Quantifying uncertainty and correlation in complex systems

Problem sheet 2

(counts $30/50$ homework marks)

2.1 (For credit) Curve fitting and model selection

Download the file `curvefitting2013.dat` from the module website, containing a noisy signal (t, X_t). We propose the following generative model for the data:

$$X_t = \langle \boldsymbol{w} | \boldsymbol{\phi}(t) \rangle + \xi_t \quad \text{where} \quad \xi_t \sim N(0, \sigma^2) \text{ iid} ,$$

$\langle \boldsymbol{\phi}(t)| = (1, t, \ldots, t^{M-1})$ are polynomial basis functions and $\langle \boldsymbol{w}| = (w_0, \ldots, w_{M-1})$ are real-valued parameters.

(a) Let us first fix $M = 4$ (i.e. we wish to fit a cubic curve to the data). Numerically compute the Maximum Likelihood Estimate for $\langle \boldsymbol{w}|$ and σ^2 and plot the fitted curve together with the original data. Generate two samples from the fitted model and plot them (in the same plot with different symbols) to check that you have obtained a plausible fit.

(b) What is the 'best' value of m? Use cross-validation or another model-selection technique to make this decision in a principled way.

(c) Use your code to detrend the preprocessed version of the FTSE100 data set (which you generated in Q2.4 from the file `ftse.dat` on the module website) and the global temperature anomaly data contained in the file `temperatureanomaly.dat` of the website. Comment on the number of parameters used to extract the trend and the stationarity of the resulting detrended data.
(Note: this is not as clear-cut as in the curve-fitting example due to correlations in the data). Save the detrended data for later use.

2.2 (For credit) Autoregressive models

An autoregressive model of degree q (AR(q) for short) is given by

$$X_t = c + \phi_1 X_{t-1} + \ldots + \phi_q X_{t-q} + \xi_t \quad \text{where} \quad \xi_t \sim N(0, \sigma^2) \text{ iid} ,$$

and $c, \phi_1, \ldots, \phi_q$ are real-valued parameters. In the following we focus on $q = 2$.

(a) Generate and plot a sample of length $N = 500$ of the AR(2) process with $X_0 = X_1 = 0$ and $c = 0$, $\phi_1 = 3/2$, $\phi_2 = -3/4$ and $\sigma^2 = 1/4$. Calculate and plot the sample autocorrelation function.
Generate and plot a sample of length $N = 500$ of the AR(2) process with $X_0 = X_1 = 0$ and $c = 0$, $\phi_1 = 1/2$, $\phi_2 = 1/3$ and $\sigma^2 = 1/4$. Calculate and plot the sample autocorrelation function and compare with the one obtained above.

(b) Assuming that the first two values of an AR(2) series are fixed, explain how to use the remaining data to set up a regression problem for the parameters c, ϕ_1 and ϕ_2 . Write down the $(i,j)^{\text{th}}$ element of the design matrix for this regression, write the regression problem as a set of linear equations and write down its formal solution.

(c) Solve this set of linear equations numerically for the datasets generated in (a) to give estimates of the parameters.
Can you give an estimate of your uncertainty in the values of the parameters you have obtained?

(d) Do you think any of the real-world time-series which you studied in Q2.4 could be modeled with an AR(2) process? Try using your code to fit an AR(2) model to one of these series and comment on the results.

2.3 (For credit) Extracting a signal from noise

Download the file `signalinnoise.dat` from the module website. This timeseries was generated by the model:

$$X_t = A\,\cos(2\pi\omega t + \phi) + \xi_t \quad \text{where} \quad \xi_t \sim N(0,\sigma^2) \text{ iid}\,,$$

and A, ϕ are real-valued parameters. The value of σ is sufficiently large that the periodic component of the signal is mostly buried in the noise. Suppose we know $\omega = 1/50$. We can use regression to extract the amplitude and phase of the periodic signal.

