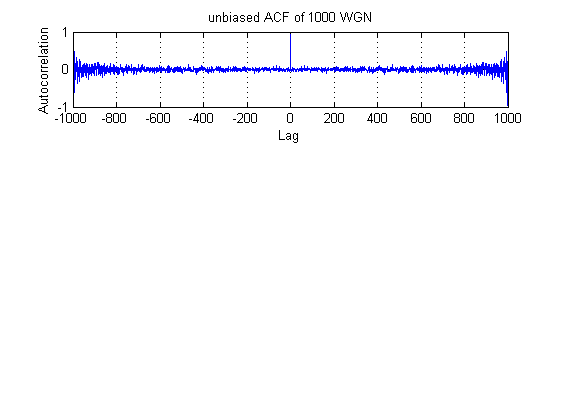
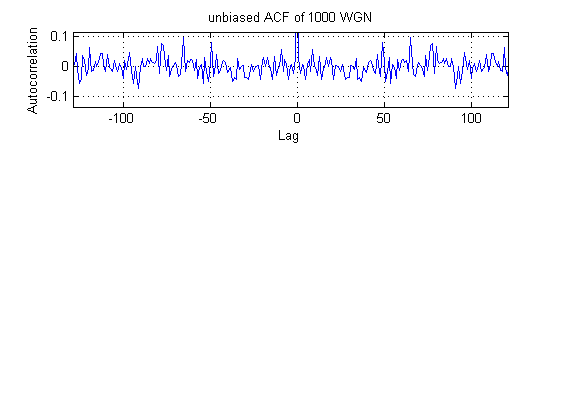
**2 Linear stochastic modelling**

**2.1 ACF of uncorrelated sequence**

**2.1.1**

**Figure 1.** Unbiased autocorrelation function for a 1000-sample realisation of white Gaussian noise

It is observed that the ACF with lag zero has the highest peak in diagram since for a random signal, sample taken each time is expected to be uncorrelated with samples taken at different time. Note that ACF here are discrete values calculated from the function: . Also, figure 1 shows obvious symmetric property of autocorrelation due to the fact that when calculating autocorrelation at - or lags, essentially same samples are used to do calculation.

**2.1.2**Figure 2 is a zoomed version to show the ACF when. The magnitude of ACF is almost always within -0.1 to 0.1. Comparing the figure1, autocorrelation magnitude gradually increases as absolute values of lags increase. This is due to the discrete ACF: . For higher lags, there are less samples available to calculate autocorrelation, leading to larger variance. Regarding small lags, sample number is large. Value calculated cancel out each other and approaches zero in average.

**Figure 2.** Zoomed version of figure 1

**2.1.3**

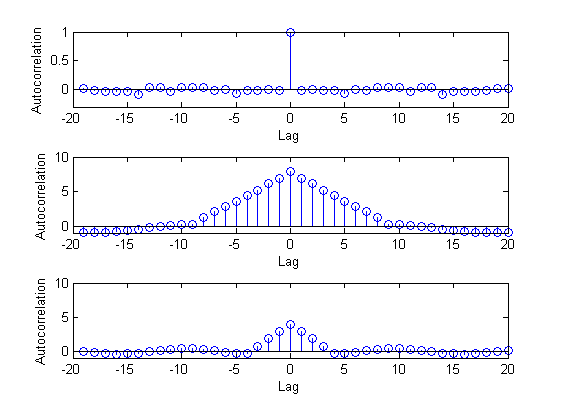
As is explained in 2.1.1 and 2.1.2, a large autocorrelation lag results in larger variance in general. Considering . Total number of samples used to calculate autocorrelation is . For small lags, more samples available and autocorrelation values cancel out each other and approaches zero in average. For larger lags, however, less samples are available, leading to larger variation of average result. Data at high lag numbers are therefore less reliable.

The autocorrelation value is still statistically reliable. If repeated experiments are carried out and results are taken average, correlation at high order is expected to approach zero as well. Note that at the extreme case, where only a pair of sample is available. Variance diverges but is still bounded on pdf of random sample. In our example where , maximum variance of autocorrelation is calculated as the following:.

