Time Domain

Definitions

Random Walks: No secular trend but correlated results, $\mu_t = 0$, $\gamma(s,t) = min(s,t)$
White Noise Process: $\{Z_1, \ldots, Z_n\}$ has mean 0 and variance σ^2 and are uncorrelated
IID Noise Process: White noise but each X_t is IID
Gaussian Noise Process: IID but each X_t is also normally distributed

White Noise

$\{X_1, \ldots, X_t\}$ is a white noise process if it has mean 0 and covariance σ^2. All the terms are uncorrelated.

Reducing Processes to White Noise

Subtracting mean: Remove μ_t. Check if residuals white noise.
Parametric Detrending: Assume parametric form for mean function μ_t. Linear, quadratic/higher order, sin and cos for seasonal or sinusoidal. Get $\hat{\mu}_t$ for each t and obtain residuals from $X_t - \hat{\mu_t}$.
Moving Average Smoothing: Take local averages in a window of q s.t. $\mu_t = \frac{1}{2q+1}\sum_{j=-q}^{q} X_{t+j}$. Use this over parametric modeling when you have local smoothness assumption vs. global linear assumption.
Filtering: Moving average smoothing is a special case of filtering. Filtering is when we apply linear transformation $\mu_t = \sum_{j=-q}^{s} a_j X_{t+j}$.
In MA smoothing, each $a_j = \frac{1}{q+s+1}$.
Kernel Smoothing: For all datapoints, each μ_t is defined as a weighted sum of the datapoints where weight a_j is determined by kernel function $K(z)$
Lowess Smoothing: Combination of regression, MA, kernel smoothing. Use fixed window, weight each point and apply weighted regression to predict $\hat{\mu_t}$.
Exponential Smoothing: Use past values only and make a_j geometrically decreasing for older values, making sure weights sum up to 1. Only uses info from past, good for forecasting, trend line lags behind major movements.
Differencing: $\nabla X_t = X_t - X_{t-1}$ where $\nabla = 1 - B$.

Seasonality

Mean has oscillating seasonal trend: $\mu_t = s_t$ where $S_{t+d} = s_t$ for all t.
$Acos(\phi + w\pi\omega t) = A_1 cos(2\pi\omega t) + A_2 sin(2\pi\omega t)$

Stationarity

Stationarity - Stochastic process $\{X_t\}$ is weakly stationary:

1. $E[X_t]$ is the same for all times t.
2. $Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h})$ for every s, t, h. Also, an equivalent definition is $Cov(X_t, X_{t+h})$ is the same for all t. This implies $Cov(X_t, X_s)$ is only a function of $|t-s|$ so $Cov(X_t, X_s) = \gamma(|t-s|)$ for all s, t.

$Var(X_t) = \gamma(0)$ for a weakly stationary process and all X_t have the same variance.

Autocorrelation Function

Sample ACF is a random vector: $R_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$

$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

ARMA Processes

Properties

Not redundant $\leftrightarrow$ $\phi(z)$ and $\theta(z)$ share no roots
Stationary $\leftrightarrow$ $\phi(z)$ has no roots with magnitude one
causal stationary Solution $\leftrightarrow$ $\phi(z)$ has no roots with magnitude less than or equal to one
invertible solution $\leftrightarrow$ $\theta(z)$ has no roots magnitude less than or equal to one

Calculating the ACVF

1. Explicit solution $X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$
2. Compute ACVF: $\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$
3. ACF: $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Coefficient Matching

$\phi(z)\psi(z) = \theta(z) \leftrightarrow$
$(1 - \phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \ldots) = (1 + \theta_1 z + \cdots + \theta_q z^q)$
$\psi_0 = 1,\ \psi_1 - \phi_1\psi_0 = \theta_1,\ \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 = \theta_2,\ \ldots$

Yule-Walker Equations

For an $ARMA(p,q)$ model $\phi(B)X_t = \theta(B)W_t$:

$$\phi(B)\gamma(k) = \begin{cases} \sigma^2 \sum_{j=0}^{q-k} \psi_j \theta_{k+j} & k \le q \\ 0 & o.w. \end{cases}$$

$$\phi(B)\rho(k) = \begin{cases} \frac{\sigma^2}{\gamma(0)} \sum_{j=0}^{q-k} \psi_j \theta_{k+j} & k \le q \\ 0 & o.w. \end{cases}$$

First Order Difference Equations

$\mu_k - \alpha\mu_{k-1} = 0$ for $k = 1, 2, \ldots$
Solution: $\mu_k = \alpha^k b_0$ where $\mu_0 = b_0$.

