Supervision Assignments in MLBI
Lent Term 2018
Set 4

Supervisor : Dionysis Manousakas
dm754@cam.ac.uk

May 25, 2019

1. *(Bayesian linear and Gaussian process regression.)*

   Plot the time series of monthly mean global CO2 concentrations obtained from the file `co2.txt` (original data obtained from `http://www.esrl.noaa.gov/gmd/ccgg/trends`).

   We will apply Bayesian linear and Gaussian process regression to predict the CO_2 concentration $f(t)$ as a function of time t, where $t = \text{Year} + (\text{Month} - 1) = 12$.

   (a) First we model the function using linear regression, that is, using the functional form

   $$f(t) = at + b + \epsilon(t)$$

   with i.i.d. noise residual $\epsilon(t) \sim N(0, 1)$ and prior $a \sim N(0, 10^2)$, $b \sim N(360, 100^2)$. Compute (e.g. using MATLAB) the posterior mean and covariance over a and b given the CO_2 data.

   (b) Let a_{MAP}, b_{MAP} be the MAP estimate in the question above. The residual is the difference between the observed function values and the predicted mean function values

   $$g_{\text{obs}}(t) = f_{\text{obs}}(t) - (a_{\text{MAP}} t + b_{\text{MAP}}),$$

   where $f_{\text{obs}}(t)$ is the observed value of the CO_2 concentration at time t. Plot $g_{\text{obs}}(t)$. Do you think these residuals conform to our prior over $\epsilon(t)$? State, with justifications, which characteristics of the residual you think do or do not conform to our prior belief.

   (c) Write a MATLAB function to generate samples drawn from a GP. Specifically, given a covariance kernel function $k(\cdot, \cdot)$ and a vector of input points $\mathbf{x}$, return a function $f(\mathbf{x})$ evaluated on the input points $\mathbf{x}$ drawn randomly from a GP with the given covariance kernel and with zero mean.

   (d) Test your function by plotting sample functions drawn from the following kernel, for various settings of the hyperparameters

   $$k(s, t) = \theta^2 \left(\exp\left(-\frac{2\sin^2(\pi(s-t)/\tau)}{\sigma^2} \right) + \phi^2 \exp\left(-\frac{(s-t)^2}{2\eta^2} \right) \right) + \zeta^2 \delta_{s=t}$$

   Describe the characteristics of the drawn functions, and how the characteristics of the functions depend on the parameters.

   (e) Suppose we were to consider modelling the residual function $g(t)$ using a zero mean GP with the covariance kernel above. Based on the plot of $g(t)$ and your explorations in the preceding part, what do you think will be suitable values for the hyperparameters of k?

(f) Extrapolate the CO_2 concentration levels to 2020 using the GP with covariance kernel k of the equation in (d), and your chosen parameter values. Specifically, compute the predictive mean and variance of the residual $g(t)$ for every month between September 2007 and December 2020 given the observed residuals $g_{\text{obs}}(t)$. Plot the means and one standard deviation error bars of the extrapolated CO_2 concentration levels

$$f(t) = a_{\text{MAP}}t + b_{\text{MAP}} + g(t)$$

along with the observed CO_2 levels. Does the behaviour of the extrapolation conform to your expectations? How sensitive are your conclusions to settings of the kernel hyperparameters?

(g) Why is the above procedure not Bayesian? How would we go about modelling $f(t)$ in a Bayesian framework?

(a) To do this we map t_i to an input vector $\mathbf{x}_i = \begin{pmatrix} t_i \\ 1 \end{pmatrix}$ and set $\mathbf{w} = \begin{pmatrix} a \\ b \end{pmatrix}$ such that $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ and perform linear regression to find the hyper-parameters for the posterior weight distribution, given a prior of $\mathbf{w} \sim \mathcal{N}(\mu_0, C), \mu_0 = \begin{pmatrix} 0 \\ 360 \end{pmatrix}, C = \begin{pmatrix} 10^2 & 0 \\ 0 & 100^2 \end{pmatrix}$. Since $y_i \sim \mathcal{N}(w^\top x_i, \sigma^2)$, ($\sigma^2 = 1$ here from the noise) the posterior on $\mathbf{w}$ will also be a normal distribution with variance Σ_n given by $\Sigma_n = (\frac{XX^\top}{\sigma^2} + C^{-1})^{-1}$ and mean $\mu_n = \Sigma_n(XY^\top + C^{-1}\mu_0)$. Both X and Y are row vectors with column vectors as each $\mathbf{x}_i, \mathbf{y}_i$ respectively.
Performing the calculations yields:

$$\mu_n = \begin{pmatrix} 1.57203354 \\ -2775.58645 \end{pmatrix} \text{ and } \Sigma_n = \begin{pmatrix} 4.17453539 \times 10^{-05} & -8.32106418 \times 10^{-02} \\ -8.32106418 \times 10^{-02} & 1.65865938 \times 10^{+02} \end{pmatrix}$$

(b) Since here the distribution is Gaussian, the mean coincides with the mode of the distribution so $\begin{pmatrix} a_{\text{MAP}} \\ b_{\text{MAP}} \end{pmatrix} = \mu_0$ found above.
The sinusoidal pattern implies that there is another process at work yearly cycles (10 cycles per decade). Accounting for that should give a $g_{\text{ons}}(t)$ closer to the $\epsilon(t)$ postulated. As is, the pattern does not conform to the $\epsilon(t)$ function postulated.

(c) See code below regarding how kernel is used as the covariance matrix of the GP with 0 mean to draw samples from a multivariate gaussian distribution. The covariance matrix, $K(X, X')$ is built from the inputs X such that $K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$.

