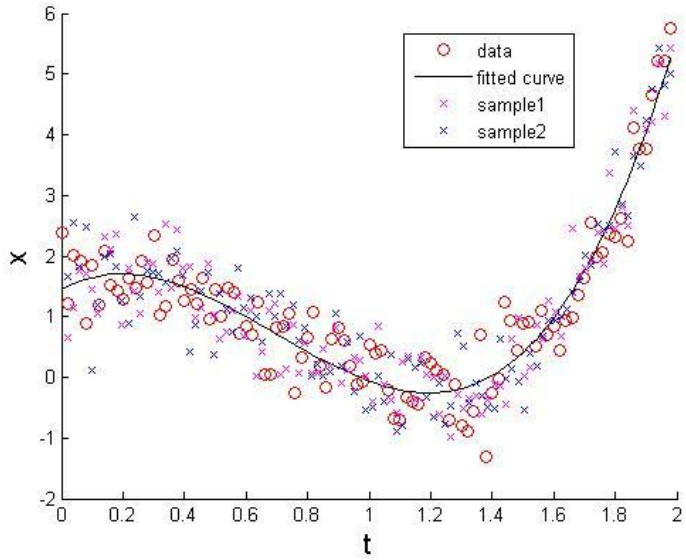
CO907 Quantifying uncertainty and correlation in complex systems

Problem sheet 2

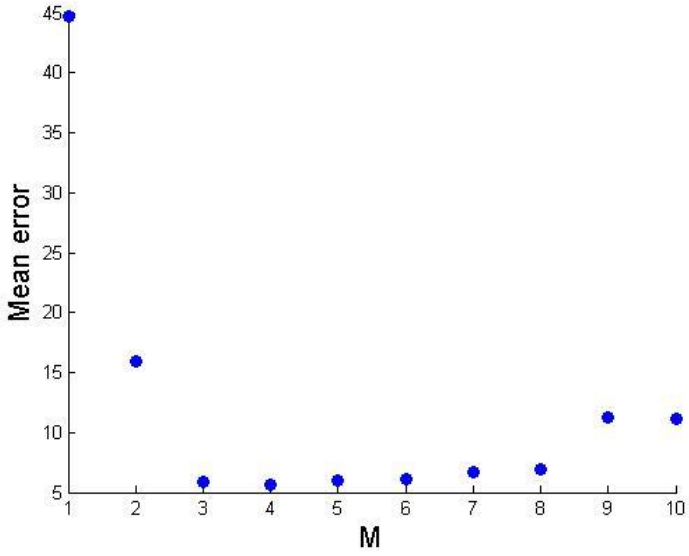
2.1

(a) M=3 (cubic curve)

$\omega = \{3.92\ -8.15\ 2.72\ 1.45\}$

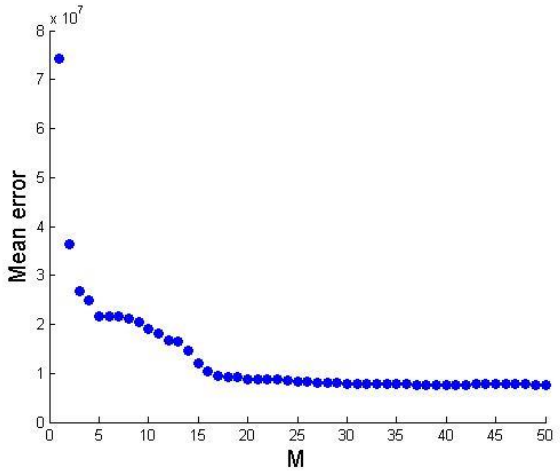


(b) Using cross validation we can prove that M=4 is the best value of M.

