

Maths Stats Formulas and Definitions

Boyd Kane

Last updated: June 26, 2021

Contents

1	Distributions, tricks, and required knowledge	5
1.1	Common Distributions	5
1.1.1	Poisson Distribution	5
1.1.2	Negative Binomial Distribution	5
1.1.3	Geometric Distribution	6
1.1.4	Binomial Distribution	6
1.1.5	Beta Distribution	6
1.1.6	Normal Distribution	6
1.1.7	Gamma Distribution	7
1.1.8	Exponential Distribution	9
1.1.9	Student's t-distribution	9
1.1.10	Chi-Squared Distribution	9
1.2	Generating Functions	11
1.2.1	Probability Generating Functions	11
1.2.2	Moment Generating Functions	11
1.3	Probability	12
1.4	Expectation and Variance	12
1.5	Random Vectors, Covariance, Correlation	13
1.6	MAM1000W things	14
1.6.1	Taylor Expansions	14
1.6.2	Calculus	14
1.6.3	Euler's Constant	14
1.7	Conditioning	15
1.8	Sum Trickery	15
2	Chains, Processes, and Forecasting	17
2.1	3041F: Chains	17
2.1.1	Chapman-Kolmogorov Equations	17

2.2	3041F: Branching Processes	17
2.3	3041F: Probability of Extinction	18
2.4	3041F: First Passage Probabilities	19
2.5	3041F: Discrete Time	19
2.5.1	Calculating the Count at a given Time	19
2.5.2	Calculating the Time until a given Count	20
2.6	3041F: Continuous Time	21
2.6.1	Calculating the Count at a given Time	21
2.6.2	Calculating the Time until a given Count	22
3	Poisson Processes	23
3.1	3041F: Poisson Processes	23
3.1.1	3041F: Probability Integral Transform	23
3.1.2	3041F: Estimation	23
3.2	3041F: Basic Poisson Process	24
3.2.1	3041F: Poisson Postulates	24
3.2.2	Poisson Process Information	24
3.2.3	3041F: Modelling Poisson Processes	25
3.2.4	3041F: Estimation of Poisson Processes	25
3.3	3041F: Compound Poisson Processes	26
3.4	3041F: Non-homogeneous Poisson Processes	27
3.4.1	3041F: Mean Value Function	27
3.4.2	3041F: Simulation of a non-homogeneous Poisson Process	28
3.4.3	3041F: Estimation of non-homogeneous Poisson Pro- cessess	29
4	Time Series I and Transformations	31
4.1	3041F: Time Series Analysis	31
4.1.1	Trend	33
4.1.2	Seasonality	33
4.1.3	Cycles	33
4.1.4	Transformations	33
4.2	3041F: Transformations	34
4.2.1	3041F: Box-Cox Transformation	35
4.2.2	3041F: Fitting Linear Filters	36
4.2.3	3041F: Moving Averages	38
4.2.4	Exponential Moving Averages	39

5	Time Series II	42
5.1	3041F: Time Series II Intro	42
5.1.1	Backshift Operator	42
5.2	Stationarity	44
5.2.1	Strict Stationarity	44
5.2.2	Autocovariance	45
5.2.3	Weak Stationarity	45
5.2.4	Some examples of Weak stationarity	46
5.2.5	More examples, introducing the autoregressive process	46
5.3	Autocorrelation	47
5.4	Significance testing of Autocorrelation	47
5.4.1	Hypothesis tests with Autocorrelation	48
5.5	Linear Models for Time Series	49
5.5.1	White Noise Process	49
5.5.2	The Random Walk Process	49
5.5.3	The Random Walk Process with Drift	50
5.6	First Order Moving Average process	50
5.7	Second Order Moving Average Process	51
5.8	Q-th Order Moving Average Process	51
5.9	General Linear Process	52
5.10	Infinite Moving Average Process	53
5.10.1	Forecasts for Moving Average Processes	54
5.11	Autoregressive Process	55
5.11.1	The Characteristic Equation	56
5.11.2	P-th order Autoregressive Process	57
5.11.3	Finding the Order of an Autocorrelation Model	57
5.11.4	Calculating lag-3 partial autocorrelations of a AR(1) process	59
5.11.5	Using Real Data to make a model and then forecast	60
5.11.6	Checking the validity of a AR model	61
5.11.7	Forecasting with an AR process	61
5.11.8	Sections in the notes that will not be covered	62
5.11.9	AR(2) Process forecasting	63
5.11.10	Forecasting with a AR(p) process	64
5.11.11	Stationarity and Invertibility	64
5.12	The Autoregressive Moving Average Process	65
5.12.1	Intro to ARMA(p,q)	65
5.12.2	Identifying the order of an ARMA process	66

5.12.3	Forecasting with an ARMA Process	67
5.13	The ARIMA Process	67
5.13.1	Difference Operator	68
5.13.2	Making a ARIMA process stationary	68
5.13.3	Finding the order of an ARIMA Process	68
5.14	Box-Jenkins Algorithm	69
6	Terms, Definitions, R trickery, and side-notes	71
6.1	Terms and Definitions	71
6.2	R trickery	73
6.2.1	runif	73
6.2.2	cumsum	73
6.2.3	plot	73
6.2.4	ts	74
6.3	Notes	74

Chapter 1

Distributions, tricks, and required knowledge

1.1 Common Distributions

1.1.1 Poisson Distribution

$$\begin{aligned} X &\sim Poi(\lambda) & f_X(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ \mathbb{E}[X] &= \lambda & G_X(s) := \mathbb{E}[s^X] &= e^{\lambda(s-1)} \\ \mathbb{V}[X] &= \lambda & \left(\sum_{i=0}^n Poi(\lambda_i) \right) &\sim Poi\left(\sum_{i=0}^n \lambda_i \right) \end{aligned}$$

1.1.2 Negative Binomial Distribution

$$\begin{aligned} X &\sim NegBin(p, r) & f_X(x) &= \binom{k-1}{r-1} p^r q^{k-r} \\ \mathbb{E}[X] &= \frac{r}{p} & \mathbb{E}[e^{sX}] &= \left(\frac{pe^s}{1-qe^s} \right)^r \\ \mathbb{V}[X] &= \frac{rq}{p^2} \end{aligned}$$

1.1.3 Geometric Distribution

$$\begin{aligned} X &\sim \text{Geo}(p) \\ \mathbb{E}[X] &= \frac{1}{p} & f_X(x) &= pq^{x-1} \\ \mathbb{V}[X] &= \frac{q}{p^2} & \mathbb{E}[e^{sX}] &= \frac{e^t p}{1 - e^t q} \end{aligned}$$

1.1.4 Binomial Distribution

$$\begin{aligned} X &\sim \text{Binomial}(n, p) \\ \mathbb{E}[X] &= np & f_X(x) &= \binom{n}{x} p^x q^{n-x} \\ \mathbb{V}[X] &= npq & \mathbb{E}[e^{sX}] &= (1 - p + pe^s)^n \end{aligned}$$

See figure 1.1.

1.1.5 Beta Distribution

$$0 < x < 1$$

$$\begin{aligned} X &\sim \text{Beta}(a, b) \\ \mathbb{E}[X] &= \frac{a}{a+b} & f_X(x) &= \frac{(a+b-1)!}{(a-1)!(b-1)!} \cdot x^{a-1}(1-x)^{b-1} \\ \mathbb{V}[X] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

See figure 1.2

1.1.6 Normal Distribution

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \\ \mathbb{E}[X] &= \mu & f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ \mathbb{V}[X] &= \sigma^2 & \mathbb{E}[e^{sX}] &= e^{\mu s + \frac{1}{2}s^2\sigma^2} \end{aligned}$$

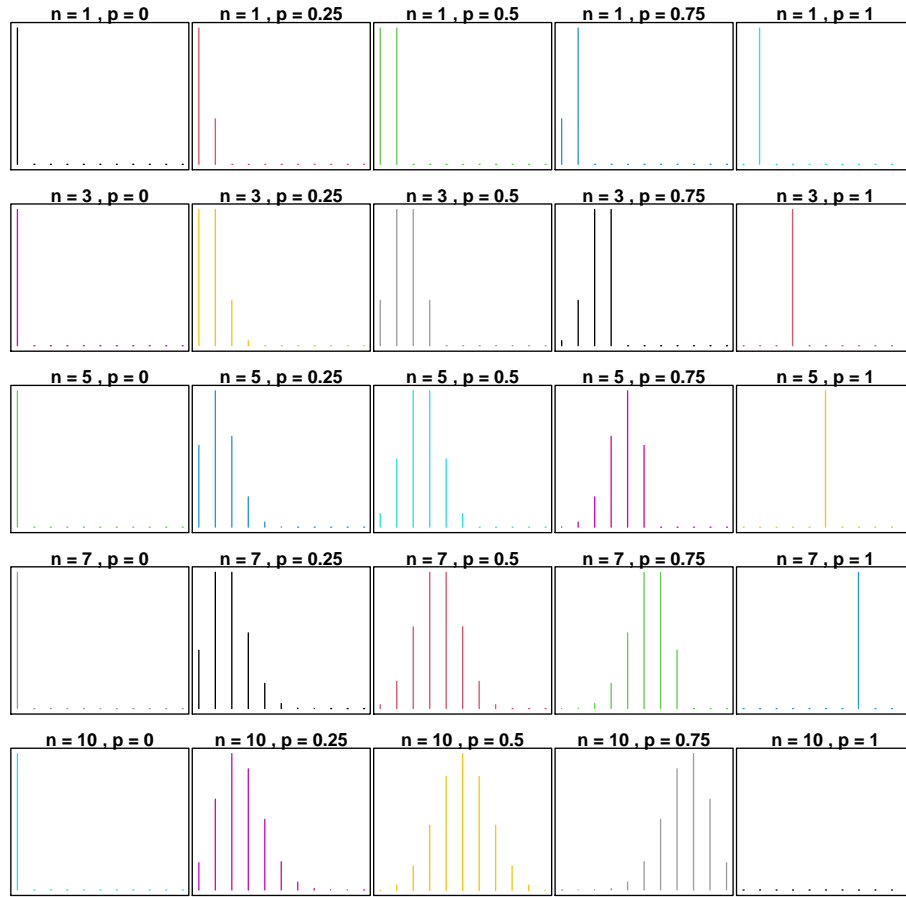


Figure 1.1: The binomial distribution with parameters n and p

1.1.7 Gamma Distribution

$\lambda > 0, \alpha > 0$

$$\begin{aligned}
 X &\sim \text{Gamma}(\lambda, \alpha) \\
 \mathbb{E}[X] &= \frac{\alpha}{\lambda} \\
 \mathbb{V}[X] &= \frac{\alpha}{\lambda^2}
 \end{aligned}
 \qquad
 \begin{aligned}
 f_X(x) &= \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{(\alpha-1)!} \\
 \mathbb{E}[e^{sX}] &= \left(\frac{\lambda}{\lambda-s} \right)^\alpha
 \end{aligned}$$

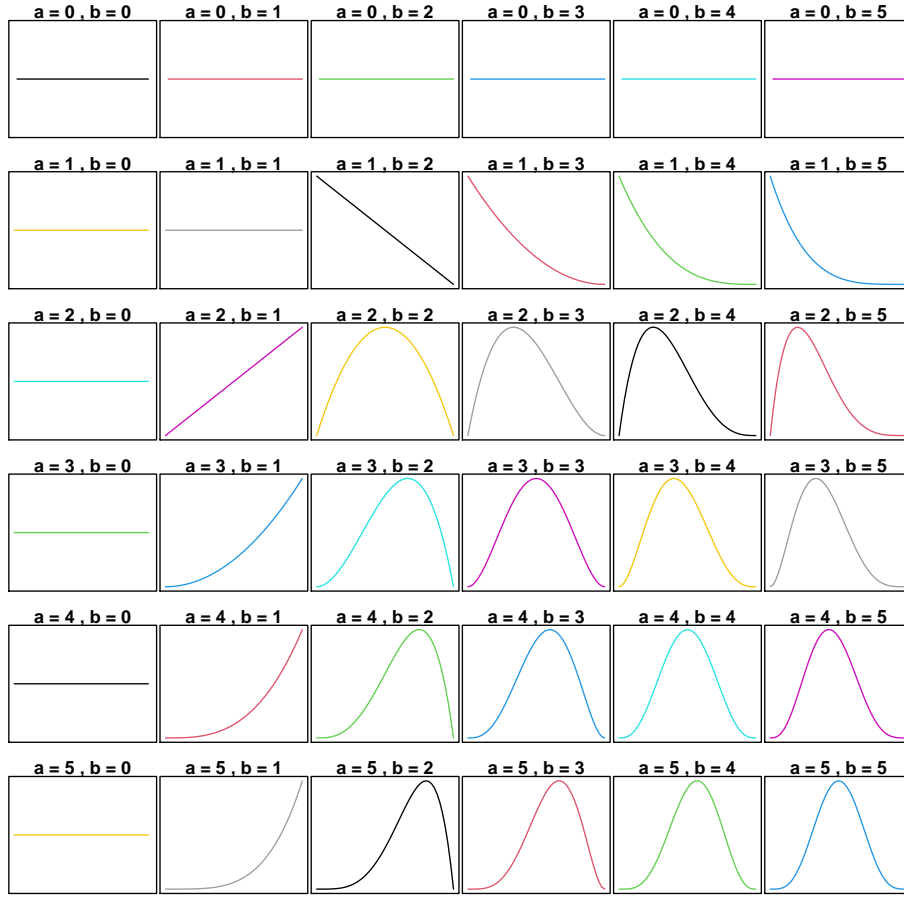


Figure 1.2: The Beta distribution with parameters a and b.

Notes

Sum of independant Exponentials is Gamma distributed:

$$\left(\sum^n Exp(\lambda) \right) \sim Gamma(\lambda, n)$$

$$Exp(\lambda) \equiv Gamma(\lambda, 1)$$

1.1.8 Exponential Distribution

$$\begin{aligned} X &\sim \text{Exp}(\lambda) & f_X(x) &= \lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0 \\ \mathbb{E}[X] &= \frac{1}{\lambda} & F_X(x) &= 1 - e^{-\lambda x} \\ \mathbb{V}[X] &= \frac{1}{\lambda^2} & \mathbb{E}[e^{sX}] &= \frac{\lambda}{\lambda - s} \quad s < \lambda \\ \hat{\lambda}_{MLE} &= \frac{1}{\bar{x}} & \mathbb{P}[\text{Exp}(\lambda_1) < \text{Exp}(\lambda_2)] &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \\ & & \min[\text{Exp}(\lambda_1), \text{Exp}(\lambda_2)] &\sim \text{Exp}(\lambda_1 + \lambda_2) \end{aligned}$$

1.1.9 Student's t-distribution

Work in progress.

