Model Specification

Basic Definitions and Notation

1. We use $y_1, y_2, \ldots, y_n$ to denote OBSERVED values and $Y_1, Y_2, \ldots, Y_n$ to denote a sequence of RVs
2. A time series model for observed data is a joint distribution of two or more random variables where the observed data are realizations
    1. A time index series set T is often integers or nonnegative integers ($\mathbb{Z}$)
    2. The random variable $Y(\omega, t)$ is a continuous distribution
3. We characterize time series by the following:
    1. Means
        1. For any t, define $\mu_t = \mathbb{E}[Y_t]$
    2. Autocovariances (ACVF)
        1. For any two time indexes t, s define the ACVF as $\gamma_{t,s} = Cov(Y_t, Y_s)$
    3. Variances
        1. A special case of ACVF where $t = s$, thus $\gamma_{t,t} = Var(Y_t)$
    4. Autocorrelation (ACF)
        1. $\rho_{t,s} = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}}$; $\rho_{t,t} = 1$
        2. Strong Linear Relation: $\rho_{t,s} = 1$
        3. Weak Linear Relation: $\rho_{t,s} \approx 1$
        4. No Linear Relation: $\rho_{t,s} = 0$
            1. If Y_t and Y_s are independent events, they are uncorrelated assuming both second moments are finite and non-zero.
4. Properties
    1. Symmetry: $\gamma_{t,s} = \gamma_{s,t}$ and $\rho_{t,s} = \rho_{s,t}$
    2. Bilinearity: for any positive integers n, m , any real numbers $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$, $\text{Cov}\left(\sum_{i=1}^n a_i Y_{t_i}, \sum_{j=1}^m b_j Y_{s_j}\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \, \text{Cov}(Y_{t_i}, Y_{s_j})$
5. Examples of Time Series
    1. Linear Regression

        1. Assume a simple linear model $Y_t = a + b_t + e_t$ where e_t is the "error/innovation/noise" term
        2. $\mu_t = a + bt$
        3. $\gamma_{t,s} = \begin{cases} 0, & \text{if } t \neq s \\ \sigma_e^2, & \text{if } t = s. \end{cases}$
        4. $\rho_{t,s} = \begin{cases} 0, & \text{if } t \neq s \\ 1, & \text{if } t = s. \end{cases}$
    2. Random Walk
        1. Assume a "random only" model: $Y_{t+1} = Y_t + e_{t+1}$; $Y_0 = 0$
        2. $\mu_t = 0$
        3. $\gamma_{t,s} = s\sigma_e^2$ (V/CV do not depend on time)
        4. $\rho_{s,t} = \frac{min\{s,t\}}{\sqrt{st}}$
    3. Moving Average
        1. Define $Y_t = \frac{1}{2}(e_t + e_{t-1})$
        2. $\mu_t = 0$
        3. $\gamma_{t,t} = \frac{1}{2}\sigma_e^2$, $\gamma_{t,t-1} = \frac{1}{4}\sigma_e^2$
        4. Both ACF and ACVF depend on lag $t - s$
6. Stationarity
    1. A MA model is example of a stationary time series. Stationary loosely means probability laws do not change with time. Note that for the MA model, ACF and ACVF do not depend on time if the lag is fixed
    2. Strong stationarity: finite dimensional joint distributions are time invariant.
    3. Weak stationarity (second order stationarity)

Stationarity

1. A time series is strictly stationary if all finite dimensional joint distributions are time invariant $F_{Y_{t_1},...,Y_{t_n}} = F_{Y_{t_1-k},...,Y_{t_n-k}}, \quad \forall n,\ \forall t_1, \dots, t_n,\ \forall k$
    1. This means that:
        1. All Y_t are identically distributed
        2. if the distribution has a finite second moment, then the mean and variance do not depend on t
            1. Mean: $\mu_t = \mu,\ \forall t \in \mathbb{Z}$
            2. Variance: $\mathrm{Var}(Y_t) = \gamma_{t,t} = \gamma_0 = \sigma^2,\ \forall t \in \mathbb{Z}.$
            3. ACVF: $\gamma_k = \mathrm{Cov}(Y_t, Y_{t-k}) = \mathrm{Cov}(Y_{t-k}, Y_t) = \gamma_{-k},\ \forall k, t \in \mathbb{Z}.$
            4. ACF: $\rho_k = \frac{\gamma_k}{\gamma_0} \quad (\text{assuming } \gamma_0 \neq 0).$
2. A time series is weakly stationary if it satisfies these three conditions:

   1. $\mu_t = \mu$ for some finite constant $\mu \forall t \in \mathbb{Z}$
   2. $\sigma^2_t = \sigma^2$ for some finite constant $\sigma^2 \forall t \in \mathbb{Z}$
   3. $\sigma^2_{Y_t, Y_{t-k}} = \gamma_k$ for some function γ_k that only depends on lag k and not t
   4. Other names: weakly stationary/ stationary / second order stationary / covariance stationary
3. Remarks on some concepts
   1. IID Noise: $e_t \sim IID(0, \sigma^2_e)$
   2. White Noise: $e_t \sim WN(0, \sigma^2_e)$; all e_t are pairwise uncorrelated
   3. Under the normality assumption (that the time series is a gaussian process), these two are the same. Outside of normality, IID noise implies WN but WN does NOT imply IID noise

q-Dependent CLT

1. Assume a stationary time series. We can find a sample mean easily if the time series is IID, but if the series is NOT IID, this can get harder.
2. For a time series Y_t:
   1. If Y_s and Y_{s+k} are independent for any s and any $k > q$, then Y_t is q-dependent
      1. "Dependent up to lag q"
   2. If Y_s and Y_{s+k} are uncorrelated for any s and any $k > q$, then Y_t is q-correlated
      1. "Covariated up to lag q"
3. Loosely stated, $\bar{Y} \approx N(\mu, Var(\bar{Y}))$, the variance decays to 0 at a rate of o(n^{-1}) sufficiently fast

Moving Average MA(q)

1. An $MA(q)$ of the time series Y_t is defined as $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$
   1. The coefficient of the error term is always 1 but theta is unrestricted
   2. Take note of the notation, some documentation uses (+) instead of (-)
   3. MA(0) is the same thing as IID noise
   4. This process is always q-Dependent, so the matrix is always truncated up to q
   5. if a process is finitely correlated, the process is an MA process

General Linear Process

1. Let error terms be IID. A process is a GLP (mean 0) if
    1. $Y_t = \cdots + \psi_{-2}e_{t+2} + \psi_{-1}e_{t+1} + \psi_0 e_t + \psi_1 e_{t-1} + \cdots$
    2. ψ is a finite constant
    3. Mean: $\mu_t = 0$
    4. Variance: $\sigma_e^2 \left(\sum_{j=-\infty}^{+\infty} \psi_j^2\right)$
    5. ACVF: $\gamma_k = \sum_{j=-\infty}^{+\infty} \psi_{k+j}\psi_j\sigma_e^2$
    6. ACF: $\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\sum_{j=-\infty}^{+\infty} \psi_{k+j}\psi_j}{\sum_{j=-\infty}^{+\infty} \psi_j^2}$
        1. If we want a GLP process with a mean, just add it as an intercept
2. Causality
    1. if $\psi_j = 0 \forall j < 0$ then Y_t is a causal / future-independent process
    2. If ψ_j is nonzero, then Y_t is a non-causal / future-dependent process
3. Operators
    1. A backshift operator is defined as $B^k Y_t = Y_{t-k} \forall k \geq 0$
    2. A forwardshift operator is the inverse: $B^{-k} Y_t = Y_{t+k}$
    3. A linear filter is an operator defined by $\Psi(B) = \sum_{j=-\infty}^{+\infty} \psi_j B^j$.
        1. Therefore the GLP Y_t is $Y_t = \psi(B)e_t$

Autoregressive AR(p)