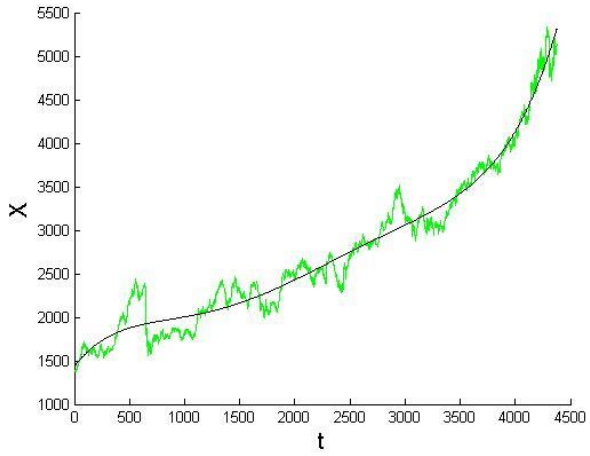


(c) First, using cross – validation, trying M from 1 to 100, choose parameter M to fit the data.

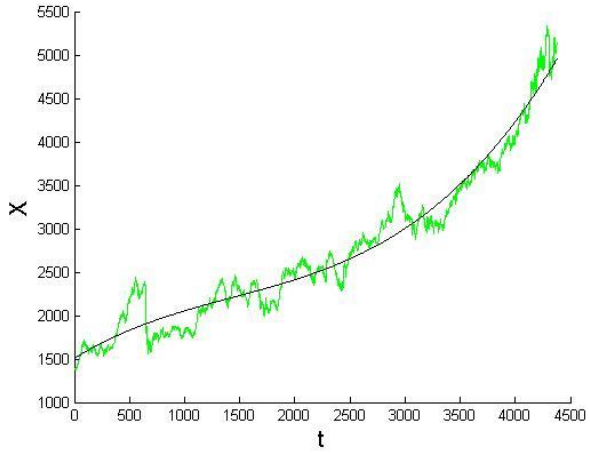
1. FTSE data



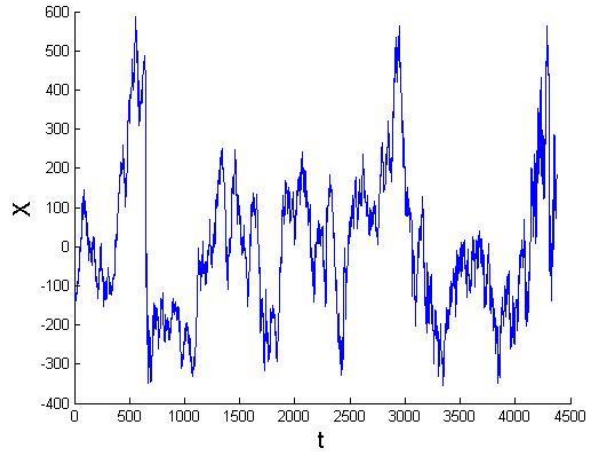
Algorithm suggests using M = 50, but as we can see from the plot above, the difference between mean errors is negligible from M=15 and the biggest 'jump' mean error does from M=1 to M=5.



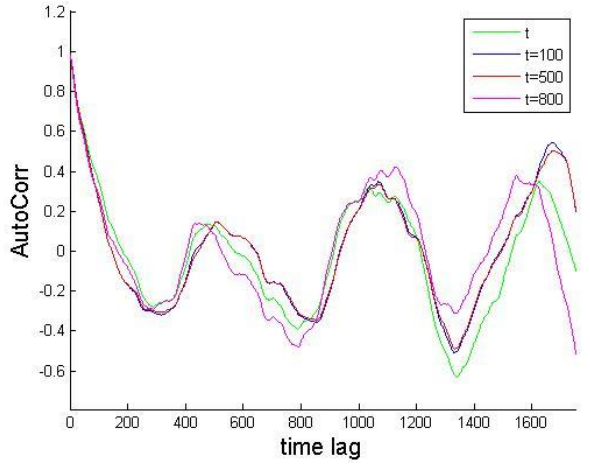
Using common sense, it seems that to fit the data polynomial of 3rd degree is sufficient.



