definitions

let

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

let

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix}$$

let

$\vec{x}_1 \in \mathbb{R}^m$

$\vec{x}_2 \in \mathbb{R}^n$

let

$$Q(\vec{x}_1, \vec{x}_2) = \begin{bmatrix} (\vec{x}_1 - \vec{\mu}_1) \\ (\vec{x}_2 - \vec{\mu}_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (\vec{x}_1 - \vec{\mu}_1) \\ (\vec{x}_2 - \vec{\mu}_2) \end{bmatrix}$$

$=$

$(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{11} (\vec{x}_1 - \vec{\mu}_1) +$

$(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{12} (\vec{x}_2 - \vec{\mu}_2) +$

$(\vec{x}_2 - \vec{\mu}_2)^T \Sigma^{21} (\vec{x}_1 - \vec{\mu}_1) +$

$(\vec{x}_2 - \vec{\mu}_2)^T \Sigma^{22} (\vec{x}_2 - \vec{\mu}_2)$

$=$

$(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{11} (\vec{x}_1 - \vec{\mu}_1) +$

$2(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{12} (\vec{x}_2 - \vec{\mu}_2) +$

$(\vec{x}_2 - \vec{\mu}_2)^T \Sigma^{22} (\vec{x}_2 - \vec{\mu}_2)$

[the immediately previous equals sign is due to:

1) $A = A^T \implies A^{-1} = (A^T)^{-1} = (A^{-1})^T$

2) $\vec{x}^T \vec{y} = \vec{y}^T \vec{x}$]

i

The marginal of x_1 immediately follows from the result in iii) so I'm going to assume iii).

let iii) hold without proof (yet)

$\vec{x} = \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix} \sim N(\vec{x}_1, \mu_1, \Sigma_{11})\ N(\vec{x}_2, b, A)$

$b := \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1)$

$A := \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$

$p(\vec{x}_1, \vec{x}_2) = \int_{\vec{x}_2 \in \mathbb{R}^n} N(\vec{x}_1, \mu_1, \Sigma_{11})\ N(\vec{x}_2, b, A) =$

$N(\vec{x}_1, \mu_1, \Sigma_{11})(1)$

because the integral of a pdf is 1

ii)

$\Sigma^{11} =$

$(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{T})^{-1} =$

$\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^{T}\Sigma_{11}^{-1}$

$\Sigma^{22} =$

$(\Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12})^{-1} =$

$\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^{T}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{T})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$

$\Sigma^{12} =$

$-\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12})^{-1} =$

$(\Sigma^{21})^{T} = \Sigma^{12}$

$\mathrm{A} = \Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}$

$\mathrm{b} = \mu_2 + \Sigma_{12}^{T}\Sigma_{11}^{-1}(x_1 - \mu_1)$

$\vec{q_i} = \vec{x}_i - \vec{\mu}_i$

$Q(\vec{x}_1, \vec{x}_2) =$

$(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{11} (\vec{x}_1 - \vec{\mu}_1) +$

$2(\vec{x}_1 - \vec{\mu}_1)^T \Sigma^{12} (\vec{x}_2 - \vec{\mu}_2) +$

$(\vec{x}_2 - \vec{\mu}_2)^T \Sigma^{22} (\vec{x}_2 - \vec{\mu}_2)$

$=$

$\vec{q}_1^T \left[\Sigma_{11}^{-1} + \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^T \Sigma_{11}^{-1} \right] \vec{q}_1 +$

$2\vec{q}_1^T \left[- \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} \right] \vec{q}_2 +$

$\vec{q}_2^T \left[(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} \right] \vec{q}_2$

$=$

$\vec{q}_1^T \Sigma_{11}^{-1} \vec{q}_1 +$

$\vec{q}_1^T \Sigma_{11}^{-1} \Sigma_{12} (A)^{-1} \Sigma_{12}^T \Sigma_{11}^{-1} \vec{q}_1 -$

$2\vec{q}_1^T \Sigma_{11}^{-1} \Sigma_{12} (A)^{-1} \vec{q}_2 +$

$\vec{q}_2^T (A)^{-1} \vec{q}_2$

$=$

$\vec{q}_1^T \Sigma_{11}^{-1} \vec{q}_1$

$+$

$(\Sigma_{12} \Sigma_{11}^{-1} q_1)^T A^{-1} (\Sigma_{12} \Sigma_{11}^{-1} q_1) - 2(\Sigma_{12} \Sigma_{11}^{-1} q_1)^T A^{-1} q_2 + q_2^T A^{-1} q_2$

let $U = \Sigma_{12} \Sigma_{11}^{-1} q_1$

note that A is symmetric, and thus A^{-1} is too

$=$

$U^T A^{-1} U - 2U^T A^{-1} q_2 + q_2^T A^{-1} q_2$

which is exactly the expansion of the quadratic form

$(q_2 - U)^T A^{-1} (q_2 - U)$

=

$(x_2 - \mu_2 - \Sigma_{12}\Sigma_{11}^{-1} q_1)^T A^{-1} (x_2 - \mu_2 - \Sigma_{12}\Sigma_{11}^{-1} q_1)$

=

$(x_2 - b)^T A^{-1} (x_2 - b)$

Thus

$Q(x_1, x_2) = \vec{q_1}^T \Sigma_{11}^{-1} \vec{q_1} + (x_2 - b)^T A^{-1} (x_2 - b)$

pdf of multivariate normal is :

$p(x_1, x_2) =$

$$\frac{1}{(2\pi)^{\frac{m+n}{2}} \mid \Sigma \mid^{1/2}} \exp\{-\tfrac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \mu)\} dx_2$$