1. Let the time series be a stationary soultion $Y_t = \phi Y_{t-1} + e_t \forall |\phi| < 1$
2. Assume the time series is independent of future error terms for t>s. We can write AR(1) as. GLP:
    1. $Y_t = \sum_{j=0}^{n} \phi^j e_{t-j} + \phi^{n+1} Y_{t-n-1}$
3. this GLP is causal. It is not q-dependent for any finite q (for generic ψ), this is called ∞-dependent. Moreover, this GLP can be seen as an MA(∞)
4. A stationary solution to AR(p) with $|\phi| < 1$ is a causal GLP as MA(∞)
5. This process is similar to a linear regression, just replace X_t with Y_t and drop the intercept.
6. Finding Functions for AR(1)
    1. Mean: $\mu_t = 0$
    2. Variance: $\sigma_t^2 = \frac{\sigma_e^2}{1-\phi^2}$
    3. ACVF: $\gamma_k = \frac{\sigma_e^2 \phi^k}{1-\phi^2}$
    4. ACF: $\rho_k = \phi^k$
        1. AR(1) is not q-dependent for any finite q, and all ARs have this behavior because ACVF and ACF decay geometrically
7. If $|\phi| > 1$, it will not converge: non-causal GLP, stationary

8. If $|\phi| < 1$, it will converge: causal GLP, stationary
9. If $|\phi| = 1$, Non-stationary
10. Wold's Decomposition Theorem
    1. Any stationary time series can be represented as a sum of a general linear process and a deterministic component.
11. Yule-Walker Method
    1. Yule-Walker method is useful for finding the ACVF/ACF of a time series from its definition equation.
    2. in general cases, to apply Y-W method, we need this time series to be mean zero, stationary, and causal
    3. Steps:
        1. Write the AR() Equation
        2. Multiply by Y_{t-k}
        3. Find the expectation
        4. Solve for ACVF (It is a linear system)
    4. If k is large, solving it recursively is hard, so use the property: If the roots of the AR Polynomial are distinct, then for A_1, A_2, A_n there is a solution:
        1. $\gamma_k = A_1 z_1^{-k} + A_2 z_2^{-k} + \cdots + A_p z_p^{-k}, \quad \forall k \geq 0.$
12. An AR(p) with mean correction: $Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$.
    1. Also can be written as $Y_t = c + \phi Y_{t-1} + e_t$ where $c = \mu(1 - \phi)$
13. AR Polynomial
    1. the AR polynomial (for the AR(p) above) is defined as the following p-th order polynomial (i.e., polynomial with order/degree p):
        1. $\Phi(x) = 1 - \phi_1 x^1 - \phi_2 x^2 - \cdots - \phi_p x^p.$
        2. with the backshift operator, we can formulate as $\Phi(B) Y_t = e_t$
        3. Causality condition: All (complex) roots of the AR polynomial are strictly greater than 1 in absolute value (the modulus of complex number).
    2. If all roots of the AR polynomial are outside the unit disc ($|z_i| > 1 \forall i$), then Y_t is causal and stationary
    3. If at least one of the roots is inside the unit disc ($|z_i| < 1$) and none of the roots is on the unit circle ($|z_i| \neq 1 \forall i$), Y_t is noncausal and stationary
    4. If at least one root is a unit root ($|z_i| = 1$), Y_t is not stationary
14. Textbook Remarks
    1. The textbook's stationary GLP means causal and stationary GLP in our notations
    2. a stationary GLP can be either causal or non-causal in our setting
15. Finding the coefficients ψ in the GLP representation fo AR(p)
    1. Method 1: Convolution
        1. We can use the formula $\psi_n = \sum_{j_1+j_2+\cdots+j_p=n} z_1^{-j_1} z_2^{-j_2} \cdots z_p^{-j_p}$

            1. $\psi_0 = z_1^0 z_2^0 = 1$
            2. $\psi_1 = z_1^0 z_2^{-1} + z_1^{-1} z_2^0 = 1 \cdot \frac{1}{2} + \frac{1}{-3} \cdot 1 = \frac{1}{2} + \left(-\frac{1}{3}\right) = \frac{1}{6}$
            3. $\psi_2 = z_1^0 z_2^{-2} + z_1^{-1} z_2^{-1} + z_1^{-2} z_2^0 = \frac{1}{4} + \frac{1}{-3} \cdot \frac{1}{2} + \frac{1}{9} = \frac{7}{36}$

16. An Observation
    1. For a causal AR(p), the ACVF has the asymptotic rate:
        1. $\gamma_k \approx c[\min\{|z_1|, \ldots, |z_p|\}]^{-k}, \quad \text{as } k \to \infty.$
        2. it decays exponentially, and never stays at zero for a "long time"
        3. EX: for an AR(3) polynomial with three roots, $\gamma_k \approx 2^{-k}$ for large k
17. ACVF for AR(2)
    1. For the AR Polynomial, there are three possible cases for the two roots
        1. Two distrinct real roots ($z_1, z_2 \in \mathbb{R}, z_1 \neq z_2$)
            1. $\gamma_k = A_1 z_1^{-k} + A_2 z_2^{-k}, \quad \forall k \geq 0.$
            2. We can use γ_0 and γ_1 to determine (A_1, A_2)
        2. Two repeated roots ($z_1 = z_2$) due to fact that coefficients are real and $\bar{z}_1 = z_2$
            1. $\gamma_k = (A_1 + A_2 k) z_1^{-k}, \quad \forall k \geq 0.$
            2. The causality condition holds and decays exponentially (goes to zero)
            3. $\rho_k = \frac{\gamma_k}{\gamma_0} = \left(1 + \frac{1+\phi_2}{1-\phi_2} k\right)\left(\frac{\phi_1}{2}\right)^k$
        3. Two distinct complex roots ($z_1, z_2 \in \mathbb{C}, z_1 \neq z_2$) a vibrating string
            1. $\gamma_k = A_1 z_1^{-k} + A_2 z_2^{-k}, \quad \forall k \geq 0.$
            2. $\rho_k = \frac{\gamma_k}{\gamma_0} = R^k \cdot \frac{\sin(k\Theta + \Phi)}{\sin(\Phi)}, \quad \forall k \geq 0$,
                1. $\cos\Theta = \frac{\phi_1/2}{\sqrt{-\phi_2}}$ (Frequency)
                2. $R = \sqrt{-\phi_2}$ (Amplitude)
                3. $\tan\Phi = \frac{\sqrt{-\phi_1^2 - 4\phi_2}}{\phi_1} \cdot \frac{1-\phi_2}{1+\phi_2}$ (Shift)
18. Various notes
    1. For AR(p), if it has roots z_n with multiplicity r, then the form of the ACVF is $\gamma_k = \left(A_1 + A_2 k + \cdots + A_r k^{r-1}\right) z_1^{-k} A_{r+1} z_{r+1}^{-k} + \cdots + A_p z_p^{-k}, \quad \forall k \geq 0$. which is a combination of (b) and one of (a) or (c)
    2. For a generic AR(p) that has multiple multicplicities, it can be a combination of any (a), (b), or (c)
    3. For an MA(q) process, $\gamma_k = 0 \forall k \geq q + 1$ (q-Dependency)
19. Invertibility
    1. A GLP Y_t is invertible is there exists finite coefficients of the MA polynomial $e_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j} = \pi_0 Y_t + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots$
    2. All roots must be outside the unit disk
    3. Relationship to causal: π becomes ψ, and e_t swapped with Y_t

    4. Will be used in forecasting, invertibility allows us to recover e_t which can be used for prediction

Mixed Models ARMA(p, q)

1. We say a model is mixed if:
2. $Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$
3. Via backshift operator, it can be simplified to $\Phi(B)Y_t = \Theta(B)e_t$ MA=AR
4. Causality condition: roots of the AR polynomial are all outside of the unit disk.
5. Invertibility condition: roots of the MA polynomial are all outside of the unit disk.
6. if $\phi = \theta$ then the model becomes $Y_t = e_t$ (white noise model) ie overparametrization
7. The causal GLP representation is
$$\begin{cases} \psi_0 = 1 \\ \psi_1 = \phi_1 - \theta_1 \\ \psi_2 = \phi_1\psi_1 + \phi_2 - \theta_2 \\ \psi_3 = \phi_1\psi_2 + \phi_2\psi_1 + \phi_3 - \theta_3 \\ \quad\vdots \\ \psi_j = \phi_1\psi_{j-1} + \phi_2\psi_{j-2} + \cdots + \phi_p\psi_{j-p}, & \text{for large } j \text{ such that } j \geq p \text{ and } j > q \end{cases}$$
8. If $\theta = 0$ it becomes an AR(1)
9. The invertible representation is symmetric, so Y_t and e_t are swapped
    1. $MA(\infty)$: $e_t = Y_t + \sum_{j=1}^{\infty} \theta^{j-1}(\theta - \phi)Y_{t-j}$
    2. $AR(\infty)$: $Y_t = e_t + \sum_{j=1}^{\infty} \phi^{j-1}(\phi - \theta)e_{t-j}$
10. Yule Walker will solve for gamma 0 and gamma 1
11. A summary so far

