Lecture 16

GAUSSIANS AND GAUSSIAN PROCESSES

Last Time

- linear regression from normal model
- identifiability
- bayesian updating for linear regression
- regularization

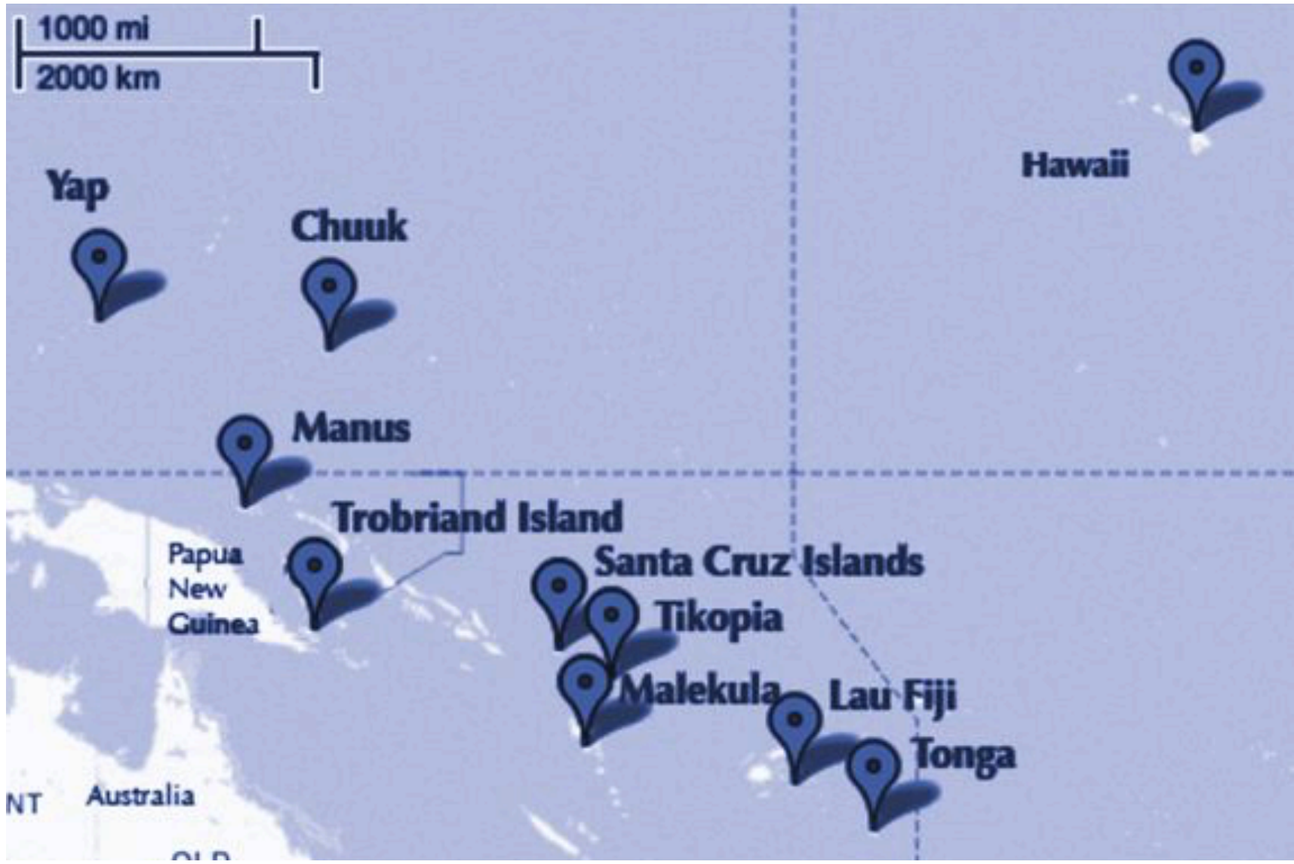
LAB ROOM CHANGE (this fri only)

NW B-101

same time, 11 am

From McIreath:

The island societies of Oceania provide a natural experiment in technological evolution. Different historical island populations possessed tool kits of different size. These kits include fish hooks, axes, boats, hand plows, and many other types of tools. A number of theories predict that larger populations will both develop and sustain more complex tool kits. So the natural variation in population size induced by natural variation in island size in Oceania provides a natural experiment to test these ideas. It's also suggested that contact rates among populations effectively increase



MVN Primer

$$p(x \mid \mu, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

JOINT: $p(x, y) = \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y\end{bmatrix}\right)$

MARGINAL: $p(x) = \int p(x, y)dy = \mathcal{N}(\mu_x, \Sigma_x)$

CONDITIONAL: $p(x \mid y) = \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T)$

Modeling correlation

General expectation of continuity as you move from one adjacent point to another. In the absence of significant noise, two adjacent points ought to have fairly similar f values.

$$k(x_i, x_j) = \sigma_f^2 exp(\frac{-(x_i - x_j)^2}{2l^2})$$

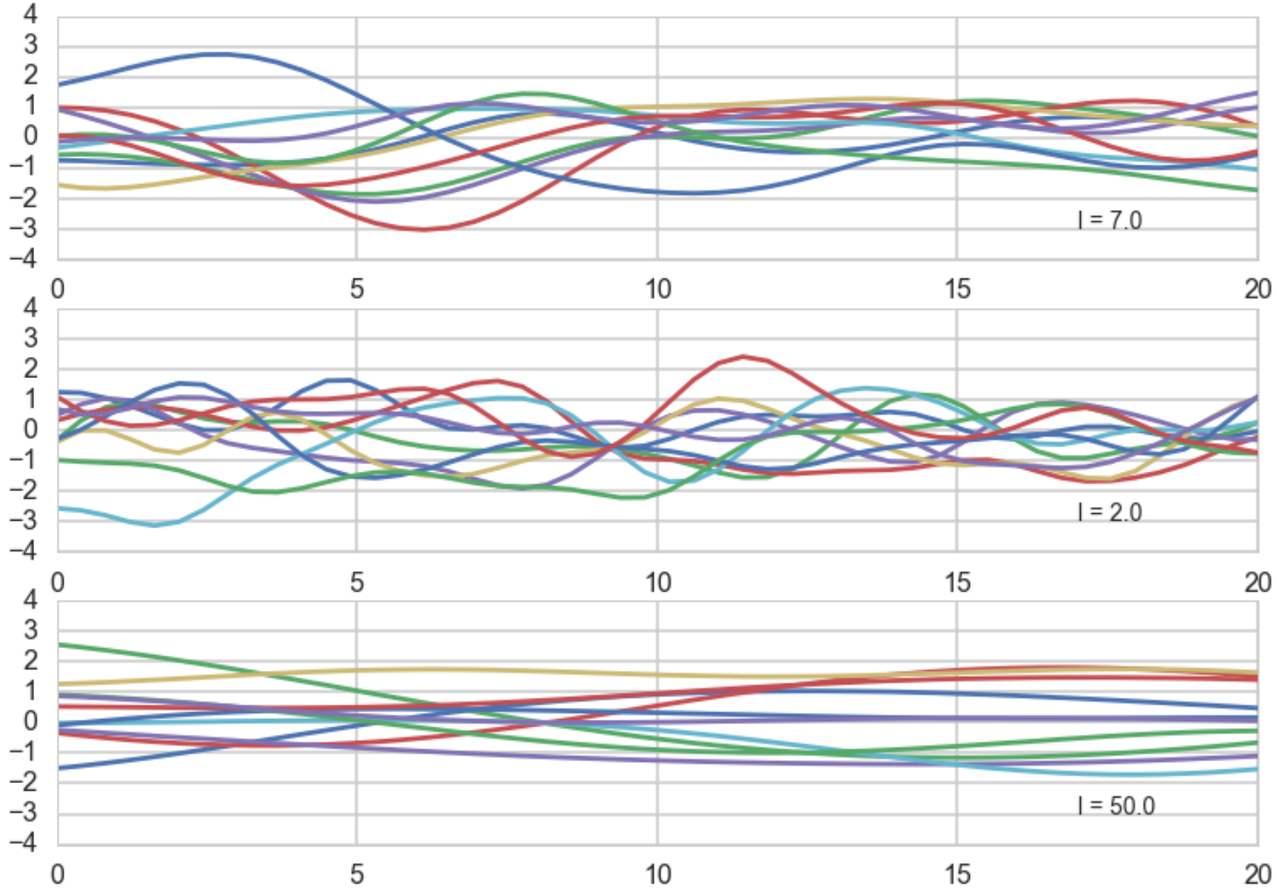
l is correlation length, σ_f^2 amplitude.

```
#Correlation Kernel
def exp_kernel(x1,x2, params):
    amplitude=params[0]
    scale=params[1]
    return amplitude * amplitude*np.exp(-((x1-x2)**2) / (2.0*scale))

#Covariance Matrix
covariance = lambda kernel, x1, x2, params: \
    np.array([[kernel(xi1, xi2, params) for xi1 in x1] for xi2 in x2])
```