Then, detrend the data:

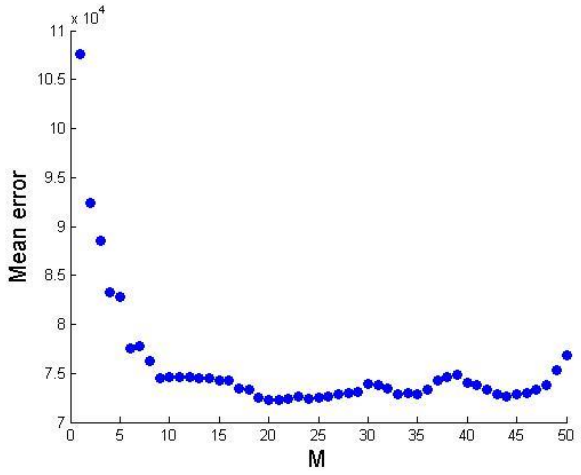


To check stationarity of the detrended data, plot autocorrelation function truncating the data:

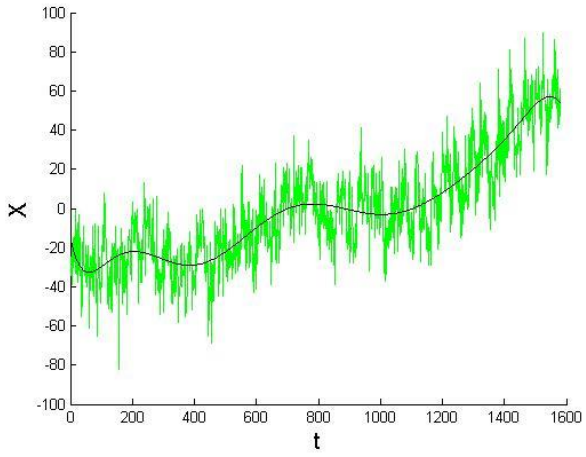


We can see, that autocorrelation functions don’t really match, so this process doesn’t look stationary.

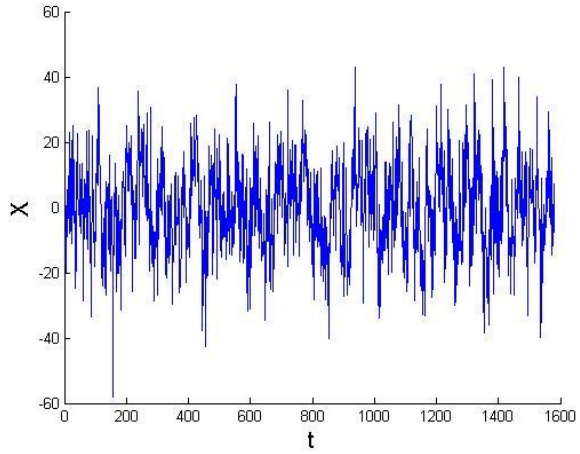
3. Temperature anomaly data



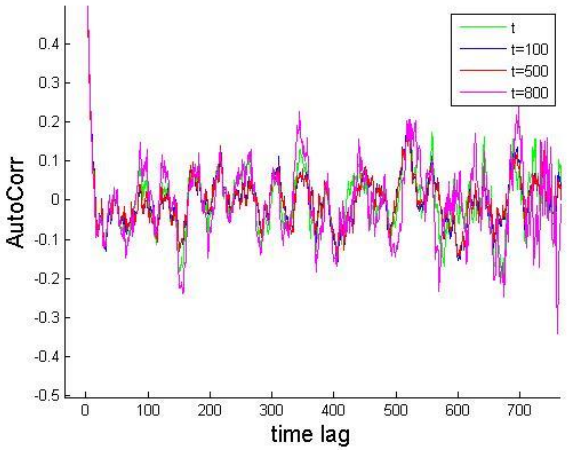
Algorithm suggests using M = 20, but we see that after M=9 mean error stops decaying rapidly.



