WARWICK
THE UNIVERSITY OF WARWICK

MA930 Data Analysis & Machine Learning

Lecture 5: Time series analysis

Haoran Ni

Recap from last week

WARWICK
THE UNIVERSITY OF WARWICK

- **Frequentist approach**
  - Parameters are fixed quantities
  - Confidence intervals
  - Hypothesis tests: Ask how likely a result at least as extreme as that seen is
  - P-values, significance
  - Probability of data given hypothesis

- **Bayesian approach**
  - Parameters characterised by probability distributions
  - Updating priors given data
  - Credible intervals

Outline

WARWICK
THE UNIVERSITY OF WARWICK

- Polynomial fitting
- Auto-regressive models
- Bayesian inference for an auto-regressive model

Preprocessing data

WARWICK
THE UNIVERSITY OF WARWICK

- Identify any bad entries in the data: NaN or similar
- Remove these, or replace with principled replacement: e.g. mean of neighbouring points, linear interpolation/extrapolation etc.
- Plot the data as a sense check
- Smooth data to remove noise (*cf.* epidemiological example from last time) and remove outliers if necessary

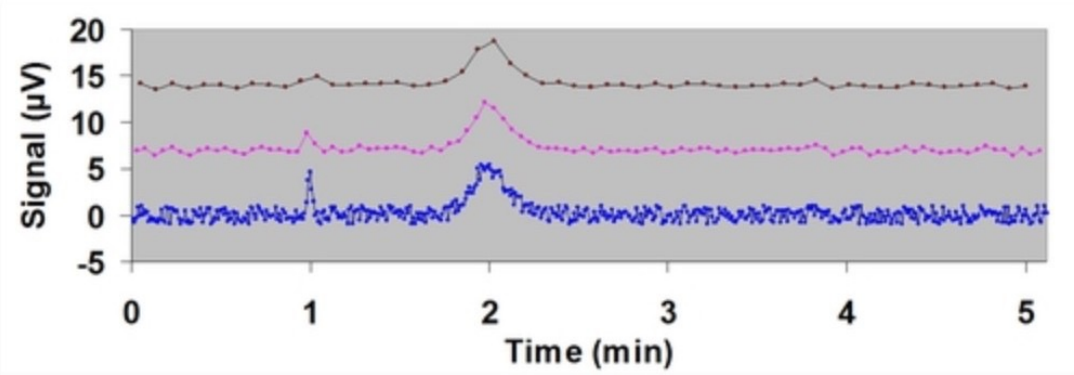
Preprocessing data

WARWICK
THE UNIVERSITY OF WARWICK

- Many different approaches exist to smooth data
- A commonly used example is *boxcar smoothing*
- Input data is a set of points $\{x_k\}$. Output is a set of (fewer) points $\{x'_k\}$
- $x'_k = \frac{1}{2n+1} \sum_{j=-n}^{n} x_{k+j}$

Preprocessing data

- Many different approaches exist to smooth data
- A commonly used example is *boxcar smoothing*
- Input data is a set of points $\{x_k\}$. Output is a set of (fewer) points $\{x'_k\}$
- $x'_k = \frac{1}{2n+1} \sum_{j=-n}^{n} x_{k+j}$



- Bottom is original data
- Middle is $n = 2$ (five-point boxcar)
- Top is $n = 4$ (nine-point boxcar)

- Remove smoothed data from original data to get at high-frequency components

Polynomial fitting

WARWICK
THE UNIVERSITY OF WARWICK

Linear regression

- Imagine *N* data points characterised by $\{t_n\}$ and $\{x_n\}$
- Want to find the best fit $f_n = k_1 t_n + k_0$ in the least-squares sense

- Define an error function $E = \frac{1}{2N}\sum_{n=1}^{N}(f_n - x_n)^2$

- What k_0 and k_1 minimise *E*? (see board)