=

$$\frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{1/2}} \exp\{-\tfrac{1}{2}(Q(x_1, x_2))\} dx_2$$

=

$$\frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{1/2}} \exp\{-\tfrac{1}{2}(\vec{q_1}^T \Sigma_{11}^{-1} \vec{q_1} + (x_2 - b)^T A^{-1} (x_2 - b))\} dx_2$$

iii

given $\mid \Sigma \mid = \mid \Sigma_{11} \mid\mid \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \mid$

previously, I showed that the pdf of mvn is :

$$\frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{1/2}} \exp\{-\tfrac{1}{2}(\vec{q_1}^T \Sigma_{11}^{-1} \vec{q_1} + (x_2 - b)^T A^{-1}(x_2 - b))\}$$

$=$

$$\frac{1}{(2\pi)^{\frac{m+n}{2}} (\mid \Sigma_{11} \mid\mid \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \mid)^{1/2}} *$$

$\exp\{-\frac{1}{2}(\vec{q_1}^T \Sigma_{11}^{-1} \vec{q_1}\} *$

$\exp\{-\frac{1}{2}(x_2 - b)^T A^{-1}(x_2 - b))\}$

$=$

$N(x_1 \mid \mu_1, \Sigma_{11}) N(x_2 \mid b, A)$

iv

$$f_{2|1}(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f(x_1)}$$

$$=$$

$$\frac{N(x_1 \mid \mu_1, \Sigma_{11})N(x_2 \mid b, A)}{N(\vec{x}_1, \mu_1, \Sigma_{11})}$$

$$=$$

$$N(x_2 \mid b, A)$$

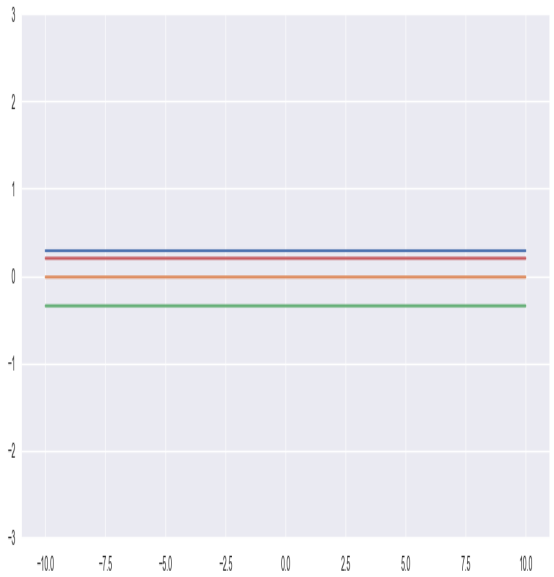
v

one sample from a multivariate Gaussian is an n vector, where n is the dimension of the Gaussian (here n=500). This vector is my function because 500 is basically infinity. So I do this 4 times to make 4 functions.

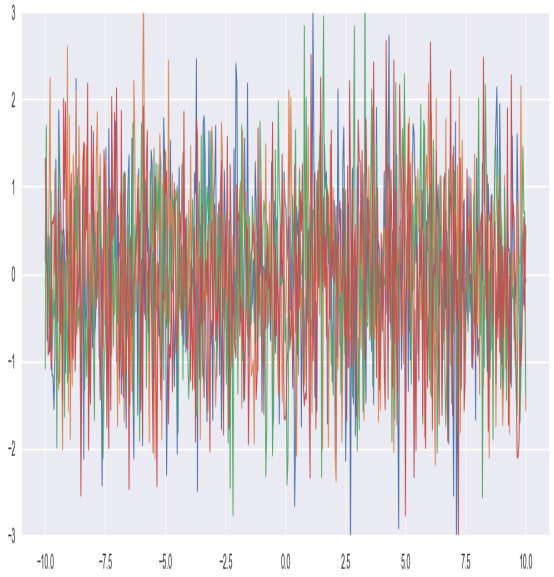
the covariance matrix tells me the relationship between the n points in my sample vector.

Σ_{ij} = is the covariance between the ith sample and the jth sample. Thus Σ is symmetric since i and j have the same covariance as j and i.

Σ = np.ones(n) means all points have the same relationship. This will result in a constant graph. Any covariance matrix with all the same number will result in constant graphs, but the bigger the number, the further apart different samples will tend to be.



Σ = identity means no covariance between the samples i.e. the functions are noise. The jth sample has no relationship with the ith sample at all because $\Sigma_{ij} = 0$.

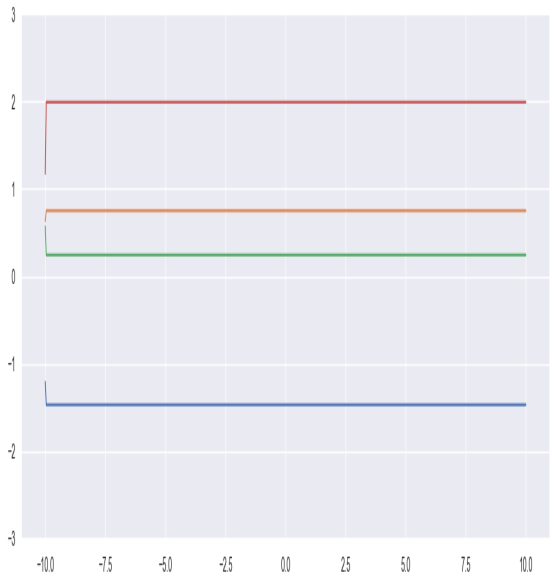


my choice of i,j for Σ_{ij} will determine how similar the ith and jth point in my n vector is. Another way to see this is to view the multivariate Gaussian as a repeated and not necessarily independent sampling from n Univariate Gaussians.The level of independence is determine by Σ (Σ_{ij} =0 means independent i,j).