Process	Causal	Invertible	(Weakly) Stationary	ACF behavior
MA(1)	Always	$\lvert\theta\rvert < 1$	Always	$\rho_1 \in [-0.5, 0.5]$, and $\rho_k = 0$ for $k \geq 2$
MA(q)	Always	Roots of MA-poly outside unit disk	Always	$\rho_k = 0$ for $k > q$
AR(1)	$\lvert\phi\rvert < 1$	Always	$\lvert\phi\rvert \neq 1$	$\rho_k = \phi^k$, exponential decay
AR(p)	Roots of AR-poly outside unit disk	Always	Roots of AR-poly outside unit circle	Exponential decay or damped sine wave

Process	Causal	Invertible	(Weakly) Stationary	ACF behavior
ARMA(1,1)	$\|\phi\| < 1$	$\|\theta\| < 1$	$\|\phi\| \neq 1$	Exponential decay
ARMA(p,q)	Roots of AR-poly outside unit disk	Roots of MA-poly outside unit disk	Roots of AR-poly outside unit circle	Exponentially decaying, possibly damped oscillations

Trends

1. All the models we looked at are stationary, which can be used in the stochastic trend
2. An observed time series can be a sum of two parts: deterministic + stochastic
    1. $Y_t = \mu_t + X_t$
    2. μ_t is usually fit as seasonality and/or trend
    3. X_t can be AR, MA, or ARMA
3. The sample mean is a good estimate for stationary trend (only holds if Y_t is IID)
4. The variance is $\operatorname{Var}(\overline{Y}) = \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n}\left(1 - \frac{k}{n}\right)\rho_k\right]$.
5. A mean-reverting time series is one where, over time, the values tend to return (revert) to a long-run average level.
    1. This means that if the process deviates from its average, there is a tendency to “pull back” toward the mean.
    2. Negative autocorrelation, usually modeled with AR(1), and stationary
    3. Shocks have temporary effect
6. A mean-avoiding time series is one that drifts away from the mean, often due to persistent shocks that are not corrected. It lacks the “pull” back toward an average.
    1. unit root or near unit root behavior (ex, random walk), non-stationary
    2. shocks have permanent effect
    3. As sample size increases, we are increasingly uncertain about random walk behavior
7. Different models of trend and stochastic components
    1. Additive model (deterministic + stochastic component)
        1. $Y_t = \mu_t + X_t$
    2. Additive model (trend + seasonality + stochastic component)
        1. $Y_t = T_t + S_t + X_t$
    3. Multiplicative model
        1. $Y_t = \mu_t X_t$ or $Y_t = T_t S_t X_t$
        2. Remark: taking the log turns it into an additive model
    4. Mixture of additive and multiplicative model

        1. $Y_t = T_t S_t + X_t$ or $Y_t = (T_t + S_t) X_t$

8. Regression Methods
    1. Suppose $Y_t = \mu_t + X_t$. Fit a regression model for μ_t to estimate the mean and solve $X_t = Y_t - \mu_t$ to estimate the stochastic component
    2. Other regression methods:
        1. Quadratic Trend
            1. $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 = [1 \quad t \quad t^2] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$
            2. Any polynomial trend can be estimated via linear regression
        2. Cosine Trend
            1. $\mu_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{f} t\right) + \beta_2 \sin\left(\frac{2\pi}{f} t\right)$
        3. Seasonal/Cyclical Trend
            1. Assume μ_t is periodic and same across months for t=month
            2. $Y_t = \beta_1 X_{\text{Jan}} + \beta_2 X_{\text{Feb}} + \cdots + \beta_{12} X_{\text{Dec}} + \varepsilon_t$
            3. where $X_{\text{Jan}} = \begin{cases} 1, & \text{if } t \text{ is January} \\ 0, & \text{otherwise} \end{cases}$, (X is a dummy/indicator variable)
    3. If the trend is polynomial, trigonometric, trigonometric polynomial, seasonal means, or a linear combination of the above, then for a stationary stochastic component, the least square estimate of the trend has the same variance as the BLUE for large sample sizes.

ARIMA(p, d, q)

1. Differencing Operator $\nabla Y_t = (1 - B) Y_t = Y_t - Y_{t-1}$
    1. Or for seasonal models, use lag-d differencing: $\nabla_d Y_t = (1 - B^d) Y_t = Y_t - Y_{t-d}$
    2. With seasonality and differencing: $\nabla_d^s Y_t = (1 - B^s)^d Y_t$
    3. If $\nabla^d Y_t$ is ARMA(p,q) then Y_t is ARIMA(p,d,q)
    4. d typically is 1 or 2. Overdifferencing leads to too much complexity and non invertibility
2. Polynomials of ARIMA models
    1. $\Phi(B)(1 - B)^d Y_t = \Theta(B) e_t$
    2. So $\Phi^*_(x) = \Phi(x)(1 - x)^d$ can be seen as an AR polynomial for Y_t. *Assume (W_t) is causal, then $\|Phi^**(x)$ has $p + d$ roots, with $z = 1$ repeated d times and the other p roots (i.e., the roots of $\Phi(x)$) are all outside the unit disk.
3. GLP-like representation
    1. an ARIMA(p,d,q) is a non stationary ARMA(p+d,q)

2. For a non-stationary time series, we cannot get a GLP representation (because GLP is stationary.)
3. Suppose $Y_t \sim ARIMA(0,1,1)$ with $\nabla Y_t = e_t - \theta e_{t-1}$
    1. The GLP-like representation is $Y_t \approx e_t + \sum_{j=1}^{\infty}(1-\theta)e_{t-j}$
    2. Has end behavior $Y_{t-m} \to 0 \quad \text{as} \quad m \to \infty$
    3. this is not exact glp because $|1-\theta|$ diverges
    4. Variance: $\text{Var}(Y_t) = \left[1+\theta^2+(1-\theta)^2(t+m)\right]\sigma_e^2$
    5. Correlation: $\rho_{t,t-k} = \frac{[1-\theta+\theta^2+(1-\theta)^2(t+m-k)]\sigma_e^2}{\sqrt{\text{Var}(Y_t)\ \text{Var}(Y_{t-k})}} \approx \frac{(1-\theta)^2(t+m)\sigma_e^2}{\sqrt{(1-\theta)^2(t+m)\cdot(1-\theta)^2(t+m-k)\sigma_e^2}} \approx 1$

Transformations of TS

1. Difference Operator
2. Variance Stabilizing Transformation

    1. This transformation has some nice property. First, using Taylor Series, we have the approximation $\log y \approx y_0 + \log'(y_0)\cdot(y-y_0)$
    2. Replace y with Y_t, and let $y_0 = \mu_t$: $\log Y_t \approx \mu_t + \frac{1}{\mu_t}(Y_t - \mu_t)$.
    3. Since μ_t is a non-random constant,
    $\text{Var}(\log Y_t) \approx \text{Var}\left[\frac{1}{\mu_t}(Y_t-\mu_t)\right] = \frac{1}{\mu_t^2}\text{Var}(Y_t) = \frac{1}{\mu_t^2}\cdot\mu_t^2\cdot\sigma^2 = \sigma^2 = \text{constant.}$
    4. So the variance of $\log Y_t$ is approximately the constant σ^2.
3. Box-Cox Transforms
    1. Transform y to $g(y)$ by $g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$
    2. λ is estimated by MLE, and as lamda approaches 0, B-C becomes VST
4. Log-differences
    1. Suppose the time series (Y_t) can be written as $Y_t = Y_{t-1} + X_t \cdot Y_{t-1} = Y_{t-1}(1+X_t)$
    2. So (X_t) is the percentage change of (Y_t) and we have: $\log Y_t = \log Y_{t-1} + \log(1+X_t)$.
    3. Then the log-difference (or the log-returns, the returns) of (Y_t) is:
    $\nabla \log Y_t = \log Y_t - \log Y_{t-1} = \log(1+X_t) \approx X_t$.