Then detrend the data:



To check stationarity of the detrended data, plot autocorrelation functions truncating the detrended data:

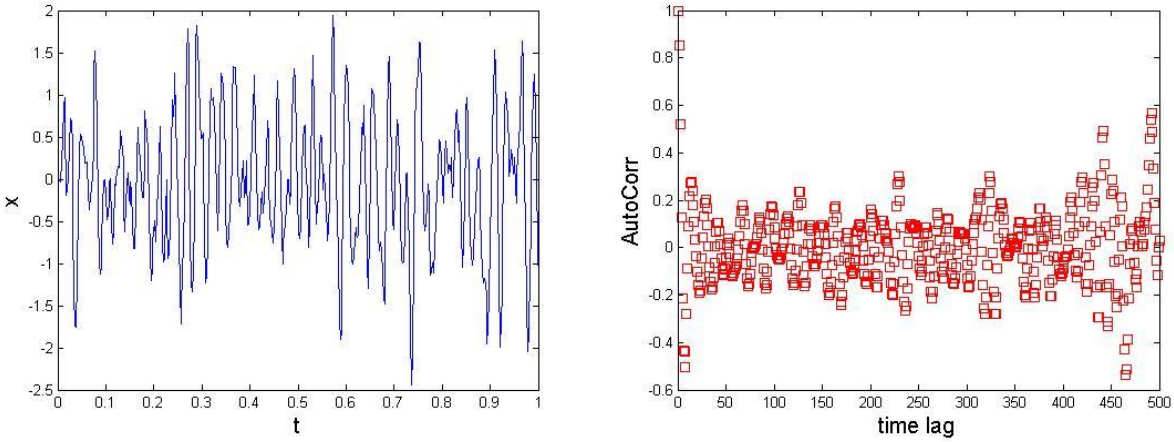


This process seems stationary from the plot we've got.

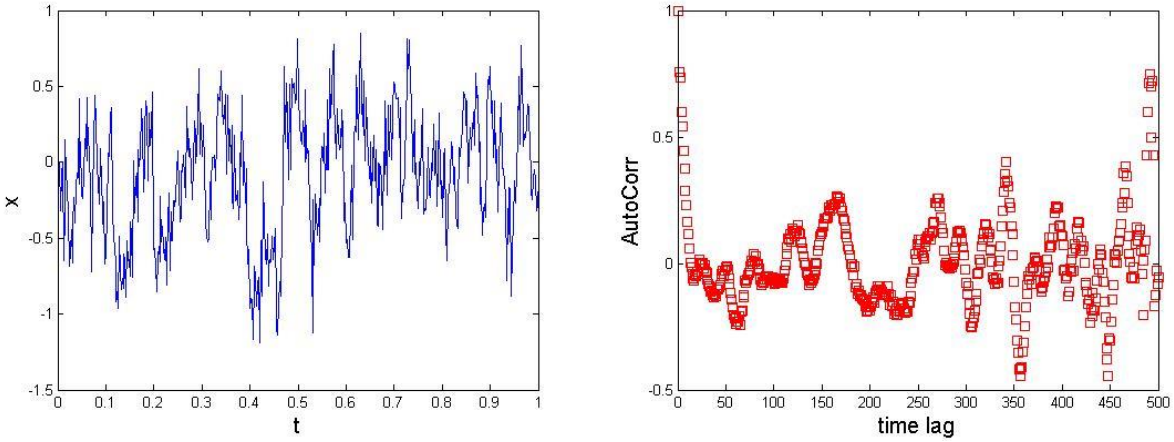
2.2

(a) $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \xi_t$, where $\xi_t \sim N(0, \sigma^2)$ and c, ϕ_1, ϕ_2 are real-valued parameters. Left plot represent sample, right plot shows autocorrelation function

Process 1. $X_0 = X_1 = 0, c = 0, \phi_1 = 3/2, \phi_2 = -3/4, \sigma^2 = 1/4$



Process 2. $X_0 = X_1 = 0, c = 0, \phi_1 = 1/2, \phi_2 = 1/3, \sigma^2 = 1/4$



(b) $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \xi_t$

Consider now writing an equation for each observation:

$X_2 = c + \phi_1 X_1 + \phi_2 X_0 + \xi_2$

$X_3 = c + \phi_1 X_2 + \phi_2 X_1 + \xi_3$

...