The smoothness of this function is very not smooth. Smooth means that a small change in x corresponds to a small change in y. Here, the x's have no relationship, so you might have a huge difference in y with a small change in x.

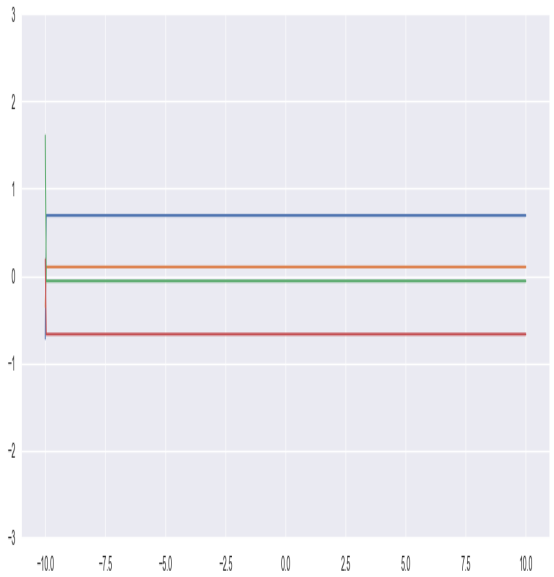
Here is a graph of covariance matrix when it's all ones except for $\Sigma_{1j} = \Sigma_{i1} = .9$ i.e. all the samples have the same relationship, except for the

first sample. The first sample is slightly less than 1 covaraince with all other samples.



as we can see, it is constant, except for the first point, which is slightly different.

if instead of .9 I use $\Sigma_{1j} = \Sigma_{i1} = .1$ I get:

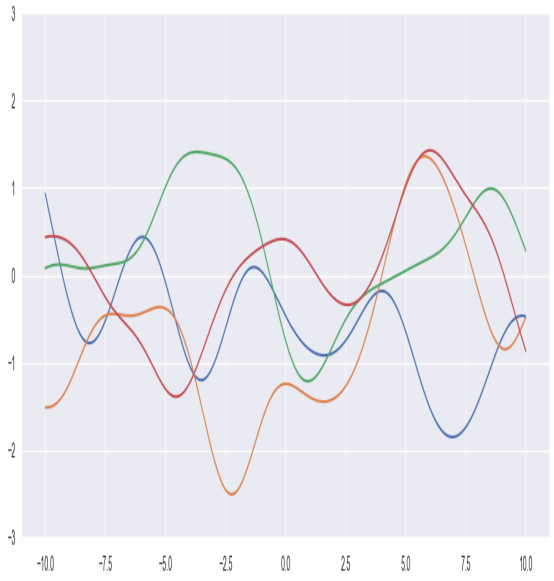


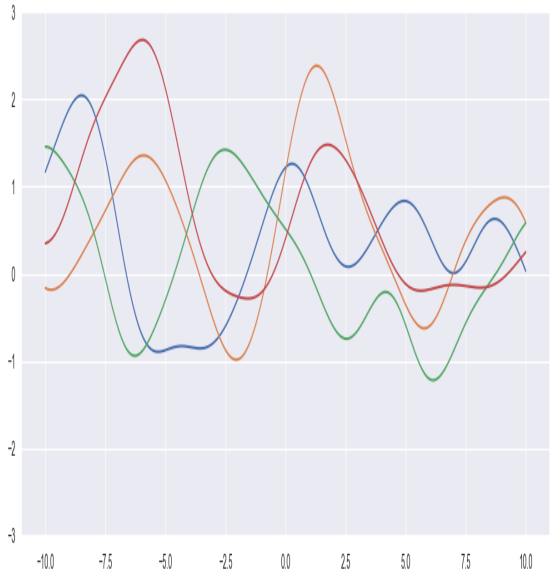
the difference between the first sample and all other samples is much larger, because their covariance is much smaller.

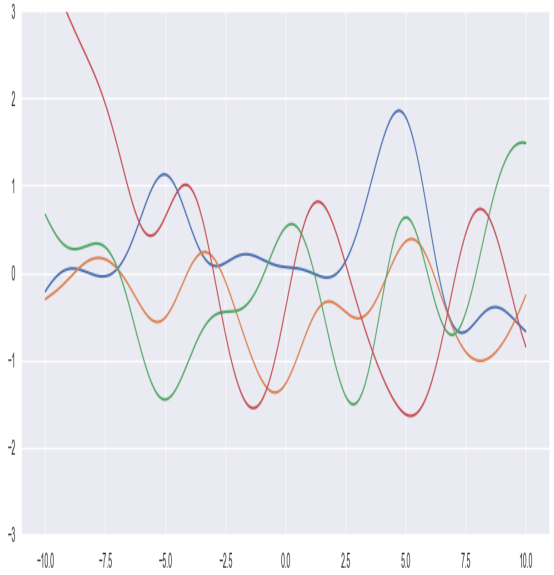
vi

now we use a gaussian kernel. This kernel says that the ith and jth sample have higher covariance when they are closer together. This results in a smooth function, because now the difference in y values for ith and jth sample is directly related to the difference in x values. I have included three different iterations of the 4 random functions to show that they are all different, but always smooth. It's random sampling, but in a way that guarantees smoothness.

Of course, it's not actually smooth since it's really a bunch of line segments connected. However, as the difference between x's get's smaller, we get smoother. The real number line is when that difference is infinitesimal, so we approximate that.