Second Order Difference Equations

$\mu_k - \alpha_1\mu_{k-1} - \alpha_2\mu_{k-2} = 0$ for $k = 2, 3, \ldots$

1. $z_1 \ne z_2$ and real
$\mu_k = c_1 z_1^{-k} + c_2 z_2^{-k}$
2. $z_1 = z_2$ and real
$\mu_k = z_1^{-k}(c_1 + c_2 k)$
3. $z_1 = \bar{z_2}$ and imaginary
$\mu_k = c_1 z_1^{-k} + \bar{c_1}\bar{z_1}^{-k} = 2a|z_1|^{-k}(cos(h\theta + b)) = 2|z_1|^{-h}(a_1 cos(h\theta) + a_2 sin(h\theta))$

Bartlett's Formula

Approximate the distribution of R_k approximately centered around $\rho(k)$ when $K \ll n$. The covariance matrix is defined as $\frac{W}{n}$ where W is composed of w_{ij} defined as follows:
$Cov(R_i, R_j) = w_{ij} =$
$\sum_{m=1}^{\infty}\big(\rho(m+i) + \rho(m-i) - 2\rho(i)\rho(m)\big)\big(\rho(m+j) + \rho(m-j) - 2\rho(j)\rho(m)\big)$

$$Cov(R_{1:k}) = \frac{1}{n}\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & & \vdots \\ \vdots & & \ddots & \\ w_{K1} & \cdots & & w_{KK} \end{bmatrix}$$

Each R_k approximately follows the normal distribution centered around $\rho(k)$: $R_k$follows $\mathcal{N}\big(\rho(k),\ \sqrt{\frac{w_{kk}}{n}}\big)$
Correlations between sample autocorrelations:
$corr(R_i, R_j) = \frac{Cov(R_i, R_j)}{\sqrt{Var(R_i, R_j)}} \approx \frac{w_{ij}/n}{\sqrt{w_{ii}/n * w_{jj}/n}} = \frac{wij}{\sqrt{w_{ii}w_{jj}}}$

Best Linear Predictor

$\hat{Y} = a_1 Z_1 + \cdots + a_m Z_m = a^T Z$ s.t. $Cov(Y - a^T Z, Z_i) = 0$
In other words, the difference between the actual value Y and predicted value $\hat{Y} = a^T Z$ is uncorrelated to every Z_i. The covariance is 0 because all linearly representable information about Y in Z is incorporated into the predictor. We use the fact that $Cov(Y - a^T Z, Z_i) = 0$ to determine a
$\zeta = [\zeta_1 \quad \ldots \quad \zeta_m]^T\ s.t.\ \zeta_i = Cov(Y, Z_i),\ \Delta\ s.t.\ \delta_{ij} = Cov(Z_i, Z_j)$
$Cov(Z, Y) - (Cov(Z, Z))a = \zeta - \Delta a = \mathbf{0} \to a = \Delta^{-1}\zeta,\ \hat{X_t} = \zeta^T \Delta^{-1} X$

Matrix Representation of a

$$a = \Delta^{-1}\zeta = \begin{bmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(m-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(m-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(m-1) & \gamma(m-2) & \ldots & \gamma(0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(m) \end{bmatrix}$$

