Gaussian Process Tutorial

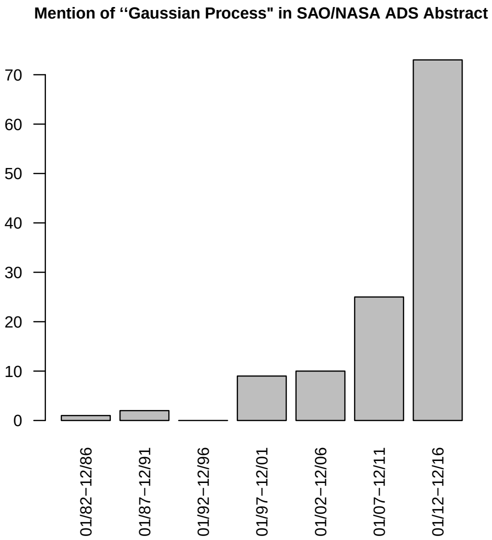
David Jones

Duke University and SAMSI

9th April 2018

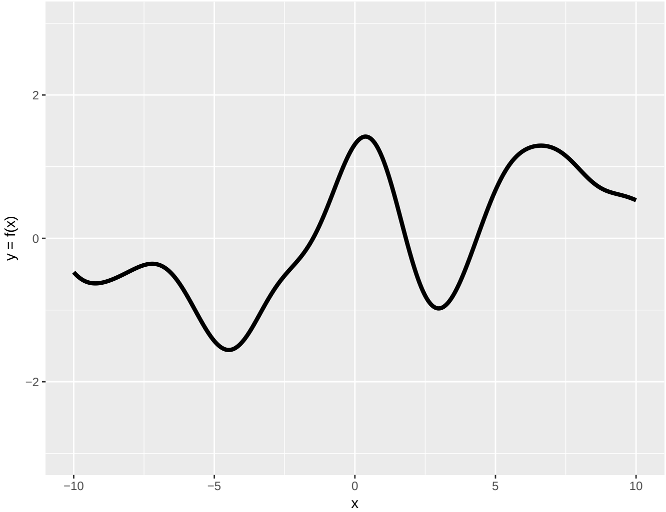
Acknowledgements: David Stenning (Imperial College London) contributed some of these slides.