3
2
1
0
-1
-2
-3
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0


3
2
1
0
-1
-2
-3
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0

vii

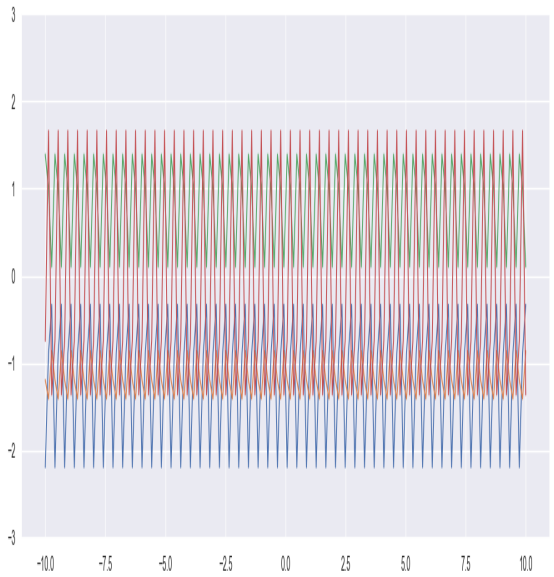
to make a periodic kernel with period T, we need the Tth samples to be highly correlated. To make a period of 3 points, we need every third entry in the ith row to be 1. We also want to preserve smoothness, so retain the quality that close x implies close y.

the value of μ only makes the y-value the same plus μ. So I choose μ=0, but the shape of the graph won't change for different mu values. If μ is different for each sample, then you get a noisier periodic function

here is an example of a periodic covariance matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

here are four random function generated by this covariance matrix for n=150 (b/c n=500 is too hard to see).



a great way to achieve having 1's in every Tth spot is to use

$k(x, x') = \exp\{-\sin^2(\pi(x-x')/T))\}$ so when x_i and x_j are T apart, you get e^0 i.e. 1. Also, this has the property of giving close x's a higher covariance, so we achieve smoothness. I square it so that matrix is positive definite, since squaring always gives positive entries.

to make it repeat every third point, I use

$T = p* \mid x_{max} - x_{min} \mid /(n-1)$

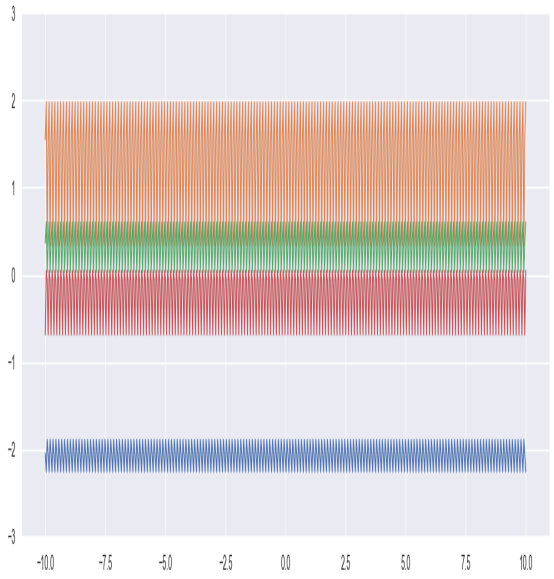
where T = period = 3, $x_{max} = 10 =$ -x_{min} and n= 500

This has an advantage over a different periodic function, tan, because it doesn't blow up to infinity. So if you want to have a large period, sin will give you a very smooth function.

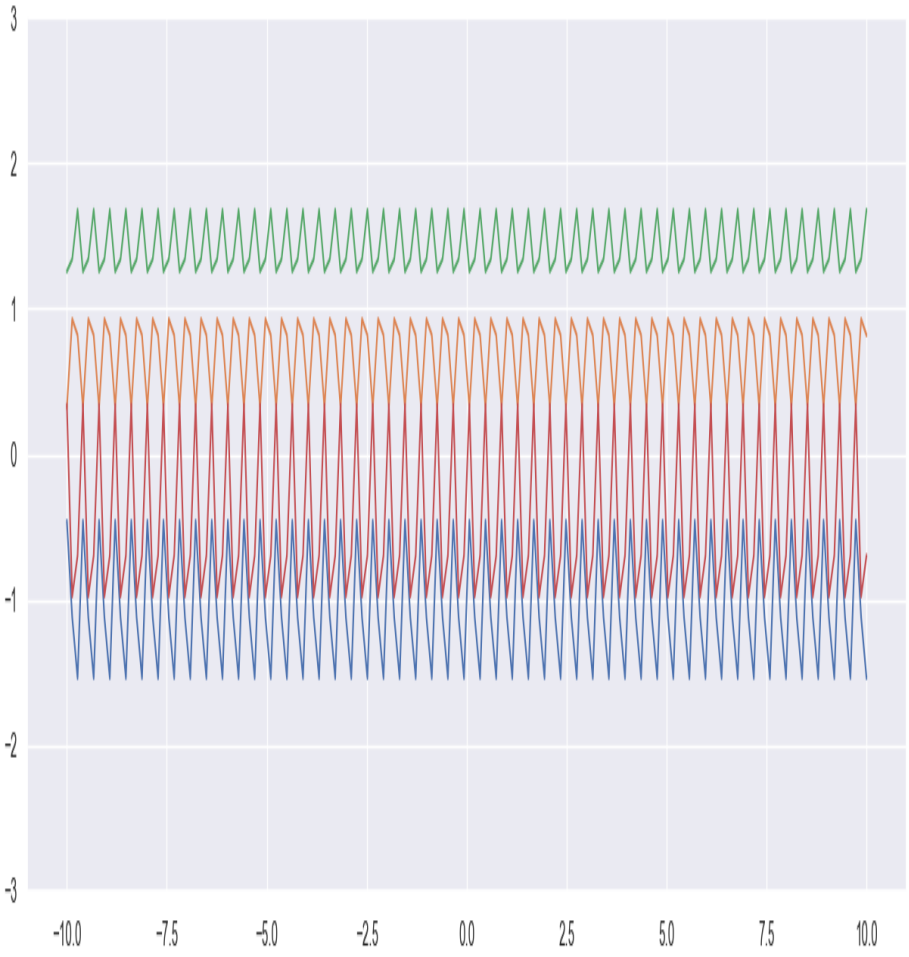
So I can get my points, here is the graph of