In figure1, variance of autocorrelation is observed to increase gradually as lag number changes. It is reasonable to set the empirical boundary as a ratio of total sample number. Here, I select half of total sample number as an empirical boundary because for lags within the boundary, variance is not noticeable compared to large lags. Also the empirical boundary is easy to apply to other sample number, even in the case that sample number is small.

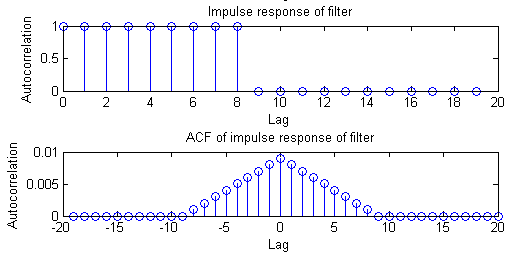
**2.2 ACF of correlated sequences**

**2.1.1**



**Figure 3.** (Top) ACF after 9th order filter, (Bottom) ACF after 4th order filter

Figure 3 shows the ACF of 1000-sample WGN after moving average filters. A triangular shape is seen in both filters showing a linear relation with negative gradient when lag changes from 0 to N (N is order number). For a moving average filter implemented above, , each value is calculated by averaging current and past N-1 samples. The smaller the lag number, the more overlapped samples used in averaging. If lag is larger than order number, outputs are no longer correlated to each other since there is no sample used as filter input. Hence the size of this triange is positively correlated to filter order.



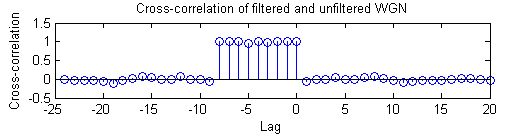
**Figure 4.** (Top) impulse response of 9th order filter, (Bottom) ACF of the 9th order filter

Here I plotted the impulse response of the filter which is a rectangular function and its ACF, which is a perfect triangle, whose size depends on filter order. As a result, ACF of filtered WGN is caused and dominated by the filter response with some noise added on. Theory is explained in the latter section. The filter function is essentially calculating the local average of the input Gaussian noise.

**2.2.2**

The autocorrelation of filtered signal is given in the equation: . where RX is the autocorrelation of the input, RX is the autocorrelation of the impulse response. Ry here represents the ACF of the filtered signal. In our example, Rx is shown in Figure 1, Rh is shown in Figure 4. The convolution of these two functions is plotted in figure 3, which agrees with our prediction. Rx is composed of an impulse at zero lag and noise in the rest part. After convoluting with the triangular function, result is a triangular with noise components added on.

**2.3 Cross-correlation function**

**2.3.1**

**Figure 5.** MATLAB plot ofCCF of filtered and unfiltered WGN

It is observed from Figure 5 that the cross-correlation of signal before and after filter is close to a square wave with width N-1, N is order number 9 in this case.

Considering the cross-correlation equation: . Ry is expected to the convolution between impulse response, which is a rectangular function and ACF of WGN, which is composed of an impulse at time zero and noise. The resultant is expected to be a rectangular function with noise, which is proved in Figure 5.

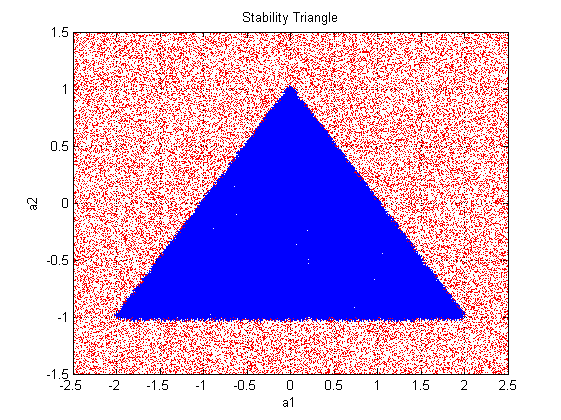
**2.3.2**

After passing WGN through a LTI system and calculated the cross-correlation of the output signal. System response can be approximated through the equation: . RXY is a good approximation of system impulse response if the variance of WGN is in a reasonable range.

Hence, filter order determines the effective range of CFF, which shows the shape of impulse response. In our example, larger filter order simply leads to the larger width of impulse response. After convolution, this effect is shown as extending the rectangular shape along the x-axis with the direction to negative x.