Gaussian Processes in Astronomy

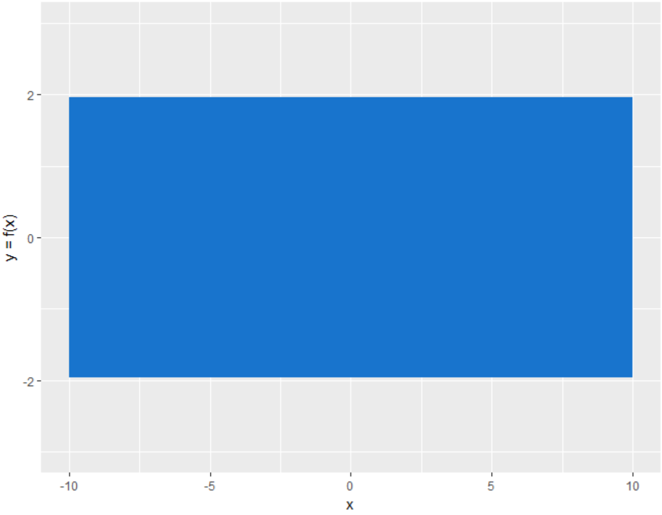


- Source: http://adsabs.harvard.edu/abstract_service.html

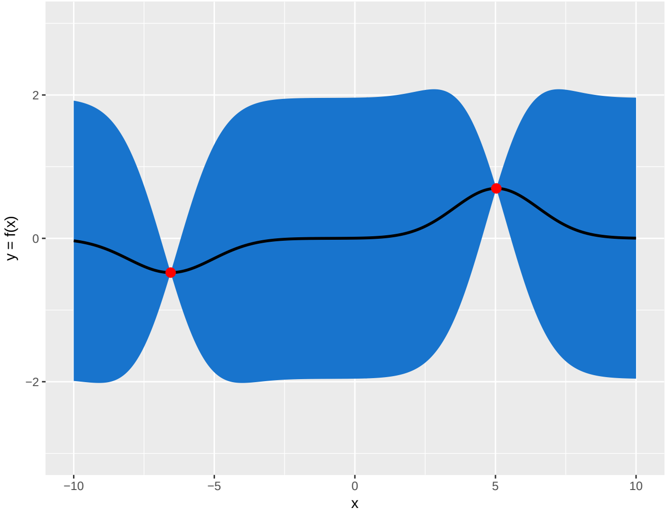
Example: function



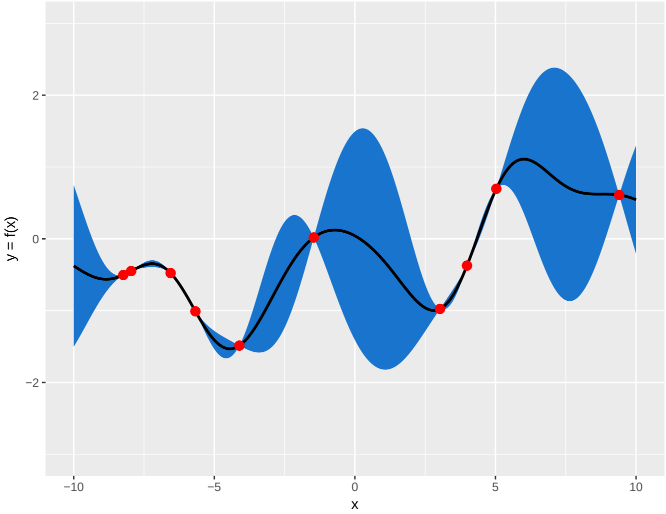
Example: no data



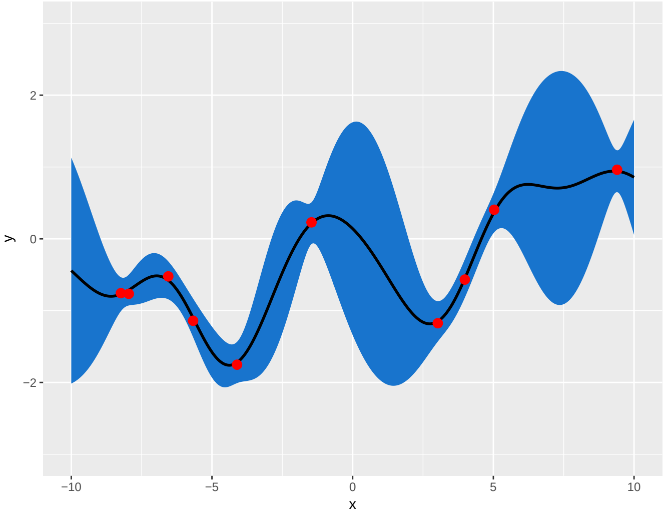
Example: function estimation



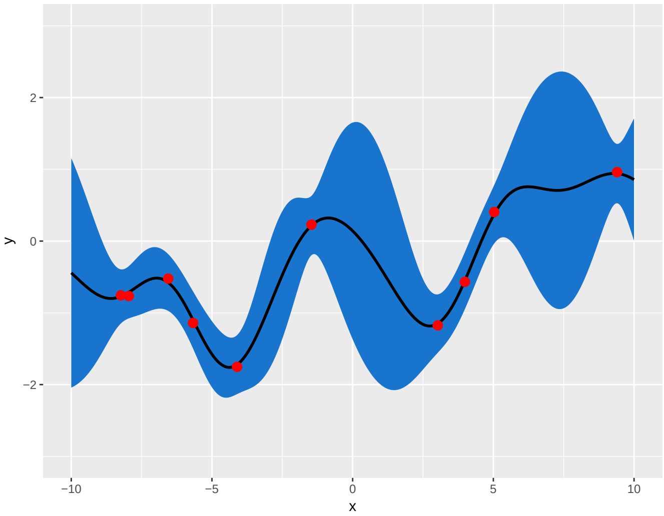
Example: function estimation



Example: noisy observations



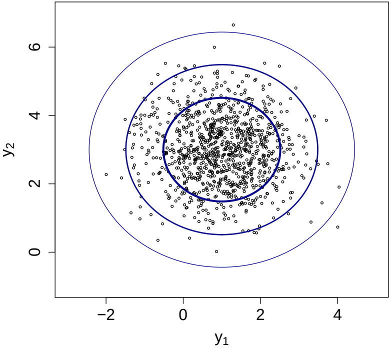
Example: prediction

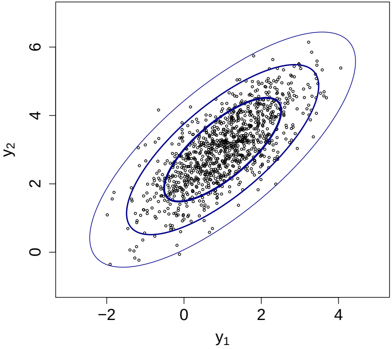


What is a Gaussian Process?