$X_n = c + \phi_1 X_{n-1} + \phi_2 X_{n-2} + \xi_n$

$$\begin{bmatrix} X_2 \\ . \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_0 \\ . & . & . \\ 1 & X_{n-1} & X_{n-2} \end{bmatrix} \begin{bmatrix} c \\ \phi_1 \\ \phi_2 \end{bmatrix} + \begin{bmatrix} \xi_2 \\ . \\ \xi_n \end{bmatrix}$$

Design matrix:

$$\Phi = \begin{pmatrix} 1 & X_1 & X_0 \\ . & . & . \\ 1 & X_{n-1} & X_{n-2} \end{pmatrix}$$

Vector of parameters $\mathbf{f}$, vector of error terms ξ: $\mathbf{f} = \begin{bmatrix} c \\ \phi_1 \\ \phi_2 \end{bmatrix}, \xi = \begin{bmatrix} \xi_2 \\ . \\ \xi_n \end{bmatrix}$

$$\mathbf{X} = \mathbf{\Phi f} + \boldsymbol{\xi}$$

We have to minimize sum of squared residuals:

$$\frac{1}{2}\sum_{t=2}^{n}(X_t - c - \phi_1 X_{t-1} - \phi_2 X_{t-2})^2 = (\mathbf{X} - \mathbf{\Phi f})^2 = (\mathbf{X} - \mathbf{\Phi f})^T (\mathbf{X} - \mathbf{\Phi f})$$

$$\frac{\partial}{\partial \mathbf{f}} (\mathbf{X} - \mathbf{\Phi f})^2 = -2\mathbf{\Phi}^T (\mathbf{X} - \mathbf{\Phi f})$$

$$\mathbf{\Phi}^T (\mathbf{X} - \mathbf{\Phi f}) = 0$$

$$\mathbf{\Phi}^T \mathbf{X} = \mathbf{\Phi}^T \mathbf{\Phi f}$$

$\mathbf{f} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{X}$ - formal solution

$$\mathbf{\Phi}^T \mathbf{\Phi} = \begin{pmatrix} 1 & . & 1 \\ X_1 & . & X_{n-1} \\ X_0 & . & X_{n-2} \end{pmatrix} \begin{pmatrix} 1 & X_1 & X_0 \\ . & . & . \\ 1 & X_{n-1} & X_{n-2} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n-1} X_i & \sum_{i=0}^{n-2} X_i \\ \sum_{i=1}^{n-1} X_i & \sum_{i=1}^{n-1} X_i^2 & \sum_{i=1}^{n-1} X_i X_{i-1} \\ \sum_{i=0}^{n-2} X_i & \sum_{i=1}^{n-1} X_i X_{i-1} & \sum_{i=0}^{n-2} X_i^2 \end{pmatrix}$$

(c)

a) Numerical implementation of formal solution gives the following parameters in average from 100 realizations of the process (1): $c = 0.0001, \phi_1 = 1.4962, \phi_2 = -0.7499$; variance $\sigma^2 = 0.5345$. Numerical implementation of formal solution gives the following parameters in average from 100 realizations of the process (2): $c = 0.0002, \phi_1 = 0.4991, \phi_2 = 0.3207$; variance $\sigma^2 = 0.1531$.

This fit is very good, because obtained parameters are very close to those, used for generating the data in both process (1) and process (2).

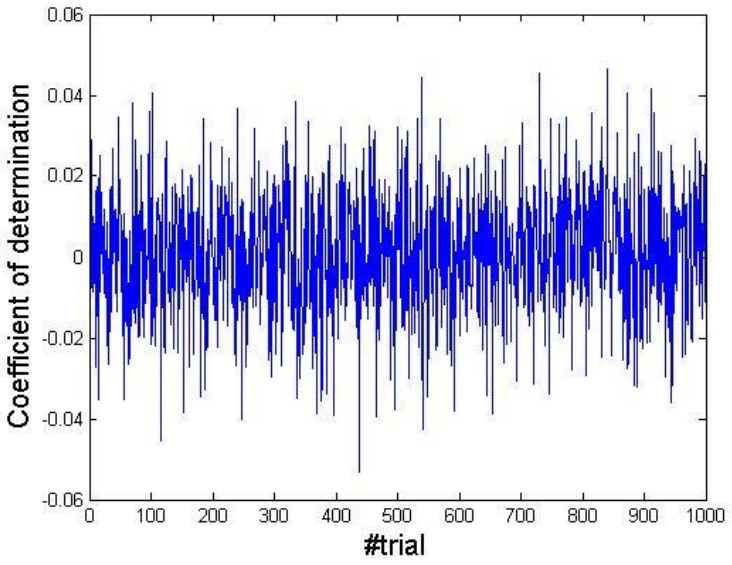
(d)