Bartlett's Theorem

1. Sample ACF: $\hat{\rho}_k = r_k = \frac{\sum_{t=k+1}^{n}(Y_t-\overline{Y})(Y_{t-k}-\overline{Y})}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}$
2. The theoretical ACF is a random number and can be unknown. The sample ACF is a random variable if we think of the TS model as a random realization, and thus follows a sampling distribution

3. Theorem:
    1. For a fixed m, the sampling (joint) distribution approaches a multivariate normal as n approaches infinity:
        1. $\vec{r} \sim \mathrm{MVN}(\vec{\rho}, \frac{1}{n}C), \quad \text{as } n \to \infty, \text{ where } \quad \vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}, \quad \vec{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_m \end{bmatrix}.$
    2. The matrix C is a quare matrix of rank m, usually non diagonal
    3. The diagonal entries are: $c_{ii} = \sum_{k=-\infty}^{+\infty} \left(\rho_{k+i}^2 + \rho_{k-i}\rho_{k+i} - 4\rho_i\rho_k\rho_{k+i} + 2\rho_i^2\rho_k^2 \right)$
    4. And the theorem implies that as $n \to \infty$, $r_i \sim \mathcal{N}\left(\rho_i, \frac{1}{n}c_{ii}\right)$
        1. so r_i is an unbiased estimator of ρ_i

Hypothesis Testing for MA(q)

1. We want to test H0: Series is MA(q) vs H1: Series is not MA(q)
    1. We move up in order, ie MA(0) up to q, testing each sample ACF
        1. MA(0) Rejection Region: $|r_i| > \frac{2}{\sqrt{n}}$
        2. MA(1) Rejection Region: $|r_i| > \frac{2}{\sqrt{n}}\sqrt{1 + 2r_1^2}$
        3. MA(2) Rejection Region: $|r_i| > \frac{2}{\sqrt{n}}\sqrt{1 + 2r_1^2 + 2r_2^2}$
        4. repeat this process until we fail to reject a hypothesis that the series is an MA(q)
2. Assume we observe a sample (n=100) with sample ACFs: $r_1 = 0.5$, $r_2 = 0.4$,$r_3 = 0.4$, $r_4 = 0.3$.
3. We want to determine an MA(q) model for this data.
    1. Start by testing if it is white noise MA(0):
        1. We reject that it is MA(0) because everything is greater than 0.2 (2/sqrt(n))
        2. We reject that it is MA(1) because everything is greater than 0.245
        3. We reject that it is MA(2) because everything is greater than 0.269
        4. We reject that it is MA(3) because everything is greater than 0.292
        5. We fail to reject that it is MA(4) because 0.3 < 0.304
        6. The model is at least MA(4) (behaves like MA(4) or higher)

Partial Autocorrelation Function

1. We saw that MA order q can be obtained through testing sample ACFs. Now we look at PACF to determine the order p for AR(p)

2. The PACF at lag k (ϕ_{kk}) is the conditional correlation of Y_t and Y_{t-k} conditional given all intermediate values
3. Definition 1: $\phi_{kk} \overset{\text{def}}{=} \text{corr}\left(Y_t, Y_{t-k} \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}\right).$
    1. PACF at lags 1 through p can be nonzero but for lags $k \geq p+1$ are all zero
    2. For $k = 0$, $\phi_{00} = 1$ by definition
    3. for $k = 1$, $\phi_{11} = \rho_1 = \text{Corr}(Y_t, Y_{t-1})$ since there is no condition
4. Definition 2: $\phi_{kk} \overset{\text{def}}{=} \text{corr}(\text{Res}_t, \text{Res}_{t-k})$
    1. Res_t is obtained from regressing Y_t on Y_{t-k+1}
    2. Res_{t-k} is obtained from regressing Y_{t-k} on Y_{t-k+1}
    3. These are the unexplained variation in Y_t and Y_{t-k} after controlling for intermediate effects
    4. We have the summary table

	MA(q)	**AR(p)**	**ARMA(p, q)**
ACF	cuts off after q	exponential decay	exponential decay
PACF	exponential decay	cuts off after p	exponential decay

5. Definition 3:

    1. ϕ_{kk} is the last ϕ_{kj} term in the AR(k) approximation to Y_t

$$\Gamma_k\,\vec{\phi}_k = \vec{\gamma}_k, \quad \text{where} \quad \Gamma_k = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k-1} & \gamma_{k-2} & \cdots & \gamma_0 \end{bmatrix}, \quad \vec{\phi}_k = \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix}, \quad \vec{\gamma}_k = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix}.$$

    2. Inversion can be expensive so we instead use Durbin-Levinson Recursion to directly compute the entries
    3. Intuitively, fitting the regression model $Y_t = \phi_{k1}Y_{t-1} + \phi_{k2}Y_{t-2} + \cdots + \phi_{kk}Y_{t-k} + \epsilon$ for an AR(p) as $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$ gives the fitted values ϕ_{kk}
    4. When $k \geq p$ this reduces to Yule-Walker equations

6. Durbin-Levinson Recursion
    1. utilizing the special form of the matrix Γ_k, define $\phi_{00} = 1$, for $l \geq 0$,
    2. $\begin{cases} \phi_{l+1,\,l+1} = & nbsp; \dfrac{\gamma_{l+1} - \sum_{j=1}^{l} \phi_{l,\,j}\gamma_{l+1-j}}{\gamma_0 - \sum_{j=1}^{l} \phi_{l,\,j}\gamma_j} = \dfrac{\rho_{l+1} - }{1 - } \\ \phi_{l+1,\,j} = \phi_{l,\,j} - \phi_{l+1,\,l+1}\,\phi_{l,\,l+1-j} \quad \text{for } 1 \leq j \leq l \end{cases}$
    3. EX: finding PACF for AR(1)
        1. Let AR(1): $X_t = \phi X_{t-1} + e_t,\ e_t \sim WN(0, \sigma^2)$, ACF: $\rho_k = \phi^k$
        2. Initialize: $\phi_{11} = \rho_1$, $v_1 = \gamma_0(1 - \phi_{11}^2)$

3. For $m \geq 2$:
   $\phi_{mm} = \frac{\rho_m - \sum_{j=1}^{m-1} \phi_{m-1,j}\rho_{m-j}}{1-\sum_{j=1}^{m-1} \phi_{m-1,j}\rho_j}, \qquad \phi_{m,j} = \phi_{m-1,j} - \phi_{mm}\phi_{m-1,m-j} \ (j = 1, \ldots, m-1).$
4. m=1: $\kappa_1 = \phi_{11} = \rho_1 = \phi.$
5. m=2: $\kappa_2 = \phi_{22} = \frac{\rho_2 - \phi_{11}\rho_1}{1-\phi_{11}\rho_1} = \frac{\phi^2 - \phi \cdot \phi}{1 - \phi \cdot \phi} = 0.$
6. m=3: $\kappa_3 = \frac{\rho_3 - \phi_{2,1}\rho_2}{1-\phi_{2,1}\rho_1} = \frac{\phi^3 - \phi \cdot \phi^2}{1 - \phi \cdot \phi} = 0,$
7. Therefore, PACF: $\boxed{\alpha_1 = \phi, \qquad \alpha_k = 0 \text{ for } k \geq 2}$

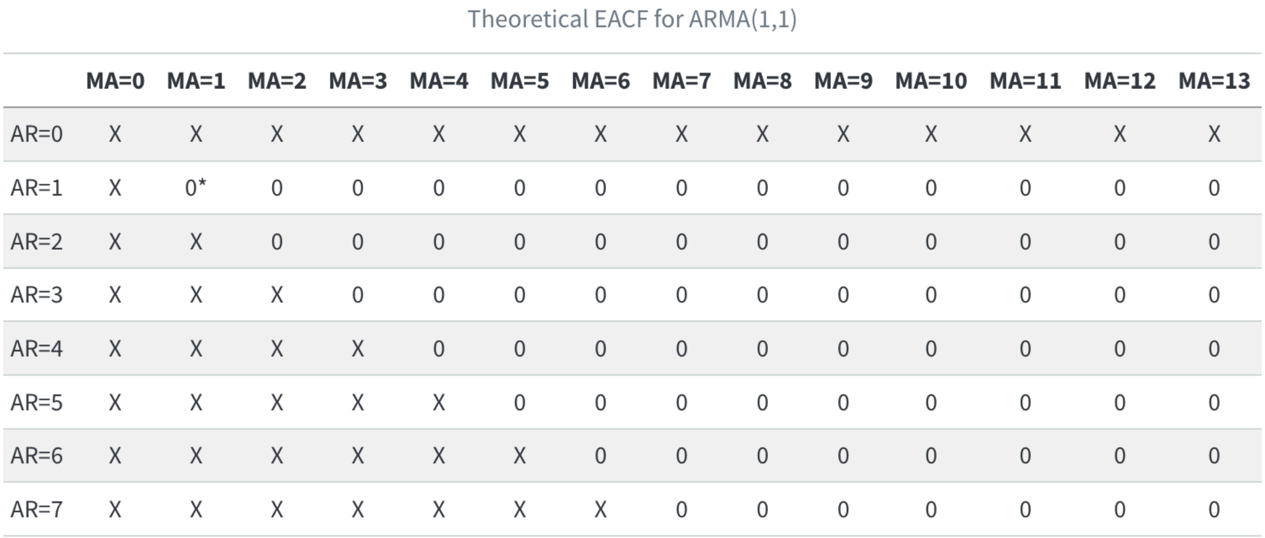
Sample PACF

1. We can get $\hat{\phi}_{kk}$ by doing DLR with the sample ACF instead of the theoretical ACF
2. The theoretical PACF can be an (unknown) random number, but the sample PACF is a random variable and we can use this value to test for AR(p) using the same algorithm as Bartlett's Test