- A GP on the real line is a random real-valued function $f(t)$, which is completely determined by its mean function $m(t)$ and covariance function $C_{tt'} = \mathrm{Cov}(f(t), f(t'))$.
- Any finite sample $(f(t_1), \ldots, f(t_n))$ has a multivariate Gaussian distribution with mean $\vec{\mu} = (m(t_1), \ldots, m(t_n))$ and covariance matrix Σ, with $\Sigma_{ij} = C_{t_i t_j}$
- An excellent reference: Rasmussen and Williams (2006): http://www.gaussianprocess.org/gpml/chapters/

Bivariate Normal Distribution

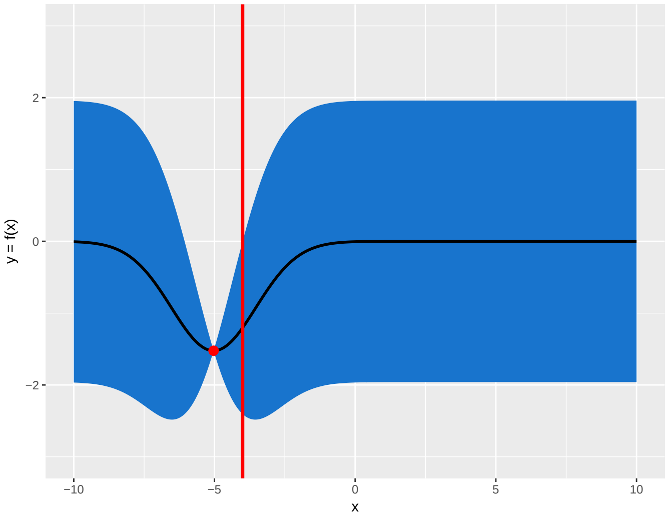




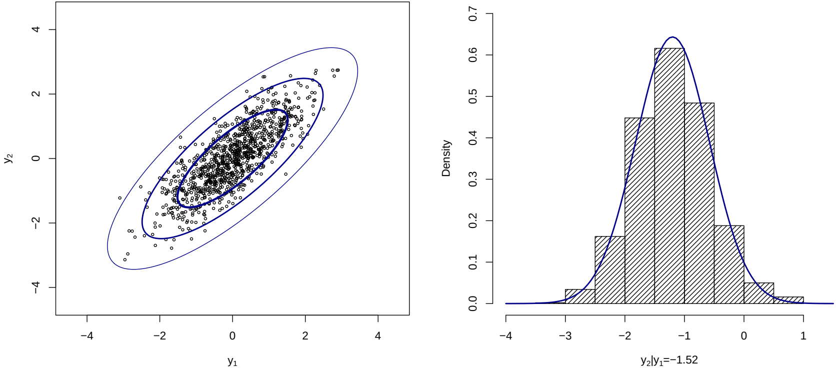
- **Left:** $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} f(t_1) \\ f(t_2) \end{pmatrix} \sim \mathrm{N}\left(\vec{\mu} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

- **Right:** $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} f(t_1) \\ f(t_2) \end{pmatrix} \sim \mathrm{N}\left(\vec{\mu} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}\right)$

Conditional distributions



Conditional distributions



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} f(t_1) \\ f(t_2) \end{pmatrix}$$

$$y_2 | y_1 = -1.52 \sim N(-1.2, 0.62^2)$$

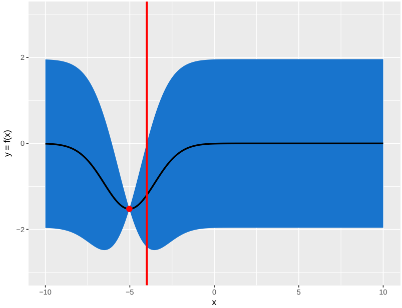
Multivariate Normal Distributions

Suppose $\vec{y}_1$ are values we observe, and $\vec{y}_2$ are values we want to predict, then:

$$\begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} =, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$$\vec{y}_2 \mid \vec{y}_1 \sim \mathrm{N}\left(\Sigma_{21} \Sigma_{11}^{-1} \vec{y}_1, \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right)$$