1.1.10 Chi-Squared Distribution

The χ_m^2 Distribution is defined as having m degrees of freedom, and is usually defined in terms of a sum of standard normal distributions. ie if

$Z_i \sim N(0, 1) \quad \forall i \in 1 \dots k$ and all Z_i independent. then:

$$\sum_{n=1}^k (Z_i)^2 \sim \chi_{k-1}^2$$

$$\mathbb{E}[\chi_k^2] = k$$

$$\mathbb{V}[\chi_k^2] = 2k$$

$$\mathbb{M}_{\chi_k^2}(t) = (1 - 2t)^{-\frac{k}{2}} \quad \text{for } t < \frac{1}{2}$$

$$\chi_2^2 \equiv \text{Exp}\left(\frac{1}{2}\right)$$

$$\begin{aligned} -2 \ln U(0, 1) &\equiv \chi_2^2 \\ &\equiv \text{Exp}\left(\frac{1}{2}\right) \end{aligned}$$

And if $X_1 \sim \chi_a^2, X_2 \sim \chi_b^2, X_1 \perp\!\!\!\perp X_2$

$$\frac{X_1}{X_1 + X_2} \equiv \text{Beta}\left(\frac{a}{2}, \frac{b}{2}\right)$$

$$\frac{\frac{\chi_a^2}{a}}{\frac{\chi_b^2}{b}} \equiv F_{a,b}$$

and if X_i are all independent with degrees of freedom f_i

$$\sum_{i=1}^k \chi_{f_i}^2 \equiv \chi_{\sum_{i=1}^k f_i}^2$$

1.2 Generating Functions

1.2.1 Probability Generating Functions

If X is a non-negative, integer-valued, and discrete, and $0 \leq s \leq 1$, then the probability generating function of X is given by:

$$\begin{aligned}
 \mathbb{G}_X(s) &:= \mathbb{E}[s^X] = \sum_{x=0}^{\infty} s^x \cdot p_X(x) \\
 \mathbb{G}_X(0) &= f_X(0) = \mathbb{P}[X = 0] \\
 \mathbb{G}_X(1) &= \sum_{x=0}^{\infty} 1^x \cdot f_X(x) = 1 \\
 \mathbb{V}[X] &= \mathbb{G}_X''(1) + \mathbb{G}_X'(1) - (\mathbb{G}_X'(1))^2 \\
 \mathbb{E}[X] &= \mathbb{G}_X'(1) \\
 \mathbb{E}[X(X-1)] &= \mathbb{G}_X''(1) \\
 \frac{d^r}{ds^r} \mathbb{G}_X(0) &= r! \mathbb{P}[X = r] \\
 \mathbb{P}[X = r] &= \frac{1}{r!} \frac{d^r}{ds^r} \mathbb{G}_X(0) \\
 \underbrace{\mathbb{E} \left[\frac{\Gamma(X+1)}{\Gamma(X+1-r)} \right]}_{\substack{\text{If } x \in \mathbb{Z}^+ \text{ then} \\ \Gamma(x+1) = x!}} &= \mathbb{E}[X(X-1) \dots (X-r+1)] = \lim_{s \rightarrow 1^-} \mathbb{G}_X^{(r)}(s)
 \end{aligned}$$

1.2.2 Moment Generating Functions

The Moment Generating Function of X is defined by:

$$M_X(t) := \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x) dx$$

But the MGF might not always exist. We can also define its complex-numbered version, the Characteristic Function, which will always exist:

$$\begin{aligned}
 \varphi(t) &:= \mathbb{E}[e^{itX}] \\
 &= \int_{-\infty}^{\infty} e^{itx} \cdot f_X(x) dx
 \end{aligned}$$

The r -th moment of X can be retrieved by taking the r -th derivative of the MGF and evaluating it at zero. A similar process also works for the characteristic function:

$$\begin{aligned} r\text{-th Moment of } X &= \mathbb{E}[X^r] \\ &= \frac{d^r}{dX^r} M_X(0) \\ &= (-i)^r \frac{d^r}{dX^r} \varphi_X(0) \end{aligned}$$

1.3 Probability

The Law of Total Probability is given as:

$$\mathbb{P}[A] = \sum_{n=1}^{\infty} \mathbb{P}[A|B_n] \mathbb{P}[B_n]$$

Bayes Theorem:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \mathbb{P}[A]}{\mathbb{P}[B]}$$

1.4 Expectation and Variance

Expectation is always taken with respect to a random variable (a bit like you always differentiate with respect to a certain variable in an equation).

If the random variable is not obvious, then the Expectation of some function $g(X)$ with respect to X (which has probability density function given by $f_X(x)$) is written like $\mathbb{E}_X[g(X)]$ and is defined as:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx$$

So for example, if $g(x) = x^2$ then we can calculate the Expectation of X^2 as:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x) dx$$

And in the special case where X is non-negative and absolutely continuous, we have the identity:

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}[\{X > x\}] dx$$

Note that expectation is a linear transformation, such that the following is true (Assuming that both expectations are finite):

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Variance can be calculated as:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

This often is shown as something like:

$$\mathbb{E}[X^2] = \mu^2 + \sigma^2$$

1.5 Random Vectors, Covariance, Correlation

So as to not get lost in a sea of brackets, define the following:

$$\begin{aligned}\mu_X &= \mathbb{E}[X] \\ \mu_Y &= \mathbb{E}[Y] \\ \sigma_X^2 &= \mathbb{V}[X] \\ \sigma_Y^2 &= \mathbb{V}[Y] \\ \sigma_{XY} &= \mathbb{Cov}[X, Y] \\ \rho_{XY} &= \mathbb{Corr}[X, Y]\end{aligned}$$

With the prior definitions, the following are useful identities, definitions, or shortcuts relating to Covariance and Correlation:

$$\begin{aligned}\mathbb{Cov}[X, Y] &= \mathbb{E}[(X - \mu_x) \cdot (Y - \mu_y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \\ \mathbb{Cov}[X, a] &= 0 \quad a \in \mathbb{R} \\ \mathbb{Cov}[X, Y] &= \mathbb{Cov}[Y, X] \\ \mathbb{Corr}[X, Y] &= \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \\ &= \frac{\mathbb{Cov}[X, Y]}{\sqrt{\mathbb{V}[X] \cdot \mathbb{V}[Y]}} \\ \mathbb{Cov}[X_1, X_2] &= 0 \quad \text{if } X_1 \text{ is independent of } X_2\end{aligned}$$

Also, we can use FOIL for covariance:

$$\mathbb{Cov}[A - B, C - D] = \mathbb{Cov}[A, C] - \mathbb{Cov}[A, D] - \mathbb{Cov}[B, C] + \mathbb{Cov}[B, D]$$

1.6 MAM1000W things

1.6.1 Taylor Expansions

$$\forall x \in (-1, 1) \quad \frac{1}{1-x} = 1 + x + x^2 + \dots = \sum_{n=0}^{\infty} x^n$$

$$\forall x \in (-1, 1] \quad \ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n}$$

1.6.2 Calculus

$$\frac{d}{dx} a^x = a^x \ln(a)$$

$$\frac{d}{dx} e^x = e^x$$

$$\frac{d}{dx} x^n = nx^{n-1}$$

$$\frac{d}{dx} \ln(x) = x^{-1}$$

$$\frac{d}{dx} f \cdot g = f'g + fg'$$

$$\frac{d}{dx} \frac{f}{g} = \frac{f'g - fg'}{g^2}$$

$$\frac{d}{dx} f(g) = f'(g) \cdot g'$$

$$\frac{d}{dx} \ln(f) = \frac{f'}{f}$$

1.6.3 Euler's Constant

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots$$

$$= \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

$$e^{i\theta} = \cos \theta + i \sin \theta$$

1.7 Conditioning

The definition of a conditional distribution is given by:

$$f_{X|Y}(x|y) = \frac{f_{xy}(x, y)}{f_Y(y)}$$

And conditional probability by:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Conditional Expectation identities and definitions:

$$\begin{aligned}\mathbb{E}[X|Y = y] &= \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) dx \\ \mathbb{E}[g(X)|X = x] &= g(x) \\ \mathbb{E}[c|X] &= c \\ \mathbb{E}[aX + bY|Z] &= a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z] \\ \mathbb{E}[g(X, Y)|Y = y] &= \mathbb{E}[g(X, y)|Y = y] \\ \underbrace{\mathbb{E}[g(X_1, X_2)|X_2 = x_2]}_{\text{If } X_1 \perp\!\!\!\perp X_2} &= \mathbb{E}[g(X_1, x_2)|X_2 = x_2]\end{aligned}$$

Expectation can be re-written as:

$$\begin{aligned}\mathbb{E}_X[X] &= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] \\ &= \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] \cdot f_Y(y) dy\end{aligned}$$

Variance can be re-written as:

$$\mathbb{V}_X[X] = \mathbb{E}_Y[\mathbb{V}_{X|Y}[X|Y]] + \mathbb{V}_Y[\mathbb{E}_{X|Y}[X|Y]]$$

1.8 Sum Trickery

If you've got a random variable defined like $S_N := \sum_{i=1}^N X_i$ where:

- N is a non-negative Random Variable

- $\{X_1, X_2, \dots\}$ are all iid.
- $N \perp\!\!\!\perp X_i$ for all i

Then these identities often comes up:

$$\begin{aligned}\mathbb{E}[S_N] &= \mathbb{E}[N] \cdot \mathbb{E}[X] \\ \mathbb{V}[S_N] &= \mathbb{E}[N] \cdot \mathbb{V}[X] + \mathbb{E}[X]^2 \cdot \mathbb{V}[N] \\ \mathbb{G}_{S_N}(t) &= \mathbb{G}_N(\mathbb{G}_X(t))\end{aligned}$$

And finally, some non-random variable based sums that often come up:

$$\sum_{i=n}^{\infty} g(x)^i = \frac{g(x)^n}{1 - g(x)} \quad \text{if } |g(x)| < 1$$

And, if $0 < r < 1$:

$$\begin{aligned}\sum_{i=1}^{\infty} ar^i &= \frac{a}{1 - r} \\ a \sum_{i=0}^{\infty} ir^i &= \frac{ar}{(1 - r)^2} \\ \sum_{i=0}^n i &= \frac{1}{2}n(n + 1) \\ \sum_{i=0}^n i^3 &= \left(\frac{1}{2}n(n + 1)\right)^2\end{aligned}$$

Chapter 2

Chains, Processes, and Forecasting

2.1 3041F: Chains

2.1.1 Chapman-Kolmogorov Equations

Basically, for Markov chains we can just do matrix multiplication to go from time i to time $i + 1$. For $0 \leq v \leq t$, and $p_{ij}(s, t) := \mathbb{P}[X_{s+t} = j | X_s = i]$:

$$p_{ij}(s, t) = \mathbb{P}[X_{s+t} = j | X_s = i] = \sum_{k \in \mathcal{U}} p_{ik}(s, v) \cdot p_{kj}(s + v, t - v)$$

Or, in matrix notation:

$$\mathbf{P}^{(t)}(s) = \mathbf{P}^{(v)}(s) \cdot \mathbf{P}^{(t-v)}(s + v)$$

And under time homogeneity

$$\mathbf{P}^{(t)} = \mathbf{P}^v \cdot \mathbf{P}^{t-v} = \mathbf{P}^t$$

2.2 3041F: Branching Processes

Let Z_m be the number of individuals in Generation m . Let X_{im} be the number of offspring produced by individual i in generation m . Assume all $X_{im} = X$, so all X_{im} are identically distributed. Assume $Z_0 = 1$ which implies $Z_1 = X$.

$$\mathbb{E}[Z_m] = \mathbb{E}[X]^m$$

Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{V}[X]$.

$$\begin{aligned} \mathbb{V}[Z_m] &= m\sigma^2 && \text{if } \mu = 1 \\ \mathbb{V}[Z_m] &= \sigma^2 \cdot \mu^{m-1} \left(\frac{\mu^m - 1}{\mu - 1} \right) && \text{if } \mu \neq 1 \end{aligned}$$

2.3 3041F: Probability of Extinction

To calculate the probability that the process is extinct by generation m :

$$\mathbb{P}[\text{Extinction by generation } m] = \mathbb{P}[Z_m = 0] = \mathbb{G}_m(0)$$

But to calculate the probability that the process is extinct at exactly generation m :

$$\mathbb{P}[\text{Extinction at generation } m] = \mathbb{G}_m(0) - \mathbb{G}_{m-1}(0)$$

Define the probability of extinction:

$$\begin{aligned} \eta &:= \mathbb{P}[\text{Eventual Extinction}] = \mathbb{P} \left[\bigcup_{m=1}^{\infty} \{Z_m = 0\} \right] \\ &= \text{smallest non-negative integer root of } \mathbb{G}_X(s) = s \end{aligned}$$

2.4 3041F: First Passage Probabilities

$$\begin{aligned}
p_{ij}^{(n)} &:= \mathbb{P}[X_{t+n} = j | X_t = i] \\
P &= [p_{ij}] \\
P_{ij}(s) &:= \sum_{n=0}^{\infty} p_{ij}^{(n)} \cdot s^n \\
\mathbf{P}(s) &:= [P_{ij}(s)] \\
&= (\mathbf{I} - s \cdot P)^{-1} \\
&= (\mathbf{I} - s \cdot [p_{ij}])^{-1} \\
f_{ij}^{(k)} &= \mathbb{P}[\text{a } k\text{-step passage from } i \text{ to } j] \quad f_{ij}^{(0)} = 0 \forall i, j \\
F_{ij}(s) &= \sum_{n=0}^{\infty} f_{ij}^{(n)} \cdot s^n
\end{aligned}$$

Don't forget the test comes with identities for converting Probability matrices to first return probabilities.