(d) From the equation, the function produced is a superposition of the periodic and squared exponential functions, with ϕ^2 controlling the relative amplitude of the latter to the first one, and the resulting superposition scaled by an amplitude θ^2. Parameters τ and σ control the oscillation period and noise in the oscillation respectively. Parameter η determines the form of the function over larger spans and so can be used to reflect the increase observed for higher t in $g_{\text{obs}}(t)$. The parameter ζ determines how bigger the diagonal of the kernel is than the rest of the kernel – the bigger the elements in the diagonal the less strongly successive input vectors are correlated. Also $\zeta > 0$ required to produce a non-singular (positive semi-definite) kernel.

(e) See code below.

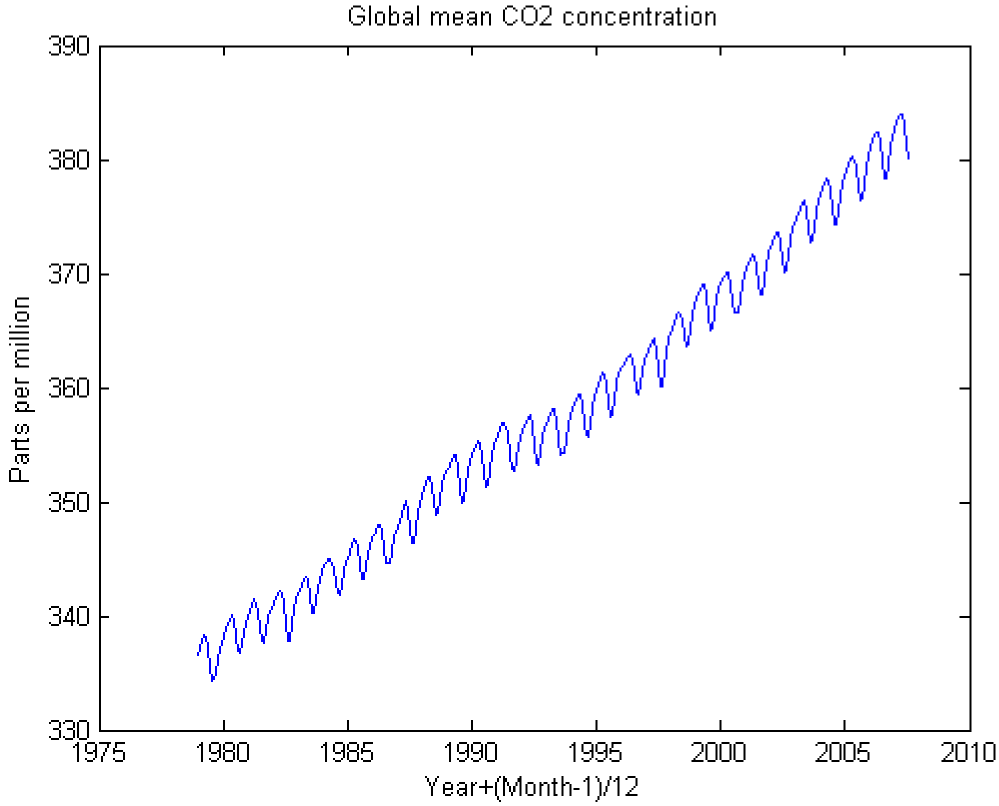
(f) See code below.

