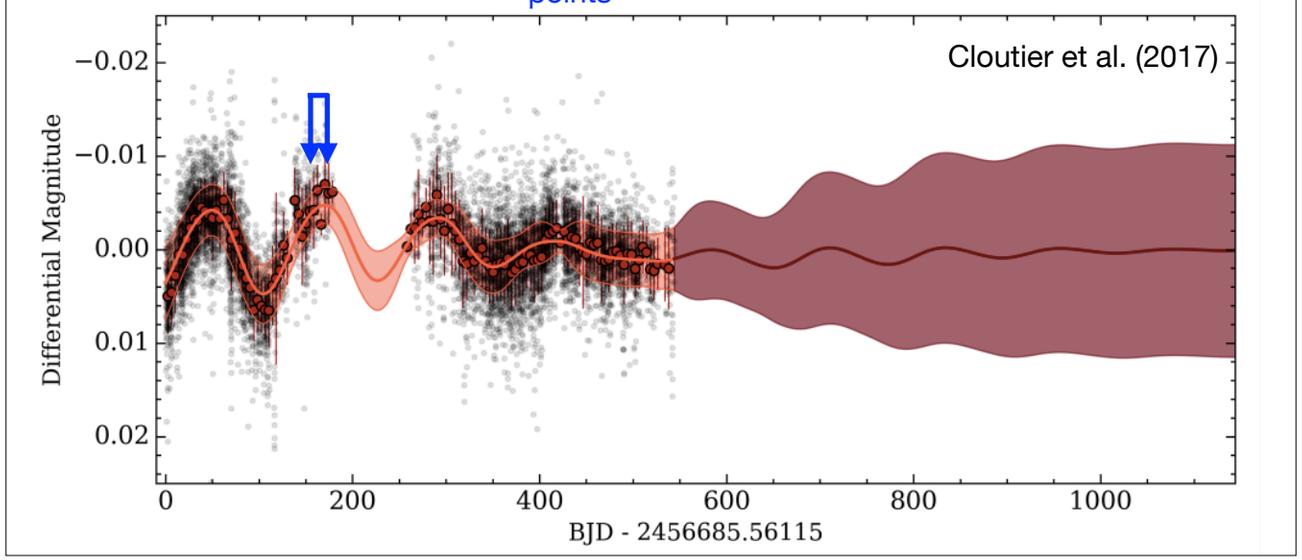


-First term is the standard exponential kernel that states nearby points are more correlated than far points.

Where (are) GPs (used)

$$\mathrm{cov} = a^2 \exp\left[-\frac{(x-x')^2}{2\lambda^2} - \Gamma^2 \sin^2\left(\frac{\pi|x-x'|}{P_{rot}}\right)\right]$$

Nearby points more correlated than far points

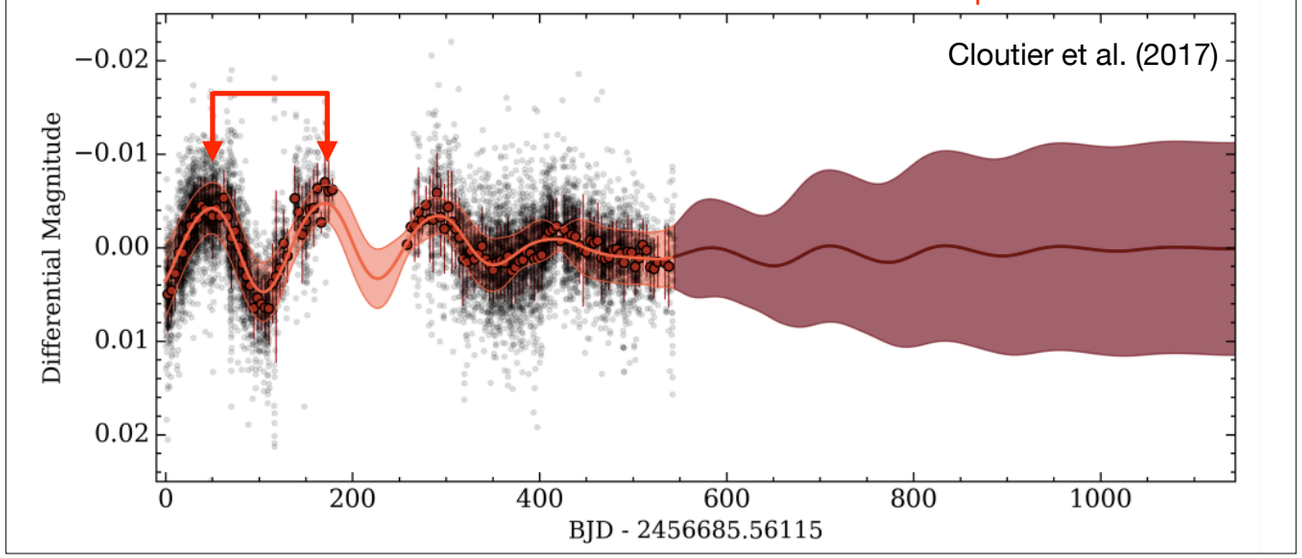


-First term is the standard exponential kernel that states nearby points are more correlated than far points.

Where (are) GPs (used)

$$\mathrm{cov} = a^2 \exp\left[-\frac{(x-x')^2}{2\lambda^2} - \Gamma^2 \sin^2\left(\frac{\pi|x-x'|}{P_{rot}}\right)\right]$$

Far terms correlated according to some characteristic period



-Second term states that far terms are correlated according to some characteristic period.

Where (can I learn more about) GPs

Learning about GPs:

- Rasmussen & Williams - http://www.gaussianprocess.org/gpml/chapters/
- Gaussian Processes for Timeseries Modelling - http://www.robots.ox.ac.uk/~sjrob/Pubs/philTransA_2012.pdf
- Daniel Foreman-Mackey's Python code George - http://dfm.io/george/current/
- Daniel Foreman-Mackey's live demo - http://dfm.io/gp.js/
- Blog post - http://katbailey.github.io/post/gaussian-processes-for-dummies/

GPs in Astro:

- https://arxiv.org/pdf/1610.09667.pdf
- https://arxiv.org/pdf/1501.00369.pdf
- https://arxiv.org/pdf/1609.07617.pdf
- https://arxiv.org/pdf/1506.07304.pdf