**2.4 Autoregressive modelling**

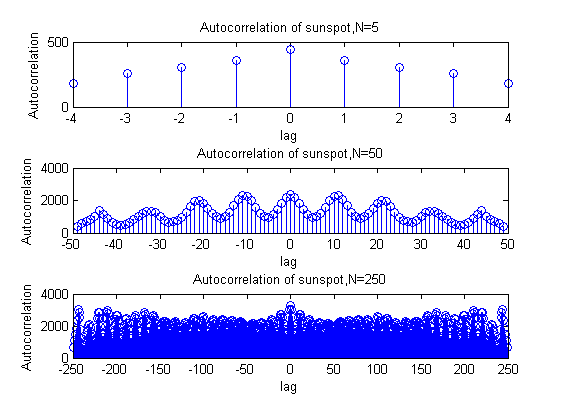
**2.4.1**



**Figure 6. (**Blue) variable pairs preserve stability, (Red) variable pairs cause non-stability

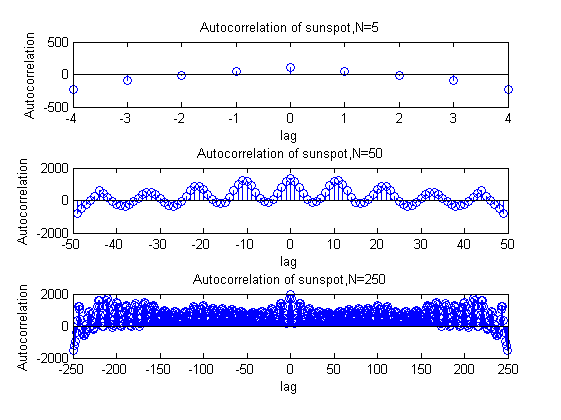
Figure 6 shows the MATLAB plot of stable and unstable variable pairs from 10000 samples, which gives a clear estimate than 100 samples. Resultant stable variable pairs form a triangle shape.

For an AR(2) process: , stationary requires a1 and a2 to satisfy the following conditions:, which is proved in the following. In z-domain, . Due to stability, the root should be inside the unit circle, apply this condition to gives us the constraints of filter coefficients.

**2.4.2**

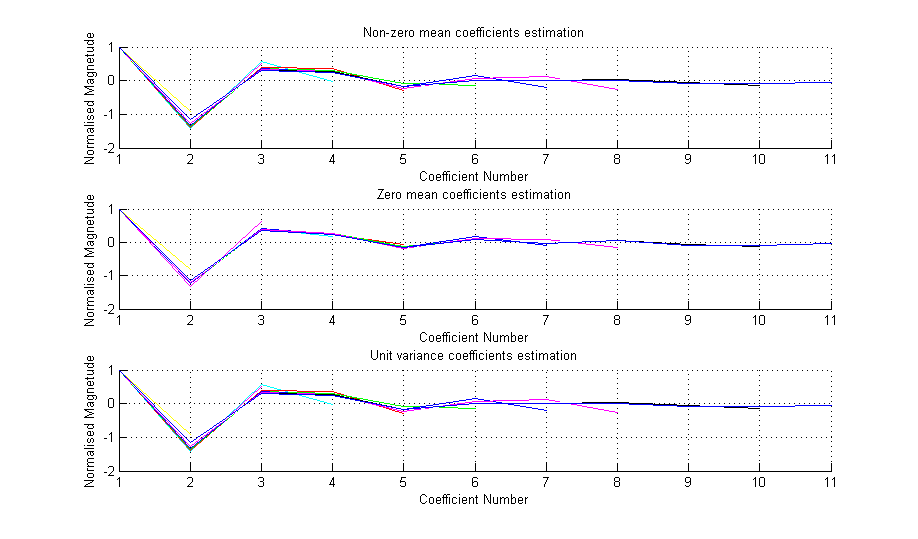
**Figure 7.** ACF of sunspots for N= 5, 50, 250 years

Figure 7 is the ACF of sunspots data for various periods. When N=5, some correlation is shown, but cannot summarised due to limited sample number. When N =50, data number is sufficient to illustrate the periodicity of 13 years of sunspots. When N=250, clear periodicity is still observed after zooming in. However, as year lag increases, variance of ACF rises and becoming less variable.