Extended ACF

1. EACF can help find the (p,q) order for ARMA
2. to test whether a process is ARMA given its observations, we can:
   1. first try to fit an AR regression using the observed data,
   2. then test if the residuals follow an MA process
3. The algorithm tests different pairs of (p,q) and repeats the two steps
4. Outer loop: fit an AR(p) model for Y_t and find the residuals $W_t = Y_t - \hat{Y}_t$
5. Inner loop: fir an MA(q) model for W_t and find the residuals $e_t = W_t - \hat{W}_t$
6. Hypothesis test for e_t :
   1. if white noise: indicate 0, else indicate X

7. the output looks like



Theoretical EACF for ARMA(1,1)

	MA=0	MA=1	MA=2	MA=3	MA=4	MA=5	MA=6	MA=7	MA=8	MA=9	MA=10	MA=11	MA=12	MA=13
AR=0	X	X	X	X	X	X	X	X	X	X	X	X	X	X
AR=1	X	0*	0	0	0	0	0	0	0	0	0	0	0	0
AR=2	X	X	0	0	0	0	0	0	0	0	0	0	0	0
AR=3	X	X	X	0	0	0	0	0	0	0	0	0	0	0
AR=4	X	X	X	X	0	0	0	0	0	0	0	0	0	0
AR=5	X	X	X	X	X	0	0	0	0	0	0	0	0	0
AR=6	X	X	X	X	X	X	0	0	0	0	0	0	0	0
AR=7	X	X	X	X	X	X	X	0	0	0	0	0	0	0

   1. The triangle starts at (1,1) so this is likely our order
   2. The zeros appear because:
      1. we can always overfit an ARMA(p,q') if q' > q
      2. we can always overfit with ARMA(p+1,q+1)
   3. if multiple triangles exist, choose model with least parameters, assuming no other information

Augmented Dickey-Fuller Test (Unit Root Test)

1. the unit root test of an AR polynomial tests for the stationarity of a time series
2. We test:
   1. H0: AR polynomial has a unit root (nonstationary)
   2. HA: AR polynomial doesn't have a unit root (stationary)
3. A small p-value of this test implies stationarity, and a large p-value implies non-stationarity.

Parameter Estimation

Parameter Estimation

1. Using all the tools we have seen (sample ACF/PACF/EACF, transformations, ADF test, ARMA subsets, etc.) we arrive at a few candidate models

2. Now we want to estimate the thetas, phis, and maybe mean of the time series and variance of the noise
3. Method of Moments (MOM)
   1. we set the theoretical parameter equal to the sample parameter and solve the system
   2. Suppose Y_t is IID Normal and the first and second moment are known $E(X) = \mu, E(X^2) = \mu^2 + \sigma^2$
   3. $\begin{Bmatrix} \mu_1 = m_1 \\ \mu_2 = m_2 \end{Bmatrix} \implies \begin{Bmatrix} \mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} Y_i \\ \mathbb{E}[Y^2] = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \end{Bmatrix} \implies \begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y} \\ \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \end{cases}$
   4. $\begin{cases} \hat{\mu}_{\text{MOM}} = \overline{Y} \\ \hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \end{cases}$
   5. Remark: MoM doesn't always exist, since solutions must be real
   6. need to estimate p+q+2 parameters (theta, phi, sigma2, mu)
   7. MoM for the Mean μ: $\hat{\mu}_{\text{MOM}} = \frac{1}{n}\sum_{t=1}^{n} Y_t = \overline{Y}$.
   8. MoM for AR(2) mean 0:
      1. Use YW method
         1. The ACVFS are
            1. $\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1$
            2. $\gamma_2 = \phi_1\gamma_1 + \phi_2\gamma_0$.
            3. For any k$\geq$ 3, $\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2}$.
         2. Let $\rho_k = \frac{\gamma_k}{\gamma_0}$ (so that $\rho_0 = 1$)
            1. $\begin{cases} \rho_1 = \phi_1 + \phi_2\rho_1, \\ \rho_2 = \phi_1\rho_1 + \phi_2. \end{cases}$
      2. Solve for the phis
         1. $\phi_1 = \rho_1(1 - \phi_2)$.
         2. $\rho_2 = \rho_1[\rho_1(1 - \phi_2)] + \phi_2 \Rightarrow \rho_2 = \rho_1^2(1 - \phi_2) + \phi_2 \Rightarrow \rho_2 = \rho_1^2 - \rho_1^2\phi_2 + \phi_2$.
         3. $\boxed{\phi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}}$.
         4. now plug into $\phi_1 = \rho_1(1 - \phi_2)$
         5. $\boxed{\phi_1 = \rho_1\frac{1 - \rho_2}{1 - \rho_1^2}}$.
      3. With sample moments
         1. $\boxed{\hat{\phi}_1 = r_1\frac{1 - r_2}{1 - r_1^2}, \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}.}$
      4. Now estimating variance
         1. $\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma_e^2$.

      2. divide by gamma 0: $1 = \phi_1\rho_1 + \phi_2\rho_2 + \frac{\sigma_e^2}{\gamma_0}$.
      3. so $\boxed{\hat{\sigma}_e^2 = \hat{\gamma}_0(1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2)}$,
   9. MoM for noise σ_e^2
      1. Express γ_0 in terms of $\phi_i, \theta_i, \rho_i, \sigma_e^2$
      2. solve for σ_e^2
      3. plug in: $\phi_i = \hat{\phi}_i, \theta_i = \hat{\theta}_i, \rho_i = r_i, \sigma_e^2, \gamma_0 = \hat{\gamma}_0 = s^2$
4. Conditional Least (Sum of) Squares (CSS)
   1. The idea is to construct an objective (loss) function that is a sum of squares, then obtain the estimated parameters by minimizing this sum of squares.
   2. Assume an AR(1) with trend:
      1. $(Y_t - \mu) = \phi(Y_{t-1} - \mu) + e_t, \quad e_t \sim \mathrm{iid}(0, \sigma_e^2)$.
   3. We set the objective function (c for conditional):
      1. $S_c(\phi, \mu) = \sum_{t=2}^n e_t^2 = \sum_{t=2}^n \left[(Y_t - \mu) - \phi(Y_{t-1} - \mu)\right]^2$
   4. LS minimizes objective:
      1. $(\widehat{\phi}_{\mathrm{LS}}, \widehat{\mu}_{\mathrm{LS}}) = \arg\min_{\phi,\mu} S_c(\phi, \mu)$.
   5. Take partial derivatives and set to 0:
      1. $0 = \frac{\partial S_c}{\partial \mu} = 2\sum_{t=2}^n (Y_t - \mu - \phi(Y_{t-1} - \mu))(-1 + \phi)$
      2. $0 = \frac{\partial S_c}{\partial \phi} = 2\sum_{t=2}^n ((Y_t - \mu) - \phi(Y_{t-1} - \mu))(-Y_{t-1} + \mu)$
   6. Solves except for end effects (residuals from incomplete observations)
   7. called conditional because of negating end effects
      1. ARMA assumes $e_p = e_{p-q+1} = 0$
      2. conditional $S_c(\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q) = \sum_{t=p+1}^n e_t^2$
5. Maximum Likelihood Estimation (MLE) (must check if model is normal)
   1. Pros:
      1. Use all the “data/information”
      2. Relevant for small datasets. We have distributional results on estimates
   2. Cons:
      1. No closed form solution.
      2. Numerical optimization is hard.
   3. define the likelihood function of the parameters as the joint pdf of the observed data
   4. Using the same AR(1):
      1. $(Y_t - \mu) = \phi(Y_{t-1} - \mu) + e_t, \quad e_t \sim \mathrm{iid}(0, \sigma_e^2)$.
   5. We derive:
      1. Likelihood: $\mathcal{L}(\mu, \phi, \sigma_e^2 \mid Y_1, Y_2, \dots, Y_n) = (2\pi\sigma_e^2)^{-\frac{n}{2}}(1 - \phi^2)^{\frac{1}{2}} \exp[-\frac{1}{2\sigma_e^2} S(\mu, \phi)]$
      2. Conditional SS: $S_c(\mu, \phi) = \sum_{t=2}^n (Y_t - \mu - \phi(Y_{t-1} - \mu))^2$
      3. Unconditional SS: $S(\mu, \phi) = S_c(\mu, \phi) + (1 - \phi^2)(Y_1 - \mu)^2$