2.5 3041F: Discrete Time

2.5.1 Calculating the Count at a given Time

Forecasting

$$\mathbb{P}[N_{n+k} = a + i | N_n = a] = \binom{k}{i} \cdot p^i \cdot q^{k-i} \quad (2.1)$$

$$(N_{n+k} - N_n) \sim \text{Binomial}(k, p) \quad (2.2)$$

$$(N_{n+k} - N_n) \perp\!\!\!\perp N_n \quad (2.3)$$

$$\mathbb{E}[N_{n+k} | N_n = a] = k \cdot p + a \quad (2.4)$$

$$\mathbb{V}[N_{n+k} | N_n = a] = k \cdot p \cdot a \quad (2.5)$$

Backcasting

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] = \frac{\binom{n}{a} \binom{k}{i}}{\binom{n+k}{a+i}} \quad (2.6)$$

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] \sim \text{HyperGeo}(n + k, n, a + i) \quad (2.7)$$

$$\mathbb{E}[N_n | N_{n+k} = a + i] = \frac{(a + i) \cdot n}{n + k} \quad (2.8)$$

$$\mathbb{V}[N_n | N_{n+k} = a + i] = \frac{n + k - a - i}{n + k - 1} \cdot \frac{(a + i) \cdot n}{n + k} \cdot \left(1 - \frac{n}{n + k}\right) \quad (2.9)$$

2.5.2 Calculating the Time until a given Count

$$T_a \sim \text{NegBinomial}(a, p) \quad (2.10)$$

$$\mathbb{E}[T_a] = \frac{a}{p} \quad (2.11)$$

$$\mathbb{V}[T_a] = \frac{a \cdot q}{p^2} \quad (2.12)$$

Forecasting

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] = \binom{j-1}{i-1} \cdot p^i \cdot q^{j-i} \quad (2.13)$$

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] \sim \text{NegBinomial}(i, p) \quad (2.14)$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i}{p} + t \quad (2.15)$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i \cdot q}{p^2} \quad (2.16)$$

Backcasting

$$\mathbb{P}[T_a = t | T_{a+i} = t + j] = \frac{\binom{t-1}{a-1} \cdot \binom{j-1}{i-1}}{\binom{t+j-1}{a+i-1}} \quad (2.17)$$

$$\mathbb{E}[T_a | T_{a+i} = t + j] = \frac{a(t+j)}{a+i} \quad (2.18)$$

$$\mathbb{V}[T_a | T_{a+i} = t + j] = \frac{ai(t+j)(t+j-a-i)}{(a+i)^2(a+i+1)} \quad (2.19)$$

2.6 3041F: Continuous Time

With the assumptions that $\lambda_i = \lambda$ and $t_i = t$ for all i :

$$X_i \sim \text{Poisson}(\lambda t) \quad (2.20)$$

$$Y_n = \sum_{i=1}^n X_i \quad (2.21)$$

$$Y_n \sim \text{Poisson}(n\lambda t) \quad (2.22)$$

$$Y_{n+k} - Y_n \sim \text{Poisson}(k\lambda t) \quad (2.23)$$

$$\text{Cov}[Y_n, Y_m] = \lambda \cdot t \cdot \min(n, m) \quad (2.24)$$

$$\text{Corr}[Y_n, Y_m] = \frac{\min(n, m)}{\sqrt{nm}} \quad (2.25)$$

2.6.1 Calculating the Count at a given Time

Forecasting

$$\mathbb{P}[Y_{n+k} = a + i | Y_n = a] = \frac{e^{-k\lambda t} \cdot (k\lambda t)^i}{i!} \quad (2.26)$$

$$\mathbb{E}[Y_{n+k} | Y_n = a] = k\lambda t + a \quad (2.27)$$

$$\mathbb{V}[Y_{n+k} | Y_n = a] = k\lambda t \quad (2.28)$$

Backcasting

$$\mathbb{P}[Y_n = a | Y_{n+k} = a + i] = \binom{a+i}{a} \left(\frac{n}{n+k} \right)^a \left(\frac{k}{n+k} \right)^i \quad (2.29)$$

$$\mathbb{P}[Y_n = a | Y_{n+k} = a + i] \sim \text{Binomial} \left(a + i, \frac{n}{n+k} \right) \quad (2.30)$$

$$\mathbb{E}[Y_n | Y_{n+k} = a + i] = \frac{(a+i) \cdot n}{n+k} \quad (2.31)$$

$$\mathbb{V}[Y_n | Y_{n+k} = a + i] = \frac{(a+i) \cdot n \cdot k}{(n+k)^2} \quad (2.32)$$

2.6.2 Calculating the Time until a given Count

$$T_1 \sim \text{Exponential} \left(\frac{1}{\lambda} \right) \quad (2.33)$$

$$T_a \sim \text{Gamma}(a, \lambda) \quad (2.34)$$

$$\mathbb{E}[T_a] = \frac{a}{\lambda} \quad (2.35)$$

$$\mathbb{V}[T_a] = \frac{a}{\lambda^2} \quad (2.36)$$

Forecasting

$$\mathbb{E}[T_{a+i} | T_a = t] = \frac{i}{\lambda} + t \quad (2.37)$$

$$\mathbb{V}[T_{a+i} | T_a = t] = \frac{i}{\lambda^2} \quad (2.38)$$

Backcasting

$$\mathbb{E}[T_a | T_{a+i} = t + j] = \frac{a(t+j)}{a+i} \quad (2.39)$$

$$\mathbb{V}[T_a | T_{a+i} = t + j] = \frac{ai(t+j)^2}{(a+i)^2 \cdot (a+i+1)} \quad (2.40)$$

Chapter 3

Poisson Processes

3.1 3041F: Poisson Processes

3.1.1 3041F: Probability Integral Transform

If we want a random variable, X , that should have a certain CDF, $F_X()$, we can take a uniform random variable $U \sim U(0, 1)$ and pass it through the inverse of $F_X()$ in order to get the random variable X :

$$F_X^{-1}(U) \text{ has CDF } F_X()$$

From this, the exponential distribution can be simulated as

$$-\frac{1}{\lambda} \ln(U) \sim \text{Exp}(\lambda)$$

And the Gamma distribution with n iid uniform random variables as:

$$-\frac{1}{\lambda} \ln(U_1 \cdot U_2 \cdot \dots \cdot U_n) \sim \text{Gamma}(n, \lambda)$$

3.1.2 3041F: Estimation

See page 5 in the notes

3.2 3041F: Basic Poisson Process

3.2.1 3041F: Poisson Postulates

A counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate $\lambda > 0$, for small h , iff:

1. $N(0) = 0$
2. The process has independent and stationary increments
3. $\mathbb{P}[N(h) = 1] = \lambda h + o(h)$
So the Probability of one occurrence within small timeframe h is equal to the rate parameter times h , plus some terms that go to zero.
4. $\mathbb{P}[N(h) > 1] = o(h)$
So the Probability of more than one occurrence within small timeframe h goes to zero.

3.2.2 Poisson Process Information

Proof page 7 in the notes.

The number of events in an interval of length t for a Poisson process with rate parameter λ is a random variable distributed $Poi(\lambda t)$.

More explicitly, for n a non-negative integer:

$$\mathbb{P}[N(t) = n] = \frac{e^{-\lambda t} (\lambda t)^n}{n!}$$

The interarrival times $T_n, n = 1, 2, \dots$ are iid $\sim Exp(\lambda)$

Define the sequence of waiting times $S_n, n = 1, 2, \dots$ as:

$$\left(S_n := \sum_{i=1}^n T_i \right) \sim Gamma(n, \lambda)$$

The sequence of arrival times $S_i, i = 1, 2, \dots, n$ have the same distribution as an ordered sample of size n taken from $U(0, S_n)$

If we have a Poisson process $\{N(t), t \geq 0\}$ where we classify each event as type 1 with probability p and type 2 with probability $1-p$, then $\{N_1(t), t \geq 0\}$

and $\{N_2(t), t \geq 0\}$ are independent Poisson processes with parameters λp and $\lambda(1 - p)$

Let $\{N^{(1)}(t), t \geq 0\}$ and $\{N^{(2)}(t), t \geq 0\}$ be independent Poisson Processes with rates λ_1 and λ_2 . Also define the arrival times $S_n^{(1)}$ and $S_m^{(2)}$ as the arrival of the n th, m th event.

Then:

$$\mathbb{P}[S_n^{(1)} < S_m^{(2)}] = \sum_{i=n}^{n+m-1} \binom{n+m-1}{i} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n+m-1-i}$$

3.2.3 3041F: Modelling Poisson Processes

There are two forms for modelling a Poisson process. Either you decide on how many events you want and let the events dictate the timeframe, or you decide on the timeframe you want and take note of the events in that timeframe.

Generating n events Calculate the events by sampling from a uniform distribution and then calculating $T_i = -\frac{1}{\lambda} \ln(U_i(0, 1))$ for $i = 1, \dots, n$.

Generating events in timeframe Define the interarrival times as T_1, \dots, T_N where all interarrival times are in the interval $[0, T]$

3.2.4 3041F: Estimation of Poisson Processes

Again, there are two scenarios. Either generate n events, or generate events within a timeframe $[0, t]$.

If the first n events of a process are observed as the interarrival times t_1, \dots, t_n . Then the MLE of λ is:

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n t_i}$$

Alternatively, if there are N observed events in a given time interval $[0, T]$, then the MLE doesn't depend on the observed interarrival times and is:

$$\hat{\lambda}_{MLE} = \frac{N}{T}$$

Also, since $N \sim Poi(\lambda T)$, $\hat{\lambda}_{MLE}$ is an unbiased estimate of λ with variance $\frac{\lambda}{T}$.

3.3 3041F: Compound Poisson Processes

For regular poisson processes, every event is identical and increases the count by a certain amount. The events themselves aren't random, it's only the duration inbetween events which is random.

For Compound Poisson Processes, both the duration between events, and the amount by which our count will be changed is a random variable.

Let X_1, \dots, X_N be a sequence of iid random variables with PFG $\mathbb{G}_X(s)$ and N a non-negative integer valued random variable with PGF $\mathbb{G}_N(s)$.

Then the count S_N and the PGF thereof are:

$$S_N = \sum_{i=1}^N X_i$$

$$\mathbb{G}_{S_N}(s) = \mathbb{G}_N(\mathbb{G}_X(s))$$

The Expectation and Variance of S_N are given as a corollary:

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X]$$

$$\mathbb{V}[S_N] = \mathbb{E}[N]\mathbb{V}[X] + (\mathbb{E}[X])^2 \mathbb{V}[N]$$

Often the following result is useful, as we can go from a random sum of random variables S_n to just a simple counting process random variable $N(t)$:

$$\{S_n \leq t\} \iff \{N(t) \geq n\}$$

$$\mathbb{P}[S_n \leq t] = \mathbb{P}[N(t) \geq n]$$

We formally define a compound Poisson process as a stochastic process $\{X(t), t > 0\}$ that can be represented as the sum of iid random variables X_i which are also independant of the Poisson process:

$$X(t) = \sum_{i=1}^{N(t)} X_i$$

From this we can give the moments for a compound poisson process. With $X(t)$ as the compound poisson process made up of the random variables $X_i : i = 1, 2, \dots$:

$$\mathbb{E}[X(t)] = \lambda t \mathbb{E}[X]$$

$$\mathbb{V}[X(t)] = \lambda t \mathbb{E}[X^2]$$

3.4 3041F: Non-homogeneous Poisson Processes

Regular Poisson Processes don't care what time it is. The rate parameter λ is constant, it doesn't change. This means that the frequency of events is constant. Non-homogeneous Poisson processes replace the constant rate parameter λ with a function that depends on the time, called the intensity function $\lambda(t)$.

The non-homogeneous poisson postulates have to be rephrased slightly:

A counting process $\{N(t), t \geq 0\}$ is a Poisson process with intensity function $\lambda(t) > 0$ and $t > 0$ for small h , iff:

1. $N(0) = 0$
2. The process has independent increments (we drop the requirement for stationary increments)
3. $\mathbb{P}[N(t+h) - N(t) = 1] = \lambda(t)h + o(h)$

So the Probability of one occurrence within small timeframe t to $t + h$ is equal to the intensity (at time t) multiplied by h , plus some terms that go to zero.

4. $\mathbb{P}[N(t+h) - N(t) > 1] = o(h)$

So the Probability of more than one occurrence within small timeframe from t to $t + h$ goes to zero.

3.4.1 3041F: Mean Value Function

Define the mean value function $m(t)$ the integral from 0 to t of $\lambda(t)$:

$$m(t) = \int_0^t \lambda(u) du$$

And then the number of events in a time interval $(t, t + s]$ for a non-homogeneous Poisson process is $\sim Poi(m(t+s) - m(t))$, with colours added to make it clear that $m(t+s) - m(t)$ is just the rate parameter to a Poisson distribution.

In English, the number of events over a given interval is poisson distributed with a rate proportional to the total intensity over that interval. So

the probability looks like:

$$\mathbb{P}[N(t+s) - N(t) = n] = \frac{e^{-(m(t+s)-m(t))} \cdot (m(t+s) - m(t))^n}{n!}$$

$$\mathbb{P}[N(t) \geq n] = \sum_{j=n}^{\infty} \frac{m(t)^j e^{-m(t)}}{j!}$$

3.4.2 3041F: Simulation of a non-homogeneous Poisson Process

Suppose that events arrive according to a poisson process with rate λ , but we only consider some of the events, choosing events with probability $p(t)$.

This describes a non-homogeneous Poisson process with intensity function $\lambda(t) = \lambda \cdot p(t)$. We can therefore simulate a non-homogeneous Poisson process by first using our methods for homogeneous poisson processes, and then discarding events according to the probability function $p(t)$.

The Thinning algorithm

The thinning algorithm generates a non-homogeneous poisson process by finding an constant valued approximation of the rate function, generating a homogeneous poisson process according to this constant, and then discarding some events such that the approximation of the rate function becomes the actual rate function.

If we want to simulate over a time interval $[0, T]$ according to intensity function $\lambda(t) > 0$ first generate events according to a poisson process with rate λ over the interval $[0, T]$ to give the N interarrival times T_1, \dots, T_N . Then select λ_{max} such that

$$\lambda_{max} \geq \lambda(t) \quad \forall t \in [0, T]$$

Finally, create an indicator function with λ_{max} to select only those events where a Uniform random variable in the range $[0, \lambda_{max})$ is less than $\lambda(t)$:

$$U(0, \lambda_{max}) \leq \lambda(t) \quad \text{for } t = \sum_{i=1}^j T_i$$

Then all those selected events make up the required non-homogeneous poisson process.

Subintervals for simulations

The thinning algorithm will reject lots of events if the intensity function $\lambda(t)$ is very different from λ_{max} . We can fix this by dividing up the time interval into multiple smaller subintervals, and then running the thinning algorithm individually on each of these sub-intervals. Each time calculating a new and different value for λ_{max} .

Combining homogeneous Poisson Processes

Since the sum of events for two independant Poisson Processes with rates λ_1 and λ_2 is itself a Poisson Process with rate $\lambda_1 + \lambda_2$ we can use this to efficiently create piecewise Poisson processes where the intensity function $\lambda(t)$ is sometimes constant.