One Predictor Example

$Z = Z_1$, $\zeta = Cov(Y, Z_1)$, $\Delta = Var(Z_1)$
$a = \Delta^{-1}\zeta = \frac{Cov(Y, Z_1)}{Var(Z_1)}$
$\hat{Y} = a^T Z = \frac{Cov(Y, Z_1)}{Var(Z_1)} Z_1$
If we define want to predict X_t (a stationary process) from X_{t-1}:
$\hat{X}_t = \frac{Cov(X_t, X_{t-1})}{Var(X_{t-1})} X_{t-1} = \frac{\gamma(1)}{\gamma(0)} X_{t-1} = \rho(1) X_{t-1}$

We can now generalize this: $X = (X_{t-1}, \ldots, X_{t-m})$,
$\zeta_i = Cov(X_t, X_{t-i}) = \gamma(i)$, $\delta_{ij} = Cov(X_{t-i}, X_{t-j}) = \gamma(i-j)$
This gives us the following BLP: $\hat{X}_t = (\Delta^{-1}\zeta)^T X = \zeta^T \Delta^{-1} X$.

BLP Facts

We can calculate the BLP knowing only the covariance.
If process X_t is gaussian, the BLP is the best predictor.

Partial Autocorrelation Function

$PACF(h)$ is the coefficient of X_{t-h} in the best linear predictor of X_t given $X_{t-1}, \ldots, X_{t-h}$.

PACF Facts

$PACF(1) = \rho(1)$

Calculating PACF

$PACF(h)$ is coefficient of X_{t-h} in the BLP of X_t given $X_{t-1}, \ldots, X_{t-h}$
$\hat{X_t}^{(1)} = a_1^{(1)} X_{t-1},\ \hat{X_t}^{(2)} = a_1^{(2)} X_{t-1} + a_2^{(2)} X_{t-2},\ \ldots,\ \hat{X_t}^{(m)} = a_1^{(m)} X_{t-1} + \cdots + a_m^{(m)} X_{t-m}$

PACF Meaning

$pacf(h)$ is the correlation between X_t and X_{t-h} with the effect of $X_{t-1}, \ldots, X_{t-h+1}$ removed.
$pacf(h) = corr(X_t - \hat{X_t}, X_{t-h} - \hat{X_{t-h}})$ where $\hat{X_t}$ and $\hat{X_{t-h}}$ are the BLP of X_t and X_{t-1} in terms of $X_{t-1}, \ldots, X_{t-h+1}$.
By stationarity, the BLP of $\hat{X_t}$ and $\hat{X_{t-h}}$ are the ceofficent vectors are the same.
Large magnitude of $pacf(h)$ indicates that X_{t-h} has lots of information about X_t, compared to intervening time points.
Alternate definition: $pacf(h) = corr(X_t - a_1^{(h-1)} X_{t-1} - \cdots - a_{h-1}^{(h-1)} X_{t-h+1},\ X_{t-h} - a_1^{(h-1)} X_{t-h+1} - \cdots - a_{h-1}^{(h-1)} X_{t-1})$

Estimation

Method of Moments for AR(p)

Match sample moments to theoretical moments:
Estimate μ as $\frac{(x_1 + \cdots + x_n)}{n}$
For parameters $\phi_1, \ldots, \phi_p$ and σ^2, recall:
$\gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p) = \sigma^2$
$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0$
To estimate, plug in sample ACF $\gamma(\hat{k})$ and solve for $\phi(z)$ and σ^2.
$\gamma(\hat{k}) = \frac{1}{n}\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(x_t - \bar{x})$

AR(1) Example

We solve the following equations for σ^2 and $\phi(z)$. $\gamma(\hat{0}) - \phi\gamma(\hat{1}) = \sigma^2$
$\gamma(\hat{1}) - \phi\gamma(\hat{0}) = 0$
Solve:
$p\hat{h}i_1 = \frac{\gamma(\hat{1})}{\gamma(\hat{0})} = r_1\ \ \hat{\sigma^2} = \gamma(\hat{0})(1 - r_1^2)$