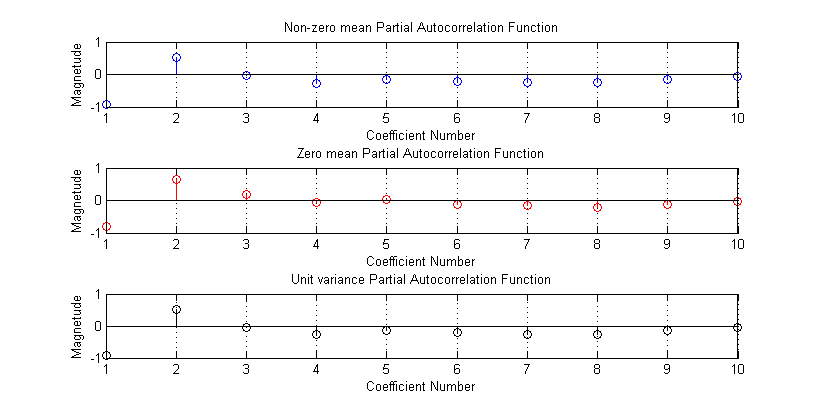


**Figure 8.** Zero-meanACF of sunspots for N=5, 50, 250 years

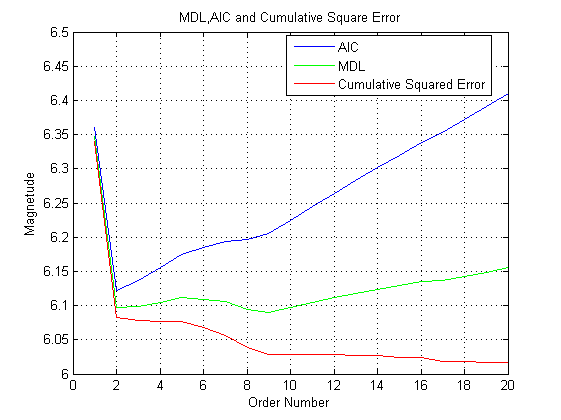
Figure 8 shows the zero-mean autocorrelation of sunspots. It is observed that, in general, the amplitude of ACF decreases or even become negative as the input sample is centred at zero after subtracting mean. In this case ACF is now becoming auto-covariance function, which shows a clear correlation between samples. Changing in ACF shape will affects the coefficients calculated using Yule-walker equations.

**2.4.3**

**Figure 9.** (Top)Non-zero meancoefficients plot. (Middle) Zero meancoefficients plot. (Bottom) Unit variancecoefficients plot.

In figure 9, coefficient estimations of Yule-Walker model using non-zero mean, zero mean and unit variance input are shown on graph. Estimated coefficients are different for zero and non-zero mean input. Intuitively, solving the Yule-walker equations with different correlation values as input should lead to different results. While variance does not affect solution to YW equations

**Figure 10.** (Top)Non-zero meanPAC. (Middle) Zero meanPAC. (Bottom) Unit variancePAC.

****Partial autocorrelation function using Yule-Walker equation is plotted in figure 10. PACs of zero-mean and unit variance input are same, while zero-mean partial autocorrelation is different. This is caused by the same reason that subtracting mean from data changes the results of Yule-Walker equations. It is observed that partial autocorrelation of first three lags have large magnitudes, while other peaks converges. As a result, AR (2) or AR (3) is a good model for sunspots.