3.4.3 3041F: Estimation of non-homogeneous Poisson Processess

The PDF of the i th interarrival time given the waiting time of the $(i-1)$ th can be written as:

$$f_{T_i|W_{i-1}}(t) = \lambda(w_{i-1} + t) \cdot e^{-(m(w_{i-1}+t)-m(w_{i-1}))}$$

Note that the arrival times are independant due to the increments being independant.

Note that the probability of the i th interarrival time being greater than some value t (given the $(i-1)$ th arrival time) is equivalent to the probability of there being zero events between the $(i-1)$ th arrival time and the $(i-1)$ th arrival time plus t . This implies:

$$\begin{aligned} \mathbb{P}[T_i > t | W_{i-1} = w_{i-1}] &= \mathbb{P}[N(w_{i-1} + t) - N(w_{i-1}) = 0] \\ &= e^{-(m(w_{i-1}+t)-m(w_{i-1}))} \end{aligned}$$

Likelihood for non-homogeneous Poisson Processes

In general, explicit expressions for the MLEs do not exist, and likelihood maximisation must be done numerically. The notes give examples, but they all boil down to two different cases. The first case is when we have data about the first n arrivals, and can be calculated as so:

1. taking a given intensity function with unknown parameters θ and observed waiting times w_1, \dots, w_n ,
2. Calculating the log-likelihood function as

$$l(\theta) = \sum_{i=1}^n \ln(\lambda(w_i; \theta)) - (m(w_n; \theta) - m(w_0; \theta))$$

3. and then maximising l with respect to θ in order to find the MLEs for θ .

The second case is where we've measured all arrivals in an interval $[0, T]$, and can be calculated as shown below. It is different, because we need to account for the probability that we saw no additional events between our last observed event and T .

1. taking a given intensity function with unknown parameters θ and observed waiting times w_1, \dots, w_n . All waiting times should be within $[0, T]$.
2. Calculating the likelihood function as

$$L(\theta) = \prod_{i=1}^N \lambda(w_i; \theta) e^{-m(T; \theta)}$$

3. and then maximising L with respect to θ will give the MLEs of θ .

The notes also go into likelihood ratio tests for seeing if a particular Poisson process is non-homogeneous, but I'm not going to write that up here.

Chapter 4

Time Series I and Transformations

4.1 3041F: Time Series Analysis

This course will only look at measurements that are equally spaced in time. We will not be looking at unequal spacing (ie when you measure each event as it comes).

Time
Series
Video 1

A time series is denoted as $X_t, t \in T$ where T can be discrete or continuous.

The index t will be discrete in this course. X_t could be continuous, discrete, or (not in this course) qualitative.

We will often plot the data, try to see if there are any patterns or trends, and then plot summary statistics which could help with forecasting.

There are several objectives of time series analysis:

- Description: Summarizing the data in useful ways like expectation, variance, etc.
- Modelling: Creating a mathematical model of the data that closely predicts what has happened.
- Forecasting: Using the created model to give confidence intervals and estimates for where the process will be at a specified point in the future.

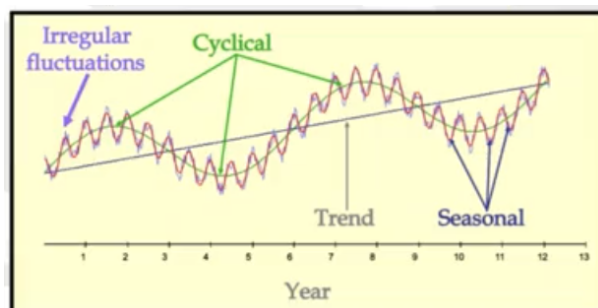


Figure 4.1: Comparison of Seasons vs Cycles

Forecasting is possible through historical data, and the existence of a pattern in that data which is expected to continue into the future.

- Control: Deciding if a given process is beyond certain safe operating bounds or if some extra factors need to be applied to the process in order to keep it within those safe operating bounds.

Time series are often modelled additively as a sequence of components:

Time
Series
Video 2

- Trend μ_t : The underlying expectation at time t of the time series. Often modelled as a simple function.
- Seasonality s_t : When there are repetitions in the data, with peaks and troughs.
- Cycles c_t : which is excluded in this course.
- Noise e_t : which represents the random fluctuations of the data.

Not every component will be included in every model. This can be modelled as

$$X_t = \mu_t + s_t + c_t + e_t$$

Or multiplicatively, although the multiplicative model will be log-transformed and then we'll just treat it as an additive model.

$$X_t = \mu_t \cdot s_t \cdot c_t \cdot e_t$$

$$\ln(X_t) = \ln(\mu_t) + \ln(s_t) + \ln(c_t) + \ln(e_t)$$

4.1.1 Trend

The long term growth/decline of a time series, modelled with a simple function. Basically just the expectation.

We can also look at the local trend, which just takes into account n prior points of the time series instead of all the data.

Some examples of Trend:

- $X_t = \epsilon_t$ with $\epsilon_t \sim N(0, a)$ all iid. Since model's expectation $\mu_t = \mathbb{E}[X_t] = 0$ doesn't depend on time, we can use the sample mean to estimate the population mean as all samples are from the same distribution.
- $X_t = t + \epsilon_t$ with $\epsilon_t \sim N(0, a)$ all iid. This model's expectation $\mu_t = \mathbb{E}[X_t] = t$ does depend on time, so the sample mean won't be a good estimate of the true mean.

4.1.2 Seasonality

When you have repeating trends in the data such as temperature increasing every summer or business flights decreasing over the weekend.

The amplitude of the seasonality is the difference between the peaks and troughs of the data. This amplitude might change over time. The amplitude will affect the variance of the overall, as a high amplitude will increase the variance.

4.1.3 Cycles

Basically seasonality on top of the existing seasonality. If seasonality repeats every year, then cycles repeat over decades.

Cycles are often used to refer to less regular oscillations in the data, such as bull and bear stock market trends which are less predictable than the annual trends produced by tax deadlines and such. We won't be covering the cyclical component in STA3041F.

4.1.4 Transformations

Because of the variability of the data, we might have to first transform the data such that our additive model is appropriate. Our additive model only

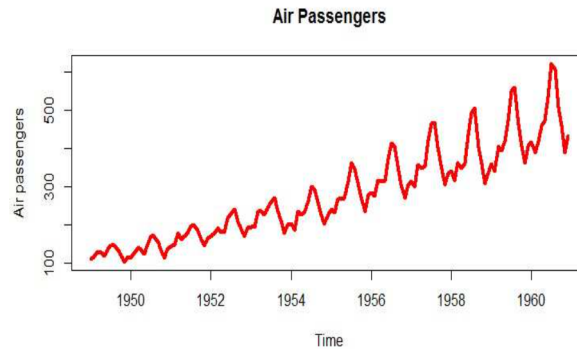


Figure 4.2: Airplane data

works with data that changes linearly in time, so if the seasonality amplitude increases exponentially then we'll have to transform the data to get rid of that exponential growth. The family of transformations we'll be using is the Box-Cox transformations.

4.2 3041F: Transformations

In figure 4.2, you can see the seasonal component in the repeating peaks and troughs. There is an upwards trend in the data and no cyclical component

There is an increasing amount of variance as time continues as shown by how the difference between the peaks and troughs. This implies that a multiplicative model might work well.

See figure 4.3. We can now subtract the long term trend (orange) from the root data (blue) in order to isolate just the error and seasonal components (assuming no cyclical component):

Sometimes you'll want to take the root data X_t and apply a transformation directly to it in an attempt to end up with linear data. Transformations might also be used to make the variance equal across the series. Possible transformations might be $\sqrt{X_t}$, $\sqrt[3]{X_t}$, $\log(X_t)$, or $-X_t^{-1}$.

However, the Box-Cox transformation can make our lives easier:

Time
Series
Video 3

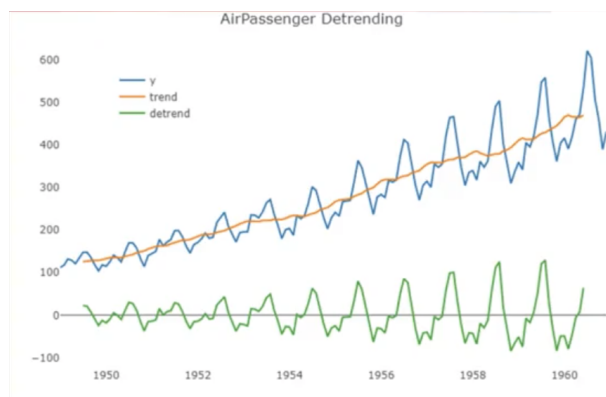


Figure 4.3: Airplane data with the trend subtracted.

4.2.1 3041F: Box-Cox Transformation

If our variance increases over time, then we want to account for this in our model. However, linear methods don't really like dealing with a changing variance. The solution is to take the original data, transform it in some way to remove/reduce how much the variance changes over time, and then proceed to use linear models on the transformed data.

Box-Cox is a class of transformations with parameter λ that tries to change reduce the amount the variance changes over time.

The original series x_1, \dots, x_n is transformed into $y_1(\lambda), \dots, y_n(\lambda)$ as follows:

$$y_t(\lambda) = \begin{cases} \frac{x_t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_t) & \text{if } \lambda = 0 \end{cases}$$

Where λ has to be estimated, and the resulting transformed data has a more stabilizer variance. The larger value of λ , the stronger the suppressing effect of the transformation. Note that taking the limit as λ trends to zero results in $\log(x_t)$, which is where the second case of the piecewise Box-Cox comes in.

Note that the Box-Cox tranformation is equivalent to many others, depending on the value of λ :

- When $\lambda = -1$: Inverse tranformation

Desmos
Inter-
active
graph
of the
 $\lambda \neq 0$
part of
Box-
Cox

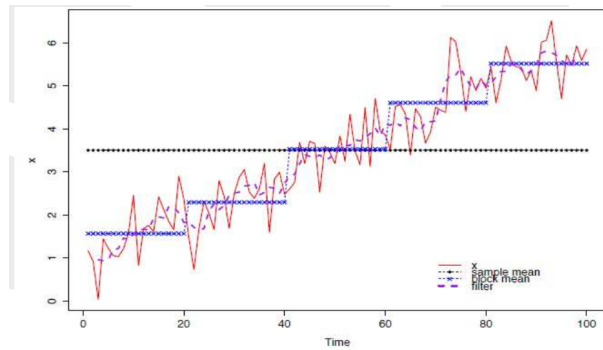


Figure 4.4: An example of moving average, block mean, and sample mean estimates of the trend of the data.

- When $\lambda = 0$: Natural Log transformation.
- When $\lambda = \frac{1}{2}$: (basically) Square root transformation.
- When $\lambda = \frac{1}{3}$: (basically) Cube root transformation.
- When $\lambda = 1$: (basically) No transformation.

4.2.2 3041F: Fitting Linear Filters

In time series analysis, we often need to remove the noise present in the data. Filtering the data is the processes of removing noise from the data, and there are different methods of filtering.

The sample mean of the data is often used to estimate the trend of that data. Although this assumes that the trend does not change over time (ie it's not good for increasing/decreasing data). An example of the sample mean is shown with the black +’s in Figure 4.4.

A better way of estimating the trend is to divide the data into smaller blocks, and calculate the sample mean for each of those blocks. Then use these sample means to be the trend, although this still isn’t great. An example of the block mean is shown with the blue x’s in Figure 4.4.

An even better way is via linear filters, one of which is the weighted average, an example of which is shown with the purple dashed line in Figure 4.4 (labelled ’filter’).

Time
Series
Video 4

You can see that the weighted average stays the closest to the data, while still ignoring some of the jagged peaks. You can think of it as smoothing the data while remaining close to it.

Linear filters can also be used to estimate seasonal components by varying various parameters that the filters take. These parameters change how smooth the filter makes the data, or how much attention it pays to local changes compared to larger scale changes.

We can transform our original time series data y_1, \dots, y_n to a filtered version $\tilde{y}_1, \dots, \tilde{y}_n$ by iterating over some of the items in y and multiplying them by a weight w .

We'll look at all the q items before the y_i we care about, as well as all the s items after that same y_i . Then we'll multiply them by a weight w_i and add them up:

$$\tilde{y}_t = \sum_{i=-q}^s w_i \cdot y_{t+i}$$

Note that the w_i 's all have to sum to 1 (so the sample mean stays the same), and usually we set $s = q$ (so we look at the same number of items behind us as we do ahead of us).

We can also show that if $\mathbb{E}[y_t] = \alpha + \beta t$ then the linear filter preserves this and $\mathbb{E}[\tilde{y}_t] = \alpha + \beta t$.

The linear filter won't always preserve a quadratic trend unless we get funky with how we choose our weights. For example, if we have an equation for y_t in the form:

$$y_t = \alpha + \beta t + \phi t^2 + e_t$$

Then the expectation of our linear filter version of y_t is given by

$$\mathbb{E}[\tilde{y}_t] = \alpha + \beta t + \phi t^2 + \phi \sum_i w_i i^2$$

So in order to preserve a quadratic trend, we'd need to choose the weights w_i such that the final term is equal to zero, ie:

$$0 = \sum_i w_i i^2$$

Estimating the weights of the linear filter can be done as follows:

1. Decide on how many weights you'll use. This is called the window size
2. Decide if your trend equation is linear, quadratic, cubic, etc.
3. Estimate the coefficients of the trend equation $y_t = \alpha + \beta t + \dots$ using least squares.
4. We can center the window such that the current y_t is at zero, which simplifies the calculations a bit.

Time Series Video 5 This video was a lot of math walk-through, which I'm not going to bother writing down.

Time
Series
Video 5
Time
Series
Video 6

4.2.3 3041F: Moving Averages

A moving average is a linear combination of current and past error terms. So we can define the moving average at time t as $\hat{\mu}_t = e_{t-n} + \dots + e_t$, where we call n the order of the moving average, often written as MA(n).