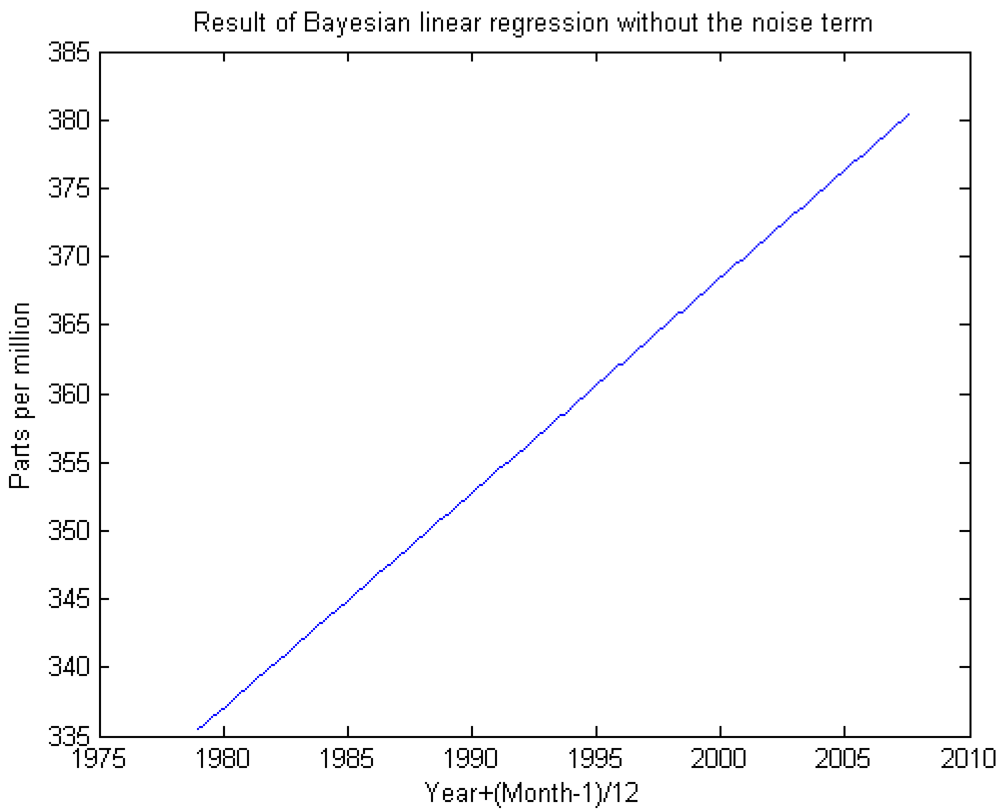
```
close all
clear all

%% Observed function plot

load co2.txt -ascii;
t=co2(:,1)+(co2(:,2)-1)/12;  % points over time
ppm=co2(:,3);   %CO2 concentration in parts per million

plot(t,ppm);
xlabel('Year+(Month-1)/12'),ylabel('Parts per million');
title('Global mean CO2 concentration');

%% Bayesian linear regression
%% (a)
% Regression function: y=ppm-360=w'x+epsilon

y=ppm-360;  % observable vector (-360 for mean zero prior on parameters)

%prior over model parameters
m_0=[0 0]';   % prior mean
cov_0=[10^2 0;0 100^2]; %prior covariance

%posterior mean and covariance
X=[t';ones(1,length(t))];

A=X*X'+inv(cov_0);
cov_pos=inv(A); % posterior covariance
m_pos=inv(A)*X*y;   % posterior mean

%% (b)
figure;
plot(t, m_pos(1)*t+m_pos(2)+360); % plot with MAP estimates of regression
xlabel('Year+(Month-1)/12'),ylabel('Parts per million');
title('Result of Bayesian linear regression without the noise term');

figure;
g_obs=ppm-(m_pos(1)*t+m_pos(2)+360); % residuals
plot(t,g_obs);
xlabel('Year+(Month-1)/12'),ylabel('Parts per million');
title('Residuals of Bayesian linear regression');


%% Gaussian process regression
%% (c)

xs = t; %input points
%% e.g. Application on squared-exponential kernel
omega=0.2; %parameters of kernel
rho=2;
% kernel function
kernel=@(x,y) omega^2*exp(-(repmat(x',size(y))-repmat(x,size(y'))).^2)/(2*rho^2);
% Gaussian Process Samples function
gps=f(kernel,xs);
figure;
plot(xs,gps);
xlabel('Year+(Month-1)/12');
```

```
title('GP samples with squared - exponential kernel');

%% (d)

xs = t; ns = size(xs,1);

%% K1
theta=1;
tau=1;
sigma=1;
phi=0;
eta=8;
zeta=0.2;

K = @(s,t) theta^2*exp(-2*sin(pi*(repmat(s',size(t))-repmat(t,size(s')))/tau).^2/sigma^2)+phi^2*exp(-(r
%gps = chol(K(xs,xs)+eps*eye(size(xs,1)))'*randn(size(xs,1),1);
gps=f(K,xs);
figure;
plot(xs,gps)
xlabel('Year+(Month-1)/12');
title('GP samples with kernel k (phi=0 -> periodic behaviour) ');


%% K2
theta=0.01;
tau=1;
sigma=1000;
phi=1/theta;
eta=8;
zeta=0.2;

K = @(s,t) theta^2*exp(-2*sin(pi*(repmat(s',size(t))-repmat(t,size(s')))/tau).^2/sigma^2)+phi^2*exp(-(r
%gps = chol(K(xs,xs)+eps*eye(size(xs,1)))'*randn(size(xs,1),1);
gps=f(K,xs)
figure;
plot(xs,gps)
xlabel('Year+(Month-1)/12');
title('GP samples with kernel k (theta<<phi -> quadratic exponential)');


%% K3
theta=0;
tau=1;
sigma=1;
phi=1;
eta=8;
zeta=0.2;

K = @(s,t) (theta^2*(exp(-2*sin(pi*(repmat(s',size(t))-repmat(t,size(s')))/tau).^2/sigma^2)+phi^2*exp(-
%gps = chol(K(xs,xs)+eps*eye(size(xs,1)))'*randn(size(xs,1),1);
gps=f(K,xs)
figure;
plot(xs,gps)
xlabel('Year+(Month-1)/12');
title('GP samples with kernel k (theta=0 ->independent noise)');

%% (e)
%% K
theta=1.2;
tau=1;
sigma=2;
phi=0.3;
eta=4;
zeta=0.01;
```

