Day 5 Basic time-series analysis

Recap of last week

Frequentist approach

- Consider all possible results of an experiment
- Ask how likely a result as extreme as that seen was
- P-values, significance and confidence intervals
- Probability of data given hypothesis

Bayesian approach

- Update prior views using data
- Resulting posterior can be used as future prior
- Direct evaluation of probability of hypothesis given data
- Credible intervals etc more naturally defined.

5.1 Curve-fitting

5.1.1 Preprocessing data

- Identify any "bad" entries in the data: NaN or similar
- Cut-out or replace with principled replacement
  for example: mean of neighbouring points, linear extrapolation, etc
- Plot out the data and take a good look
  do this before applying any statistical tests
- Smooth data and remove outliers if necessary

5.1.1 Preprocessing: smoothing

- Many different approaches to smoothing a data set x_k

Could be an issue b/c smoothed using data from the future (uses n forward data points)
- for financial data & similar need to smooth using only backward data points, purely causal

not super good

- Boxcar smoothing $x'_k = \underbrace{\frac{1}{2n+1}}_{\text{weight}} \sum_{j=-n}^{n} x_{k+j}$ assume there is some notion of sequentiality/continuity
- Weighted smoothing $x'_k = \sum_{j=-n}^{n} \underbrace{w_j}_{\text{weights change}} x_{k+j}$ where $\sum w_j = 1.$

like a discrete convolution

- Iterative $x'_k = x_k(1-2\alpha) + \alpha(x_{k-1} + x_{k+1})$ where α is small.
  Repeat many times as required: effectively diffusive smoothing. Useful for 2-D efficient

  so $w_0 = 1-2\alpha$
  $w_1 = w_{-1} = \alpha$
  takes $1-2\alpha$ of the original data point
  takes α of original data point and dump it either side of the original
- Remove smoothed data from original data to get at high-frequency components.

α	α	α
α	$1-8\alpha$	α
α	α	α

5.1.2 Linear regression

- Consider first trivial case of fitting a straight line.
- Imagine N data points data $\{t_n\}$ and $\{x_n\}$
- Want to find the best fit $f_n = \kappa_1 t_n + \kappa_0$ in the least-squares sense.

$(a+b+c)(a+b+c)$

$a^2+ab+ac+ab+b^2+bc+ac+bc+c^2$

$a^2+b^2+c^2+2ab+2ac+2bc$

- Define an error function $E = \frac{1}{2N}\sum_{n=1}^{N}(f_n - x_n)^2 = \frac{1}{2N}\overbrace{\sum_{n=1}^{N}(\kappa_1 t_n + \kappa_0 - x_n)^2}^{\text{minimize}}$
- **Question:** What κ_0 and κ_1 minimise E? $\frac{\partial}{\partial t}(\kappa_1 t_n + \kappa_0 - x_n)^2 = 2(\kappa_1 t_n + \kappa_0 - x_n)\cdot\kappa_1$
- **HINT** Introduce quantities like $\langle xt \rangle = \frac{1}{N}\sum_{n=1}^{N} x_n t_n$ etc

$\sum_{n=1}^{N}\left[\underbrace{\kappa_1^2 t_n^2}+\kappa_0^2+\underbrace{x_n^2}+\underbrace{2\kappa_0\kappa_1 t_n}\underbrace{-2\kappa_1 t_n x_n}\underbrace{-2\kappa_0 x_n}\right]$ $\quad \kappa_1^2\sum_{n=1}^{N} t_n^2 + \kappa_0^2 + \sum_{n=1}^{N} x_n^2 + 2\kappa_0\kappa_1\sum_{n=1}^{N} t_n$

5.1.3 Polynomial regression

Can also use to smooth data

- Straightforward to generalise to an Mth order polynomial.
- Now $f_n = \sum_{m=0}^{M}\kappa_m t_n^m$ with error still $E = \frac{1}{2N}\sum_{n=1}^{N}(f_n - x_n)^2$.
- Minimising E with respect to $\kappa_{m'}$ gives

$$\kappa_0\langle t^{m'}\rangle + \kappa_1\langle t^{m'+1}\rangle + \cdots + \kappa_M\langle t^{m'+M}\rangle = \langle xt^{m'}\rangle$$

- **Question:** Confirm this!
- There are $(M+1)$ of these equalities, so that

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

- Which can be solved numerically by multiplying by the inverse matrix.