Each curve in plots is generated as:

```
a = 1.0
nsamps = 50
xx = np.linspace(0,20,nsamps)

#Create Covariance Matrix
sigma = covariance(exp_kernel,xx,xx, [a,ell]) + np.eye(nsamps)*1e-06

#Draw samples from a 0-mean gaussian with cov=sigma
samples = np.random.multivariate_normal(np.zeros(nsamps), sigma)
```

The greater the correlation length, the smoother the curve.



What did we just do?

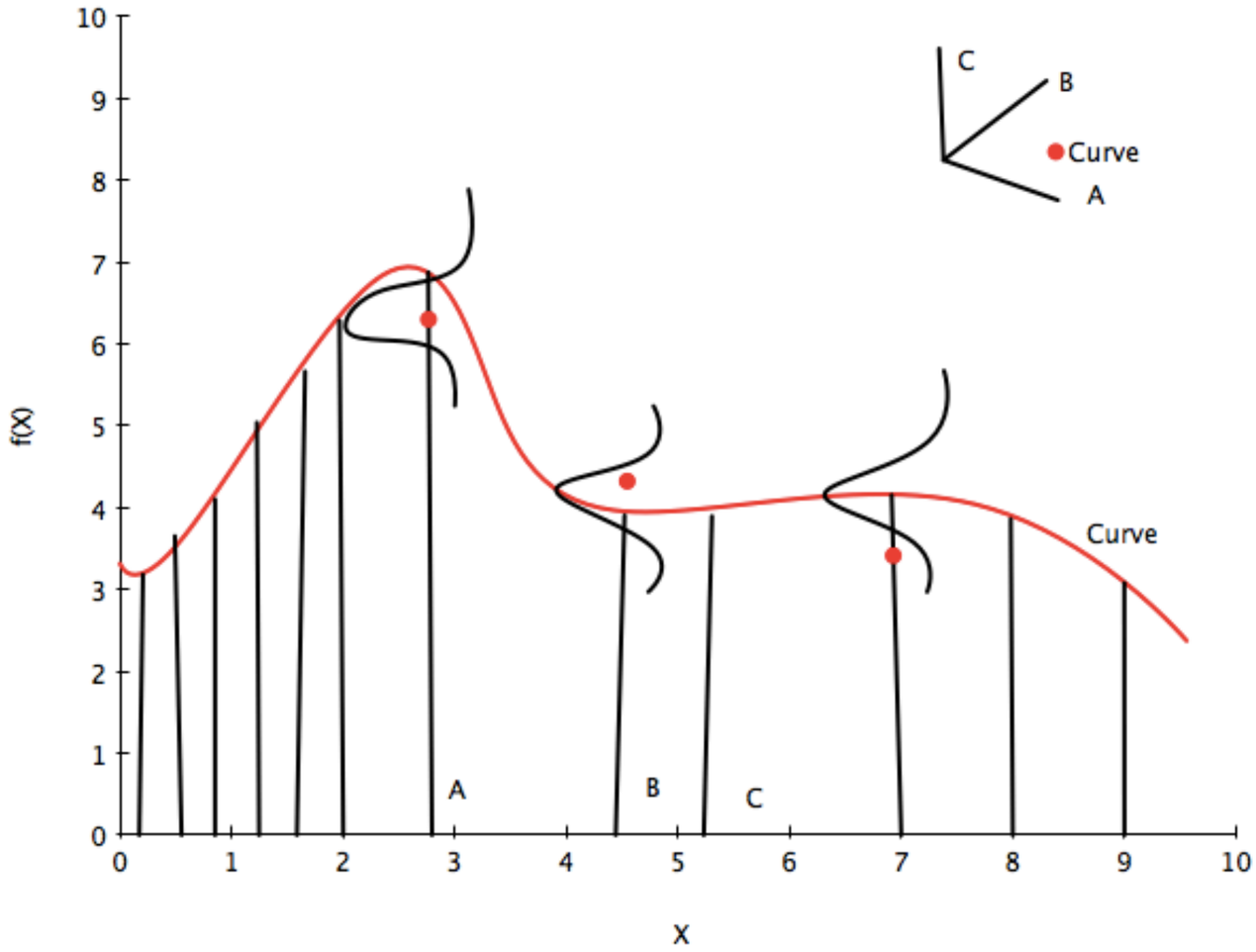
- we have not seen any data yet
- but we expect the function representing our data to have some level of continuity
- thus we considered different **PRIOR** functions that might represent our data
- as having come from MVNs with a covariance matrix based on the correlation length we think we have

The red curve represents one of these functions from the calculation above.

We have 3 red data points, so it would be seem to be one of the curves consistent with the data.

We can consider the curve as a point in a multi-dimensional space, a draw from a multivariate gaussian with as many points as points on the curve.

Consider the 3 red data points to have been generated IID from some regression function $f(x)$ (like $w \cdot x$) with some univariate gaussian noise σ^2 at each point.