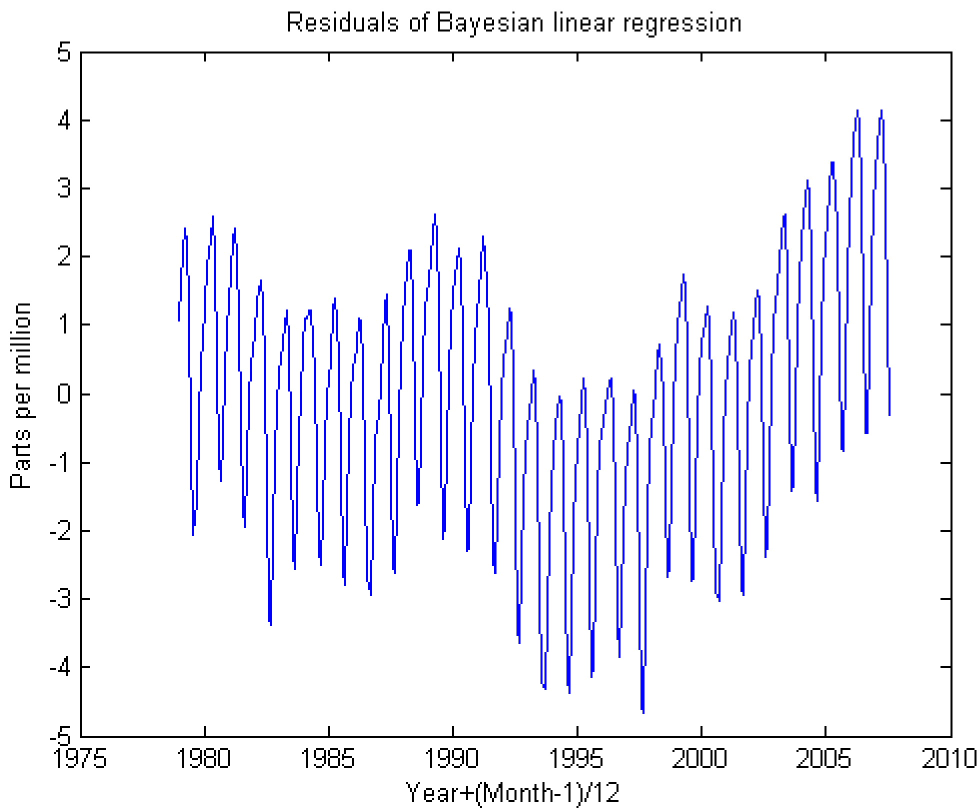
```
K = @(s,t) theta^2*exp(-2*sin(pi*(repmat(s',size(t))-repmat(t,size(s')))/tau).^2/sigma^2)+phi^2*exp(-(r
%gps = chol(K(xs,xs)+eps*eye(size(xs,1)))'*randn(size(xs,1),1);
gps=f(K,xs)
figure;
plot(xs,gps)
xlabel('Year+(Month-1)/12');
title('GP samples with kernel k (tuned parameters)');


%% (f) Extrapolate
t_=(t(end)+1/12:1/12:2020)'; % future times
X_=(1979:1/12:2020)'; % past and future times

eps=1e-8;
m=K(X_(1:length(t)),X_(length(t)+1:end))*inv(K(X_(1:length(t)),X_(1:length(t))))*g_obs; % predicted mea
cov=K(X_(length(t)+1:end),X_(length(t)+1:end))-K(X_(1:length(t)),X_(length(t)+1:end))*inv(K(X_(1:length
g_pred= m+chol(cov+eps*eye(length(m)))*randn(length(m),1); % predicted residual
figure;
plot(X_,[ppm; m_pos(1)*X_(length(t)+1:length(X_))+m_pos(2)+g_pred+360])
xlabel('Year+(Month-1)/12');
title('Extrapolation with GP');
hold on;
std=cov2corr(cov) %standard deviation
errorbar(t_,[m_pos(1)*X_(length(t)+1:length(X_))+m_pos(2)+g_pred+360],std)
hold off

figure
errorbar(t_,[m_pos(1)*X_(length(t)+1:length(X_))+m_pos(2)+g_pred+360],std)
xlabel('Year+(Month-1)/12');
title('Predicted f(t) with one standard deviation error bars');



%% K with detuned parameters
theta=1.2;
tau=1;
sigma=2;
phi=0.9;
eta=4;
zeta=0.01;

K = @(s,t) theta^2*exp(-2*sin(pi*(repmat(s',size(t))-repmat(t,size(s')))/tau).^2/sigma^2)+phi^2*exp(-(r
%gps = chol(K(xs,xs)+eps*eye(size(xs,1)))'*randn(size(xs,1),1);
gps=f(K,xs)
figure;
plot(xs,gps)
xlabel('Year+(Month-1)/12');
title('GP samples with detuned kernel k ');


m=K(X_(1:length(t)),X_(length(t)+1:end))*inv(K(X_(1:length(t)),X_(1:length(t))))*g_obs; % predicted mea
cov=K(X_(length(t)+1:end),X_(length(t)+1:end))-K(X_(1:length(t)),X_(length(t)+1:end))*inv(K(X_(1:length
g_pred= m+chol(cov+eps*eye(length(m)))*randn(length(m),1); % predicted residual
figure;
plot(X_,[ppm; m_pos(1)*X_(length(t)+1:length(X_))+m_pos(2)+g_pred+360])
xlabel('Year+(Month-1)/12');
title('Extrapolation with GP with inappropriate parameter values');
hold on;
std=cov2corr(cov) %standard deviation
errorbar(t_,[m_pos(1)*X_(length(t)+1:length(X_))+m_pos(2)+g_pred+360],std)
hold off



function GPsamples = f( k, x )
```

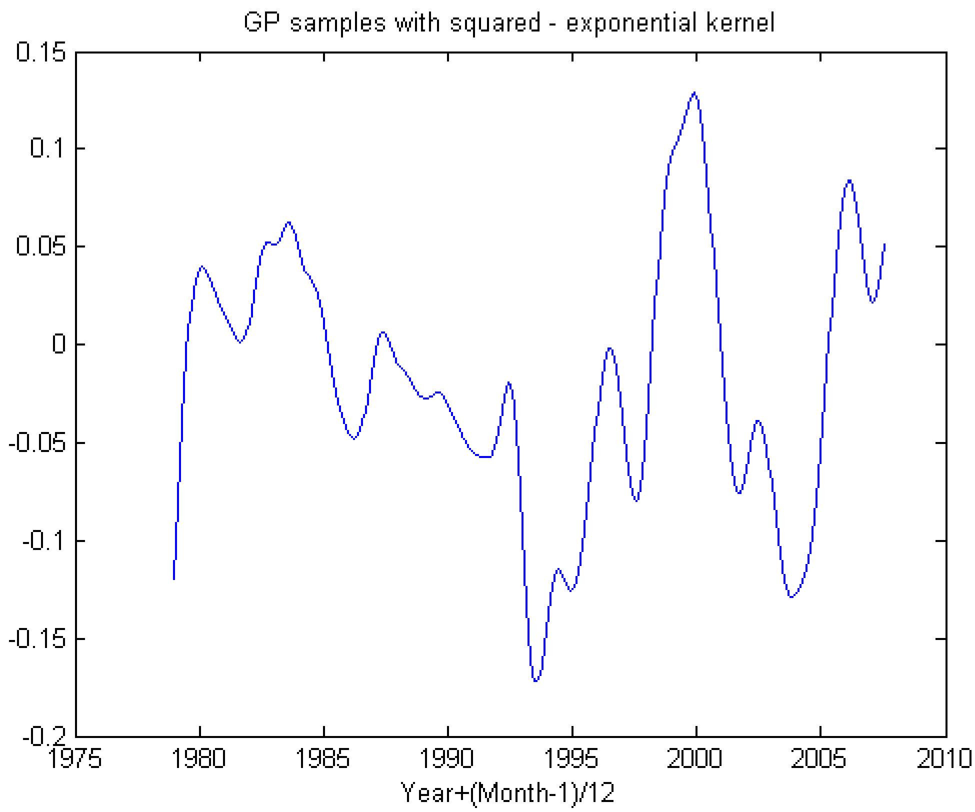
```
%f function that generates samples drawn from a GP
%INPUT
% - GP with zero mean and covariance kernel function k
% - x vector of input points
%OUTPUT
% - GPsamples on the input points x

eps = 1e-8; % small constant added for arithmetic reasons at calculation of Cholesky matrix (to ensure

% We use the Cholesky decompositon of the positive-definite covariance
% matrix: K(.,.)=L*L'. We generate a random vector z of dimension equal
% to the length of input points and obtain a vector L*z with covariance
% matrix K and zero mean
GPsamples = chol(k(x, x)+eps*eye(size(x,1)))'*randn(size(x,1),1)

end
```