Lin Reg

Quadratic

$$E = \frac{1}{2N} \sum_{n=1}^{N} \left(K_0 + K_1 t_n - x_n \right)^2$$

$$\frac{\partial E}{\partial K_0} = 0 = \frac{1}{2N} \cdot \sum_{n=1}^{N} 2 \left(K_0 + K_1 t_n - x_n \right) \cdot 1$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left(K_0 + K_1 t_n - x_n \right)$$

$$\frac{\partial E}{\partial K_1} = 0 = \frac{1}{2N} \sum_{n=1}^{N} 2 \left(K_0 + K_1 t_n - x_n \right) \cdot t_n$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left(K_0 + K_1 t_n - x_n \right) t_n$$

$$K_0 + K_1 \langle t \rangle = \langle x \rangle$$

$$K_0 \langle t \rangle + K_1 \langle t^2 \rangle = \langle x t \rangle$$

$$\begin{bmatrix} 1 & \langle t \rangle \\ \langle t \rangle & \langle t \rangle \end{bmatrix} \begin{bmatrix} K_0 \\ K_1 \end{bmatrix} = \begin{bmatrix} \langle x \rangle \\ \langle x t \rangle \end{bmatrix}$$

5.1.3 Question: Code a polynomial fit

Write functions that accept N data pairs (t_n, x_n) and return:

- The polynomial fit parameters for a $M + 1$ polynomial.
- The error of the fit
- The best fit line

Then...

- Generate some data for a given polynomial (say linear, or cubic)
- Add some zero-mean noise to the data points.
- Fit with polynomials of different order.
- Plot error as a function of polynomial order.

5.2 Stochastic time-series models

5.2.1 Auto-regressive models

- Widespread use to model stochastic data
- Model is discrete, writing current state as function of previous ones   need to take care with discretized timesteps
- An order-p model

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

- where ϵ_t are uncorrelated normal random numbers with

$\langle \epsilon \rangle = 0$ and $\langle \epsilon^2 \rangle = \sigma_\epsilon^2$.

- Here we will examine an order-1 model

like discrete O-U process

$$X_t = c + \phi X_{t-1} + \epsilon_t$$

5.2.2 First-order auto-regressive model

- First-order model, where we assume $|\phi| < 1$

$X_t = c + \phi X_{t-1} + \epsilon_t$ where $\langle \epsilon \rangle = 0$ and $\langle \epsilon^2 \rangle = \sigma_\epsilon^2$.

Questions:

$\langle X_t \rangle = \langle X_{t-1} \rangle$ true in stationary state (initial transients die away)

$\langle X_t \rangle = \langle c + \phi X_{t-1} + \epsilon_t \rangle$

$\langle X_t \rangle = c + \phi \langle X_{t-1} \rangle + \langle \epsilon_t \rangle$

- What is the mean of $\langle X \rangle$ $\langle X_t \rangle = c + \phi \langle X_t \rangle + 0$
- What is Var(X)? $\langle X \rangle = c + \phi \langle X \rangle \Rightarrow \langle X \rangle = \frac{c}{1-\phi}$
- What is the general solution, if the process has always be ongoing?

$\mathrm{Var} = \langle (X-\mu)^2 \rangle$ $X_t - \langle X_t \rangle = \phi X_{t-1} + \epsilon_t$

square it $= \mathrm{var}(X_t)$ indep $\Rightarrow = 0$

$\mathrm{Var}(X_t) = \mathrm{Var}(X_{t-1})$ in stationary state (initial transients die away)

$\mathrm{Var}(X) = \phi^2 \mathrm{Var}(X) + 2\phi \mathrm{Var}(X_{t-1}\epsilon_t) + \mathrm{Var}(\epsilon_t)^2$

$\mathrm{Var}(X) = \phi^2 \mathrm{Var}(X) + \sigma_\epsilon^2$

$\mathrm{Var}(X) = \frac{\sigma_\epsilon^2}{1-\phi^2}$

$\langle X_t^2 \rangle = \langle (c + \phi X_{t-1} + \epsilon_t)^2 \rangle$

$= \langle c^2 + \phi^2 X_{t-1}^2 + \epsilon_t^2 + 2c\phi X_{t-1} + 2c\epsilon_t + 2\phi X_{t-1}\epsilon_t \rangle$