   6. Maximizing likelihood is the same as maximizing log-likelihood
      1. $\ell(\mu, \phi, \sigma_e^2) = -\frac{n}{2}\log(2\pi\sigma_e^2) + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma_e^2}S(\mu, \phi)$
   7. Taking the partial derivative and setting to zero
      1. $\frac{\partial \ell}{\partial \sigma_e^2} = -\frac{n}{2} \cdot \frac{1}{\sigma_e^2} + \frac{1}{2(\sigma_e^2)^2}S(\mu, \phi)$
      2. $\hat{\sigma} e^2 = \frac{1}{n}S(\hat{\mu}_{\text{MLE}}, \hat{\phi}_{\text{MLE}})$
   8. Remark: This is biased, so we need to divide (n-p) if making unbiased (since we are estimating two parameters)
6. Unconditional Sum of Squares (UCSS)
   1. If we ignore the first two terms, we get UCSS instead
   2. in small sample sizes they are different, but close in large sample sizes
   3. It is a good compromise between CSS and MLE
7. Comparing the methods
   1. Recap
      1. MLE maximizes likelihood
      2. UCSS minimizes the unconditional sum of squares
      3. CSS minimizes the conditional sum of squares
   2. For large samples, the three methods MLE, UCSS, CSS have very similar results
   3. In general, UCSS is different from MLE, especially if ARMA models are close to nonstationarity
   4. If sample size is small or medium, MLE is preferred. MLE is hard to compute though, often numerical which is still hard
      1. Note: As part of this, MoM estimates (which is much easier to get) are often used as the initial guesses in the numerical computing of MLE
   5. MLE is conceptually better than the other since it uses “all information” and do not assume something is zero
   6. MOM uses only the first k moments. (And it may not exist as we have seen.)
   7. CSS throws away some information from the likelihood function
   8. UCSS is a compromise between MLE and CSS
8. Asymptotic theory
   1. Asymptotically unbiased: $\lim_{n\to\infty} E(\hat{\theta}_i) = \theta_i$ and $E(\hat{\theta}_i) = F(\theta_i, n)$
   2. Via asymptotic MLE theory, estimators are unbiased, asymptotically normal, and have some certain variance
      1. Asymptotic MLE is similar to UCSS/CSS
      2. Variances for special cases are as follows
         1. MA(2) and AR(2): $\text{Var}(\hat{\phi}_1) \approx \text{Var}(\hat{\phi}_2) \approx \frac{1-\phi_2^2}{n}$
            1. AR(1) and MA(1) will follow the same
      3. Correlation for AR/MA: $\text{corr}(\hat{\phi}_1, \hat{\phi}_2) \approx -\frac{\phi_1}{1-\phi_2}$

        4. ARMA(1,1)
            1. $\text{Var}(\hat{\phi}) \approx \frac{1-\phi^2}{n}\left(\frac{1-\phi\theta}{\phi-\theta}\right)^2$,
            2. $\text{Var}(\hat{\theta}) \approx \frac{1-\theta^2}{n}\left(\frac{1-\phi\theta}{\phi-\theta}\right)^2$,
            3. $\text{corr}(\hat{\phi}, \hat{\theta}) \approx \frac{\sqrt{(1-\phi^2)(1-\theta^2)}}{1-\phi\theta}$.
        5. For $\theta = \phi$ variance blows up because the model is white noise $Y_t = e_t$

9. Overfitting as a tool
    1. if corresponding parameter(s) have larger variance/standard error after fitting a larger model than a smaller model, then this suggests overfitting and we should go back to the smaller model.
10. Note for non-stationarity ARIMA(p,d,q): turn it into $\nabla^d Y_t \sim ARMA(p, q)$ then use an estimation method

Model Checking

Diagnostics

1. Residual Analysis
    1. For an AR(p), residuals can be found normally by $e_t = Y_t - \hat{Y}_t$
    2. For MA(q) or ARMA(p,q) we need to do this via invertible (GLP) representation
    3. Residuals should be independent, normal, mean zero and constant variance
    4. Tool 1: Sample ACF of Residuals: $\hat{r}_k = \widehat{\text{corr}}(\hat{e}_t, \hat{e}_{t-k})$.
        1. Note, this is different from the Sample ACF of Y_t (r_k) and the theoretical ACF (ρ_k)
        2. We use WN as a hueristic and assume the sample ACF is $N(0, \frac{1}{n})$
    5. Tool 2: Plot of standardized residuals
        1. we can plot $\frac{\hat{e}_t}{\widehat{\text{sd}}(\hat{e}_t)}$ and check outliers and verify constant variance and mean 0
    6. Tool 3: use QQ-plot or Shapiro-Wilk test to check for normality
    7. Tool 4: Ljung-Box Test (tests $\text{ACF}_k(\hat{e}_t) = 0$ for all k at once)
        1. H0: ACF=0 for all k>0 (suggests the fitted model is good)
        2. H1: ACF !=0 for some k>0 (we should adjust the model)
        3. Test Statistic: $Q_{\text{LB}} = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_K^2}{n-K}\right)$
            1. K is typically chosen from 5,6,...,30 such that ψ_j =0 for j>K
            2. Under the null hypothesis, Q asymptotically follows χ^2_{K-p-q}
        4. If Q>CV, reject null in favor of the alternative (errors dependent, adjust model)

        5. If Q <=CV, do not reject null (errors independent, model is fine)

2. Overfitting
    1. Suppose AR(2) is the correct model but we overfit with AR(3). we can confirm AR(2) over AR(3) if:
        1. we fit AR(3) and ϕ_3 is not significant
        2. we fit AR(2). if ϕ_1 and ϕ_2 from both models are similar (the CIs overlap significantly)
    2. Similarity we can overfit ARMA(2,1) for an AR(2)
    3. In general, for an ARMA(p,q) we fit ARMA(p+1,q) or ARMA(p,q+1). Never fit ARMA(p+1,q+1) because it causes parameter unidentifiability, formally $(1-cB)\Phi(B)Y_t = (1-cB)\Theta(B)e_t \forall c$

Forecasting

Forecasting

1. We can neutrally predict $Y_{t+1} = \mu$ using the conditional expectation $E(Y_{t+1}|Y_1, \ldots, Y_n)$
2. Define the following notation
    1. t=forecast origin
    2. h=lead time
    3. $\hat{Y}_t(h)$ is the predicted value of Y_{t+h}
    4. $e_t(h)$ is the forecast error: $e_t(h) = Y_{t+h} - \hat{Y}_t(h)$
3. If Y_t is a normal process, then the $(1-\alpha)$ 100% Prediction interval is $\left[\hat{Y}_t(h) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\operatorname{Var}(e_t(h))}\right]$ or $\left[\hat{Y}_t(h) \pm z_{1-\frac{\alpha}{2n}} \cdot \sqrt{\operatorname{Var}(e_t(h))}\right]$ with correction
4. Predicted values are therefore defined by $\hat{Y}_t(h) \overset{\text{def}}{=} \mathbb{E}[Y_{t+h} \mid Y_1, \ldots, Y_t]$ usually called the "min squared error prediction"
5. Properties of conditional expectation
    1. For a RV X, a fixed number x and a function g: $\mathbb{E}[g(X) \mid X = x] = g(x)$ (a fixed number)
    2. For a RV X and a function g: (a random variable) $\mathbb{E}[g(X) \mid X] = g(X)$
    3. If RVs X and Y are independent: $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$
6. Forecasting Examples
    1. Forecasting Trend+Noise $Y_t = \mu_t + X_t$
        1. The prediction is μ_{t+h} with error X_{t+h}.
        2. Similarily the prediction interval is $\mu_{t+h} \pm 2\sigma$