JOINT:

$$p(y, f^*) = \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_{f^*} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{yf^*} \\ \Sigma_{yf^*}^T & \Sigma_{f^*f^*} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K + \sigma^2 I & K_* \\ K_*^T & K_{**} \end{bmatrix}\right)$$

MARGINAL: $p(f^*) = \int p(f^*, y) dy = \mathcal{N}(\mu_*, K_{**})$

CONDITIONAL:

$$p(f^* \mid y) = \mathcal{N}\left(\mu_* + K_*(K + \sigma^2 I)^{-1}(y - \mu),\ K_{**} - K_*(K + \sigma^2 I)^{-1} K_*^T\right)$$

where: $K = K(x, x); K_* = K(x, x^*); K_{**} = K(x^*, x^*)$

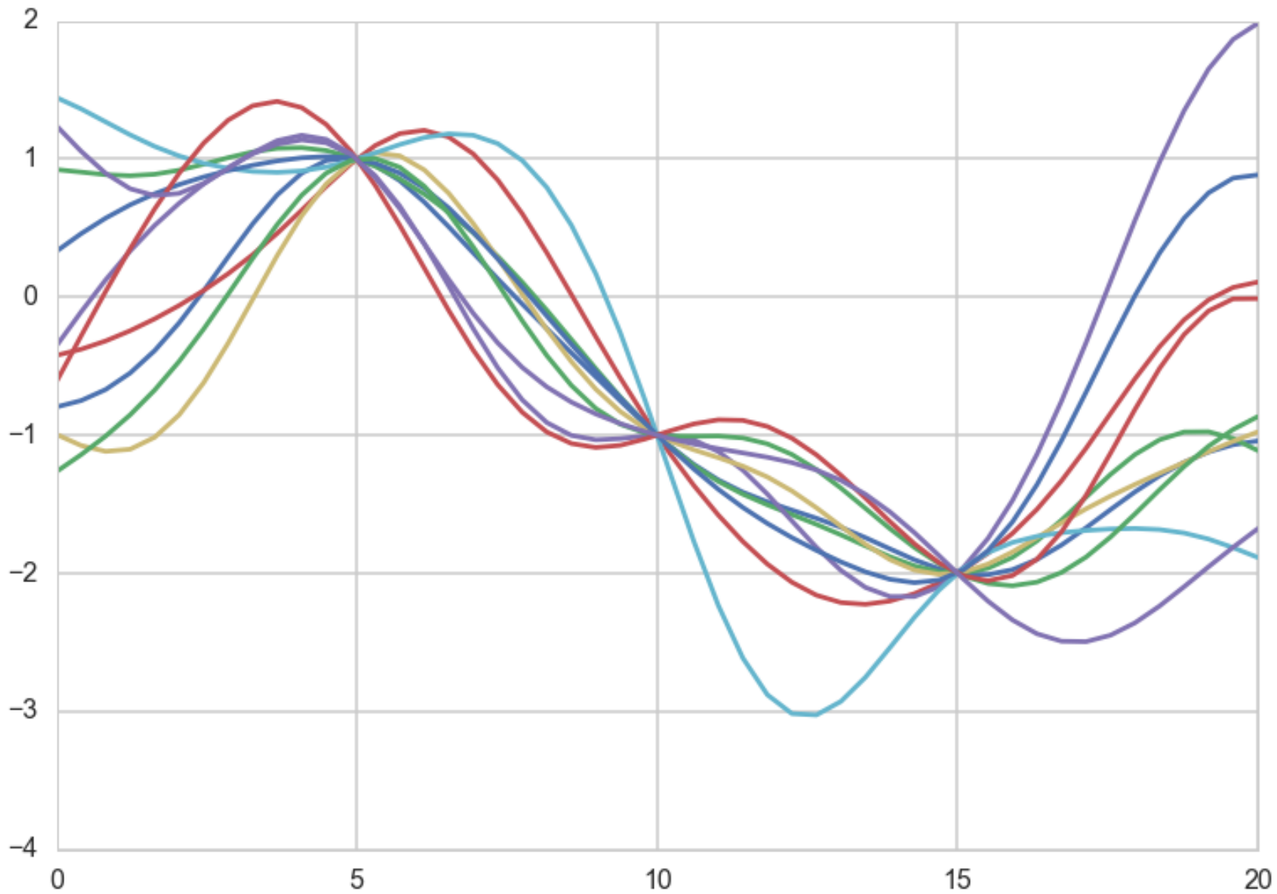
Conditional

$$p(f^* \mid y) = \mathcal{N}\left(\mu_* + K_*(K + \sigma^2 I)^{-1}(y - \mu),\ K_{**} - K_*(K + \sigma^2 I)^{-1} K_*^T\right)$$