AR(2) Example

$\gamma(\hat{0}) - \phi_1\gamma(\hat{1}) - \phi_2\gamma(\hat{2}) = \sigma^2$
$\gamma(\hat{1}) - \phi_1\gamma(\hat{0}) - \phi_2\gamma(\hat{1}) = 0$
$\gamma(\hat{2}) - \phi_1\gamma(\hat{1}) - \phi_2\gamma(\hat{0}) = 0$
Solve defining r_k as the sample autocorrelations at lag k:
$\hat{\phi_1} = \frac{r_1(1 - r_2)}{1 - r_1^2}\ \ \hat{\phi_2} = \frac{r_2 - r_1^2}{1 - r_1^2}$

Least Squares Estimation for AR(p)

Mimics the best linear prediction. Given $x_1, \ldots, x_n$, fit an $AR(p)$ model. We estimate the BLP for $X_t - \mu$ given $X_{t-1} - \mu, \ldots, X_{t-p} - \mu$ by minimizing $x^(t) = [x_{t-1} \quad \ldots \quad x_{t-p}]$. We define the conditional sum of squares formula: $S_c(\mu, a) = \sum_{t=p+1}^{n}\big((x_t - \mu) - a^T(x^{(t)} - \mu)\big)^2$.
For an $AR(p)$ model, $a_1 = \phi_1, \ldots, a_p = \phi_p$.

AR(1) Example

$S_c(\mu, \phi) = \sum_{t=2}^n \left((x_t - \mu) - \phi(x_{t-1} - \mu)\right)^2 =$
$\sum_{t=2}^n \left(x_t - \mu - \phi x^{(t)} - \phi\mu\right)^2 = \sum_{t=2}^n \left(x_t - \mu(1-\phi) - \phi x^{(t)}\right)^2$
Define $\beta_0 = \mu(1-\phi)$ and $\beta_1 = \phi$
$\sum_{t=2}^n \left(x_t - \beta_0 - \beta_1 x^{(t)}\right)^2$
This is just linear regression with parameters β_0 and β_1. We get the following results:
$\hat{\beta_1} = \frac{\sum_{t=1}^n((x_t - \bar{x_{(2)}})(x_{t-1} - \bar{x_{(1)}})}{\sum_{t=1}^n (x_{t-1} - \bar{x_{(1)}})^2}$ where $\bar{x_{(1)}} = \frac{x_1 + \cdots + x_{n-1}}{n-1}$ and $\bar{x_{(2)}} = \frac{x_2 + \cdots + x_n}{n-1}$. $\beta_2 = \bar{x_{(2)}} - \hat{\beta_1}\bar{x_{(1)}}$. This gives us the following results for $\hat{\phi}$ and $\hat{\mu}$:
$\hat{\phi} = \hat{\beta_1},\ \hat{\mu} = \frac{\bar{x_{(2)}} - \hat{\phi}\bar{x_{(1)}}}{1 - \hat{\phi}}$

Maximum Likelihood Estimation

We assume that X_t is a gaussian stationary process and the random variables $X_{(1)}, \ldots, X_{(n)}$ follow a multivariate distribution. The gaussian process has mean function μ_t and covariance function parameter $\Gamma(s,t)$. For an $AR(p)$, the dataset is distribution with mean $(\mu, \ldots, \mu)^T$ and has covariance matrix Γ with entries $\gamma_{ij} = \gamma(i-j)$. These entries are functions of σ^2 and $\phi_1, \ldots, \phi_p$. We maximize the log-likelihood function:
$L(\mu, \phi, \sigma^2) = (2\pi)^{-n/2}|\Gamma|^{-1/2} \exp\Big(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\Big)$.

MLE AR(1) Example

Decompose joint density:
$f_{\mu,\phi,\sigma^2}(x_1, \ldots, x_n) = f(x_1)f(x_2|x_1)\ldots f(x_n|x_1, \ldots, x_{n-1})$. For $i \geq 2$, the conditional distribution of x_i is normal with mean $\mu + \phi(x_{i-1} - \mu)$ and variance σ^2. For $i = 1$, normal with mean μ and variance $\frac{\sigma^2}{(1-\phi^2)}$. We get the following likelihood function:
$L(\mu, \phi, \sigma^2) = (2\pi\sigma^2)^{n/2}(1-\phi^2)^{1/2}\exp(-\frac{S(\mu,\phi)}{2\sigma^2}$ where
$S(\mu, \phi) = (1-\phi^2)(x_1 - \mu)^2 + \sum_{t=2}^n \Big(x_t - \mu - \phi(x_{t-1} - \mu)\Big)^2$. The S function is the unconditioned sum of squares function and this whole this is a nonlinear optimization problem so there is no closed form solution.