(a) Use a trigonometric identity to write the model in regression form:

$$X_t = B_1\,\cos(2\pi\omega t) + B_2\,\sin(2\pi\omega t) + \xi_t\,,$$

where B_1 and B_2 are functions of A and ϕ which you need to find.

(b) Using linear regression on the data, find the values of B_1 and B_2 and hence the values of A and ϕ.

(c) Plot the periodic signal on the same graph as the original data. Does your result look plausible?

2.4 (Not for credit) Plotting and preprocessing timeseries data

Download the files `ftse.dat`, `xray.dat` and `temperatureanomaly.dat` from the module website. These contain timeseries data of the value of the FTSE100 index, the x-ray flux from an astronomical source and the monthly global average temperature anomaly. Details of the data are contained in comment lines at the top of the individual files.

(a) Some of these timeseries contain gaps. Devise a reasonable method to deal with these gaps. Explain your choice and implement it to pre-process the data files, and save the processed version.
Produce properly formatted timeseries plots of the three signals.

(b) Which of these timeseries look stationary? Comment qualitatively on any features of the data which might be worth investigating.

2.5 (Not for credit) Simple generative models of timeseries data

(a) A Gaussian white noise process with variance σ^2 is a timeseries of iid $N(0,\sigma^2)$ random variables. Write a code to generate samples of length N from this process, and generate and plot a sample of length 200.
Plot the sample autocorrelation function (e.g. using `autocorr` in MATLAB).

(b) Generate data $(X_0, \ldots, X_N)$ for $N = 200$ from the model

$$X_t = \sin(2\pi t/N) + \xi_t \quad \text{where} \quad \xi_t \sim N(0, 1/4) \text{ iid} .$$

Use `polyfit` and `polyval` to generate a polynomial least squares estimate up to power $M - 1$. Plot the fitted curve together with the data for $M = 4$ and 5.
Compute the training error for the original sample, and the test error for an independent test sample as a function of M. What is the best choice for M?

(c) Write a code to generate samples of autoregressive AR(1) and AR(2) processes. Explore how these models behave as the parameters are varied and plot some graphs to illustrate what you discover, including sample time series and autocorrelation functions.
Fit the parameter values, e.g. using `estimate` and the `arima` class in MATLAB.

2.6 (not for credit) Spectral analysis of real data

(a) Use the Discrete Fourier Transform to calculate the sample power spectrum of the uniformly sampled and detrended FTSE100 data, the uniformly sampled xray data, and the detrended temperature anomaly data.
Plot your results. Dont forget to label your axes (including units) and to choose appropriate axes and data ranges to bring out the interesting features in the spectra.

(b) Identify any significant peaks in the power spectra and relate them to physical timescales in approriate units. Do these timescales correspond to anything you might expect?

2.7 (not for credit) Smoothing of timeseries and low-pass filtering

Download `sunspots.dat` from the module website. It contains monthly measurements $(X_i : i = 1, \ldots, N)$ of the average daily sunspot numbers observed between 1749 and 1983.

(a) Plot the power spectrum of the data and identify the physical timescales associated with any significant peaks which you find.

(b) Let us produce a smoothed timeseries from the sunspot data as follows:

$$Y_i := \sum_{j=1}^{N} w_{ij} X_i$$

where the weights $w_{ij} > 0$ are generated from a Gaussian filter with bandwidth b:

$$w_{ij} = K\Big(\frac{i-j}{b}\Big) \Big/ \sum_{j=1}^{N} K\Big(\frac{i-j}{b}\Big) ,$$

with $\qquad K(x) = \dfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2} .$

Implement this filter (a truncation might be necessary) and filter the data with bandwidths of 2, 4, 8 and 16 months. In each case, compare the filtered data to the original.

(c) We can define the fluctuating component Z_i of the series as what is left over when the smooth part is removed, $Z_i := X_i - Y_i$.
Explore how the correlation properties of the fluctuating component of the series behaves as the filter bandwidth is changed.

(d) Plot the power spectra of the 4 smoothed series. Explain why the power spectrum behaves the way it does as the bandwidth is varied.