Polynomial fitting

WARWICK
THE UNIVERSITY OF WARWICK

Linear regression

- Imagine *N* data points characterised by $\{t_n\}$ and $\{x_n\}$
- Want to find the best fit $f_n = k_1 t_n + k_0$ in the least-squares sense
- Define an error function $E = \frac{1}{2N}\sum_{n=1}^{N}(f_n - x_n)^2$
- What k_0 and k_1 minimise *E*? (see board)
- $k_0 = \frac{\langle x\rangle\langle t^2\rangle - \langle t\rangle\langle xt\rangle}{\langle t^2\rangle - \langle t\rangle^2}$, $k_1 = \frac{\langle xt\rangle - \langle x\rangle\langle t\rangle}{\langle t^2\rangle - \langle t\rangle^2}$.

Polynomial fitting

WARWICK
THE UNIVERSITY OF WARWICK

Polynomial regression

- Straightforward(ish) to generalise this to a Mth order polynomial
- Now $f_n = k_M t_n^M + \cdots + k_1 t_n + k_0$, with error still $E = \frac{1}{2N}\sum_{n=1}^{N}(f_n - x_n)^2$
- Again, minimise E with respect to each of the k_i'/s

Exercise 5.1. Polynomial regression

Verify that the resulting system is:

$$\begin{pmatrix} 1 & \langle t \rangle & \langle t^2 \rangle & \cdots & \langle t^M \rangle \\ \langle t \rangle & \langle t^2 \rangle & \cdots & & \langle t^{M+1} \rangle \\ \langle t^2 \rangle & \cdots & & & \langle t^{M+2} \rangle \\ \vdots & & & & \vdots \\ \langle t^M \rangle & \langle t^{M+1} \rangle & \langle t^{M+2} \rangle & \cdots & \langle t^{2M} \rangle \end{pmatrix} \begin{pmatrix} \kappa_0 \\ \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_M \end{pmatrix} = \begin{pmatrix} \langle x \rangle \\ \langle xt \rangle \\ \langle xt^2 \rangle \\ \vdots \\ \langle xt^M \rangle \end{pmatrix}$$

Polynomial fitting

Exercise 5.2. Code a polynomial fit

i) Generate training data: 11 points from the cubic $t(t-0.5)(t-0.8)$ evaluated at $t = 0, 0.1, \ldots, 1$ with independent Gaussian noise at each point with mean 0 and standard deviation 0.01.

ii) Generate test data: 11 more points from the same cubic, also with Gaussian noise mean 0 and standard deviation 0.01.

iii) Find least squares fits of polynomials of order $M = 1,2,\ldots,8$ to the training data, and plot the error as a function of M.

iv) Using the fitted parameters generated using the training data, find the error of the test data and plot as a function of M for the test set on the same graph as iii.

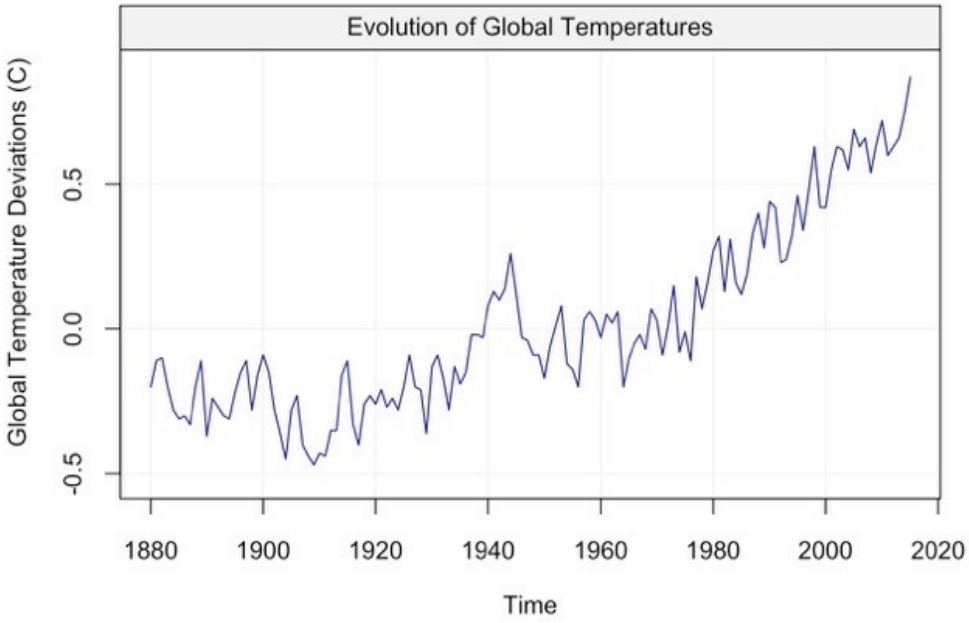
v) Generate a new plot of the polynomial fits for $M = 1,2,3$ and 4 along with the training data.