Illustration



- $$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} f(-5) \\ f(-4) \end{pmatrix} \sim \mathrm{N}\left(\vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.790 \\ 0.790 & 1 \end{pmatrix}\right)$$

- Applying the conditional Gaussian result

$$y_2|y_1 = -1.52 \sim N\left(0.790(1)^{-1}(-1.52), 1 - 0.790(1)^{-1}0.790\right)$$

$$y_2|y_1 = -1.52 \sim N(-1.2, 0.62^2)$$

Gaussian Process

- Assume $y = f(x)$ is a univariate *function* of d-dimensional x
- For a zero-mean Gaussian Process (GP), any (finite) collection $y_1, \ldots, y_m$ corresponding to $x_1, \ldots, x_m$ is distributed

$$\vec{y} \sim \mathrm{N}\left(\vec{0}, \Sigma\right)$$

  where $\Sigma_{ij} = R(x_i, x_j)$
- $R(x, x')$ is a covariance function (i.e. kernel) that we specify.
  - A common choice is the *squared exponential kernel*:

$$R_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

  - σ^2 is a scale factor (all kernels have this term)
  - The length-scale, l, controls the "wiggliness" of the function

Gaussian Process

- Assume $y = f(x)$ is a univariate *function* of d-dimensional x
- For a zero-mean Gaussian Process (GP), any (finite) collection $y_1, \ldots, y_m$ corresponding to $x_1, \ldots, x_m$ is distributed

  $$\vec{y} \sim \mathrm{N}\left(\vec{0}, \Sigma\right)$$

  where $\Sigma_{ij} = R(x_i, x_j)$
- $R(x, x')$ is a covariance function (i.e. kernel) that we specify.
  - A common choice is the *squared exponential kernel*:

    $$R_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

    - σ^2 is a scale factor (all kernels have this term)
    - The length-scale, l, controls the "wiggliness" of the function

Periodic and Locally Periodic Kernels

- A *periodic kernel* models functions that repeat (periodically):

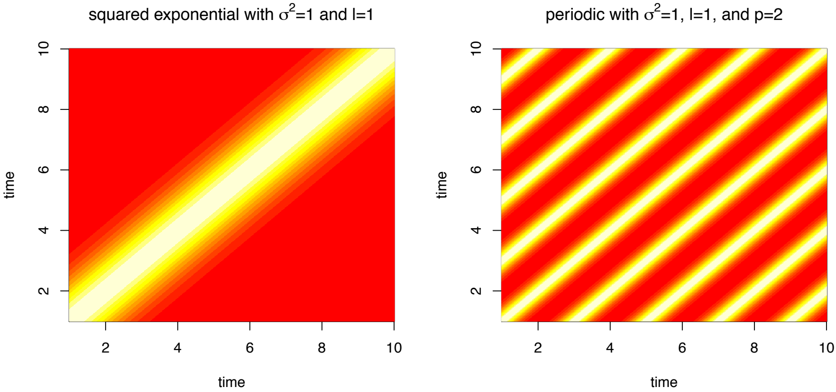
$$R_{\mathrm{Per}}(x,x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi |x-x'|/p)}{l_p^2}\right)$$

- A *locally periodic kernel* yields functions with a periodic component that may evolve over time:

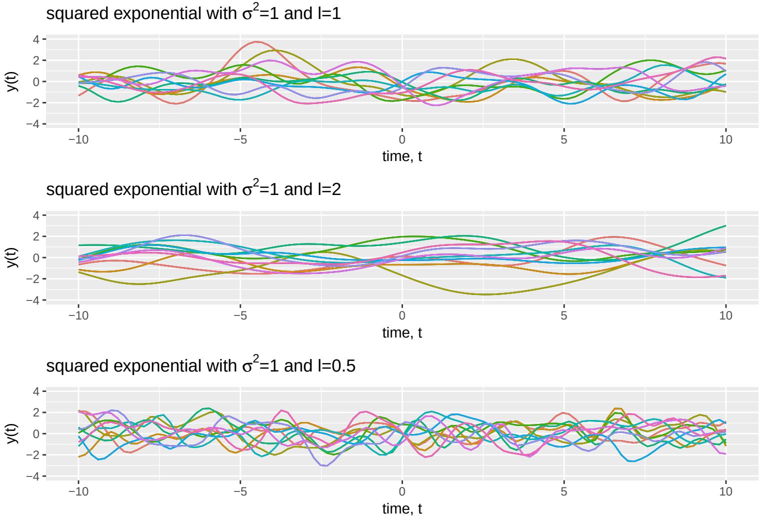
$$R_{\mathrm{LocPer}}(x,x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi |x-x'|/p)}{l_p^2}\right) \exp\left(-\frac{(x-x')^2}{2l_e^2}\right)$$