$k(x, x') = \exp\{-\sin^2(\pi(x - x')/T))\}$

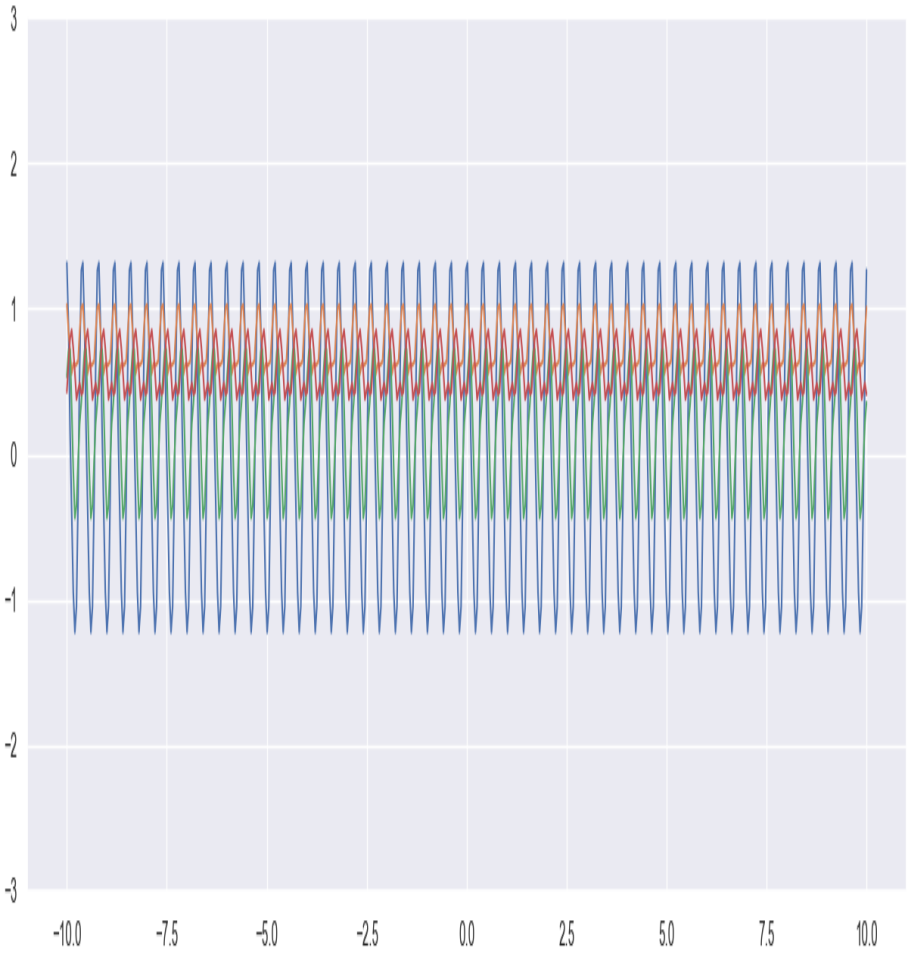
with n=500, T=3,



alright so it's periodic great. Here it is with a smaller n so we can actually see what's going on.



Hopefully you can see that it's a bit jagged. this makes sense because it has to repeat every third point, so we don't get the smooth structure brought on by close x corresponding to close y. SO let's try with a bigger T and see what we get.



so having a larger period makes it smoother because sin kernel gives us similar smoothing behavior as Gaussian kernel, but it also repeats every T.

viii

in part iv I attempted to prove that

$f_{2|1}(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f(x_1)}$

=

$N(x_2 \mid b, A)$

with

$\mathrm{A} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$

$\mathrm{b} = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1)$

against my better judgement I will assume that I am correct.

instead of

$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

we have

$\Sigma = \begin{bmatrix} k(X, X) & k(X, \bar{X}) \\ k(\bar{X}, X) & k(\bar{X}, \bar{X}) \end{bmatrix}$

$\bar{Y} \mid Y$ is analogous to $p(x_2 \mid x_1)$

since the training data is given (i.e. we know Y), it makes sense to condition on it to find the correct value for $\bar{Y}$ (the y-values we aim to predict from x). I used a mean of 0 btw.