$= \langle c^2 \rangle + \langle \phi^2 X_{t-1}^2 \rangle + \langle \epsilon_t^2 \rangle + \langle 2c\phi X_{t-1} \rangle + \langle 2c\epsilon_t \rangle + \langle 2\phi X_{t-1}\epsilon_t \rangle$

$= 0 + \phi^2 \langle X_{t-1}^2 \rangle + \sigma_\epsilon^2 + 2c\phi \langle X_{t-1} \rangle + 0 + 0$

$\langle X_t^2 \rangle = \phi^2 \langle X_t^2 \rangle + \sigma_\epsilon^2 + 2c\phi \langle X_t \rangle$

$\langle X_t^2 \rangle = \phi^2 \langle X_t^2 \rangle + \sigma_\epsilon^2 + 2\frac{c^2\phi}{1-\phi}$ $\Rightarrow \langle X^2 \rangle = \frac{\sigma_\epsilon^2 + 2\frac{c^2\phi}{1-\phi}}{1-\phi^2}$

Auto-Regressive Model

$\langle X \rangle = \frac{c}{1-\phi}$, $\langle X^2 \rangle = \frac{\sigma_\varepsilon^2 + 2\frac{c^2\phi}{1-\phi}}{1-\phi^2}$

$Var(X) = \langle X^2 \rangle - \langle X \rangle^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2} + \frac{2\frac{c^2\phi}{1-\phi}}{1-\phi^2} - \frac{c^2}{(1-\phi)^2}$

Want $\frac{\sigma_\varepsilon^2}{1-\phi^2}$

General Solution to $X_t = c + \phi X_{t-1} + \varepsilon_t$

Solve for $X_t = c + \phi X_{t-1}$; $X_t = \phi X_{t-1} + \varepsilon_t$
Then add together

$$X_t = c + \phi X_t$$
$$X_t - \phi X_t = c$$
$$X_t(1-\phi) = c$$
$$X_t = \frac{c}{1-\phi}$$

Now $X_t = \phi X_{t-1} + \varepsilon_t$

can say $X_t = \phi X_{t-1} + g(t)$

homog

$X_t = \phi X_{t-1}$

$X_t = \phi^t X_0$

inhomog

$X_t = \phi X_{t-1} + g(t)$

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + g(t-1)] + g(t) \\ &= \phi^2 X_{t-2} + \phi g(t-1) + g(t) \\ &= \phi^2 X_{t-2} + [\phi^1 g(t-1) + \phi^0 g(t)] \\ &= \phi^2(\phi X_{t-3} + g(t)) + [\phi^1 g(t) + \phi^0 g(t)] \\ &= \phi^3 X_{t-3} + [\phi^2 g(t-2) + \phi^1 g(t-1) + \phi^0 g(t)] \end{aligned}$$

$$X_t = \phi^t X_0 + \sum_{k=0}^{t} \phi^k g(t-k)$$

for steady state, let $t \to \infty$ also $g(t-i) = \varepsilon_{t-k}$
note $|\phi| < 1$

$$\lim_{t\to\infty} X_t = \lim_{t\to\infty} \phi^t X_0 + \sum_{k=0}^{t} \phi^k \varepsilon_{t-k}$$

$$X = \sum_{k=0}^{\infty} \phi^k \varepsilon_{t-k}$$

weight term

it's also also a filter
It's also a convolution

$$\phi^k = e^{k\ln(\phi)} = e^{-kq} \quad -q = \ln(\phi) \quad (\ln(\phi) < 0)$$

Add all back together

$$X = \frac{c}{1-\phi} + \sum_{k=0}^{\infty} \phi^k \varepsilon_{t-k}$$

5.2.3 Autocovariance

- First-order autoregressive model, in steady state.

$$X_t = c + \phi X_{t-1} + \epsilon_t \quad \text{where} \quad \langle \epsilon \rangle = 0 \quad \text{and} \quad \langle \epsilon^2 \rangle = \sigma_\epsilon^2.$$

- Derived mean, variance and general solution

$$\langle X \rangle = \frac{c}{1-\phi}\ , \text{Var}(x) = \frac{\sigma_\epsilon^2}{(1-\phi^2)} \quad \text{and} \quad X_t = \frac{c}{1-\phi} + \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k}$$

- Examine temporal structure given by autocovariance:

$$\langle (X_{t+n} - \langle X \rangle)(X_t - \langle X \rangle) \rangle = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \langle \epsilon_{t+n-j} \epsilon_{t-k} \rangle$$

- **Question:** Assume $n > 0$. What is the autocovariance?