- A good resource: The Kernel Cookbook (by David Duvenaud)

Another View of Kernels



GP Draws: Squared Exponential Kernel



Inference with Gaussian Processes

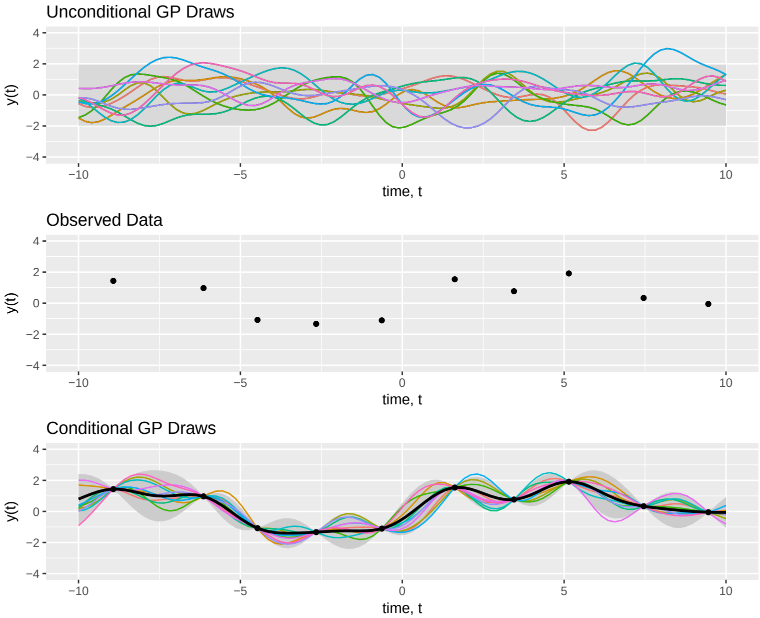
- Let $\vec{y}_1$ be some values we observe and $\vec{y}_2$ are values we want to predict. Then:

$$\begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix}, \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \right)$$

$$\vec{y}_1 \mid \vec{y}_2 \sim \mathrm{N}\left(R_{11} R_{22}^{-1} \vec{y}_2,\ R_{11} - R_{12} R_{22}^{-1} R_{21} \right)$$

  - Mean for the new points is a weighted average of the observed points
  - Mean of a new point approaches value of an observed point as the new point approaches the observed point
  - Variance of a new point goes to zero as the new point approaches an observed point

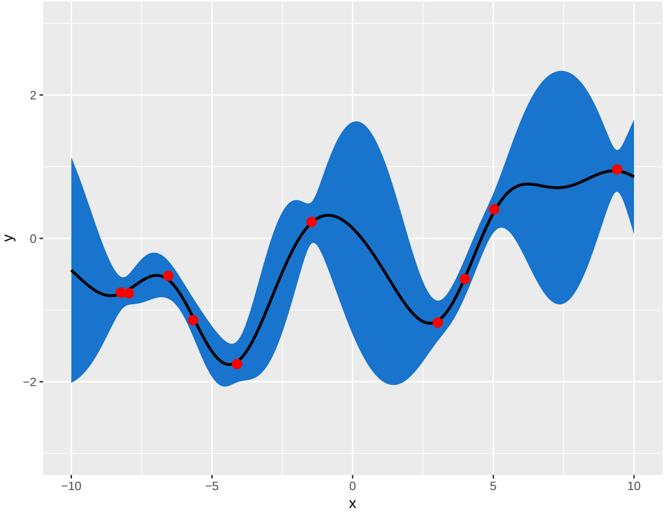
Inference with Gaussian Processes



What are we actually doing?