Time series

- The previous exercise highlights an important concept of time-series analysis
- A time series can often be separated into a “signal” and “noise”
- In the previous example, the signal was the underlying polynomial and normally distributed noise was added on top

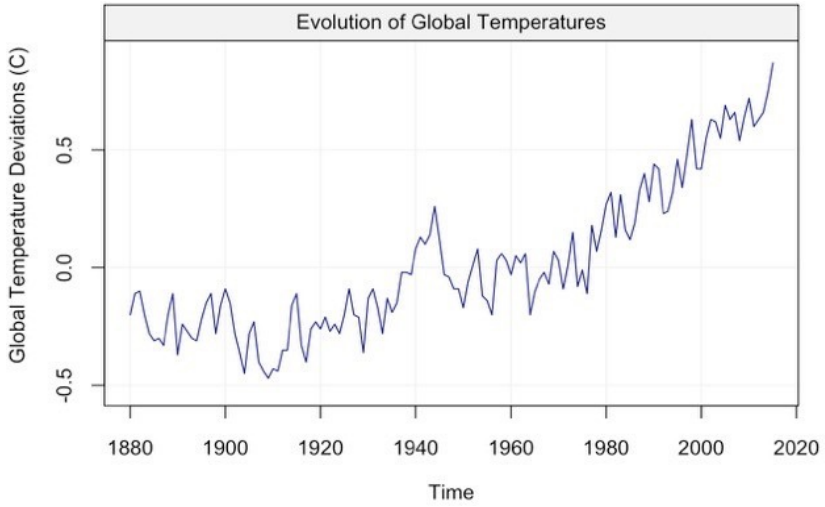
Time series

- Leaving aside technical details, Wold's decomposition theorem states that many time series can be written as: $Y_t = D_t + W_t$, where D_t is the signal (the deterministic part) and W_t is the noise (the random part)



Time series

- $Y_t = D_t + W_t$, where D_t is the signal (the deterministic part) and W_t is the noise (the random part)
- Noise terms are not usually iid. Instead, the noise in one timestep may depend on the noise in previous timesteps.
- For example, a common noise model used when modelling annual changes in mean temperature is $W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \varepsilon_t$ where the ε_t are iid Gaussians.



Autoregressive models

- Widespread use to model stochastic noise
- Model is discrete, writing current state as a function of previous ones
- An order p autoregressive model is given by: $W_t = c + \sum_{i=1}^{p} \phi_i W_{t-i} + \varepsilon_t$
- $\{\varepsilon_t\}$ are uncorrelated, normally distributed random numbers with mean 0 and variance σ_ε^2

- For a first order autoregressive model, when:
  - $\phi_t = 0$: White noise
  - $\phi_t = 1, c = 0$: Random walk
  - $\phi_t = 1, c \neq 0$: Random walk with drift

Autoregressive models

Autoregressive models are often used to represent *stationary* noise (the unconditional probability distribution for the state W_t, does not shift in time; the mean and variance are stationary). For a first order autoregressive model, this requires $|\phi_1| < 1$.

Example 5.1. First order autoregressive model

Find the mean, variance and general solution W_t, for a first order (stationary) autoregressive model, assuming the process is in steady state.