   2. AR(1) with mean $Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$
      1. Assume causality and stationarity hold true
      2. prediction is $\mu + \phi^h(Y_t - \mu)$ with error $\sigma_e^2 \frac{1-\phi^{2h}}{1-\phi^2}$ (using ϕ) or $\sigma_e^2 \sum_{j=0}^{h-1} \psi_j^2$ via GLP
      3. The prediction interval is $\left[\hat{Y}_t(h) \pm 2\sigma_e \sqrt{\sum_{j=0}^{h-1} \psi_j^2}\right]$
         1. The width of prediction intervals increase and converge to some fixed number $4\sqrt{\gamma_0}$
   3. MA(1) with mean $Y_t = \mu + e_t - \theta e_{t-1}$
      1. The prediction is $\mu - \theta e_t$ with error $e_{t+h} - \theta e_{t+h-1}$ with an interval $\left[\hat{Y}_t(h) \pm 2\sigma_e\right]$
   4. Random Walk with Drift
      1. The prediction is $\hat{Y}_t(h) = Y_t + h\theta_0$ with error $\sum_{j=1}^{h} e_{t+j}$
   5. ARMA(1,1) with mean $Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} + \theta_0$
      1. The prediction is $\hat{Y}_t(h) = \phi^h Y_t - \phi^{h-1}\theta e_t + \frac{1-\phi^h}{1-\phi}\theta_0$; note $\theta_0 = (1-\phi)\mu$
      2. $E(e_t | Y_{t-1}) = e_t$
   6. Summary of exercises
      1. RW+drift is unstationary (variance explodes as lead time approaches infinity)
      2. AR(1),MA(1), ARMA(1,1,) with mean: with stationarity assumed, the predictions approach the mean
      3. AR(1),MA(1) causes variance to approach γ_0
      4. In general:
         1. for invertible MA models, when h>q, the variance approaches γ_0
         2. for causal AR and causal+invertible ARMA, variance is increasing in lead time and converges to γ_0

7. Prediction for ARMA(p,q)
   1. for lead time h>q, the predictions are the YW equations assuming causal and invertible
   2. $Y_{t+h} = \psi_0 e_{t+h} + \psi_1 e_{t+h-1} + \cdots + \psi_{h-1} e_{t+1} + \psi_h e_t + \psi_{h+1} e_{t-1} + \cdots$ (wold) can be expressed as $Y_{t+h} = I_t(h) + C_t(h)$ = future+past
   3. Now the forecast is $\widehat{Y}_t(h) = C_t(h) = \psi_h e_t + \psi_{h+1} e_{t-1} + \cdots$
   4. The forecast error is $e_t(h) = Y_{t+h} - C_t(h) = I_t(h) = \psi_0 e_{t+h} + \psi_1 e_{t+h-1} + \cdots + \psi_{h-1} e_{t+1}$
8. Prediction for ARIMA(p,d,q)
   1. Rewrite it as ARIMA(p,d,q)=ARMA(p+d,q)
      1. $\Phi(B)(1-B)^d Y_t = \Theta(B) e_t \implies \widetilde{\Phi}(B) Y_t = \Theta(B) e_t$
   2. For a general ARIMA, if $d \geq 1$ then the ARIMA is not stationary or causal
   3. Prediction interval: $\left[\widehat{Y}_t(h) \pm z_{1-\alpha/2} \sqrt{\mathrm{Var}(e_t(h))}\right]$
      1. The concept PI and CI are different. Prediction interval is for the random variable Y_{t+h}. Confidence interval is for a parameter θ , where θ is a fixed but unknown value.

Exponentially Weighted Moving Average

1. EWMA is a quick way to generate forecasts using $\widehat{Y}_t(1) = (1-\alpha)Y_t + \alpha\widehat{Y}_{t-1}(1)$ where α is chosen ad hoc
    1. alternatively written as forecast + (α) x forecast error
    2. $\hat{Y}_t(1) = \hat{Y}_{t-1}(1) + (1-\alpha)\left(Y_t - \hat{Y}_{t-1}(1)\right)$
2. Remark: the coefficients in the invertible representation of ARIMA are generally

$$\pi_j = \begin{cases} \displaystyle\sum_{i=1}^{\min(j,q)} \theta_i \pi_{j-i} - \widetilde{\Phi}_j, & \text{if } 1 \le j \le p+d \\ \displaystyle\sum_{i=1}^{\min(j,q)} \theta_i \pi_{j-i}, & \text{if } j > p+d \end{cases}$$

    1. Special: use $\pi_0 = 1$ and $\pi_j = (\theta - 1)\theta^{j-1}$
3. Other EWMAs exist, like Holt (double exponential) or Holt-Winters (triple exponential)

Seasonal ARIMA (SARIMA)

1. Consider the model $MA(1)_{12} = Y_t = e_t - \Theta e_{t-12}$ (MA model of order 1 with seasons of 12)
    1. can also be viewed as MA(12)
    2. ACVF: $\begin{cases} \gamma_0 = (1+\Theta^2)\sigma_e^2 \\ \gamma_{12} = -\Theta\sigma_e^2 \\ \gamma_k = 0, \quad \text{if } k \ne 0, 12 \end{cases}$
    3. ACF: $\begin{cases} \rho_0 = 1 \\ \rho_{12} = \dfrac{-\Theta}{1+\Theta^2} \\ \rho_k = 0, \quad \text{if } k \ne 0, 12 \end{cases}$
2. Seasonal MA $MA(Q)_s$: $Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs}$ (backshift form)
    1. Polynomial: $\Theta(x) = 1 - \Theta_1 x^s - \Theta_2 x^{2s} - \cdots - \Theta_Q x^{Qs}$
3. Seasonal AR $AR(P)_s$: $Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t$
    1. Polynomial: $\Phi(x) = 1 - \Phi_1 x^s - \cdots - \Phi_P x^{Ps}$
4. Seasonal ARMA: combine AR and MA to form $ARMA(P,Q)_s$ where s is the same
5. Multiplicative SARMA
    1. We can combine a nonseasonal ARMA with a seasonal ARMA by multiplying the AR and MA polynomials and denote as $ARMA(p,q) \cdot (P,Q)_s$ ie $\theta(x)\Theta(x)$ and $\phi(x)\Phi(x)$
6. Multiplicative SARIMA
    1. Similar to SARMA we can denote a $ARIMA(p,d,q) \cdot (P,D,Q)_s$
    2. Recall that ARIMA is an ARMA + differencing operator: $ARIMA(p,d,q) = \nabla^d Y_t \sim ARMA(p,q)$ where $\nabla Y_t = Y_t - Y_{t-1}$ so $\nabla^d Y_t = (1-B)^d Y_t$