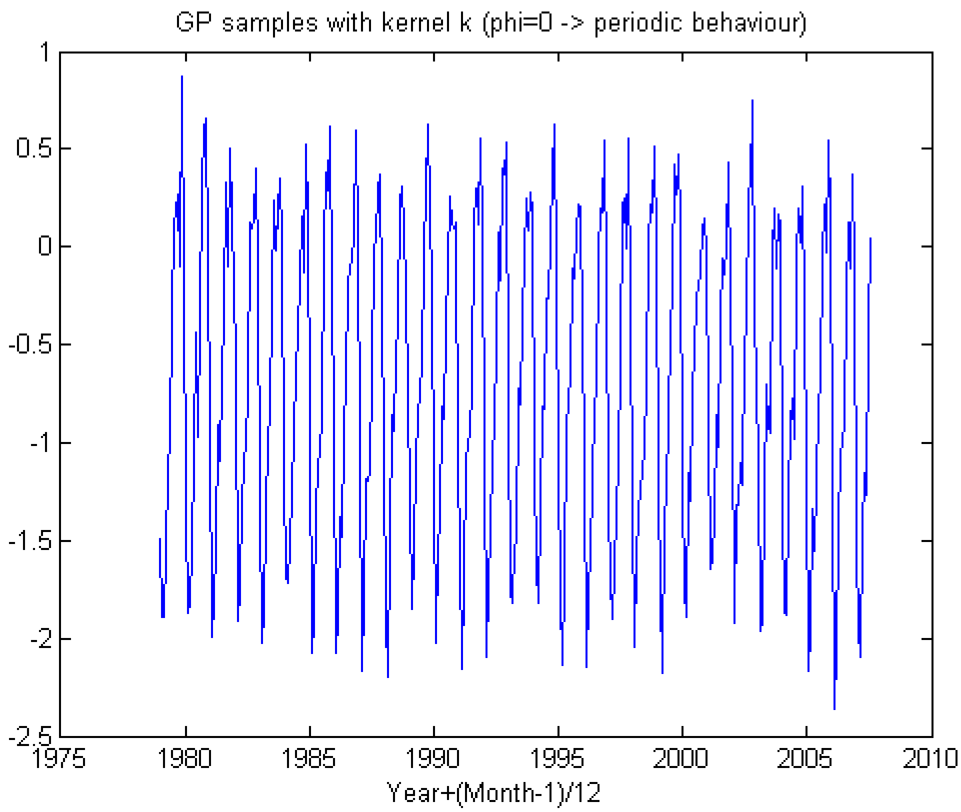


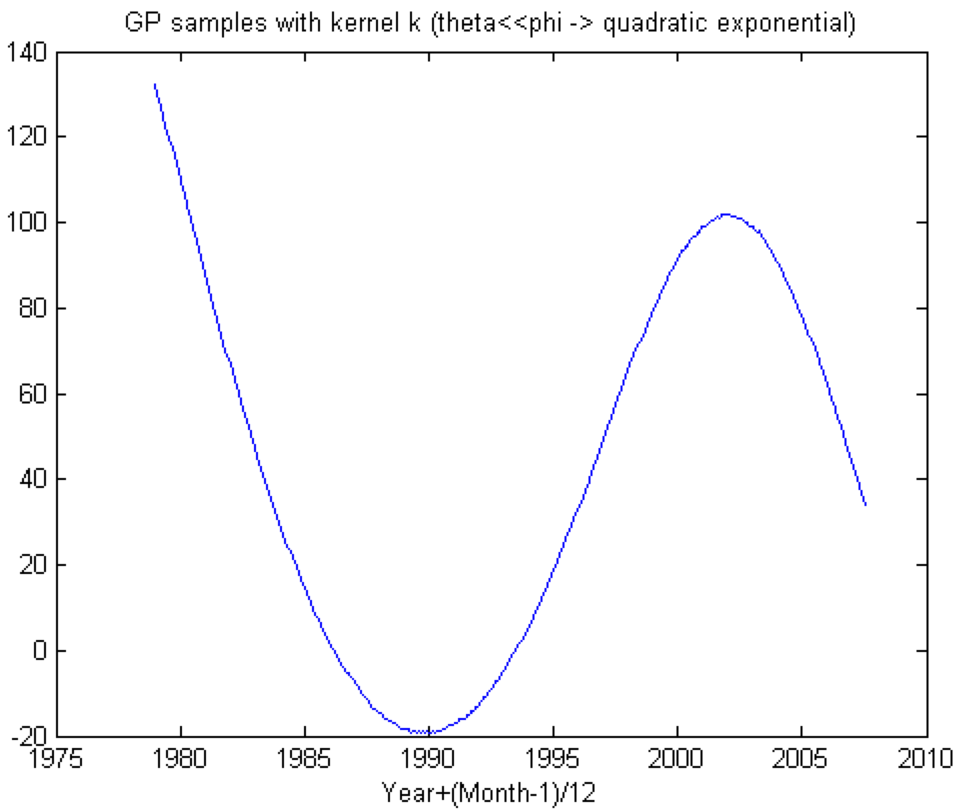


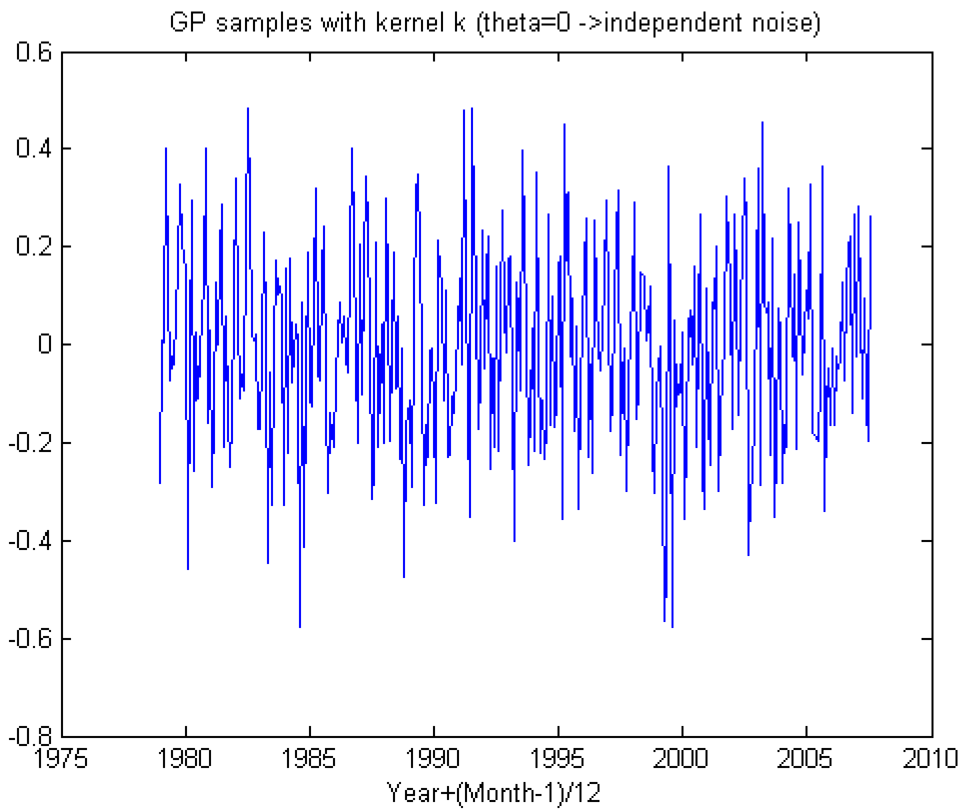
Result of Bayesian linear regression without the noise term
385
380
375
370
365
360
355
350
345
340
335
Parts per million