- When using GPs, we are specifying a prior on the *relationship* between t and $f(t)$, instead of some parameters that describe this relationship
  - i.e. "nonparametric"
- GPs especially useful for prediction; (maybe) not as useful for making inference about the relationship
  - e.g., useful for predicting sunspot cycle; less useful for learning about the cycle

Noisy observations



Gaussian Processes in the case of noisy observations

- Now $y_1, \ldots, y_m$ corresponding to $x_1, \ldots, x_m$ is distributed

$$\vec{y} \sim \mathrm{N}\left(\vec{0}, \Sigma + \tau^2 I_m\right)$$

  where $\Sigma_{ij} = R(x_i, x_j)$.

- More specifically the model is

$$\vec{y} \sim \mathrm{N}\left(\vec{f}, \tau^2 I_m\right)$$

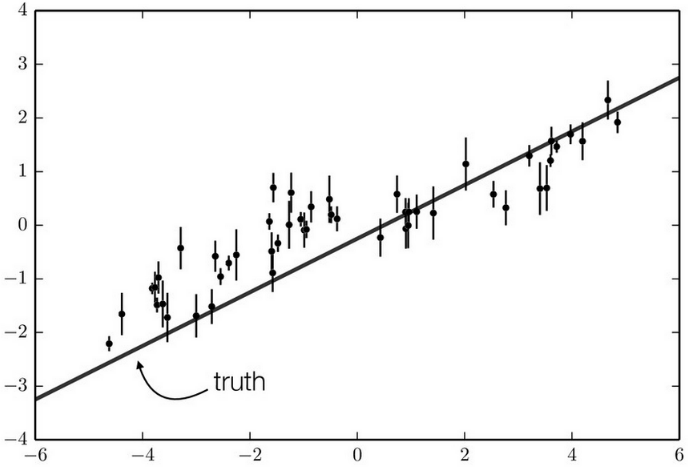
$$\vec{f} \sim \mathrm{N}\left(\vec{0}, \Sigma\right)$$

  where $\vec{f} = (f(x_1), \ldots, f(x_m))^T$

Hyper-parameters

- For real research problems, we often (always?) lack the information needed to fix the parameters of the covariance function
- Typical solutions:
  - Maximum likelihood estimation
  - Cross validation
  - Specify some prior distributions and do MCMC
- Caveat: C^{-1} is $\mathcal{O}\left(N^3\right)$; exploit sparsity if possible

Underlying Model + Correlated Noise

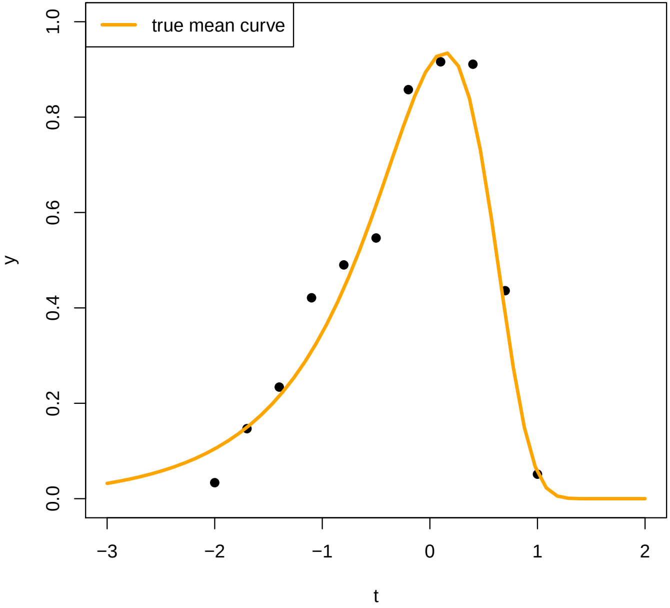


- Image: https://astrobites.org/2014/07/01/beyond-chi-squared-an-introduction-to-correlated-noise/

Toy Example: setup

- Consider the following setup:
  - Physical model: $g_\phi(t) = a_1\sqrt{10^t} + a_2\sqrt{10^t}\exp\left(\frac{-10^t}{a_3}\right)$
  - Physical parameters: $\phi = (a_1, a_2, a_3)$
  - Reality: $a_1 = 1$, $a_2 = 0.5$, $a_3 = 2$
- Have 11 observations with correlated noise
- We want to infer a_1, a_2, and a_3

Toy Example: observations



Toy Example: model formulation

Covariance Function (Kernel):