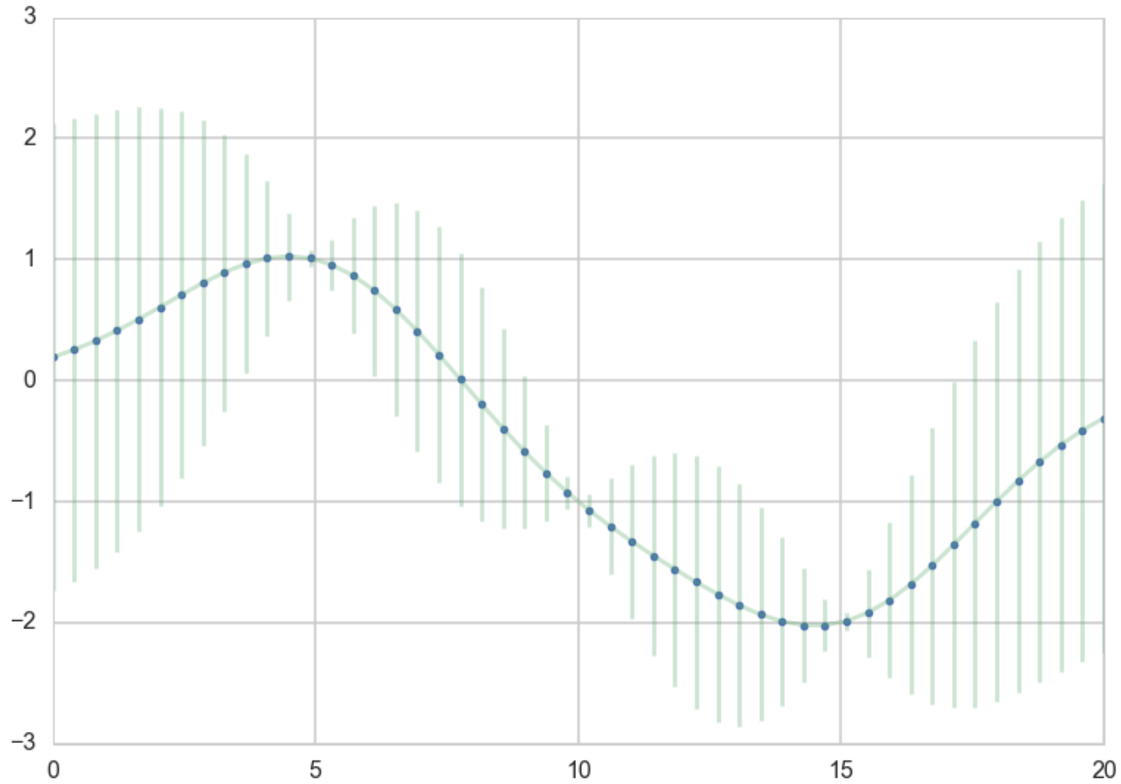
EQUALS Predictive

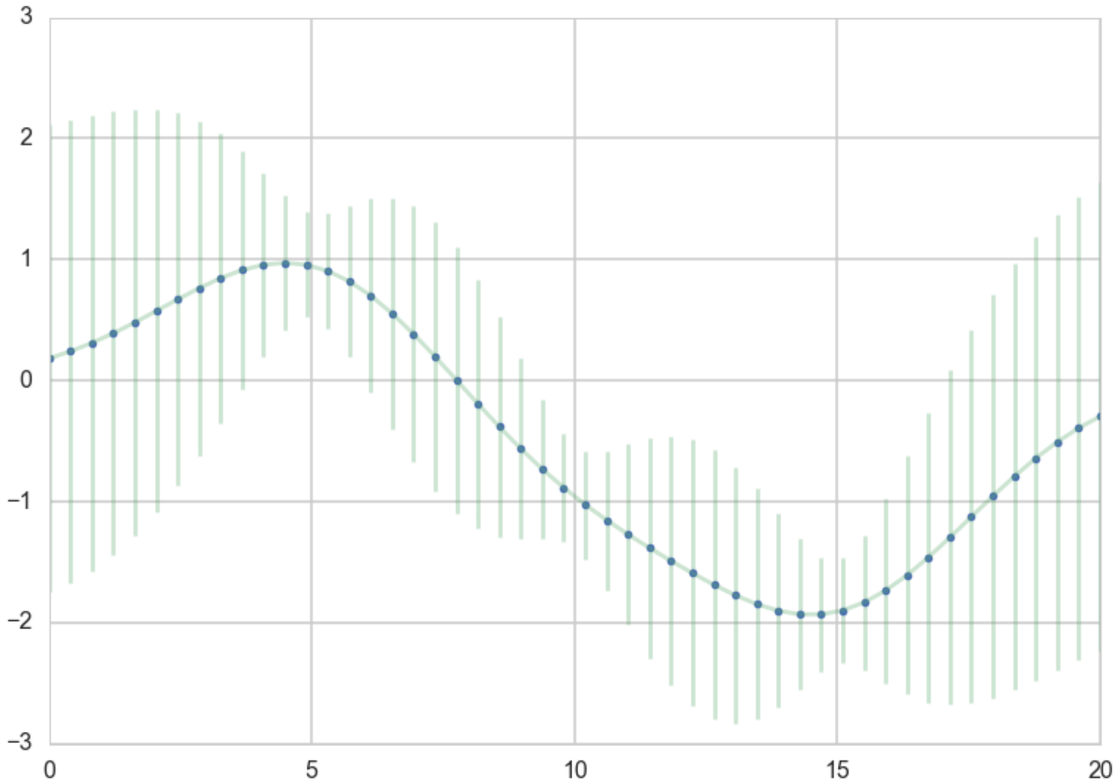
$l = 7$, added a small noise term.

```
#"test" data
x_star = np.linspace(0,20,nsamps)

# defining the training data
x = np.array([5.0, 10.0, 15.0]) # shape 3
f = np.array([1.0, -1.0, -2.0]).reshape(-1,1)


K = covariance(exp_kernel, x,x,[a,ell])
#shape 3,3

K_star = covariance(exp_kernel,x,x_star,[a,ell])
#shape 50, 3

K_star_star = covariance(exp_kernel, x_star, x_star, [a,ell])
#shape 50,50

K_inv = np.linalg.inv(K)
#shape 3,3

mu_star = np.dot(np.dot(K_star, K_inv),f)
#shape 50

sigma_star = K_star_star  - np.dot(np.dot(K_star, K_inv),K_star.T)
#shape 50, 50
```



Posterior and predictive





So far

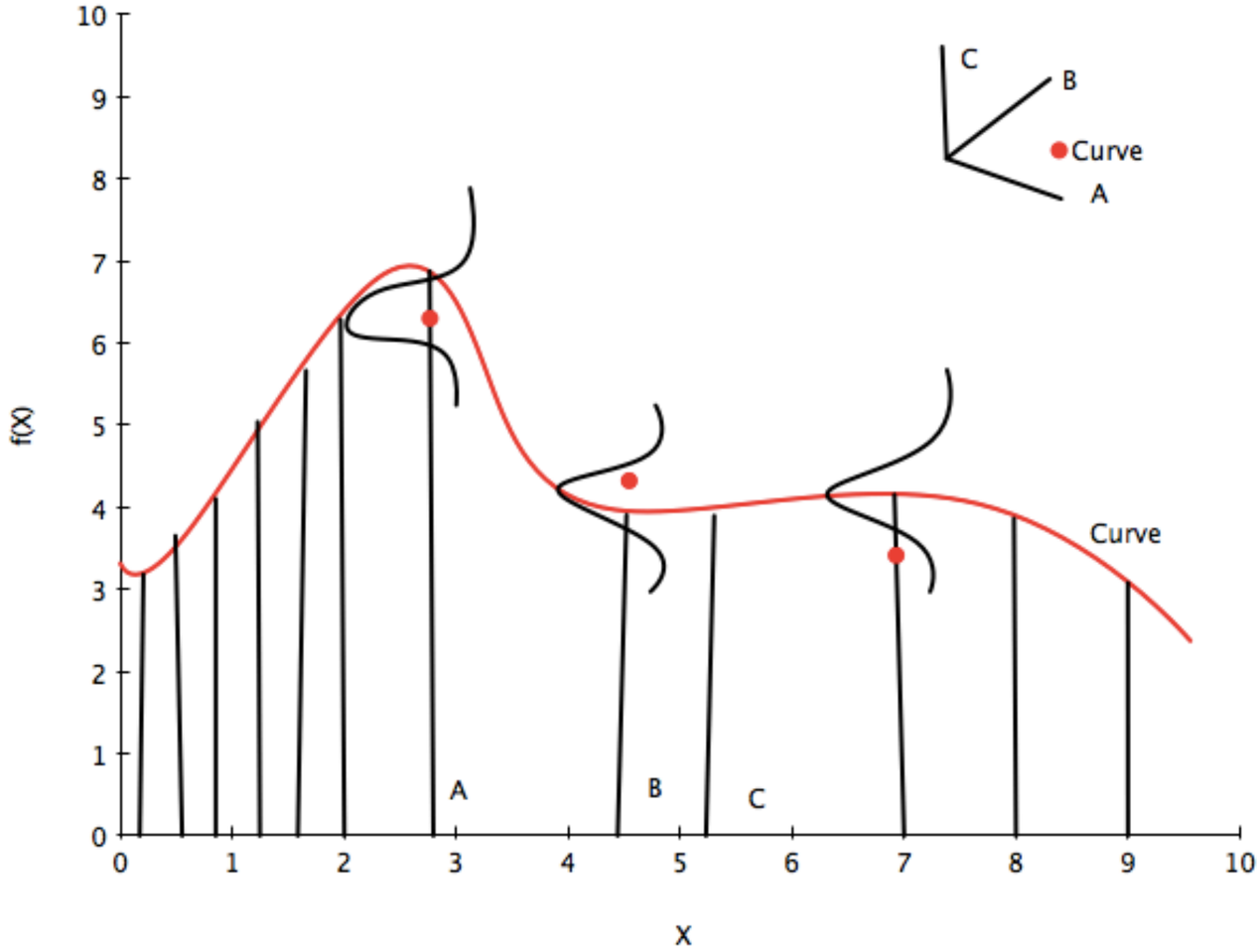
1. We built a covariance matrix from a kernel function
2. Use the covariance matrix to generate a "curve" as a point in a multi-dimensional space from a MVN
3. multiple such curves serve as prior fits for our data
4. now we bring in the data and condition on it (with noise added if needed) using normal distribution formulae
5. the conditional has the form of a predictive and we are done
6. Also notice that the marginal only has quantities from the predictive block. This means that we dont care about the size of the original block in calculating the marginal.