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12

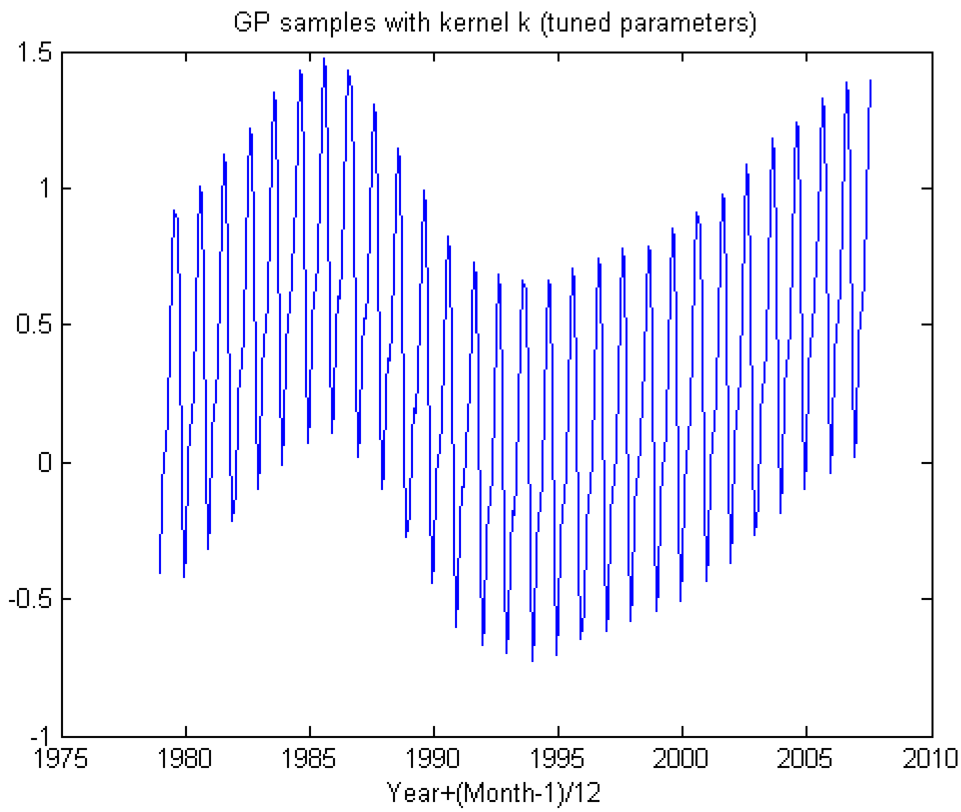

Residuals of Bayesian linear regression
Parts per million
Year+(Month-1)/12
5
4
3
2
1
0
-1
-2
-3
-4
-5
1975
1980
1985
1990
1995
2000
2005
2010