- $R(t,t') = \sigma^2 \exp\left(-\beta\,(t-t')^2\right) + \delta_{tt'}\tau^2$
  - $\delta_{tt'} = 1$ if $t = t'$ and 0 otherwise

Sampling Model:

- $\vec{y} \sim N\left(g_\phi\left(\vec{t}\right), \Sigma\right)$
  - $\Sigma_{ij} = R(t_i, t_j)$
  - $\phi = (a_1, a_2, a_3)$

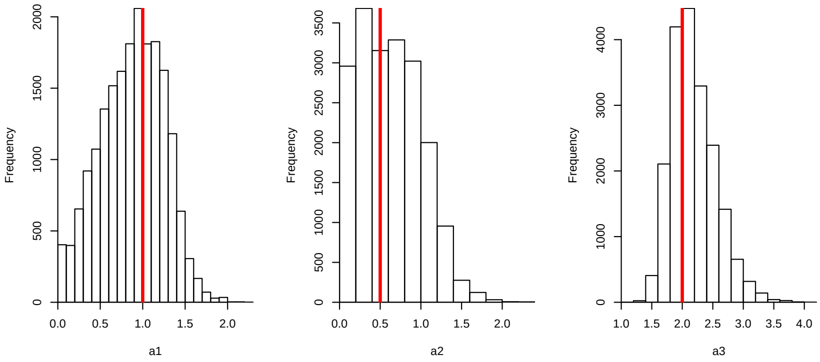
Priors:

- $\beta \sim \mathrm{Exponential}(1)$
- $\sigma^2 \sim \mathrm{Inv\text{-}Gamma}(5, 0.1)$
- $\tau^2 \sim \mathrm{Inv\text{-}Gamma}(5, 0.01)$
- Flat priors on a_1, a_2, and a_3

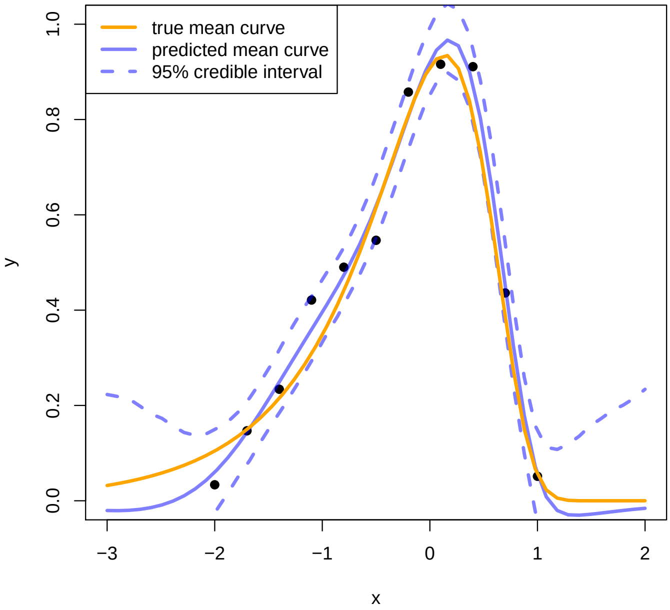
Model Fitting:

- Parameters estimated with one-at-a-time Metropolis MCMC

Toy Example: results



Toy Example: results



Real Example I: Czekala et al. 2015

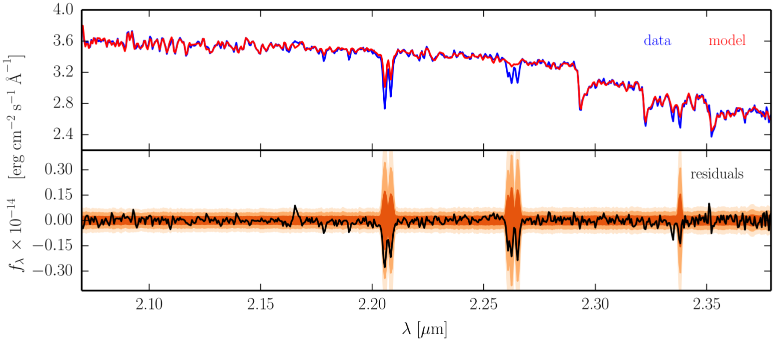
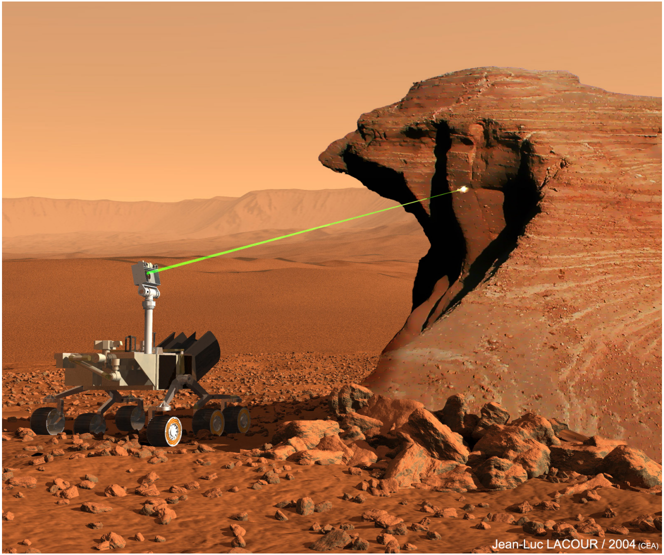


Figure 11. The K-band SPEX spectrum of Gl 51 (*blue*) compared with a PHOENIX model (*red*) generated by drawing parameters from the inferred posterior distribution. (*bottom*) The residual spectrum along with contours representing the distributions of a large number of random draws from the covariance matrix (the shading is representative of the 1, 2, and 3 σ spreads of that distribution of draws), as in Fig. 9. Note how the 'outlier' features (Na I at 2.21 μm and Ca I at 2.26 μm) are identified and treated by the local covariance kernels.

- Likelihood framework for spectroscopic inference based on synthetic model spectra and GPs
- Addresses mismatches in model spectral line strengths w.r.t. data due to intrinsic model imperfections
- https://arxiv.org/abs/1412.5177

Example II: Mars Rover ChemCam



Artistic rendering of ChemCam LIBS analyses using NASA's Mars Curiosity Rover

Example II: Mars Rover ChemCam

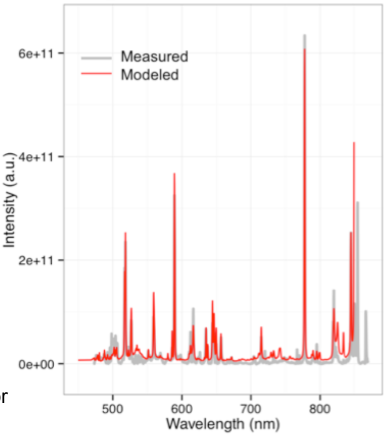
General concept: Estimate the settings of a theoretical model's input parameters $\boldsymbol{\theta}$ that are consistent with physical measurements y.

measured spectrum　　modeled spectrum

$$y = \eta(\theta) + \delta + \varepsilon$$

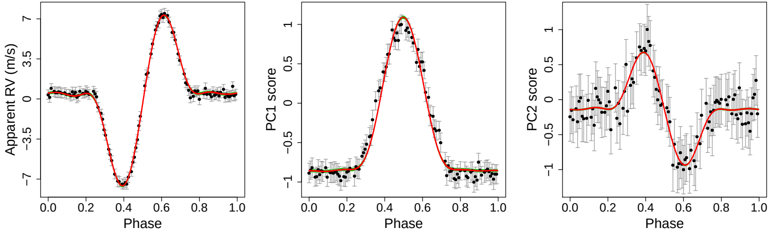
discrepancy term　　measurement error

Measured and modeled LIBS spectra of basalt.



Slide courtesy Kary Myers (LANL)

Real Example III: D. Jones, D. Stenning, et al. (under revision)



- Model the relationships between the apparent RV of a star due to a spot and proxies for stellar variability
- Use locally periodic kernel

$$R_{\mathrm{LocPer}}(t, t') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi |t - t'| / p)}{l_p^2}\right) \exp\left(-\frac{(t - t')^2}{2 l_e^2}\right)$$

GPs in Python

Packages include:

- scikit-learn
- GPflow
- PyMC3
- George
- . . .

Many good tutorials online e.g.

- https://blog.dominodatalab.com/fitting-gaussian-process-models-python/

For more information...

- Rasmussen and Williams (2006): http://www.gaussianprocess.org/gpml/chapters/
- Contact me:
  - dav.jones2000@gmail.com
  - dej17@duke.edu