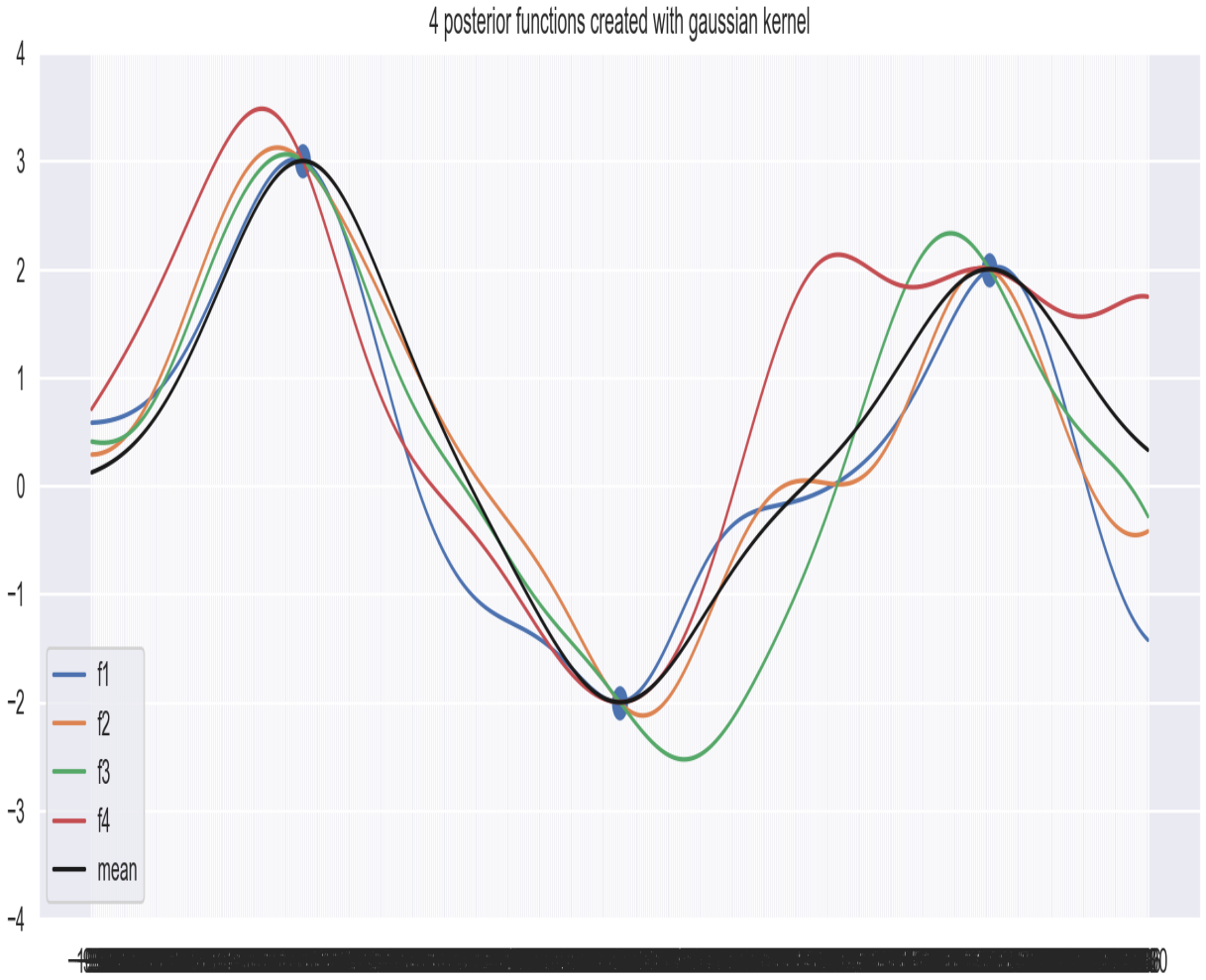
$p(\bar{Y} \mid Y) \sim N(K_{12}^T K_{11}^{-1} Y, K_{22} - K_{12}^T K_{11}^{-1} K_{12}) =$

$N(K_{21} K_{11}^{-1} Y, K_{22} - K_{21} K_{11}^{-1} K_{12}) =$

$$N(K(\bar{X}, X)K(X, X)^{-1}Y, K(\bar{X}, \bar{X}) - K(\bar{X}, X)K(X, X)^{-1}K(X, \bar{X})) =$$
$$N(y \mid b', A')$$

ix

so we now have an expression for the distribution of the test data as a function of the training data. We can now predict the test data by sampling from a multivariate normal with the parameters of that posterior distribution. So it's a Bayesian regression method. Here is the result of sampling from mvn(y, b', A')



the posterior goes through the training points, and then randomly samples for the other points. This is the reason that non-parametric regression requires many points, because the further away from the test data you are, the more of a guess you make.