These observations are the building blocks of the GP.

Use infinite gaussians!

- think of the function as an infinite vector.
- Draw $\overline{f}$ from some 'infinite' gaussian distribution with some mean and some kernel.

  This then is the Gaussian Process, which we use to set a prior on the space of functions.



Back to our formulae...

JOINT:

$$p(f, f^\infty) = \mathcal{N}\left(\begin{bmatrix} \mu_f \\ \mu_{f^\infty} \end{bmatrix}, \begin{bmatrix} \Sigma_{ff} & \Sigma_{ff^\infty} \\ \Sigma_{ff^\infty}^T & \Sigma_{f^\infty f^\infty} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_\infty \end{bmatrix}, \begin{bmatrix} K & K_\infty \\ K_\infty^T & K_{\infty\infty} \end{bmatrix}\right)$$

MARGINAL:

$$p(f) = \int p(f, f^\infty) df^\infty = \mathcal{N}(\mu_f, K)$$

where: $K = K(x, x); K_\infty = K(x, x^\infty); K_{\infty\infty} = K(x^\infty, x^\infty)$

KEY INSIGHT:

MARGINAL IS DECOUPLED

...for the marginal of a gaussian, only the covariance of the block of the matrix involving the unmarginalized dimensions matters! Thus "if you ask only for the properties of the function (you are fitting to the data) at a finite number of points, then inference in the Gaussian process will give you the same answer if you ignore the infinitely many other points, as if you would have taken them all into account!"

-Rasmunnsen

Definition of Gaussian Process

Assume we have this function vector $f = (f(x_1), \dots f(x_n))$. If, for ANY choice of input points, $(x_1, \dots, x_n)$, the marginal distribution over f:

$$P(F) = \int_{f \notin F} P(f) df$$

is multi-variate Gaussian, then the distribution $P(f)$ over the function f is said to be a Gaussian Process.

a Gaussian Process defines a prior distribution over functions!

Once we have seen some data, this prior can be converted to a posterior over functions, thus restricting the set of functions that we can use based on the data.

Since the size of the "other" block of the matrix does not matter, we can do inference from a finite set of points.

Any m observations in an arbitrary data set, $y = y_1, \ldots, y_n = m$ can always be represented as a single point sampled from some m-variate Gaussian distribution. Thus, we can work backwards to 'partner' a GP with a data set, by marginalizing over the infinitely-many variables that we are not interested in, or have not observed.

GP regression

Using a Gaussian process as a prior for our model, and a Gaussian as our data likelihood, then we can construct a Gaussian process posterior.

Likelihood: $y|f(x), x \sim \mathcal{N}(f(x), \sigma^2 I)$

where the infinite $f(x)$ takes the place of the parameters.

Prior: $f(x) \sim \mathcal{GP}(m(x) = 0, k(x, x\prime))$

Infinite normal posterior process: $f(x)|y \sim \mathcal{GP}(m_{post}, \kappa_{post}(x, x\prime))$.

The posterior distribution for f is:

$$m_{post} = k(x\prime, x)[k(x, x) + \sigma^2 I]^{-1} y$$
$$k_{post}(x, x\prime) = k(x\prime, x\prime) - k(x\prime, x)[k(x, x) + \sigma^2 I]^{-1} k(x, x\prime)$$

Posterior predictive distribution for $f(x_*)$ for a test vector input x_*, given a training set X with values y for the GP is:

$$m_* = k(x_*, X)[k(X^T, X) + \sigma^2 I]^{-1} y$$
$$k_* = k(x_*, x_*) - k(x_*, X^T)[k(X^T, X) + \sigma^2 I]^{-1} k(X^T, x_*)$$

The predictive distribution of test targets y_* : add $\sigma^2 I$ to k_*.

What did we do

- usually in a parametric model we had some m (small) number of parameters
- but here our covariance functions are NxN !
- no free lunch: calculation involves inverting a NxN matrix as in the kernel space representation of regression.
- cannot thus handle large data if no approximations are used

Parametric models

- In general, parametrization restricts the class of functions we use. If our data is not well modeled by our choices, we might underfit.
- Increasing flexibility might lead to overfitting.

INSTEAD: consider every possible function and associate a prior probability with this function. e.g. assign smoother functions higher prior probability. But how are we possibly going to calculate over an uncountably infinite set of functions in finite time?

Linear Regression

$$y = f(X) + \epsilon, \epsilon \sim N(0, \sigma^2), f(X) = X^T w, w \sim N(0, \Sigma)$$

Posterior: $p(w|X, y) \sim N(\frac{1}{\sigma^2} A^{-1} Xy, A^{-1})$ where $A = \frac{1}{\sigma^2} XX^T + \Sigma^{-1}$

Posterior predictive distribution:

$p(f(x_*)|x_*, X, y) = N(\frac{1}{\sigma^2} x_*^T A^{-1} Xy, x_*^T A^{-1} x_*).$

For posterior predictive of y_*, just add σ^2 to the variance above.