**2.4.4**

**Figure 10.** MDL, AIC and Cumulative Square Error

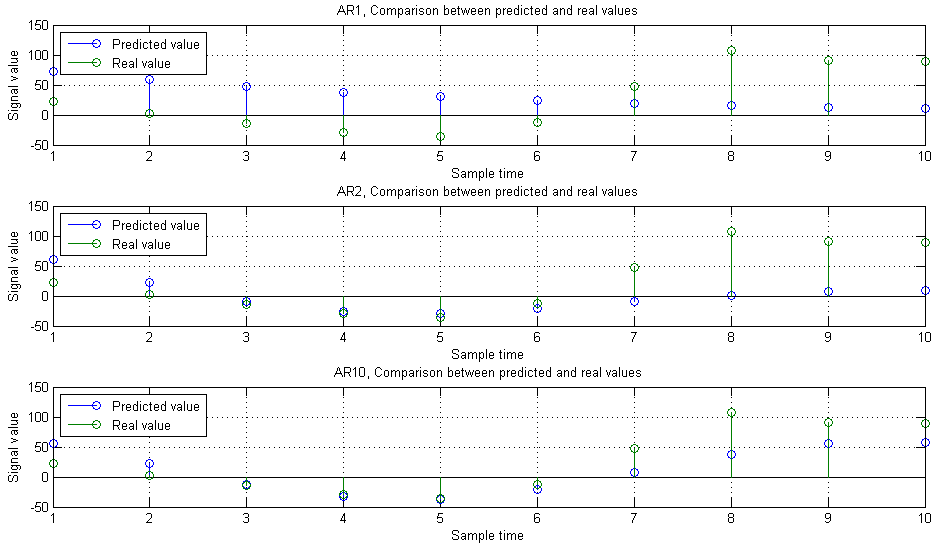
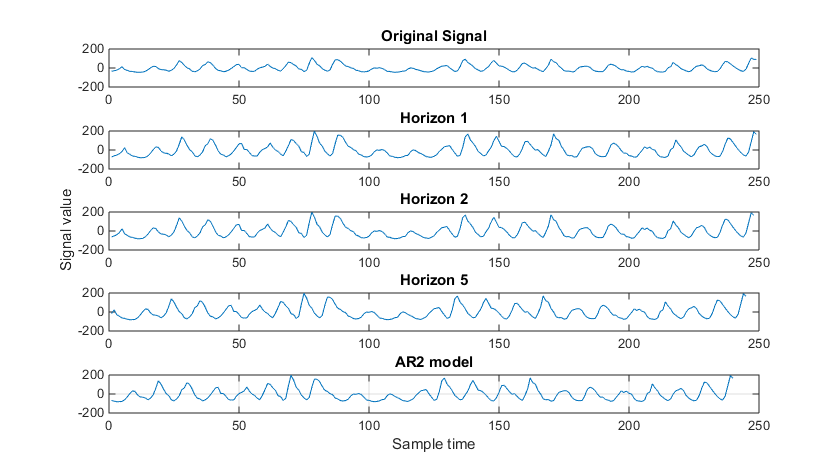
MDL and AIC are plotted using equations:

,

According to Figure 10, the best model order is the minimum value of MDL and AIC, which is three in this case. So considering both accuracy and complexity, AR (9) is the best model. AR(2) is also a reasonably good option.