3. We introduce seasonal differencing: $\nabla_s Y_t = Y_t - Y_{t-s} = (1 - B^s)Y_t$
    1. therefore Seasonal ARIMA is $ARIMA(P, D, Q)_s = \nabla_s^D Y_t \sim ARMA(P, Q)$
4. For a multiplicative model, $Y_t \sim ARIMA(p, d, q) \cdot ARIMA(P, D, Q)$ is $\nabla^d \nabla_s^D Y_t \sim ARMA(p, q) \cdot ARMA(P, Q)$ with polynomial $\phi(B)\,\Phi(B)\,(1 - B)^d\,(1 - B^s)^D\,Y_t = \theta(B)\,\Theta(B)\,e_t$
    1. ϕ is the NONSEASONAL AR poly
    2. Φ is the SEASONAL AR poly
    3. θ is the NONSEASONAL MA poly
    4. Θ is the SEASONAL MA poly
5. Example: Consider $Y_t = 0.5Y_{t-1} + Y_{t-4} - 0.5Y_{t-5} + e_t - 0.3e_{t-1}$
    1. Y_{t-1} — is a nonseasonal AR(1) term
    2. Y_{t-4} suggests seasonality with period 4, so it’s a seasonal AR(1) term
    3. $Y_{t-5} = Y_{t-1}Y_{t-4}$ is a multiplicative term, i.e., a product of lag 1 and lag 4, indicating interaction between seasonal and nonseasonal AR
    4. Error term includes e_t and e_{t-1}, so it includes a nonseasonal MA(1) structure
    5. Expressing via Backshit operator
        1. $Y_t - 0.5BY_t - B^4Y_t + 0.5B^5Y_t = e_t - 0.3Be_t$
        2. $(1 - 0.5B - B^4 + 0.5B^5)Y_t = (1 - 0.3B)e_t$ (group terms)
        3. $(1 - 0.5B)(1 - B^4)Y_t = (1 - 0.3B)e_t$ (factor AR)
        4. This shows that Nonseasonal $AR(1) = (1 - 0.5B)$ and Seasonal AR(1) with period 4: $(1 - B^4)$
    6. Specification
        1. Nonseasonal part: $ARIMA(p, d, q)$
            1. p = 1 from $0.5Y_{t-1}$
            2. d = 0: no nonseasonal differencing is done
            3. q = 1 from $-0.3e_{t-1}$
        2. Seasonal part: $ARIMA(P, D, Q)_s$
            1. P = 0: there is no explicit seasonal AR term in the factorization
            2. D = 1: since we factor $(1 - B^4)$, this acts as seasonal differencing of period 4
            3. Q = 0: there are no seasonal MA terms
            4. s = 4: the periodicity is 4, indicated by the lag 4 and B^4

Cross-Covariance and Cross-Correlation Function

1. We previously used Y to predict Y. What if we want to use the past values of X to predict Y?

2. Suppose we have a TS vector: (X_t, Y_s)
    1. The Cross-Covariance function is defined as $\gamma_{t,s}(X,Y) \overset{\text{def}}{=} \text{Cov}(X_t, Y_s)$.
    2. A vector time series is jointly weakly stationary if:
        1. $\mathbb{E}[X_t]$ is a constant μ_X for all t, $\mathbb{E}[Y_t]$ is a constant μ_Y for all t.
        2. $\text{Var}(X_t)$ is a constant for all t, $\text{Var}(Y_t)$ is a constant for all t.
        3. ACVF $\gamma_{t,s}(X) = \text{Cov}(X_t, X_s)$ only depends on the lag difference $t-s$, $\gamma_{t,s}(Y) = \text{Cov}(Y_t, Y_s)$ only depends on the lag difference $t-s$.
        4. CCVF $\gamma_{t,s}(X,Y) = \text{Cov}(X_t, Y_s)$ only depends on the lag difference $t-s$.
    3. We can do the same replacement $\gamma_{t,s} = \gamma_{t-s}$
        1. $\gamma_0(X,Y) = \gamma_{t,t}(X,Y) = \text{Cov}(X_t, Y_t), \text{ for any } t$
        2. $\gamma_1(X,Y) = \gamma_{t+1,t}(X,Y) = \text{Cov}(X_{t+1}, Y_t), \text{ for any } t$
        3. $\gamma_{-1}(X,Y) = \gamma_{t-1,t}(X,Y) = \text{Cov}(X_{t-1}, Y_t), \text{ for any } t$
        4. Remark: while for a single time series $\gamma_k = \gamma_{-k}$ for a vector time series this is NOT true
3. We can also define the Cross-Correlation function: $\rho_k(X,Y) \overset{\text{def}}{=} \text{corr}(X_t, Y_{t-k}) = \frac{\gamma_k(X,Y)}{\sqrt{\gamma_0(X)\cdot\gamma_0(Y)}}$
4. Bartlett's Theorem on Sample CCF
    1. Similar to the sample ACF we can obtain a sample CCF where the sampling distribution is $r_m(X,Y) \sim \mathcal{N}\left(\rho_m(X,Y), \frac{1}{n}\left(1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)\right)\right)$
    2. This may lead to spurious correlation, where if the theoretical CCF is small, the sample CCF can be large ($> \frac{2}{\sqrt{n}}$) implying correlation
        1. if the variance is larger than $\frac{1}{n}$ then the results from software are not reliable
        2. If we can transform one of the series into white noise, then we can get rid of spurious correlation

Pre-Whitening

1. if we apply a prewhitening filter to X_t we can theoretically get white noise e_t . We can then apply this filter to both time series in our vector
2. X_t is transformed into white noise and Y_t gets transformed into a new time series. This preserves the dependence and we can safely look at CCF with no spurious correlation
3. In practice, folllow the steps
    1. Make X_t and Y_t stationary via $\nabla^d \nabla_s^D$
    2. fit an AR(p) (choose a large p) to X_t so this creates an AR filter ($\Phi(B)$)that can be applied as a prewhitening filter ($\Pi(B)$) to Y_t
    3. Find the new time series $\tilde{Y_t} = \Phi(B)Y_t$
    4. Estimate the CCF between the new series and transformed X_t

GARCH Models

1. GARCH is Generalized AutoRegressive with Conditional Heteroskedasticity which means that they allow the variance of the error terms to change over time. (usually used in financial time series where volatility clusters exist)
2. ARCH(1) $\begin{cases} Y_t = \mu_0 + \sigma_{t|t-1} e_t, \quad e_t \sim iid(0,1) \\ \sigma^2_{t|t-1} = \alpha_0 + \alpha_1 Y^2_{t-1} \end{cases}$
    1. Conditional because the error term variance depends on the past values of the series
    2. for this model to be valid, both α need to be >0
    3. This process s stationary if $\alpha \in$ (0,1)
        1. It may seem contradictory that a stationary process can have non-constant conditional variance. But, recall that weakly stationary processes has constant unconditional variance.
    4. Financial returns: $Y_t = \nabla \log(X_t) = \frac{X_t - X_{t-1}}{X_{t-1}}$
        1. the standard deviation of Y_t is volatility
        2. if the previous time series is large, then volatility is also large, indicating high uncertainty
    5. Properties
        1. Expectation: $E(Y_t) = 0$
        2. Variance: $\mathrm{Var}(Y_t) = \frac{\alpha_0}{1-\alpha_1}$ $\alpha_1 > 1$ is a necessary condition and conditional variance is never constant
        3. Dependence: Y_t and Y_{t-h} are uncorrelated for h>0 but are dependent and the covariance of Y_t and $Y_{t-1} = 0$
        4. Y_t has heavier tails than a normal distribution (leptokurtotic): $K = 3 \cdot \frac{1-\alpha_1^2}{1-3\alpha_1^2}$
            1. Excess Kurtosis: $K - K_{normal} = K - 3 = \frac{6\alpha_1^2}{1-3\alpha_1^2} > 0$
            2. implies market crashes are more likely than what a normal model would predict
        5. Y_t is symmetric (dependent symmetric heavy-tailed white noise process)
    6. ARCH roughly follows AR(1) as non-normal, mean zero innovation process η_t which is uncorrelated but dependent (ARCH effect)
    7. If $Y_t \sim WN$ and $Y_t^2 \sim AR(1)$ then $Y_t \sim ARCH(1)$.
3. ARCH(p): $Y_t = \sigma_t e_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i Y_{t-i}^2$
    1. Stationary if $\sum_{i=0}^p \alpha_i < 1$ (sum of coefficients less than 1)
4. GARCH(p,q)
    1. GARCH extends ARCH by including lagged values of the conditional variance itself
        1. GARCH: $Y_t = \sigma_t e_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i Y_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
    2. Stationary if $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$

    1. all $\alpha, \beta > 0$

3. Fitting an ARIMA-GARCH model to the mean and variance
    1.
        1. Make X_t stationary (if needed) using differencing or transformation → W_t.
    2. Fit an ARMA(p, q) model to the stationary series W_t. Then define residuals: $Y_t = W_t - ARMA(p, q)$
    3. Check whether Y_t exhibits ARCH effects. In other words, whether it is white noise, heavy-tailed, and $Y_t^2 \approx ARMA(p_0, q_0)$
    4. Fit a GARCH(p_0, q_0) model to the residuals Y_t if ARCH effects are present.
    5. Caveats
       1. In general, $p_0, q_0 \leq 2$.
       2. Y_t^2 rarely follows a simple ARMA structure.
       3. Sometimes $|Y_t|$ exhibits a better behavior, close to the predicted ARMA structure.
       4. In reality, one would fit a low-dimensional GARCH model to Y_t and then check the residuals for GARCH effects.