We can show that we can rewrite the above posterior predictive as:

$$p(f(x_*)|x_*, X, y) = N(x_*^T \Sigma X^T (K + \sigma^2 I)^{-1} y, x_*^T \Sigma x_* - x_*^T \Sigma X^T (K + \sigma^2 I)^{-1} X \Sigma x_*)$$

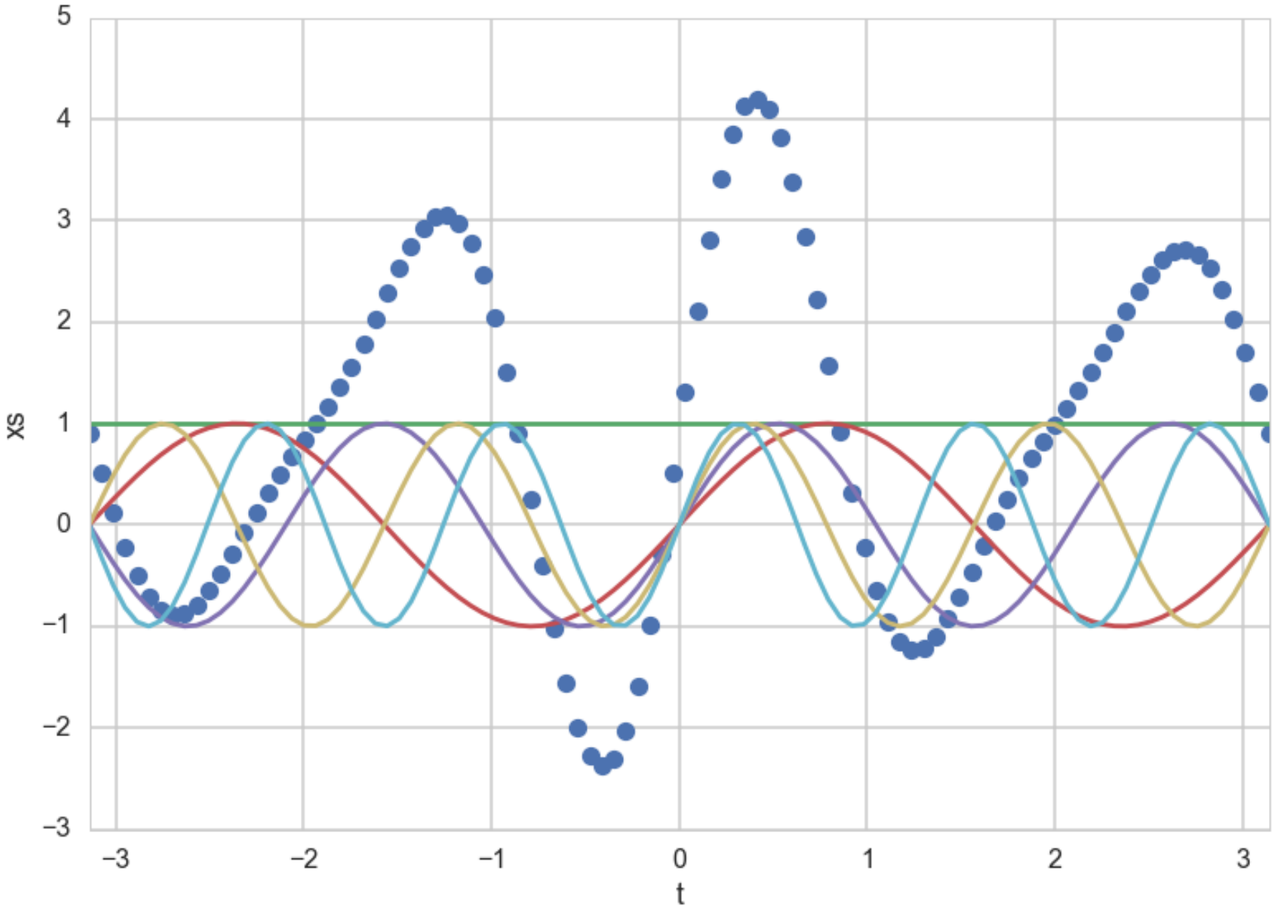
where $K = X \Sigma X^T$ which is of size N x N.

Notice now that the features only appear in the combination $\kappa(x, x') = x^T \Sigma x'$, thus, the dual representation:

$$p(f(x_*)|x_*, X, y) =$$

$$N\left(\kappa(x_*, X)\left(\kappa(X^T, X) + \sigma^2 I\right)^{-1} y,\ \ \kappa(x_*, x_*) - \kappa(x_*, X^T)\left(\kappa(X^T, X) + \sigma^2 I\right)^{-1} \kappa(X^T, x_*)\right)$$

Basis functions



A scalar x could be projected into a polynomial space: $\phi(x) = (1, x, x^2, x^3, \dots)$. So let us now have

$$f(x) = \phi(x)^T w$$

Let $\Phi(X)$ be the aggregation of columns $\phi(x)$ for all training set cases $x \in X$.

Then the posterior predictive is $p(f(x_*)|x_*, X, y) = N(\frac{1}{\sigma^2}\phi_*^T A^{-1} \Phi y, \phi_*^T A^{-1} \phi_*)$.

where $\phi_* = \phi(x_*)$ and $A = \frac{1}{\sigma^2}\Phi\Phi^T + \Sigma^{-1}$

which can as before be written as

$$N(\kappa(x_*, X)\left(\kappa(X^T, X) + \sigma^2 I\right)^{-1} y, \;\; \kappa(x_*, x_*) - \kappa(x_*, X^T)\left(\kappa(X^T, X) + \sigma^2 I\right)^{-1} \kappa(X^T, x_*)$$

where the kernel is now $\kappa(x, x') = \phi(x)^T \Sigma \phi(x')$

Then defining $\psi(x) = \Sigma^{(1/2)}\phi(x)$, we have $\kappa(x, x') = \psi(x)^T \psi(x')$

Kernel Trick

If an algorithm is defined just in terms of inner products in input space then we can make the algorithm work in higher-dimensional feature space by replacing occurrences of those inner products by $\kappa(x, x')$.

So we learn that covariance can be kernelized, and dimensions can be lifted. This might remind you of SVM.

Infinite basis sets

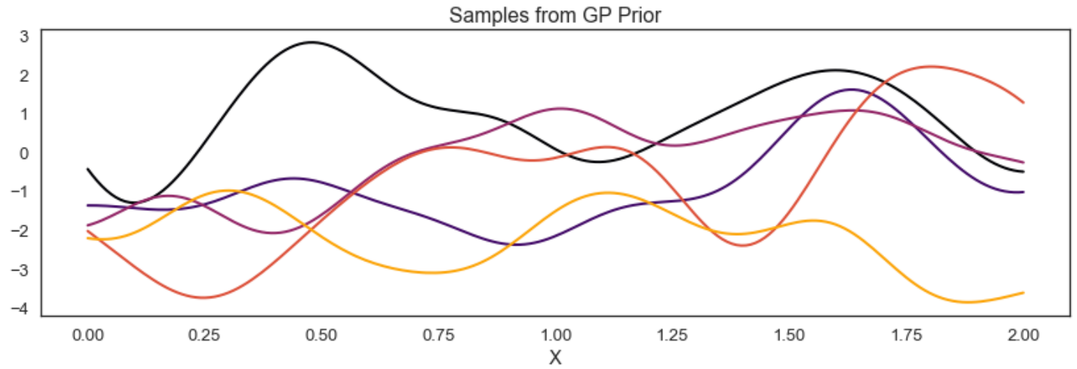
Now consider an infinite set of $\phi(x)$. Like a fourier series or a Bessel series.