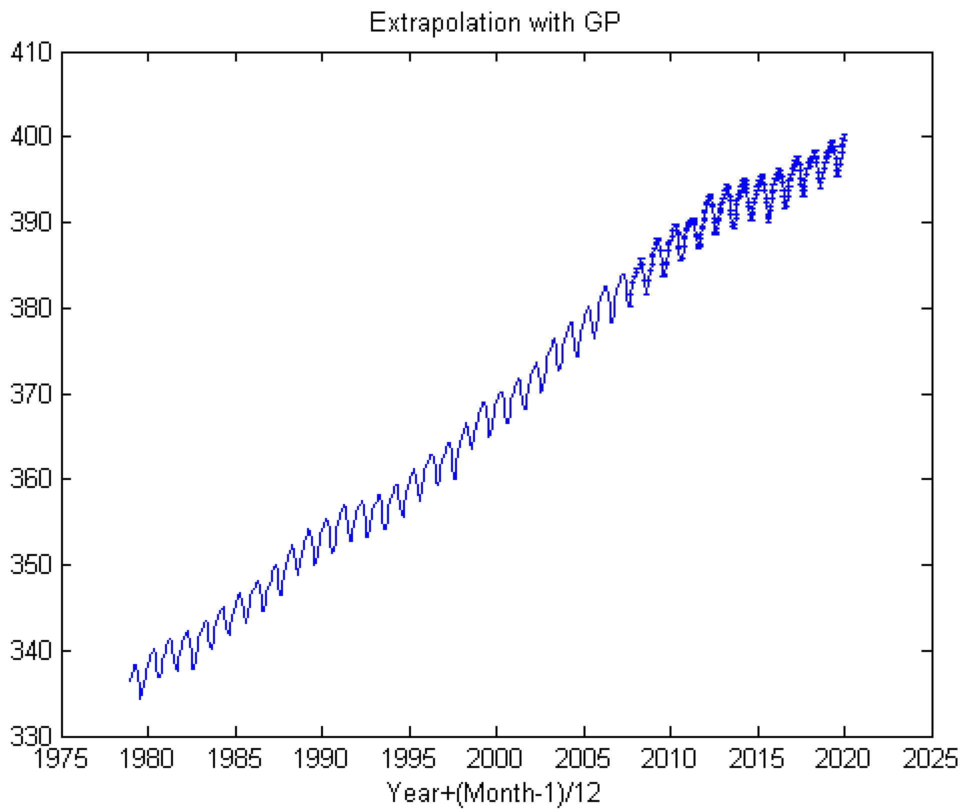

GP samples with squared - exponential kernel
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12


GP samples with kernel k (phi=0 -> periodic behaviour)
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12

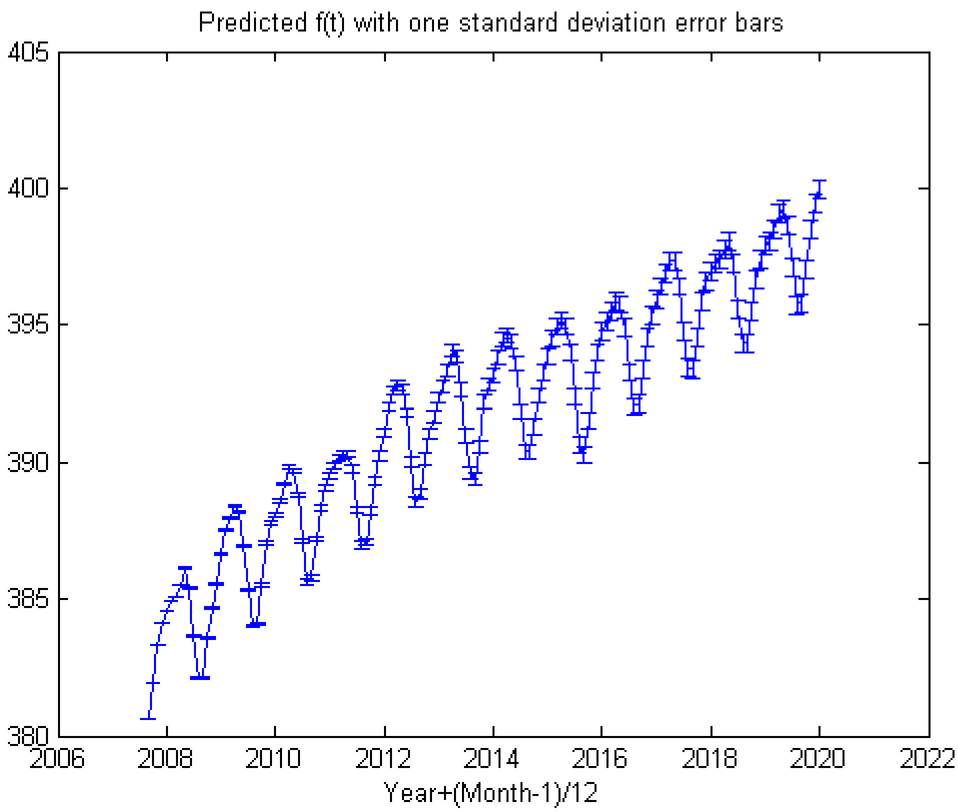

GP samples with kernel k (theta<<phi -> quadratic exponential)
140
120
100
80
60
40
20
0
-20
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12


GP samples with kernel k (theta=0 ->independent noise)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12