5.3 Summary and additional questions

Day 5 Basic time-series analysis

Questions

Make sure you have understood and done all the questions in the lectures.

The questions below are to be handed in for marking by 10am Monday 3rd December 2018.

Q5.1 Analysis of exo-planet data
Q5.2 Simulation of an autoregressive model

Autocovariance

$$\left\langle (X_{t+n} - \langle X \rangle)(X_t - \langle X \rangle) \right\rangle = \left\langle \left(\frac{c}{1-\phi} + \sum_{k=0}^{\infty} \phi^k \varepsilon_{t+n-k} - \frac{c}{1-\phi}\right)\left(\frac{c}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} - \frac{c}{1-\phi}\right) \right\rangle$$

$$= \left\langle \left(\sum_{k=0}^{\infty} \phi^k \varepsilon_{t+n-k}\right)\left(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\right) \right\rangle = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \langle \phi^k \phi^j \varepsilon_{t+n-k} \varepsilon_{t-j} \rangle$$

$$= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \phi^{k+j} \langle \varepsilon_{t+n-k} \varepsilon_{t-j} \rangle$$

Switch

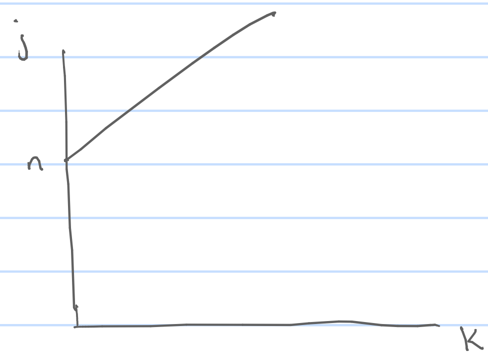
$$\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \langle \varepsilon_{t+n-j} \varepsilon_{t-k} \rangle$$

0 except when $t+n-j = t-k$

$n-j = -k \Rightarrow k = j-n$

$n = j-k$

$$\sum_{j=0}^{\infty}$$



$$\sigma_\varepsilon^2 \sum_{k=0}^{\infty} \phi^{n+k} \phi^k$$

$$\sigma_\varepsilon^2 \phi^n \sum_{k=0}^{\infty} \phi^{2k} = \sigma_\varepsilon^2 \phi^n \cdot \frac{1}{1-\phi^2}$$

$$\langle X_{t+n} X_t \rangle = \phi^n \operatorname{Var}(X) \qquad n > 0$$

for $n < 0$,

only commutes b/c X both $\qquad \langle XY \rangle \neq \langle YX \rangle$

b/c time invariant

$$\langle X_{t+n} X_t \rangle = \langle X_t X_{t+n} \rangle$$

let $s = t+n$

$$= \langle X_{s-n} X_s \rangle$$

$$= \phi^{-n} \operatorname{Var}(X) \qquad \text{remember } n<0$$

so $= \phi^{|n|} \operatorname{Var}(X)$

Q5.1 Analysis of exo-planet data

- Download the **exoplanet-data.txt** file from the course website.
- The x-axis is in days, the y-axis is raw light intensity from the star.
- The dips are caused by a planet passing in front of the star.
- It can be assumed that this is a single-planet system.

- Some cleaning of the data will be required.

- Part (a) Estimate the orbital period of the planet, with a figure showing method.

- Part (b) Assume the flux from the sum is proportional to its area.
  Estimate the ratios of the planetary and solar radii. Provide a figure showing method.

- Part (c) Estimate the transit time in days, with a figure showing method.

NOTE There is no need to provide ranges for data.

Q5.2 Simulation of an autoregressive model

- Simulate a first order autoregressive model with your parameter choice
  for c, ϕ and σ_ϵ (but with $|\phi| < 1$).

- Run the simulation for 5000 steps measure $\langle X \rangle$ and Var(X).
  How do they compare to theory?

- Extract the autocovariance from the data.
  Provide a plot comparing it to the theoretical prediction.

In [ ]:
```

```