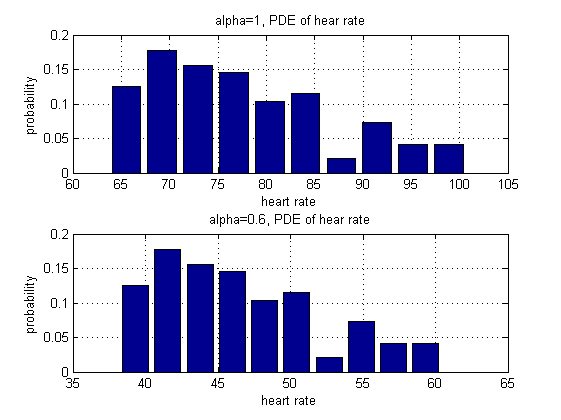
**2.4.5**

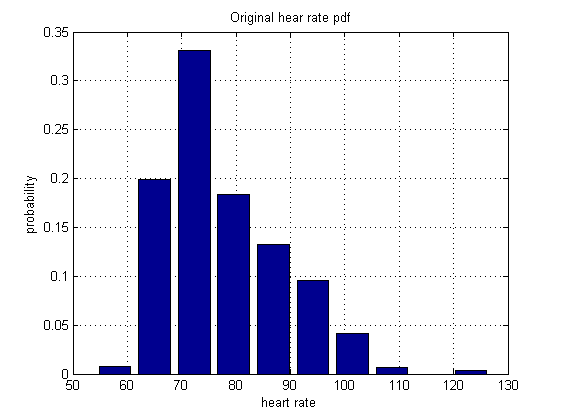
Figure 11 shows the comparison between predicted values and real values using the past 250 data for various model order. 10 future values are predicted. In general, estimating further future time leads to larger inaccuracy. Regarding the relation between model order number and estimation accuracy, larger order number usually means better estimation of future data, at the cost of complexity. However, it is observed that AR2 model is sufficient to show the general shape of the curve. If we compare the coefficients generated using AR2 and AR10, they are pretty much similar. As the order number becomes higher, the difference between coefficients are smaller, so the improvement of result is tiny compared to cost. As a result, AR2 is a suitable model here.

From the formula, ,that effect of increasing prediction horizon is same as shifting the prediction with horizon 1 to left. This change only makes the prediction worse. Figure 12 shows the prediction of AR2 with different horizons.

**Figure 12.** AR2 comparison with prediction horizon 1, 2,5,10

**Figure 11.** Prediction comparison of AR 1, 2 and 10

**2.5.1**

****

**Figure 13.** Averaged heart rate pdf plot

**Figure 12.** Original heart rate pdf plot

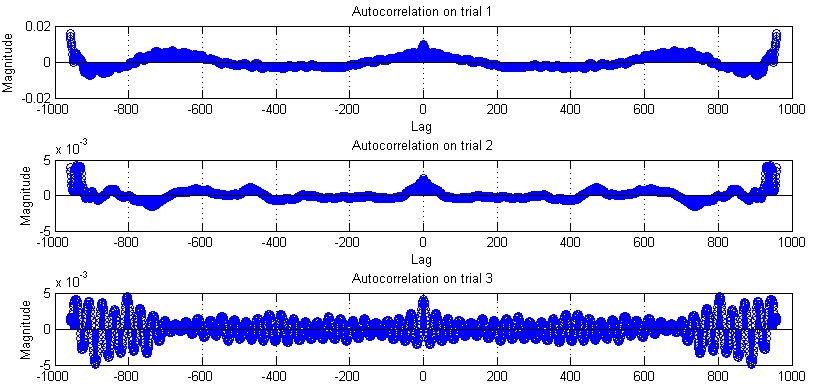
Figure 12 and Figure 13 show the PDE of the original hear rate and also the averaged heart rate.

**2.5.2**

PDE of the original heart rate is centralised around 70 but have some outliers. However, after averaging, most data are distributed in a small range between 65 and 100. Therefore, centralised data has smaller average and it is a better performance indicator over time.

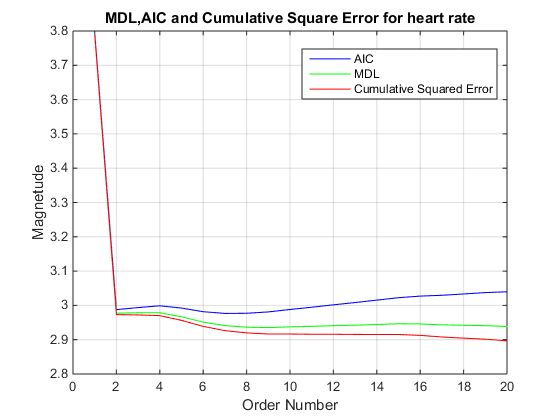
The constant alpha behaves as the offset of heart rate. It does not affect the shape of heart rate distribution but it shifts the whole graph horizontally. Also, it affect the range of the averaged distribution since the difference between heart rate samples will be either reduced or increased depending on alpha value.

**2.5.3**

****Figure 14 next page shows the autocorrelation function of three trials. It is observed that autocorrelation function tails off gradually. As a result, an autoregressive model is suggested. Since if data satisfy MA model, autocorrelation function should suddenly cut off after a certain lag.

**2.5.4**

**Figure 14.** Averaged heart rate pdf plot

****Figure 15 shows the trade-off between model accuracy and complexity. It indicates that autoregressive model can be used to describe heart rate. AR2 model is the most efficient model for trial two. Similarly, graph generated by trial three also shows the optimal order is 2.

**Figure 15.** MDL, AIC and Cumulative Squared Error of heart rate trial two