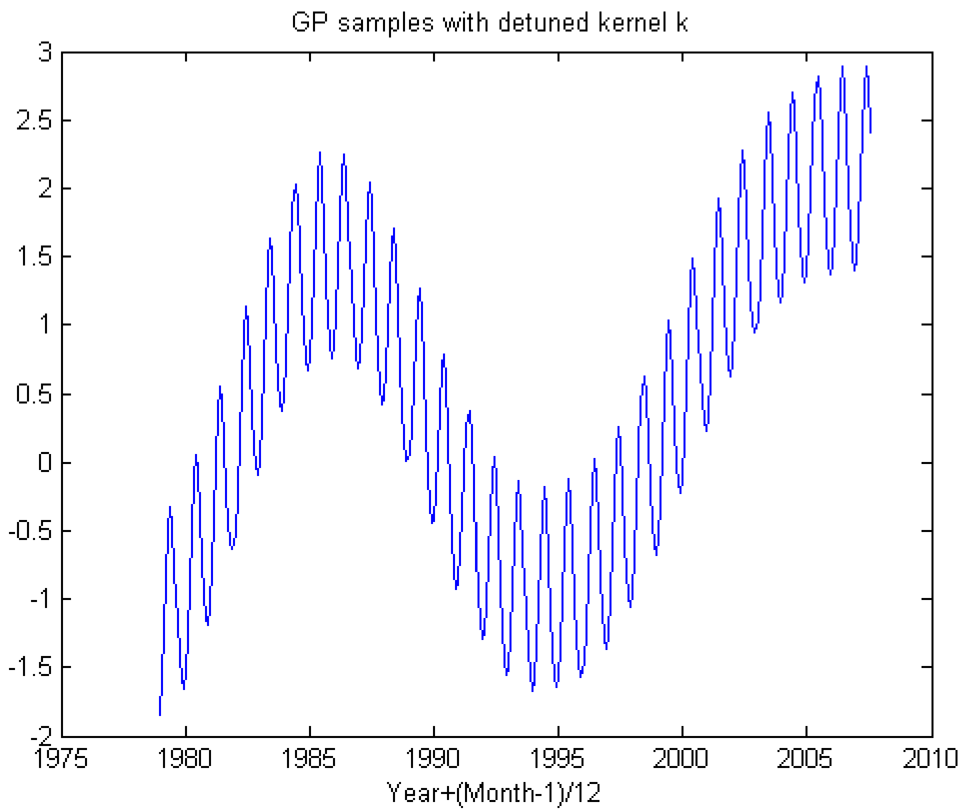

GP samples with kernel k (tuned parameters)
1.5
1
0.5
0
-0.5
-1
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12

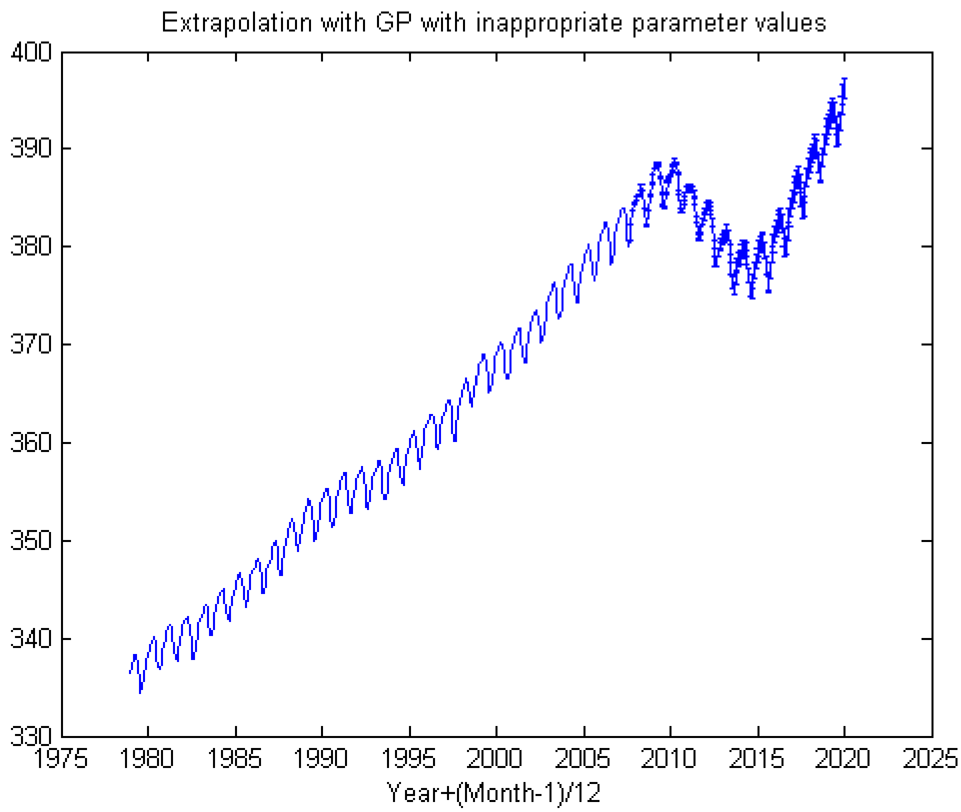


Extrapolation with GP
410
400
390
380
370
360
350
340
330
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
Year+(Month-1)/12



Predicted f(t) with one standard deviation error bars
405
400
395
390
385
380
2006
2008
2010
2012
2014
2016
2018
2020
2022
Year+(Month-1)/12


GP samples with detuned kernel k
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
1975
1980
1985
1990
1995
2000
2005
2010
Year+(Month-1)/12


Extrapolation with GP with inappropriate parameter values
400
390
380
370
360
350
340
330
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
Year+(Month-1)/12