There are certain useful properties, but in order for them to hold we need the time series to not have a seasonal component and be represented by

$$y_t = \mu_t + e_t \quad \text{where } \mu_t \text{ is the mean at time } t$$

and the error terms $e_t \sim N(0, \sigma^2)$ are all independent. Also, define the smoothened trend at time t , $\hat{\mu}_t$ as:

$$\hat{\mu}_t = \sum_{i=-s}^s w_i \cdot y_{t+i}$$

Then the following properties hold:

$$\begin{aligned} \mathbb{V}[\hat{\mu}_t] &= \sigma^2 \sum_{i=-s}^s w_i^2 \\ \text{Cov}[\hat{\mu}_t, \hat{\mu}_{t+k}] &= \sigma^2 \sum_{i=-s}^{s-k} w_i \cdot w_{i+k} \\ \text{Corr}[\hat{\mu}_t, \hat{\mu}_{t+k}] &= \frac{\text{Cov}[\hat{\mu}_t, \hat{\mu}_{t+k}]}{\mathbb{V}[\hat{\mu}_t, \hat{\mu}_{t+k}]} \\ &= \frac{\sum_{i=-s}^{s-k} w_i \cdot w_{i+k}}{\sum_{i=-s}^s w_i^2} \end{aligned}$$

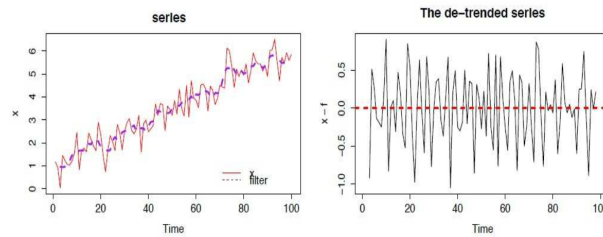


Figure 4.5: The detrended series has an expectation of zero if the estimate of the trend is unbiased.

It can also be proven that the variance of the moving average is less than the variance of the original data.

If we have a time series described by $y_t = \mu_t + e_t$ and the estimate for the trend is given by \tilde{y}_t , then we define the de-trended series as the difference $y_t - \tilde{y}_t$. If \tilde{y}_t is an unbiased estimate of the trend then $\mathbb{E}[y_t - \tilde{y}_t] = 0$.

An example of the original series and the de-trended series is shown in figure 4.5

Note that the seasonal component of a series can be removed using a moving average with a window size equal to the period of the season. This can be seen by looking at the MA(12) line in figure 4.6

Some comments about Moving Averages

- They smoothen out a time series (ie they reduce variation)
- They estimate the trend of a time series
- The transformed series will be correlated with itself, as t_i th point in the moving average is a linear combination of the t_{i-s}, \dots, t_{i+s} points in the original data.
- The above point means that even uncorrelated, completely random data will end up with correlations.
- We are unable to estimate the end points, as there are no data points in our window

4.2.4 Exponential Moving Averages

Exponential smoothing is another type of filter. We're effectively taking the average but giving a higher weighting to the most recent points. This filter

Time
Series
Video 7

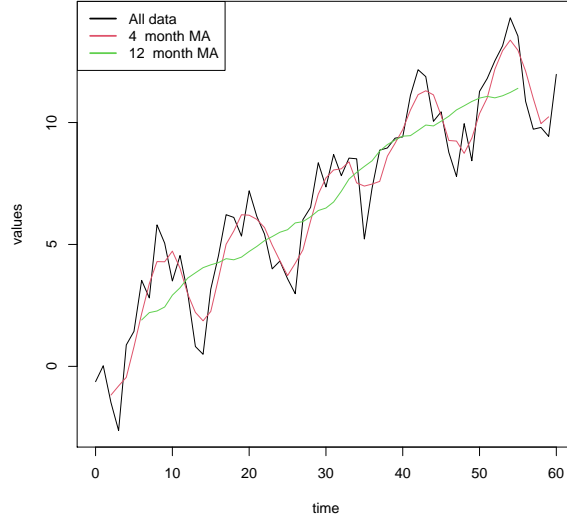


Figure 4.6: The seasonality can be removed by applying a moving average which has a window size of the period of the data.

is an asymmetric infinite filter, applying exponentially decreasing weights to values further back in time.

Explicitly, the weight w_i for the i th point in the data takes a parameter $0 < \alpha < 1$ and is calculated as:

$$w_i = \alpha^i(1 - \alpha)$$

Note that this means the sum of all weights $1 \dots \infty$ is equal to 1:

$$\sum_{i=0}^{\infty} w_i = 1$$

Single exponential smoothing is the process of using an exponentially weighted moving average. There are other exponential smoothing techniques, but we won't be looking at those.

The exponentially smoothed value at time t (s_t) is given by:

$$s_t = y_t(1 - \alpha) + \alpha \cdot s_{t-1}$$

So the smoothed value at time t depends on the smoothed value at time $t - 1$. We define the forecast for time $t + h$ as being conditional on all the points up to time t , and equal to the smoothed value s_t up until that point:

$$\hat{y}_{t+1|t} = s_t = y_t(1 - \alpha) + \alpha \cdot s_{t-1}$$

By assuming we've got an infinite number of historical values, we can get a closed expression for s_t .

$$s_t = \sum_{i=0}^{\infty} w_i \cdot y_{t-i}$$

However, in the real world we don't have infinite historical values. Often what we do is set s_0 equal to the first value of the time series, and then all following values (s_1, s_2, \dots) can be calculated off of that starting value.

About the value of alpha:

1. A large α will give lots of preference to prior values, leading to not very much smoothing. The weights decay very quickly.
2. A small α will give lots of preference to prior values, leading to more smoothing. A small α is used for very noisy data.

Basically, the choice of α will be chosen by trial and error, comparing different values according to some metric like Mean Squared Error, Mean Absolute Error, Sum of Squares Error, etc.

Chapter 5

Time Series II

5.1 3041F: Time Series II Intro

Time
Series
Video 8

We'll be covering:

- Stationarity, strict and weak variants.
- Autocorrelation: Measure the same variable but at different time points, and calculate the correlation between those timepoints
- Autocovariance
- Linear Time series models (moving average, autoregressive process, autoregressive moving average process)

5.1.1 Backshift Operator

We'll (eventually) need the **Backshift Operator**, so define it as:

1.

$$B^i y_t = y_{t-i}$$

2.

$$Bc = c$$

3.

$$(B^i + B^j)y_t = B^i y_t + B^j y_t = y_{t-i} + y_{t-j}$$

4.

$$B^i B^j y_t = B^j B^i y_t = B^i y_{t-j} = y_{t-j-i}$$

5. Also, specially define the following:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

6. For $|a| < 1$:

$$(1 + aB + a^2 B^2 + \dots) y_t = \frac{y_t}{1 - aB}$$

7. For $|a| > 1$:

$$(1 + \frac{1}{aB} + \frac{1}{a^2 B^2} + \dots) y_t = \frac{-aB y_t}{1 - aB}$$

The backshift operator becomes **really** useful when used with linear combinations of time series processes. For example, if we define V_t, W_t, Z_t as:

$$W_t := W_{t-1} - 0.16W_{t-2} + a_t \quad \text{Where } a_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$V_t := 0.8V_{t-1} + b_t \quad \text{Where } b_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$Z_t := W_t + V_t$$

Then we can find $Z_t \sim ARMA(2, 1)$ by first expressing W_t and V_t in terms of the backshift operator:

$$W_t := W_{t-1} - 0.16W_{t-2} + a_t \quad \text{Where } a_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$(1 - B + 0.16B^2)W_t = a_t$$

$$W_t = \frac{a_t}{1 - B + 0.16B^2}$$

And likewise for V_t :

$$V_t := 0.8V_{t-1} + b_t \quad \text{Where } b_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$= 0.8BV_t + b_t$$

$$(1 - 0.8B)V_t = b_t$$

$$V_t = \frac{b_t}{1 - 0.8B}$$

So now that we've gotten rid of all the $V_{t-1}, W_{t-1}, W_{t-2}$ terms, we can rewrite our original expression $Z_t := W_t + V_t$ in terms of just Z_t and the backshift

operator:

$$Z_t := W_t + V_t$$

$$Z_t = \frac{a_t}{1 - B + 0.16B^2} + \frac{b_t}{1 - 0.8B}$$

Which quickly yields to some manipulation to get a recognisable ARMA process:

$$Z_t := W_t + V_t$$

$$Z_t = \frac{a_t}{1 - B + 0.16B^2} + \frac{b_t}{1 - 0.8B}$$

$$Z_t = \frac{a_t}{(1 - 0.8B)(1 - 0.2B)} + \frac{b_t}{1 - 0.8B}$$

Multiply by $(1 - 0.8B)(1 - 0.2B)$

$$(1 - 0.8B)(1 - 0.2B)Z_t = a_t + (1 - 0.2B) \cdot b_t$$

$$(1 - B + 0.16B^2)Z_t = a_t + b_t - 0.2b_{t-1}$$

$$Z_t - Z_{t-1} + 0.16Z_{t-2} = a_t + b_t - 0.2b_{t-1}$$

$$Z_t = Z_{t-1} - 0.16Z_{t-2} + a_t + b_t - 0.2b_{t-1}$$

The last line is an ARMA(2,1) process with $a_t + b_t$ forming essentially one error term. So we can more easily manipulate combinations of time series by (ab)using the backshift operator.

5.2 Stationarity

5.2.1 Strict Stationarity

Strict stationarity is a property of a time series.

- A time series is strictly stationary iff for every offset parameter $m > 0$, all time points t_1, \dots, t_n , the joint distribution of Z_{t_1}, \dots, Z_{t_n} is equal to that of $Z_{t_1-m}, \dots, Z_{t_n-m}$.
- We also require that if there are two points, Z_t and Z_{t+m} then the distribution of the random vector (Z_t, Z_{t+m}) must be a function of the **lag** $|m|$.
- Given a timeline of a time series and an offset parameter m

- Consider Random Variables Z_{t_1}, \dots, Z_{t_n} , find the joint distribution of them all, then apply a timeshift by the offset parameter m : $Z_{t_1-m}, \dots, Z_{t_n-m}$.
- then if the joint distribution of the first set of Random variables is equal to that the joint distribution of the second set of random variables, then the time series is said to be **Strictly Stationary**.

5.2.2 Autocovariance

Define the **Autocovariance** as the covariance of the same random variable but at different time points:

$$\begin{aligned}\mathbb{Cov}[Z_t, Z_{t+m}] &:= \gamma_{|m|} \\ &= \mathbb{E}[Z_t \cdot Z_{t+m}] - \mathbb{E}[Z_t] \cdot \mathbb{E}[Z_{t+m}]\end{aligned}$$

Also note that:

$$\gamma_m = \gamma_{-m} \mathbb{V}[Z_t] = \gamma_0$$

We can use $|m|$ because Covariance doesn't depend on the order of the random variables given to it.

Define the **Autocorrelation** as the autocovariance divided by the variance of a time series. The autocorrelation (given the offset parameter m) ρ_m is defined by:

$$\begin{aligned}\rho_m &= \frac{\gamma_m}{\gamma_0} \\ \rho_0 &= \frac{\gamma_0}{\gamma_0} = 1\end{aligned}$$

5.2.3 Weak Stationarity

Weak stationarity is defined by orders (order 1, order 2, etc). So to have weak stationarity of order n , the time series must have equal n -th moments. For example, weak stationarity of order 1 requires a constant expectation.

Weak Stationarity of order 2 requires

- A constant expectation.

$$\mathbb{E}[Z_t] = \mathbb{E}[Z_{t-m}] \forall m$$

- the autocovariance to only depend on the offset parameter m . For example, where t and s are valid indices to the time series Z :

$$\mathbb{Cov}[Z_t, Z_{t-m}] = \mathbb{Cov}[Z_s, Z_{s-m}]$$

Time
Series
Video 9

Therefore Weak Stationary of order 2 implies Weak Stationarity of order 1.

5.2.4 Some examples of Weak stationarity

If we define a time series as $y_t = e_t - \frac{1}{2}e_{t-1}$ with all error terms independent and $e_t \sim N(0, \sigma^2)$ then:

Time
Series
Video
10

$$\begin{aligned}\mathbb{E}[y_t] &= \mathbb{E}[e_t - \frac{1}{2}e_{t-1}] = 0 + 0 = 0 \\ \gamma_0 &= \mathbb{E}\left[\left(e_t - \frac{1}{2}e_{t-1}\right)\left(e_t - \frac{1}{2}e_{t-1}\right)\right] - \mathbb{E}[y_t]\mathbb{E}[y_t] \\ &= \dots \\ &= \sigma^2(1 + 0.25) + 0 \cdot 0 \\ &= 1.25\sigma^2 \\ \gamma_1 &= \text{Cov}[y_t, y_{t-1}] \\ &= \mathbb{E}[(y_t)(y_{t-1})] - 0 \cdot 0 \\ &= \dots \\ &= -0.5\sigma^2 \\ \gamma_k &= \mathbb{E}[y_t \cdot y_{t-k}] - 0 \cdot 0 \\ &= \mathbb{E}\left[\left(e_t - \frac{1}{2}e_{t-1}\right)\left(e_{t-k} - \frac{1}{2}e_{t-k-1}\right)\right] \\ &= 0 - \frac{1}{2} \cdot 0 - \frac{1}{2} \cdot 0 + \frac{1}{4} \cdot 0 \\ &= 0\end{aligned}$$

5.2.5 More examples, introducing the autoregressive process

We'll introduce the autoregressive (AR) process as having parameter $|\phi| < 1$ and being a time series like:

Time
Series
Video
11

$$y_t = \phi y_{t-1} + e_t$$

Which implies that the expectation can be calculated as:

$$\mathbb{E}[y_t] = \dots = \phi \mathbb{E}[y_{t-1}]$$

So now we've got a self-referential statement where the expectation at time t depends on the expectation at time $t - 1$. This can collapse to become:

$$\begin{aligned}\mathbb{E}[y_t] &= \lim_{h \rightarrow \infty} \phi^h \mathbb{E}[y_{t-h}] \\ &= 0 \cdot \mathbb{E}[y_{t-h}] \\ &= 0\end{aligned}$$

And the variance will similarly be self-referential, and collapse to give us:

$$\mathbb{V}[y_t] = \dots = \phi^{2h} \mathbb{V}[y_{t-h}] + \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}$$

And so taking the limit as $h \rightarrow \infty$:

$$\begin{aligned}\mathbb{V}[y_t] &= \sigma^2 \sum_{i=0}^{h-1} \phi^{2i} \\ &= \frac{\sigma^2}{1 - \phi^2}\end{aligned}$$

5.3 Autocorrelation

Given a certain time series like $X_t = (x_1, \dots, x_n)$ we can calculate the autocorrelation for offset m like so:

$$\hat{\rho}_m = \text{Corr}[X_t, X_{t-m}] = \frac{\sum_{t=m+1}^n (X_t - \bar{X})(X_{t-m} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

A **Correlogram** is a plot with m on the x axis and the m -th lagged autocorrelations on the y axis.

5.4 Significance testing of Autocorrelation

Note that when you calculate autocorrelation with an offset of m , you'll lose exactly m data points. For example, in calculating the autocorrelation with $m = 1$, there is no Z_{t_0} data point to match up with the Z_{t_1} data point.

Today we look at the statistical significance of each of the m autocorrelations. If the underlying process is a moving average process, then the

Time
Series
Video
12

number of statistically significant autocorrelations is equal to the order of that moving average process.