AR(p) Estimation

Three methods give similar results. Yule Walker uses least information and MLE uses most information. MLE numerically unstable. All methods converge to the same solution asymptotically.

Estimation Uncertainty

How do we estimate the uncertainty in parameter estimates for $AR(p)$ - used to develop confidence intervals.
Asymptotic Indifference: 3 estimation methods converge to same result for infinitely long process.

Asymptotic Properties

Consistency: $\phi = (\phi_1, \ldots, \phi_p)$ For large n, $\hat{\phi} \to \phi$.
Normality: For large n, approximate distribution $\sqrt{n}(\hat{\phi} - \phi)$ is normal with mean 0 and covariance matrix $\sigma^2\Gamma_p^{-1}$ where Γ_p is the covariance matrix defining each entry $\gamma_{ij} = \gamma(i-j)$. We use this fact to define the distribution of $\hat{\phi}$ as $N(\phi, \frac{\sigma^2}{n}\Gamma_p^{-1})$. This is a multivariate normal distribution with mean vector ϕ and covariance matrix $\frac{\sigma^2}{n}\Gamma_p^{-1}$.

Nested Models

We show that if we use an $AR(2)$ to approximate and $AR(1)$, we lose precision in the model.
For $AR(2)$, $Var(\phi_1) = \frac{(1-\phi_2^2)}{n} = 0$
For $AR(1)$, $Var(\phi_2) = \frac{(1-\phi_1^2)}{n}$
It seems we lose precision when we use nested models.

Estimation for ARMA

X_t follow $\phi(B)X_t = \theta(B)W_t$. From observed parameters, x_i, estimate ϕ, θ, μ, σ^2. Don't use Yule-Walker method - inaccurate and solution does not necessarily exist unlike AR. We discuss first conditional least squares. Treat first p points as constants since we cannot find residuals. They are initial condition. Explain as much observed variance as possible with parameters.

Conditional Least Squares MA(1)

Isolate white noise component and minimize variance of white noise. For $MA(1)$, $X_t - \mu = W_t - \theta W_{t-1}$. This gives us the following equations:
$w_1 = x_1 - \mu - \theta w_0$
$w_2 = x_2 - \mu - \theta w_1$
...
$w_n = x_n - \mu - \theta w_{n-1}$
We assume $w_0 = 0$ (conditional part) and for any guessed μ, θ, we can recover the $w_1, \ldots, w_n$, giving us minimization equation:
$S_c(\mu, \theta) = \sum_{i=1}^n w_i^2$. We plug in implicit data w_i:
$S_c(\mu, \theta) = \sum_{i=1}^n \Big((x_i - \mu) - \theta(x_{i-1} - \mu)\Big)^2$

Conditional Least Squares ARMA(1, 1)

$X_t - \mu - \phi(X_{t-1} - \mu) = W_t + \theta W_{t-1}$. Assume $w_1 = w_0 = 0$,
$w_2 = x_2 - \mu - \phi(x_1 - \mu) - \theta w_1,\ \ldots,\ w_n = x_n - \mu - \phi(x_{n-1} - \mu) - \theta w_{n-1}$.
Then we minimize $S_c(\mu, \theta, \phi) = \sum_{i=2}^n w_i^2$

Conditional Least Squares ARMA(p, q)

We now generalize to $ARMA(p,q)$.
$X_t - \mu - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}$
If we know that $w_1, \ldots, w_p = 0$ solve
$w_t = x_t - \mu - \phi_1(x_{t-1} - \mu) - \cdots - \phi_p(x_{t-p} - \mu) - \theta_1 w_{t-1} - \cdots - \theta_q w_{t-q}$
for $t \geq p+1$.
We then set $w_a, \ldots, w_p = 0$ where $a = \min(0, p-q)$ treating $x_1, \ldots, x_p$ are constants. For guessed values of μ, ϕ, θ, can get $w_{p+1}, \ldots, w_n$.
Minimize $S_c(\mu, \phi, \theta) = \sum_{i=p+1}^n w_i^2$. Non-linear optimization problem so no closed from solution.