1. FTSE

Trying to fir AR(2) model to FTSE data, we obtain following parameters:

$c = 0.182, \phi_1 = 1.059, \phi_2 = -0.06$; variance $\sigma^2 = 429.127$

In order to check goodness of fit we generate a sample using parameters from above 1000 times and each time calculate coefficient of determination, then plot coefficients for each try.



From this we can conclude that FTSE data cannot be fitted correctly with AR(2).

2.3

(a) $X_t = A\cos(2\pi\omega t + \phi) + \xi_t = A(\cos\phi\cos 2\pi\omega t - \sin\phi\sin 2\pi\omega t) + \xi_t =$

$= (A\cos\phi)\cos(2\pi\omega t) + (-A\sin\phi)\sin(2\pi\omega t) + \xi_t = B_1\cos(2\pi\omega t) + B_2\sin(2\pi\omega t) + \xi_t$

(b) Using *Curve Fitting Toolbox* in MATLAB:

Coefficients (with 95% confidence bounds):

B_1 = -0.63 (-1.07, -0.19)

B_2 = -2.155 (-2.59, -1.72)

$$\begin{cases} A\cos\phi = B_1 \\ -A\sin\phi = B_2 \end{cases}$$

$$\begin{cases} tg\phi = -\dfrac{B_2}{B_1} \\ B_2^2 + B_1^2 = A^2 \end{cases}$$

$$\phi = \arctan(-\frac{B_2}{B_1}) = \arctan(-3.42) \approx -1.29$$

$$A = \sqrt{B_2^2 + B_1^2} \approx 2.25$$

(c) Goodness of fit:

Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2 = 25040$. A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction.

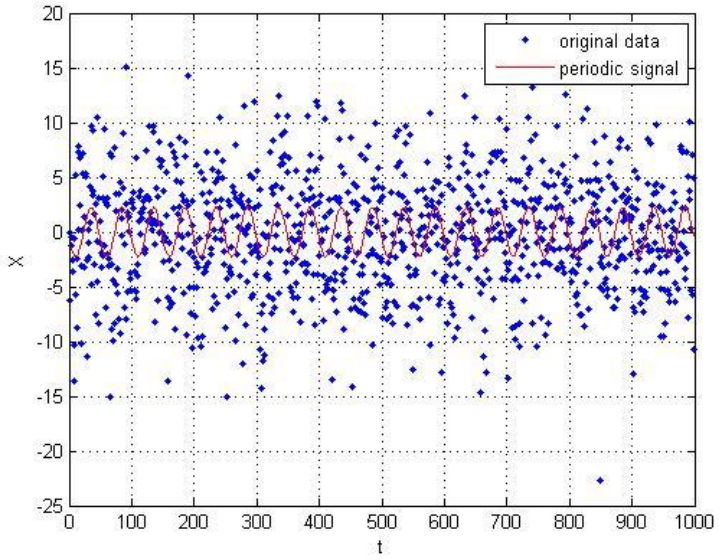
$$\sigma^2 = \frac{RSS}{n} = 25.040$$

Coefficient of determination: $R^2 = 1 - \dfrac{RSS}{TSS} = 0.0907$, $TSS = \sum_{i=1}^{n}(x_i - \hat{\mu})^2$, $\hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. For example, an R-square value of 0.0907 means that the fit explains 9.07% of the total variation in the data about the average.

Adjusted coefficient of determination: $\bar{R}^2 = 1 - \frac{RSS}{TSS}\frac{n-1}{n-m} = 0.0898$, m is a number of fitted coefficients. The adjusted R-square statistic can take on any value less than or equal to 1, with a value closer to 1 indicating a better fit.

Root Mean Squared Error: $RMSE = \sqrt{\frac{RSS}{n-m}} = 5.009$



As we can see linear regression provided not a ‘good fit’, because of high level of noise.