Recall that the m -th autocorrelation is defined as:

$$\hat{\rho}_m = \frac{\sum_{t=m+1}^n (X_t - \bar{X})(X_{t-m} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

Now we can form a standard normal random variable by subtracting the mean and dividing by the sample variation. We have to find the distribution of $\hat{\rho}_m$ numerically, by simulating the histogram many many times (No easy analytical solution exists). After doing this, we find:

$$\hat{\rho}_m \sim N(0, \frac{1}{n})$$

And so we can construct a standard normal like so:

$$Z = \frac{\hat{\rho}_m - 0}{\frac{1}{\sqrt{n}}}$$

5.4.1 Hypothesis tests with Autocorrelation

We know that $\hat{\rho}_m \sim N(0, \frac{1}{n})$ (as $n \rightarrow \infty$). Thus if $|\hat{\rho}_m| > \frac{t_{\alpha/2}}{\sqrt{n}}$ we would reject $H_0 : \rho_m = 0$ at the α significance level (where $t_{\alpha/2}$ is the usual value from student's t-distribution).

The **Bartlett's Test** is the same as above, except we reject H_0 if $|\hat{\rho}_m| > \frac{2}{\sqrt{n}}$ because fuck accuracy.

The **Portmanteau test** allows us to also test more complicated hypotheses such as:

$$\begin{aligned} H_0 : \hat{\rho}_i &= 0 \quad \forall i \in \{1, 2, \dots, k\} \\ H_1 : \hat{\rho}_i &\neq 0 \quad \text{For at least one } i \in \{1, 2, \dots, k\} \end{aligned}$$

Recalling that the sum of squared independent standard normals is chi-square distributed, we define the Portmanteau test for offset m as:

$$Q(m) = n \underbrace{\sum_{i=1}^m \hat{\rho}_i^2}_{\sim \chi_m^2}$$

So now we've got a chi-squared distribution, we can look up the value of $Q(m)$ in the tables.

5.5 Linear Models for Time Series

5.5.1 White Noise Process

This process is defined as only having an error term e_t like:

$$y_t = e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

So we can calculate the moments as:

$$\begin{aligned}\mathbb{E}[y_t] &= 0 \\ \mathbb{V}[y_t] &= \sigma^2 \\ \gamma_1 &= \mathbb{Cov}[y_t, y_{t-1}] = \dots = 0 \\ \gamma_k &= \dots = 0\end{aligned}$$

5.5.2 The Random Walk Process

This is defined as being dependant on the previous time step:

$$y_t = y_{t-1} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

And we (usually) assume y_0 as:

$$y_0 = 0$$

Note that the errors are cumulative:

$$\begin{aligned}y_t &= y_0 + \sum_{i=1}^t e_i \\ &= \sum_{i=1}^t e_i\end{aligned}$$

Recall that if $X_1 \perp\!\!\!\perp X_2$ then:

$$\mathbb{E}[X_1 \cdot X_2] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2]$$

So the moments are given as:

$$\begin{aligned}\mathbb{E}[y_t] &= \dots = y_0 = 0 \\ \mathbb{V}[y_t] &= \dots = t\sigma^2 \\ \gamma_1 &= \mathbb{Cov}[y_t, y_{t-1}] = \dots = (t-1)\sigma^2 \\ &= \text{Note in general, if we have } s < t: \\ \mathbb{Cov}[y_t, y_s] &= s\sigma^2 \\ \gamma_k &= \dots = 0\end{aligned}$$

5.5.3 The Random Walk Process with Drift

This is defined as being dependant on the previous time step, in addition to some accumulative term:

$$y_t = y_{t-1} + \mu + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

And we (usually) assume y_0 as:

$$y_0 = 0$$

So the moments are given as:

$$\mathbb{E}[y_t] = \dots = y_0 + t\mu$$

$$\mathbb{V}[y_t] = \dots = t\sigma^2$$

$$\gamma_1 = \text{Cov}[y_t, y_{t-1}] = \dots = (t-1)\sigma^2$$

= Note in general, if we have $s < t$:

$$\text{Cov}[y_t, y_s] = s\sigma^2$$

Define the **difference operator** as having symbol ∇ and being like:

$$\nabla y_t := y_t - y_{t-1} = \underbrace{\mu + e_t}_{\substack{\text{For a} \\ \text{ran-} \\ \text{dom} \\ \text{walk} \\ \text{with} \\ \text{drift}}}$$

Often the difference operator can help create a stationary process from a non stationary one.

5.6 First Order Moving Average process

We'll start talking about first order, then second order, and then we'll abstract to talking about the q -th order moving average process. For now we talk about the first order moving average process and define it as:

$$Z_t = e_t - \theta e_{t-1} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Where θ is some weight that we can choose. Note that the sign between the two terms is irrelevant, as we can choose θ to counteract whatever sign we choose.

Time
Series
Video
14

We can then find the moments to be:

$$\begin{aligned}
\mathbb{E}[Z_t] &= 0 \\
\gamma_0 &= \mathbb{V}[Z_t] = \sigma^2(1 + \theta^2) \\
\gamma_k &= \mathbb{E}[Z_t \cdot Z_{t-k}] \\
&= \begin{cases} \sigma^2(1 + \theta^2) & \text{For } k = 0 \\ -\theta\sigma^2 & \text{For } k = 1 \\ 0 & \text{For } k > 1 \end{cases} \\
\rho_k &= \begin{cases} 1 & k = 0 \\ -\frac{\theta}{1+\theta^2} & k = 1 \\ 0 & \text{For } k > 1 \end{cases}
\end{aligned}$$

5.7 Second Order Moving Average Process

$$Z_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

We can find the moments to be:

$$\begin{aligned}
\mathbb{E}[Z_t] &= 0 \\
\gamma_k &= \mathbb{E}[Z_t \cdot Z_{t-k}] \\
&= \begin{cases} \sigma^2(1 + \theta_1^2 + \theta_2^2) & \text{For } k = 0 \\ \sigma^2\theta_1(\theta_2 - 1) & \text{For } k = 1 \\ -\theta_2\sigma^2 & \text{For } k = 2 \\ 0 & \text{For } k > 2 \end{cases} \\
\rho_k &= \begin{cases} 1 & k = 0 \\ \frac{\theta_1(\theta_2-1)}{1+\theta_1^2+\theta_2^2} & k = 1 \\ \frac{-\theta_2}{1+\theta_1^2+\theta_2^2} & k = 2 \\ 0 & k > 2 \end{cases}
\end{aligned}$$

5.8 Q-th Order Moving Average Process

$$Z_t = e_t - \sum_{i=1}^q \theta_i e_{t-i} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Time
Series
Video
15

We can find the moments to be:

$$\begin{aligned}\mathbb{E}[Z_t] &= 0 \\ \mathbb{V}[Z_t] &= \sigma^2 + \sigma^2 \sum_{i=1}^q \theta_i^2 \\ \gamma_k &= \mathbb{E}[Z_t \cdot Z_{t-k}] \\ &= \begin{cases} -\sigma^2 \theta_k + \sigma^2 \sum_{i=1}^q \theta_{k+i} \theta_i & \text{For } 1 \leq k \leq q \\ 0 & \text{For } k > q \end{cases} \\ \rho_k &= \begin{cases} \frac{-\theta_k + \sum_{i=1}^q \theta_{k+i} \theta_i}{1 + \sum_{i=1}^q \theta_i^2} & \text{For } 1 \leq k \leq q \\ 0 & \text{For } k > q \end{cases}\end{aligned}$$

Time
Series
Video
16

5.9 General Linear Process

General Linear Process: A weighted infinite sum of error terms, where we call the i th weight ψ_i .

$$\begin{aligned}\sum_{i=1}^{\infty} \psi_i^2 &< \infty \\ |\psi_i| &< 1 \\ \psi_0 &= 1 \\ Z_t &= \sum_{i=0}^{\infty} \psi_i \cdot e_{t-i}\end{aligned}$$

For example, we can set $\psi_i = (\frac{1}{2})^i$. This obeys the requirements.

With a question like **Identify this process**, you just need to say if it's a General Linear Process, a moving average process, etc.

With the example linear process like: $\psi_i = \phi^i$ and with $\phi_i \in (0, 1)$ we've got a fully define process given by:

$$\begin{aligned}
Z_t &= \sum_{i=0}^{\infty} \phi^i e_{t-i} \\
\gamma_0 &= \mathbb{E}[Z_t^2] \\
&= \sigma^2 \sum_{i=0}^{\infty} \phi^{2i} \\
&= \frac{\sigma^2}{1 - \phi^2} \gamma_k = \frac{\phi^k \sigma^2}{1 - \phi^2} \quad \forall k = 0, 1, \dots \\
\rho_k &= \phi^k \quad \forall k = 0, 1, \dots
\end{aligned}$$

An example of calculating the autocovariance γ_k :

$$\begin{aligned}
\gamma_k &= \text{Cov}[Z_t, Z_{t+k}] = \mathbb{E}[Z_t \cdot Z_{t+k}] \\
&= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \phi^i e_{t-i}\right) \left(\sum_{j=0}^{\infty} \phi^j e_{t+k-j}\right)\right] \\
&\quad \text{let } t - i = t + k - j \iff j = i + k \\
\gamma_k &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \phi^i e_{t-i}\right) \left(\sum_{i=0}^{\infty} \phi^{i+k} e_{t-i}\right)\right] \\
&= \phi^k \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \phi^i e_{t-i}\right)^2\right] \\
&= \frac{\phi^k \sigma^2}{1 - \phi^2}
\end{aligned}$$

Note that general linear processes are weakly stationary of order 2 since neither the mean nor the covariance depends on time t .

5.10 Infinite Moving Average Process

Very similar to the above General Linear Process:

$$Z_t = \sum_{i=0}^{\infty} \omega_i e_{t-i} \quad \omega_0 = 1$$

$$\mathbb{E}[Z_t] = 0$$

$$\gamma_0 = \lim_{T \rightarrow \infty} \sigma^2 \sum_{i=0}^T \omega_i^2$$

$$\gamma_k = \lim_{T \rightarrow \infty} \sigma^2 \sum_{i=0}^T \omega_i \omega_{k+i}$$

$$\rho_k = \lim_{T \rightarrow \infty} \frac{\sum_{i=0}^T \omega_i \omega_{k+i}}{\sum_{i=0}^T \omega_i^2}$$

Have previously looked at the moving average process. Will now be looking at forecasts for moving average processes

Time
Series
Video
17

5.10.1 Forecasts for Moving Average Processes

With the moving average process defined by:

$$Z_t = e_t - \phi e_{t-1} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

And now let's define the forecast value as \hat{Z}_{t+1} and we set it equal to be:

$$\begin{aligned} \hat{Z}_{t+1} &= \mathbb{E}[Z_{t+1} | \underbrace{Z_t, Z_{t-1} \dots Z_1}_{\text{Often called just } I_t, \text{ the information at } t}] \\ &= \mathbb{E}[e_{t+1} - \phi e_t | I_t] \\ &= \underbrace{\mathbb{E}[e_{t+1} | I_t]}_{\substack{e_{t-1} \text{ is} \\ \text{still a} \\ \text{random} \\ \text{variable}}} - \phi \underbrace{\mathbb{E}[e_t | I_t]}_{\substack{e_t \text{ has a} \\ \text{measured} \\ \text{value}}} \\ &= 0 - \phi e_t = -\phi e_t \end{aligned}$$

So the forecast for a moving average process is $\hat{Z}_{t+1} = -\phi e_t$.

We often also want to know **the margin of error**, so we define the one

step ahead forecast error $\hat{e}(1)$ as:

$$\begin{aligned}\hat{e}(1) &= Z_{t+1} - \hat{Z}_{t+1} \\ &= (e_{t+1} - \phi e_t) - (-\phi e_t) \\ &= e + t + 1\end{aligned}$$

Note that we can also calculate:

$$\mathbb{V}[\hat{e}(1)] = \sigma^2$$

Now, for a **MA(q) model**, we have:

$$\begin{aligned}\hat{Z}_{t+l} &= - \sum_{i=l}^q \phi_i e_{t-i+l} \quad \forall 1 \leq l \leq q \\ \hat{e}(l) &= Z_{t+l} - \hat{Z}_{t+l} \\ \mathbb{V}[\hat{e}(l)] &= \begin{cases} \sigma^2 & \text{for } l = 1 \\ \sigma^2 \left(1 + \sum_{i=1}^{l-1} \phi_i^2\right) & \text{for } 2 \leq l \leq q \\ \sigma^2 \left(1 + \sum_{i=1}^q \phi_i^2\right) & \text{for } l > q \end{cases}\end{aligned}$$

5.11 Autoregressive Process

A first order autoregressive process (AR(1)) can be defined as a weighted linear sum of the previous value in the time series Z_{t-1} and an error term e_t :

$$Z_t = \phi Z_{t-1} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Since Z_t depends on Z_{t-1} , this can be expanded into an infinite moving average process (assuming $|\phi| < 1$):

$$Z_t = \sum_{i=0}^{\infty} \phi^i e_{t-1}$$

So some properties are given by:

$$\begin{aligned}\mathbb{E}[Z_t] &= 0 \\ \mathbb{V}[Z_t] &= \frac{\sigma^2}{1 - \phi^2} \\ \gamma_k &= \frac{\phi^k \sigma^2}{1 - \phi^2} = \phi^k \gamma_0 \\ \rho_k &= \phi^k\end{aligned}$$

Time
Series
Video
18

We can also write the AR(1) process as:

$$\begin{aligned} e_t &= Z_t - \phi Z_{t-1} = (1 - \phi B)Z_t \\ &= \phi(B)Z_t \end{aligned}$$

Now we'll write the AR(2) process as:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Or alternatively:

$$(1 - \phi_1 B - \phi_2 B^2)Z_t = e_t$$

And if we assume the time series is weakly stationary of order 2, and that $\sum_{i=1}^2 \phi_i \neq 1$: (**WS(2)**) then:

$$\mathbb{E}[Z_t] = 0$$

$$\gamma_0 = \left(\frac{1 - \phi_2}{1 + \phi_2} \right) \frac{\sigma^2}{1 - \phi_2^2 - \phi_1^2}$$

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}$$

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}$$

Where $\rho_{-1} = \rho_1$ and $\rho_0 = 1$.

The above equations for γ_k and ρ_k are called the **Yule-Walker** equations.