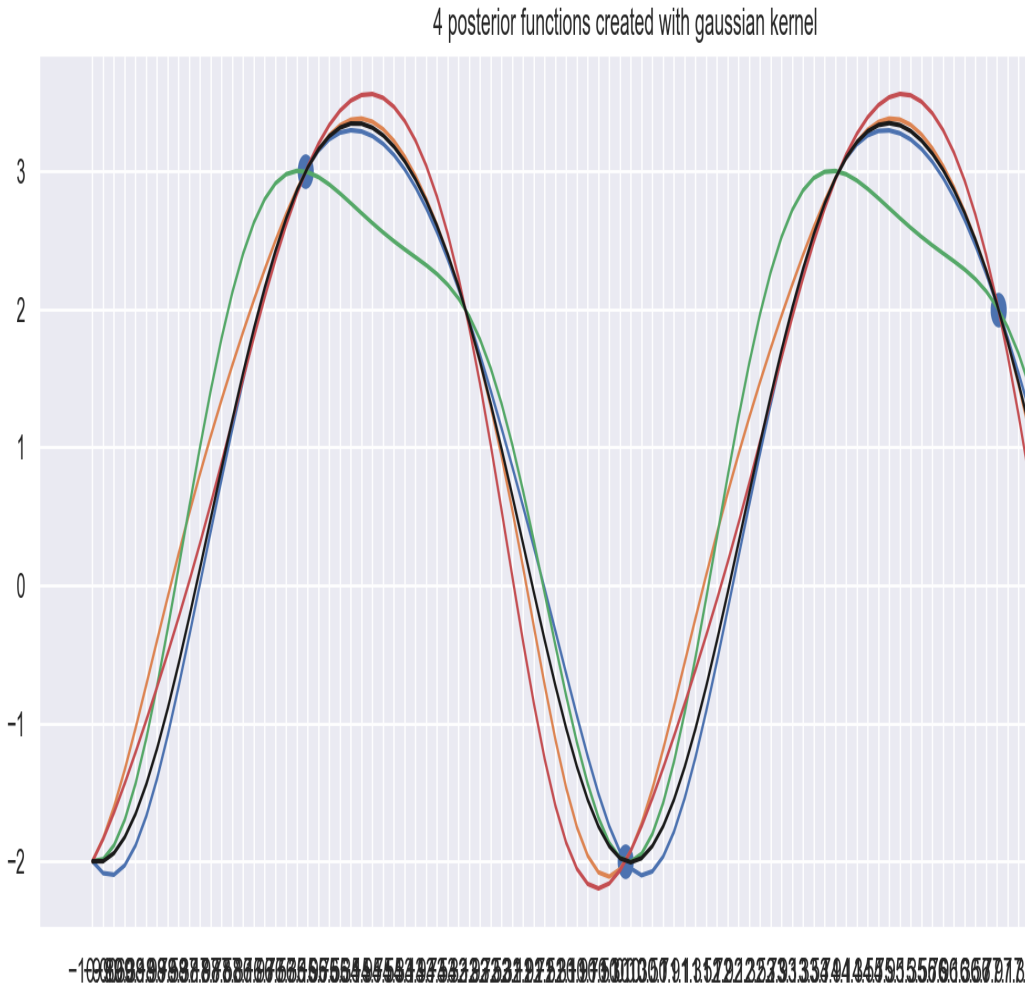
This isn't a problem in linear regression, where we embed some prior

belief about linearity into the prediction which dictates our shape during sparse areas. Here, our only prior belief is the test data, so in a sparse area it's a random guess (based on the kernel). In a known area (x associated with training data) we know exactly what to guess - the y associated with the training data (for more than 1 y for a single x, it might be an average).

The problem with linear regression is: what if your data isn't linear. So here, we rely on more data to get a good fit, but we don't have to make any assumptions.

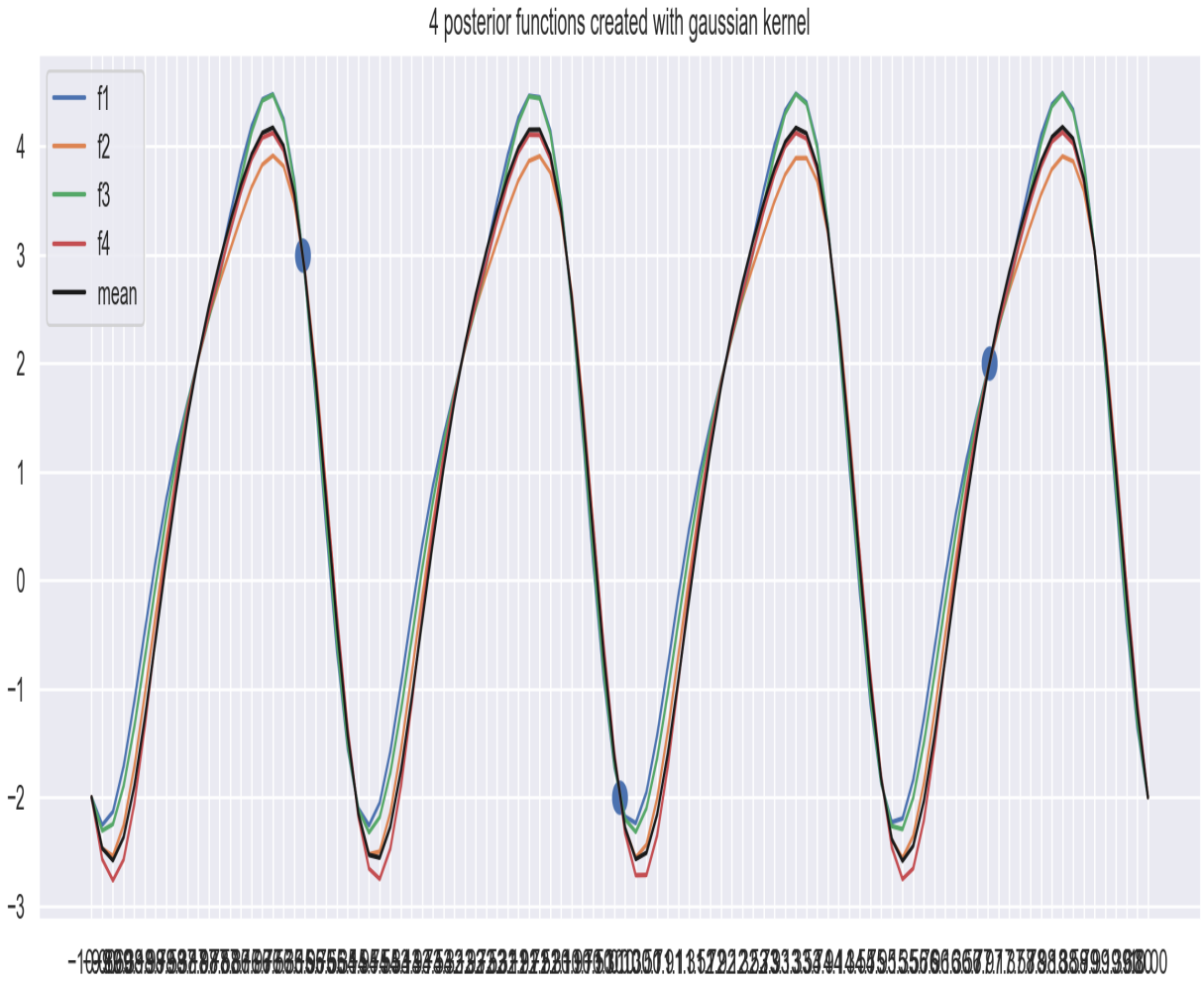
x

using a periodic covariance matrix will give a periodic shape to the graph. The same way that a linear regression embeds some information into our graph, namely that it's linear, here we embed that it is periodic. So in between training points, we see a periodic graph.