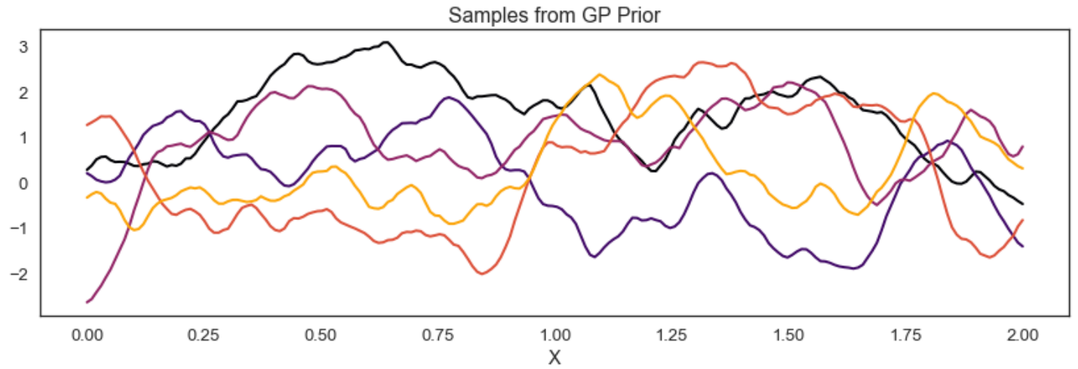
We can construct an infinitely parametric model.

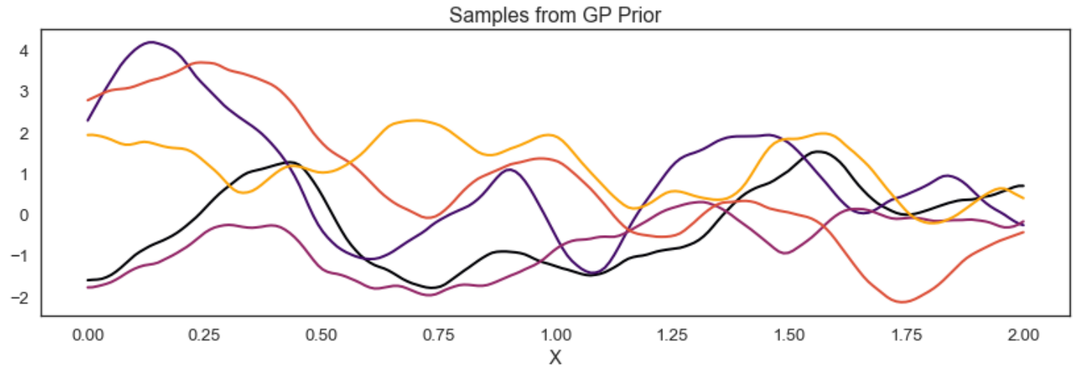
This is called a non-parametric model.

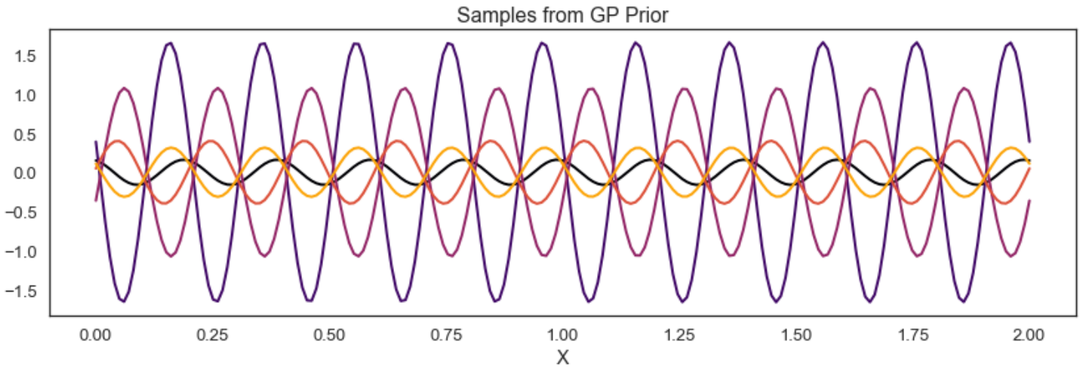
We just need to be able to define a finite kernel $\kappa(x, x') = \psi(x)^T \psi(x')$!!

Kernels (**row1**: exp, m32 **row2**: m52, cos)









Universal Approximation

- Exact correspondence between the gaussian process (direct usage of gaussians in space) to the basis function regression (in feature space with gaussians for prior parameters) in the kernelized representation, as long as we identify the GP covariance function k with the kernel function $\kappa(x, x') = \phi(x)^T \Sigma \phi(x')$. (Mercer's theorem)
- We have seen such universal approximation in NN
- there is a connection for both single layer and deep NN

Setting up the model

```
with pm.Model() as model:
    # priors on the covariance function hyperparameters
    l = pm.Uniform('l', 0, 10)
    # uninformative prior on the function variance
    s2_f = pm.HalfCauchy('s2_f', beta=10)
    # uninformative prior on the noise variance
    s2_n = pm.HalfCauchy('s2_n', beta=5)
    # covariance functions for the function f and the noise
    f_cov = s2_f**2 * pm.gp.cov.ExpQuad(1, l)
    mgp = pm.gp.Marginal(cov_func=f_cov)
    y_obs = mgp.marginal_likelihood('y_obs',
        X=xtrain.reshape(-1,1), y=ytrain, noise=s2_n,
        is_observed=True)
```

Back to the rat tumor model

Posterior-predictive distribution, as a function of upper level parameters $\eta = (\alpha, \beta)$.

$$p(y^*|D, \eta) = \int d\theta \, p(y^*|\theta) \, p(\theta|D, \eta)$$

A likelihood with parameters η and simply use maximum-likelihood with respect to η to estimate these η using our "data" y^*

Empirical Bayes or Type-2 MLE

- MLE with respect to η
- involves an optimization
- unlike cross-validation, θs not-yet estimated on training set.
- indeed we marginalize over θs so can use training set.
- in practice often match moments of predictive or posterior

Levels of Bayes

Method	Definition
Maximum Likelihood	$\hat{\theta} = argmax_\theta p(D\|\theta)$
MAP estimation	$\hat{\theta} = argmax_\theta p(D\|\theta)p(\theta\|\eta)$
ML-2 (Empirical Bayes)	$\hat{\eta} = argmax_\eta \int d\theta\, p(D\|\theta)p(\theta\|\eta) = argmax_\eta p(D\|\eta)$
MAP-2	$\hat{\eta} = argmax_\eta \int d\theta\, p(D\|\theta)p(\theta\|\eta)p(\eta) = argmax_\eta p(D\|\eta)p(\eta)$
Full Bayes	$p(\theta, \eta\|D) \propto p(D\|\theta)p(\theta\|\eta)p(\eta)$

INFERENCE

Use the marginal likelihood:

$$p(y|X) = \int_f p(y|f,X)p(f|X)df$$

The Marginal likelihood given a GP prior and a gaussian likelihood is:

$$\log p(y|X) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|K+\sigma^2 I| - \frac{1}{2}y^T(K+\sigma^2 I)^{-1}y$$