5.11.1 The Characteristic Equation

Taking the formula for a AR(2) process, we can write it as:

$$\begin{aligned} Z_t &= \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + e_t \\ e_t &= Z_t \left(\underbrace{1 - \phi_1 B - \phi_2 B^2}_{\substack{\text{Equate this to} \\ \text{zero}}} \right) \end{aligned}$$

$$1 - \phi_1 B - \phi_2 B^2 = 0$$

$$\text{Let } x = \frac{1}{B}$$

$$x^2 - \phi_1 x - \phi_2 = 0$$

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

Time
Series
Video
19

Time
Series
Video
20

In general, an **AR(q) process is weakly stationary of order 2** iff the characteristic roots lie inside the unit circle in the complex plane, ie:

$$\begin{aligned}\phi_1 + \phi_2 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ |\phi_2| &< 1\end{aligned}$$

5.11.2 P-th order Autoregressive Process

The p-th order autoregressive process AR(p) can be written as a linear combination of the previous p time points, plus some random error.

$$Z_t = e_t + \sum_{i=1}^p \phi_i Z_{t-i} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Then the lag-k autocovariance γ_k can be found by:

$$\begin{aligned}\gamma_k &= \mathbb{E}[Z_t \cdot Z_{t-k}] \\ Z_t \cdot Z_{t-k} &= Z_{t-k} \cdot e_t + Z_{t-k} \cdot \sum_{i=1}^p \phi_i Z_{t-i} \\ \gamma_k &= \mathbb{E}[Z_{t-k} \cdot e_t] + Z_{t-k} \cdot \sum_{i=1}^p \phi_i \mathbb{E}[Z_{t-i}] \\ \gamma_k &= \mathbb{E}[Z_{t-k} \cdot e_t] + \mathbb{E} \left[Z_{t-k} \cdot \sum_{i=1}^p \phi_i Z_{t-i} \right]\end{aligned}$$

And now we can label the following equations as the Yule-Walker Equations:

$$\begin{aligned}\gamma_k &= 0 + \sum_{i=1}^p \phi_i \gamma_{k-i} \\ \rho_k &= \sum_{i=1}^p \phi_i \rho_{k-i}\end{aligned}$$

5.11.3 Finding the Order of an Autocorrelation Model

The aim is to use data we've measured to find the best value for p , and to evaluate how good this best value is. We also define

- **ACF**: Autocorrelation Function and
- **PACF**: Partial Autocorrelation Function

And then the lag- i partial autocorrelation is defined as ϕ_{ii} with the double subscript to differentiate from the coefficients of the autoregressive process, ϕ_i .

The partial autocorrelation between Z_t and Z_{t-k} is the correlation between Z_t and Z_{t-k} but with all the in between correlations from $Z_{t-1}, \dots, Z_{t-k-1}$ removed. This is done by subtracting a linear combination of all the in between time series points, with the coefficients are assigned as $\alpha_1, \dots, \alpha_{k-1}$. To make this explicit:

$$\phi_{ii} = \text{Corr} \left[Z_t - \sum_{i=1}^{k-1} \alpha_i Z_{t-i}, \quad Z_{t-k} - \sum_{i=1}^{k-1} \alpha_i Z_{t-i} \right]$$

Mean Square Error estimation of ϕ_{ii}

To solve for ϕ_{ii} , note how we can use FOIL when calculating covariance:

$$\text{Cov} [A - B, C - D] = \text{Cov} [A, C] - \text{Cov} [A, D] - \text{Cov} [B, C] + \text{Cov} [B, D]$$

But before we can use the above formula, we need estimates for α_i .

We can define the MSE as the minimum value with respect to α of:

$$\begin{aligned} \text{MSE} &= \mathbb{E} [(Z_t - \alpha Z_{t-1})^2] \\ &= \mathbb{E} [Z_t^2 - 2\alpha Z_t Z_{t-1} + \alpha^2 Z_{t-1}^2] \\ &= \gamma_0 - 2\alpha \gamma_1 + \alpha^2 \gamma_0 \end{aligned}$$

Taking the derivative with respect to α :

$$\begin{aligned} \frac{d\text{MSE}}{d\alpha} &= 0 \\ -2\gamma_1 + 2\alpha\gamma_0 &= 0 \\ \alpha &= \frac{2\gamma_1}{2\gamma_0} \\ &= \rho_1 \end{aligned}$$

So we now can calculate ϕ_{22} as:

$$\begin{aligned}
 \phi_{22} &= \text{Corr}[Z_t - \alpha Z_{t-1}, Z_{t-2} - \alpha Z_{t-1}] \\
 &= \frac{\gamma_2 - \rho_1 \gamma_1 - \rho_1 \gamma_1 + \rho_2^2 \gamma_0}{\mathbb{V}[Z_t - \rho_1 Z_{t-1}]} \\
 &= \frac{\gamma_2 - 2\rho_1 \gamma_1 + \rho_1^2 \gamma_0}{\gamma_0 - 2\rho_1 \gamma_1 + \rho_1^2 \gamma_0} \\
 \text{Note } \rho_1 &= \phi, \text{ and } \rho_2 = \phi \rho_1 &= \frac{\rho_1 - 2\rho_1^2 + \rho_1^2}{1 - 2\rho_1^2 + \rho_1^2} \\
 &= \frac{\phi^2 - 2\phi^2 + \phi^2}{1 - 2\phi^2 + \phi^2} \\
 &= 0
 \end{aligned}$$

So the lag-2 partial autocorrelation is equal to zero. This is because the autoregressive process is of order 1, **AR(1)** so any lag-n partial autocorrelation with n greater than 1 will be zero. 31 minutes Today:

Time
Series
Video
21

1. calculating ϕ_{33}
2. Forecasting for an AR Process
3. How to identify an AR Process via EDA with measured data.

Recall that the number of statistically significant partial autocorrelations is equal to the order of the AR process. Then the parameters of the AR process can be estimated via least squares.

5.11.4 Calculating lag-3 partial autocorrelations of a **AR(1)** process

With a model like:

$$Z_t = \phi Z_{t-1} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Theorem: Any partial autocorrelation of order greater than the order of the AR process will be zero

So ϕ_{11} is the only non zero of $\phi_{11}, \phi_{22}, \dots$

Where α_i and β_i are unknown constants of the model that we will estimate via the values that minimise the mean squared error.

$$\phi_{33} = \text{Corr} [Z_t - \alpha_1 Z_{t-1} - \alpha_2 Z_{t-2}, Z_{t-3} - \beta_1 Z_{t-1} - \beta_2 Z_{t-2}]$$

We define the MSE as the minimum value with respect to α of:

$$\begin{aligned} \text{MSE} &= \min_{\alpha} [\mathbb{E} [(Z_t - \alpha_1 Z_{t-1} - \alpha_2 Z_{t-2})^2]] \\ &= \text{Expand the trinomial...} \\ &= \min_{\alpha} [\gamma_0 + \alpha_1^2 \gamma_0 + \alpha_2^2 \gamma_0 - 2\alpha_1 \gamma_1 - 2\alpha_2 \gamma_2 + 2\alpha_1 \alpha_2 \gamma_1] \end{aligned}$$

Setting the derivative to zero

$$\begin{aligned} \frac{d\text{MSE}}{d\alpha_1} &= 0 \\ &= 2\alpha_1 \gamma_0 - 2\gamma_1 + 2\alpha_2 \gamma_1 \\ &= 2\alpha_1 \gamma_0 + \gamma_1 (-2 + 2\alpha_2) \end{aligned}$$

$$\begin{aligned} \frac{d\text{MSE}}{d\alpha_2} &= 0 \\ &= 2\alpha_2 \gamma_0 - 2\gamma_2 + 2\alpha_1 \gamma_1 \end{aligned}$$

Use the two equations to simultaneously find values for α_1 and α_2

Then we can find ϕ_{33} as :

$$\phi_{33} = \text{Corr} [Z_t - \alpha_1 Z_{t-1} - \alpha_2 Z_{t-2}, Z_{t-3} - \beta_1 Z_{t-1} - \beta_2 Z_{t-2}]$$

Now note that we can calculate correlation using FOIL like:

$$\begin{aligned} \text{Corr} [a_1 X + b_1 Y, a_2 Z + b_2 T] &= \frac{\text{Cov} [a_1 X + b_1 Y, a_2 Z + b_2 T]}{\sqrt{\mathbb{V} [a_1 X + b_1 Y]}} \\ &= \frac{a_1 a_2 \text{Cov} [X, Z] + a_1 b_2 \text{Cov} [X, T] + b_1 a_2 \text{Cov} [Y, Z] + b_1 b_2 \text{Cov} [Y, T]}{\gamma_0} \end{aligned}$$

for the second line, we assume that the two variables $a_1 X + b_1 Y$ and $a_2 Z + b_2 T$ have equal variance (ie the process is **WS(2)**).

5.11.5 Using Real Data to make a model and then forecast

If you have measured data points z_1, \dots, z_n . Via some glossed over EDA, God tells us that this data comes from an autoregressive process of order 3. So

we can write a model for the data like:

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + e_t$$

Now to find values for the parameters, we'll do regression analysis and the Ordinary Least Squares analysis.

Note that because the model dictates the data points are not completely independent of each other, we require the error term to be a white noise process for the OLS analysis to hold mathematically. We also won't be able to use the first and last few data points as the model requires the $(t - 3)$ th data point to estimate the t -th data point.

5.11.6 Checking the validity of a AR model

We'll be looking at the residuals and checking our assumption that they come from a white noise process. We'll calculate the autocorrelations of these white noise processes and assert that there are no statistically significant autocorrelations.

Time
Series
Video
22

5.11.7 Forecasting with an AR process

With the model given by:

$$Z_t = \phi Z_{t-1} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

The **K-step ahead forecast** is defined as:

$$\hat{Z}_{t+k} = \mathbb{E}[Z_{t+k}|I_t]$$

1-step ahead forecast

For example, the 1-step ahead forecast is given by:

$$\hat{Z}_{t+1} = \mathbb{E}[Z_{t+1}|Z_t, Z_{t-1}, \dots, Z_1]$$

So now we'll calculate Z_{t+1}

$$Z_{t+1} = \phi Z_t + e_{t+1}$$

$$\hat{Z}_{t+1} = \mathbb{E}[\phi Z_t + e_{t+1}|I_t]$$

$$= \phi Z_t + \underbrace{\mathbb{E}[e_{t+1}]}_{\text{equal to zero}}$$

$$= \phi Z_t$$

So now that we've got \hat{Z}_{t+1} , let's find the variance of it. We'll define the **one-step ahead forecast error** $\hat{e}(1)$ as:

$$\begin{aligned}\hat{e}(1) &= Z_{t+1} - \hat{Z}_{t+1} \\ &= \phi Z_t + e_{t+1} - \phi Z_t \\ &= e_{t+1}\end{aligned}$$

And we can calculate the variance of the one-step ahead forecast error as:

$$\begin{aligned}\mathbb{V}[\hat{e}(1)] &= \mathbb{V}[e_{t+1}] \\ &= \sigma^2\end{aligned}$$

2-step ahead forecast

$$\begin{aligned}\hat{Z}_{t+2} &= \mathbb{E}[Z_{t+2}|I_t] \\ &= \mathbb{E}[\phi Z_{t+1} + e_{t+2}|I_t] \\ &= \phi \underbrace{\mathbb{E}[Z_{t+1}|I_t]}_{\substack{\text{1-step} \\ \text{ahead} \\ \text{forecast}}} + \underbrace{\mathbb{E}[e_{t+2}|I_t]}_{\substack{\text{R.V. is} \\ \text{zero}}} \\ &= \phi Z_{t+1} + 0\end{aligned}$$

So the two step ahead forecast is just the one-step ahead forecast times ϕ . The error of the forecast, and the variance thereof is given by:

$$\begin{aligned}\hat{e}(2) &= Z_{t+2} - \hat{Z}_{t+2} \\ &= \phi Z_{t+1} + e_{t+2} - \phi \hat{Z}_{t+1} \\ &= \phi (Z_{t+1} - \hat{Z}_{t+1}) + e_{t+2} \\ &= \phi \hat{e}(1) + e_{t+2} \\ \mathbb{V}[\hat{e}(2)] &= \mathbb{V}[\phi \hat{e}(1) + e_{t+2}] \\ &= \phi^2 \sigma^2 + \sigma^2 \\ &= \sigma^2(1 + \phi^2)\end{aligned}$$

So the variance of the 1-step forecast error was σ^2 and then for the 2-step forecast error, the variance grew to $\sigma^2(1 + \phi^2)$ (where $0 < \phi^2 \leq \phi < 1$)

5.11.8 Sections in the notes that will not be covered

- The unit root process.

5.11.9 AR(2) Process forecasting

We'll be doing 1, 2, and 3 step forecasting for a AR(2) process given by:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

So we'll calculate the 1 step forecast by:

$$\begin{aligned}\hat{Z}_{t+1} &= \mathbb{E}[Z_{t+1}|I_t] \\ &= \mathbb{E}[\phi_1 Z_t + \phi_2 Z_{t-1} + e_{t+1}|I_t] \\ &= \phi_1 Z_t + \phi_2 Z_{t-1} + 0 \\ \hat{e}(1) &= Z_{t+1} - \hat{Z}_{t+1} = e_{t+1} \\ \mathbb{V}[\hat{e}(1)] &= \sigma^2\end{aligned}$$

And the 2-step ahead forecast as:

$$\begin{aligned}\hat{Z}_{t+2} &= \mathbb{E}[Z_{t+2}|I_t] \\ &= \mathbb{E}[\phi_1 Z_{t+1} + \phi_2 Z_t + e_{t+2}|I_t] \\ &= \phi_1 \hat{Z}_{t+1} + \phi_2 Z_t + 0 \\ \hat{e}(2) &= Z_{t+2} - \hat{Z}_{t+2} \\ &= \phi_1 \hat{e}(1) + e_{t+2} \\ \mathbb{V}[\hat{e}(2)] &= \sigma^2 (1 + \phi_1^2)\end{aligned}$$

And the 3-step ahead forecast as:

$$\begin{aligned}Z_{t+3} &= \phi_1 Z_{t+2} + \phi_2 Z_{t+1} + e_{t+3} \quad \hat{Z}_{t+3} = \mathbb{E}[Z_{t+3}|I_t] \\ &= \phi_1 \hat{Z}_{t+2} + \phi_2 \hat{Z}_{t+1} + 0 \\ \hat{e}(3) &= Z_{t+3} - \hat{Z}_{t+3} \\ &= \phi_1 \hat{e}(2) + \phi_2 \hat{e}(1) + e_{t+3} \\ \mathbb{V}[\hat{e}(3)] &= \phi_1^2 \sigma^2 (1 + \phi_1^2) + \phi_2^2 \sigma^2 + \sigma^2 \\ &= \sigma^2 (1 + \phi_1^2 + \phi_1^4 + \phi_2^2)\end{aligned}$$

So (intuitively) as we try to forecast further into the future, the variance increases.