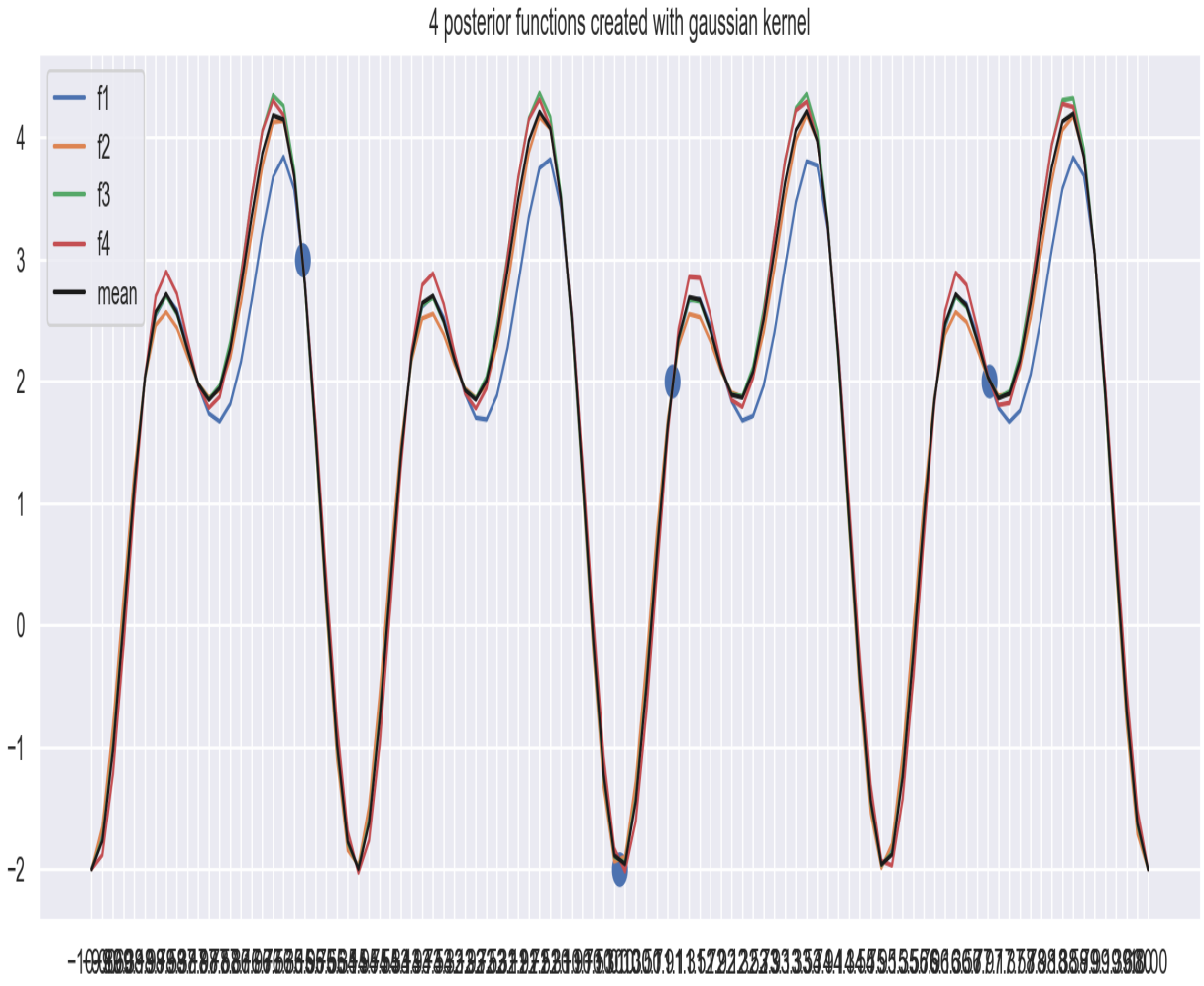
here is T=10

here is T=5



so the one with a lower period still follows the pattern, just with more bumps inbetween the training data. If we had more training data, it would still fit it, but it would have to change it's beatiful shape. These training data happen to be perfect for sin. Here is the graph when we

add an extra training point.



so as you can see it messes up the shape a bit when we have a point not on perfectly on our kernel. Instead of ignoring it, or slightly adjusting to it like in a linear regression, we fit it, and change the shape of the output. This the essence of what it means to embed knowledge into

the model versus having the prior be your source of knowledge. Non-parametric must fit the training data since it's all it has to go on; it can't just ignore or slightly adjust to poorly fitted points.

my question for the TA's: in lecture Prof Verma said that when we have two y-training values for the same x, we take the average. When I tried that, I got a non-semidefinite matrix and the result is super wacky. It didn't just average out the y-values as I expected. Did I misunderstand, or am I bad at coding?

x

$K(\bar{X}, X)K(X, X)^{-1}Y$

is the mean derived for the posterior in part viii

xi

the black line in the graphs is the mean