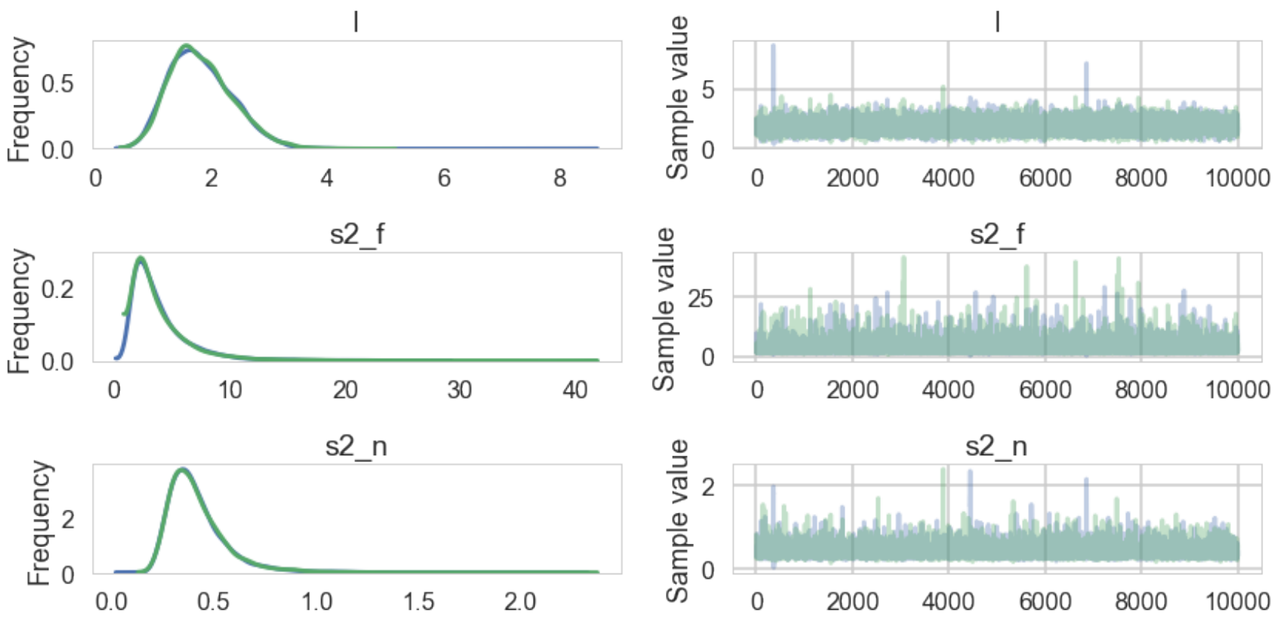
MAP-2 Fitting in pymc3

```
with model:
    marginal_post = pm.find_MAP()
```

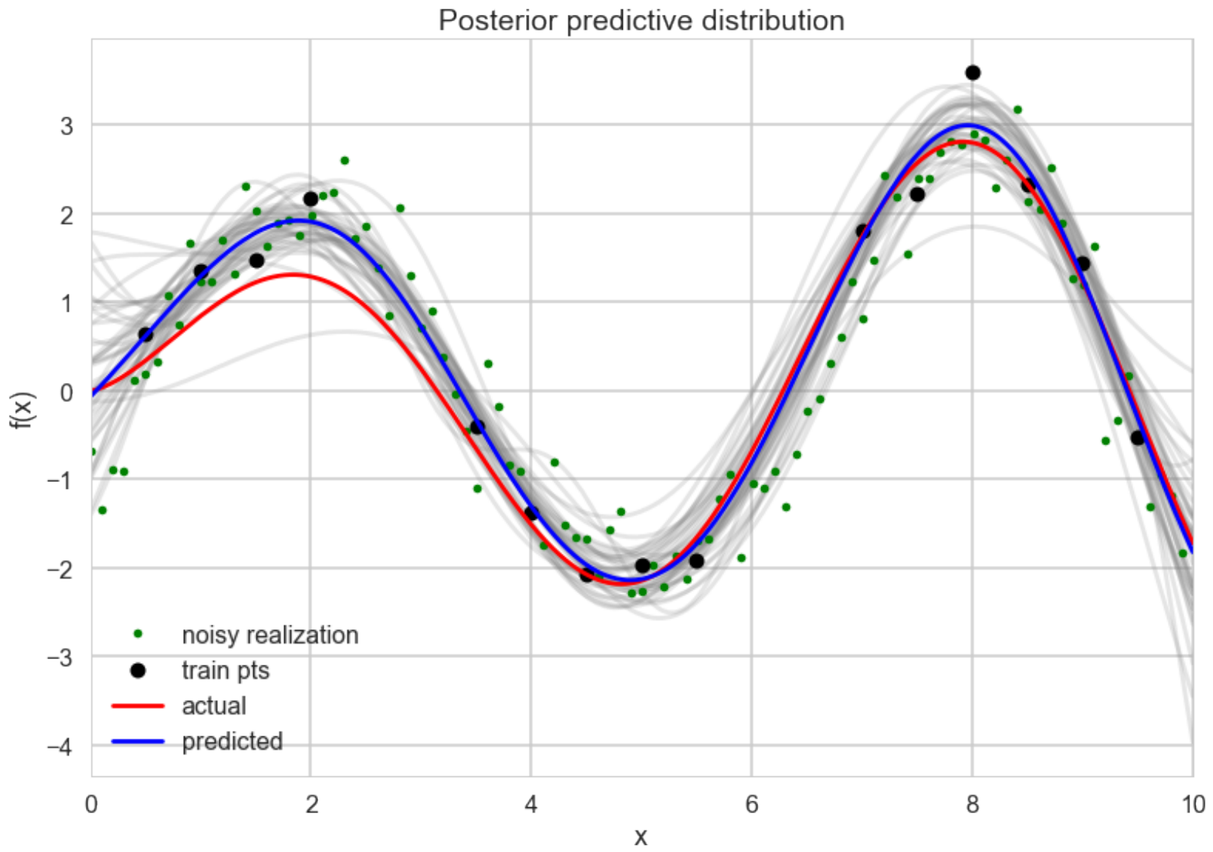
```
{'l': array(1.438132008790354),
 'l_interval__': array(-1.7839733426164 66),
 's2_f': array(2.047500439200898),
 's2_n': array(0.3465300514941838)}
```

MCMC

```
with model:
     trace = pm.sample(10000, tune=2000,
        nuts_kwargs={'target_accept':0.85})
 with model:
     fpred = mgp.conditional("fpred",
        Xnew = x_pred.reshape(-1,1),
        pred_noise=False)
     ypred = mgp.conditional("ypred",
        Xnew = x_pred.reshape(-1,1),
        pred_noise=True)
     gp_samples = pm.sample_ppc(trace,
         vars=[fpred, ypred],
         samples=200)
```



Posterior (predictive) curves



Where are GPs used?

- geostatistics with kriging, oil exploration
- spatial statistics
- as an interpolator (0 noise case) in weather simulations
- they are equivalent to many machine learning models such as kernelized regression, SVM and neural networks (some)
- ecology since model uncertainty is high
- they are the start of non-parametric regression
- time series analysis (see cover of BDA)
- because of the composability of kernels, in automates statistical analysis (see the automatic statistician)