MLE for ARMA(p, q)

Likelihood of observed data vector $x = (x_1, \ldots, x_n)$ is
$L(\mu, \phi, \theta, \sigma^2) = (2\pi)^{-n/2}|\Gamma|^{-1/2}\exp\Big(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\Big)$.
$\Gamma(i,j) = \gamma(i-j)$ where $\gamma(h)$ is the theoretical ACVF.

Asymptotic Properties

For large n, $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$, $\hat{\beta} \to \beta$.
Distribution: $\sqrt{n}(\beta - \hat{\beta}) \to N(0, \sigma^2\Gamma_{p,q}^{-1})$ where $\Gamma_{p,q}$ is a $(p+q) \times (p+q)$ matrix of the form $\Gamma_{p,q} = \begin{bmatrix}\Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta}\end{bmatrix}$ Computing $\Gamma_{p,q}$: define two AR processes A_t and B_t with same W_t. $\phi(B)A_t = W_t$ and $\theta(B)B_t = W_t$. We define the Γ matrices as follows:
$\Gamma_{\phi\phi}(i,j) = \gamma_A(i-j)$
$\Gamma_{\theta\theta}(i,j) = \gamma_B(i-j)$
$\Gamma_{\phi\theta}(i,j) = \Gamma_{\theta\phi}(j,i) = Cov(A_i, B_j)$
Distribution of $\hat{\beta}$ is $N(\beta, \frac{\sigma^2}{n}\Gamma_{p,q}^{-1})$.

ARIMA Models

Process Y_t is said to be ARIMA if $X_t = (I-B)^d Y_t$ is ARMA(p, q) with mean μ. Equation: $\phi(B)(\nabla^d Y_t - \mu) = \phi(B)(X_t - \mu) = \theta(B)W_t$

Forecasting

Predict X_{n+m} given $X_1, \ldots, X_m$ and ARIMA model. Use BLP for $\hat{X_{n+m}}$.
$\hat{X_{n+m}} - \mu = a_1^{(n,m)}(X_n - \mu) + \cdots + a_n^{(n,m)}(X_1 - \mu)$ where
$a^{(n,m)} = (a_1^{(n,m)}, \ldots, a_n^{(n,m)})$ satisfies $Cov(X_{n+m} - \hat{X_{n+m}}, X_i) = 0$.
We can also define Δ and ζ where
$\Delta(i,j) = Cov(X_{n-1}, X_{n-j}) = \gamma(i-j)$ and
$\zeta_i^{(m)} = Cov(X_{n+m}, X_{n-i}) = \gamma(m+i)$.
The coefficients satisfy: $\Delta a^{(n,m)} = \zeta^{(m)}$

Prediction Interval

For a stationary time series, the mean squared prediction error:
$P^{(m)} = E[(X_{n+m} - \hat{X_{n+m}})^2] = \gamma(0) - \zeta^{(m)}\Delta^{-1}\zeta^{(m)}$
$\hat{X_{n+m}} = E(X_{n+m}|X_n, \ldots, X_1)$
$P^{(m)} = Var(X_{n+m}|X_n, \ldots, X_1)$
95 percent confidence interval: $\hat{X_{n+m}} \pm 1.96\sqrt{P^{(m)}}$

Standized Residual ACF

Let $r_e(h)$ be the sample ACF of standardized residuals.

Ljung-Box-Pierce Test

For $h \ll n$: $Q = n(n+2)\sum_{h=1}^H \frac{r_e^2(h)}{n-h}$. Q follows χ^2_{H-p-q} Reject is Q is too large.