Autoregressive models

- For an autoregressive model, the "*autocovariance"* of the model is defined as

$$\gamma(k) = \text{cov}(W_t, W_{t+k})$$

- This tells us about the general relationship between the state at two timepoints that are k timepoints apart (i.e. with a *lag* of k) – i.e. tells us about temporal structure

Example 5.2. First order autoregressive model

Show that the autocovariance of the first order autoregressive model is given by

$$\gamma(k) = \phi_1^k \frac{\sigma_\varepsilon^2}{1-\phi_1^2}, \text{assuming } k > 0$$

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 5.3. First order autoregressive model
Simulate a first order autoregressive model with your parameter choice for c, $|\phi_1|$ and σ_ε (but with $|\phi_1| < 1$).

Run the simulation for 5000 steps and demonstrate that the mean and variance are comparable to the theoretical results.

Extract the autocovariance from the data, and provide a plot comparing it to the theoretical prediction.

Autoregressive models

Exercise 5.4. Sample means of a first order autoregressive model
Simulate a first order autoregressive model with $c = 0$, $\phi_1 = 0.9$ and $\sigma_\varepsilon = 2$ for 100,000 steps.

Calculate the distribution of sample means in two ways. First, take 10,000 random samples of 10 adjacent data points. Second, take 10,000 random samples of 10 data points each picked from any step of the autoregressive process.

Plot the histograms of sample means for these two ways, and see if they agree with what would be expected from a naïve application of the CLT (you can assume knowledge of the true (theoretical) mean and variance used to generate the autoregressive process). Comment on the differences (if any) between the two ways of calculating the sample means.

Exercise 5.5. Analysis of exoplanet data

Download the data file “exoplanet-data.txt”

The x-axis (first column) is in days, the y-axis (second column) is raw light intensity from the star. The dips are caused by a planet passing in front of the star (and it can be assumed that this is a single-planet system).

1) Estimate the orbital period of the planet
2) Assume that the relative reduction in light measured when the planet is blocking
3) the telescope is proportional to the area of the telescope lense blocked out.
4) Estimate the ratio of the planetary to solar radii (as visible from the telescope).
5) Use the figure to estimate the transmit time in days.

WARWICK
THE UNIVERSITY OF WARWICK

Motivation

- Consider the specific autoregressive model: $x_k = ax_{k-1} + bs_k$, where $-1 < a < 1$ and $0 < b < 2$ (for this example).
- The noise terms $\{s_k\}$ are independent Gaussians with zero mean and unit variance

- Given an observed path (a dataset), can we estimate a and b?

Bayesian inference and autoregressive models

WARWICK
THE UNIVERSITY OF WARWICK

- $x_k = ax_{k-1} + bs_k$

- Conditional on x_{k-1}, then x_k follows a normal distribution with mean ax_{k-1} and variance b^2 [Note: for any $RV\ X$, easy to show that: $\mathrm{Var}(kX) = k^2\mathrm{Var}(\mathrm{X})$].

- The distribution for x_k given x_{k-1} is therefore given by the pdf

$$f(x_k|x_{k-1}) = \frac{\exp(-\frac{1}{2}(\frac{x_k - ax_{k-1}}{b})^2)}{\sqrt{2\pi b}}$$

- Given the initial value x_1, the likelihood of a chain of points x_1, x_2, x_3, x_4 is then given by

$$L(a,b) = \Pr(x_1, x_2, x_3, x_4 | a, b) = f(x_4|x_3) f(x_3|x_2) f(x_2|x_1).$$

N.b. It is often numerically better to sum the log likelihoods and then exponentiate at the end.

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 5.6. Bayesian inference and autoregressive models
Choose values of a and b within the range for the motivatory example, and generate $N = 30$ data points starting from the initial point $x_1 = bs_1$.

Write computing code to calculate the total log-likehood for the *N* data points, conditional on a and b over their possible ranges. This will involve a sum of log-likelihoods for each step in the chain of points. At the end you can exponentiate to give the total likelihood.

Plot the density on the a-b plane and plot the marginal distributional estimates. If your result is considered as the posterior estimate of a-b, then what is the underlying assumption you have made about the prior?