5.11.10 Forecasting with a AR(p) process

The model looks like:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$= e_t + \sum_{i=1}^p \phi_i Z_{t-i}$$

And the forecast value will look like:

$$\hat{Z}_{t+1} = \mathbb{E}[Z_{t+1}|I_t]$$

$$= \mathbb{E}\left[e_{t+1} + \sum_{i=1}^p \phi_i Z_{t-i+1} | I_t\right]$$

$$= \sum_{i=1}^p \phi_i Z_{t-i+1}$$

This equation is given without much explanation really.

$$Z_{t+l} = e_{t+l} + \sum_{i=1}^p \phi_i Z_{t+l-i} \hat{Z}_{t+l} = \sum_{i=1}^p \phi_i Z_{t+l-i}$$

5.11.11 Stationarity and Invertibility

Consider the AR process:

$$Z_t = \phi Z_{t-1} + e_t \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

If $|\phi| < 1$ then we can rewrite the model with the backshift operator like:

$$Z_t = \phi B Z_t + e_t$$

$$\underbrace{(1 - \phi B)Z_t}_{\text{The characteristic polynomial}}$$

And so we can get the **Characteristic equation** by substituting $B = \frac{1}{x}$:

$$0 = 1 - \phi B$$

$$1 - \frac{\phi}{x} = 0$$

$$x - \phi = 0$$

So our root is $x = \phi$. If this root lies inside the unit circle in the complex plane, then the time series model is stationary.

Invertible: we can take the process and rewrite it as an infinite moving average process.

In this case, our AR process is invertible. Let's write it as an infinite moving average process:

$$\begin{aligned}
 (1 - \phi B)Z_t &= e_t \\
 Z_t &= \underbrace{\frac{e_t}{1 - \phi B}}_{\substack{\text{Note} \\ |\phi| < 1}} \\
 Z_t &= \sum_{i=0}^{\infty} (\phi B)^i e_t \\
 Z_t &= \sum_{i=0}^{\infty} \phi^i B^i e_t \\
 Z_t &= \sum_{i=0}^{\infty} \phi^i e_{t-i}
 \end{aligned}$$

So to check invertibility:

1. Get the characteristic roots
2. Check they lie inside the unit circle
3. Therefore they are stationary
4. Then check if you can write the process as an infinite moving average process.

5.12 The Autoregressive Moving Average Process

5.12.1 Intro to ARMA(p,q)

The ARMA process has both Autoregressive components and moving average components. And ARMA(p,q) process is shown below, where p is the order

Time
Series
Video
24

of the order of the autoregressive component. q is the order of the moving average component:

$$Z_t = \phi Z_{t-p} + e_t - \theta e_{t-q} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Assuming the model is weakly stationary of order 2, we find the various properties by:

$$\begin{aligned} Z_t Z_{t-1} &= \phi Z_{t-1}^2 + e_t Z_{t-1} - \theta e_{t-1} Z_{t-1} \\ \mathbb{E}[Z_t Z_{t-1}] &= \gamma_1 = \phi \gamma_0 + 0 - \theta \sigma^2 \\ &= \phi \gamma_0 - \theta \sigma^2 \\ \gamma_2 &= \text{calculated in a similar way to } \gamma_1 \\ \mathbb{V}[Z_t] &= \mathbb{E}[Z_t^2] \\ &= \mathbb{E}[\phi^2 Z_{t-1}^2 + e_t^2 + \theta^2 e_{t-1}^2 + 2\phi Z_{t-1} e_t - 2\theta \phi e_{t-1} Z_{t-1} - 2\theta e_t e_{t-1}] \\ &= \phi^2 \underbrace{\mathbb{E}[Z_{t-1}^2]}_{\substack{\text{due to} \\ \text{WS}(2)}} + \sigma^2 + \theta^2 \sigma^2 - 2\theta \phi \sigma^2 \\ \mathbb{V}[Z_t] - \phi^2 \mathbb{V}[Z_t] &= \sigma^2 + \theta^2 \sigma^2 - 2\theta \phi \sigma^2 \\ \mathbb{V}[Z_t] &= \frac{\sigma^2(1 + \theta^2 - 2\theta \phi)}{1 - \phi^2} \end{aligned}$$

Exact expressions for variance, autocovariance, and autocorrelation are left as an exercise for the reader.

5.12.2 Identifying the order of an ARMA process

It's not super simple to identify the order(s) of an ARMA process. In general we tentatively find the parameters of the ARMA process, often splitting it up into the AR and MA processes and then solving them separately. Finally we'd join them together and iteratively evaluate our proposed model against a (possibly better) alternative model.

When identifying the order(s) of an ARMA process, we often want to try fit the data to an AR process and MA process separately, and then use those values as a starting point for fitting the ARMA process.

Time
Series
Video
25

5.12.3 Forecasting with an ARMA Process

With the first order ARMA process like:

$$Z_t = \phi Z_{t-1} + e_t - \theta e_{t-1} \quad \text{Where } e_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

So the 1-step forecast can be found like:

$$\begin{aligned} \hat{Z}_{t+1} &= \mathbb{E}[Z_{t+1}|I_t] \\ &= \mathbb{E}[\phi Z_t + e_{t+1} - \theta e_t | I_t] \\ &= \phi Z_t + 0 - \theta e_t \\ \hat{e}(1) &= Z_{t+1} - \hat{Z}_{t+1} = e_{t+1} \\ \mathbb{V}[\hat{e}(1)] &= \sigma^2 \end{aligned}$$

And the 1-step forecast can be found like:

$$\begin{aligned} \hat{Z}_{t+2} &= \mathbb{E}[Z_{t+2}|I_t] \\ &= \mathbb{E}[\phi Z_{t+1} + e_{t+2} - \theta e_{t+1} | I_t] \\ &= \phi \mathbb{E}[Z_{t+1}|I_t] + \mathbb{E}[e_{t+2}|I_t] - \theta \mathbb{E}[e_{t+1}|I_t] \\ &= \phi \hat{Z}_{t+1} + 0 + 0 \\ \hat{e}(2) &= Z_{t+2} - \hat{Z}_{t+2} = \phi \hat{e}(1) + e_{t+2} - \theta e_{t+1} \\ \mathbb{V}[\hat{e}(2)] &= \phi^2 \sigma^2 + \sigma^2 + \theta^2 \sigma^2 \\ &= \sigma^2 (\phi^2 + 1 + \theta^2) \end{aligned}$$

Note that it can be shown that as we forecast into the future, the forecast value converges onto the expectation of the time series:

$$\lim_{l \rightarrow \infty} \hat{Z}_{t+l} = \mathbb{E}[Z_t]$$

And there's a similar result for the variance of the error:

$$\lim_{l \rightarrow \infty} \mathbb{V}[\hat{e}(l)] = \mathbb{V}[Z_t]$$

5.13 The ARIMA Process

This is the first of our models which is non-stationary. It stands for an Autoregressive Integrated Moving Average Process. Our overall tactic here

will be to transform ARIMA processes into stationary time series so that we can analyse it more easily. It is parametrised by ARIMA(p, d, q) where the d is degree of integration.

In practice, many time series models are non-stationary. The ARIMA process is denoted by

5.13.1 Difference Operator

Define the ∇ operator as:

$$\begin{aligned}\nabla Z_t &:= Z_t - Z_{t-1} \\ \nabla^n Z_t &:= \nabla^{n-1} (\nabla Z_t) \\ \nabla^2 Z_t &= Z_t - 2Z_{t-1} + Z_{t-2}\end{aligned}$$

5.13.2 Making a ARIMA process stationary

We'll be using the ∇ operator to convert a non-stationary time series into a stationary time series. If the time series W_t given by:

$$W_t = \nabla Z_t = Z_t - Z_{t-1}$$

is stationary, then Z_t is from an ARIMA(p, 1, q) process.

Or in general, if you have to use the ∇ difference operator d times in order to convert the non-stationary process Z_t into a stationary process W_t , then we say $W_t \sim \text{ARIMA}(p, d, q)$.

Note that sometimes the MA or AR component will fall away in the process of taking the difference:

- If the differenced, now stationary process W_t only has an MA component (and no AR component) then we call Z_t an IMA(0,d,q) process, Integrated Moving Average process.
- If the differenced, now stationary process W_t only has an AR component (and no MA component) then we call Z_t an IAR(p,d,0) process, Integrated Autoregressive process.

5.13.3 Finding the order of an ARIMA Process

The ARIMA process has three orders: p, d, q. We can look at the autocorrelation function (ACF) and partial autocorrelation function (PACT), and if

the time series is stationary then a graph of lag vs correlation will show that as the lag increases, the autocorrelation goes to zero.

- **Stationary Series:** If the autocorrelations go to zero as the lag increases, then the series is probably stationary.
- **Bartlett's test:** $\frac{2}{\sqrt{n}}$ can tell you how many autocorrelations are statistically significant.

The above will determine if the current time series is stationary or not. If it is stationary, then all is good and you know that the order of d is 1. If it's not stationary, then you need to use the difference operator ∇Z_t to get a new time series, and then perform the same tests on this differenced time series. The number of times you have to use the difference ∇ operator is the order d of the data.

Once we've got a stationary process, then we can try find p and q by treating it as an ARMA(p,q) model and examining it as in the ARMA section above.

5.14 Box-Jenkins Algorithm

This is a high level method for processing a given time series Z_t and attempting to build a model for it.

1. If Z_t is not stationary, difference ∇ it until you've got a stationary process.
2. Examine the ACF and PACF of the (now stationary) time series using the Bartlett's test to test for statistical significance.
3. From the number of statistically significant ACFs and PACFs, tentatively model the data with the ARMA(p,q) process. p is the number of statistically significant ACFs, q is the number of statistically significant PACFs,
4. Evaluate the tentative model with Ordinary Least Squares, Method of Maximum Likelihood, or Method of Moments estimates.
5. Then check for **serial autocorrelation** to make sure that the error terms actually are independent and identically distributed.

6. If the errors aren't iid, then we need to adjust the parameters for p and q until they are.
7. Also check that the parameters of the model (ϕ_i, θ_i , etc) are statistically significant with the usual t-tests.
8. Now that the errors are iid, we can use the model to forecast.

Note that differencing too many times can lead to unwanted correlations. For instance, for the model $Z_t = Z_{t-1} + e_t$ then the difference is given by:

$$\nabla Z_t = Z_t - Z_{t-1} = e_t$$

This is a stationary series, and we should stop here.

But if we continue:

$$\nabla(\nabla Z_t) = e_t - e_{t-1}$$

we get a MA process looking like:

$$Z_t = e_t + \theta e_{t-1}$$

So over differencing can erroneously introduce parameters (like θ above) which aren't actually part of the underlying model.

Chapter 6

Terms, Definitions, R trickery, and side-notes

6.1 Terms and Definitions

Absorbing Once you're in an absorbing state, it's impossible to leave.

Accessible State j is accessible from state i if there exists n greater than 0 such that the probability of going from j to i in n steps is greater than zero.

Autocovariance Covariance of a given random variable with itself, but at different points in time.

Closed Sets A subset of the set of all possible states, which entered cannot be exited.

Communicate States i and j communicate if they are both accessible from each other. By convention, every state communicates with itself.

Cross-sectional Data All features are collected at a single period in time. All measurements are independent

Decomposition Theorem The state space can be partitioned uniquely into one set of transient states (although this need not be an equivalence class), and several closed irreducible sets of recurrent states.

Equivalence Class A set of states in which all states communicate.

Ergodic A class that's both aperiodic and positively persistent.

Interarrival Time The duration between event i and $i+1$, or the sequence of all such events.

Irreducible Chain When the state space of the chain is an irreducible class.

Irreducible Class A set that's both closed and an equivalence class

Limiting Distribution If a Markov Chain has a limiting distribution, then after enough time steps the chain will end up at this distribution regardless of starting state.

Periodic A state i is periodic with period k if the probability of first passage back to i in some number of steps n is equal to zero for all $n \% k \neq 0$.

Persistent, null A recurrent state in a chain with infinite total states and therefore the mean recurrence time infinite.

Persistent, positively A recurrent state in a chain with finite total states and therefore the mean recurrence time is less than infinity.

Recurrent State Starting at a recurrent state i , and given infinite time, you will return to state i with probability 1.

Recurrent An equivalence class with 1 or more recurrent states.

Regular Chain Given enough steps, you can visit every state, regardless of which state you start out at. This implies W^n only has positive entries.

Reversibility A stochastic matrix is reversible with respect to a given distribution if stepping forwards or backwards in time has no effect on the distribution.

Stationary Distribution When the marginal distribution doesn't change over time.

Time Series Data Features are measured at multiple points in time. Measurements in the future are often dependant on the measurements in the past.

Time Series A particular realisation of a stochastic process.

Transient State Starting at transient state i , the average time until first passage back to i is infinite, and we say the passage is uncertain.

6.2 R trickery

Useful functions:

6.2.1 runif

`runif(n)`: list of 'n' Random uniform distribution variables

6.2.2 cumsum

`cumsum(sequence)`: cumulative sum of the given 'sequence'

6.2.3 plot

`plot(x y)`: Plot a plot of x vs y

- `type=""`, type of plot should be drawn. Possible types are
 - "p" for *p*oints,
 - "l" for *l*ines,
 - "b" for *b*oth,
 - "c" for the lines part alone of "b",
 - "o" for both *o*verplotted',
 - "h" for '*h*istogram' like (or 'high-density') vertical lines,
 - "s" for stair *s*teps,

- "S" for other *s*teps, see 'Details' below,
- "n" for no plotting.
- 'main' an overall title for the plot: see 'title'.
- 'sub' a sub title for the plot: see 'title'.
- 'xlab' a title for the x axis: see 'title'.
- 'ylab' a title for the y axis: see 'title'.
- 'asp' the y/x aspect ratio, see 'plot.window'.

6.2.4 ts

`ts(data = NA, start = 1, end = numeric(), frequency = 1, deltat = 1, ts.eps = getOption("ts.eps"), class = , names =)`

The function 'ts' is used to create time-series objects. 'as.ts' and 'is.ts' coerce an object to a time-series and test whether an object is a time series.

6.3 Notes

These should eventually all be categorised, but for now they're all lumped together here.

An irreducible chain implies it has a unique stationary distribution.

Trace of a matrix is equal to the sum of it's eigenvalues

A Markov Matrix always has 1 as an eigenvalue

Diagonalization: $D = PAP^{-1}$