Seasonal ARMA

$ARMA(P,Q)_s$ satisfies $\Phi(B^s)X_t = \Theta(B^s)W_t$.
ACF and PACF are non zero only at seasonal lags $h = 0, s, 2s, \ldots$.

Multiplicative Seasonal ARMA

$ARMA(p,q) \times (P,Q)$ satisfies $\phi(B)\Phi(B^s)X_t = \Theta(B^s)\theta(B)W_t$.
ARMA(0, 1)x(0, 1) Properties
ACF is 0 for all except the following:
$\rho(1) = \frac{\theta}{1+\theta^2}$ and $\rho(12) = \frac{\Theta}{1+\Theta^2}$
$\rho(11) = \rho(13) = \frac{\theta\Theta}{(1+\theta^2)(1+\Theta^2)}$
ARMA(0, 1)x(1, 0) Properties
ACF is 0 except the following:
$\rho(12h) = \Phi^h$ and $\rho(12h-1) = \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h$

SARIMA Models

$ARMA(p,d,q) \times (P,D,Q)_s$ satisfies
$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \delta + \Theta(B^s)\theta(B)W_t$.

Hierarchy of Heuristics

Internal Validity

Is the model internally consistent? Goodness of fit tests, if we re-predict the data, does it do good job

Local External Validity

Does model predict well in identical replications? AIC, BIC, standard cross-val, if we repeat experiment, predict outcomes well
AIC - Akaike Information Criterion:
$AIC = -2\log(\text{maximum likelihood}) + 2k$
BIC - Bayesian Information Criterion:
$BIC = -2\log(\text{maximum likelihood}) + k\log n$
Want to choose model with the smallest score. We want a model that captures structure that persists between replications and ignore noise. Penalize overfitting. AIC good if truth complex and BIC good if truth simple.

General External Validity

Does model predict well in non-identical replications? well designed cross-validation, extrapolate to new contexts and still predict outcomes well, prediction is extrapolation.
Cross-Validation Experiment

1. Exclude datapoints $x_{n-m+1}, \ldots, x_n$
2. Fit using $x_1, \ldots, x_{n-m}$.
3. Construct estimates $\hat{x_{n-m+1}}, \ldots, \hat{x_n}$
4. Compute error measurement.

Frequency Domain

Discrete Fourier Transform

Fit regression with terms for n Fourier frequencies. For data $x_0, \ldots, x_{n-1}$, derive coefficients b_j for each x_t:
$x_t = \frac{1}{n}\sum_{j=0}^{n-1} b_j exp(\frac{2\pi ijt}{n})$
Coefficient b_j represent the DFT:
$b_j = \sum_{t=0}^{n-1} x_t exp(-\frac{2\pi ijt}{n})$ for $j = 0, \ldots, n-1$
Symmetry:
$b_{n-j} = \sum_t x_t exp(-\frac{2\pi i(n-j)t}{n}) = \sum_t x_t exp(\frac{2\pi ijt}{n})exp(-2\pi it) = \bar{(b_j)}$

Periodogram

Representing the contribution of a frequency in terms of the magnitude of DFT.
$I(j/n) = \frac{1}{n}|b_j|^2$ for $j = 0, \ldots, \text{floor}(n/2)$
Spikes represent contributions of fourier frequencies. For signal + noise data, large magnitude spikes for signal, smaller magnitude spikes for noise.

Periodogram from Sample ACVF

$I(j/n) = \frac{|b_j|}{n} = \sum_{|h|<n} \hat{\gamma}(h) exp\big(-\frac{2\pi ijh}{n}\big)$ for $j = 1, \ldots, n-1$

Proof

$\sum_{t=0}^{n-1} exp(-\frac{2\pi ijt}{n}) = 0$ for $j = 1, \ldots, n-1$ because it's a geometric formula and $z = exp(-\frac{2\pi ijt}{n}$ so the sum is equal $\frac{1-z^n}{1-z}$. Numerator is 0. Basically $exp(-\frac{2\pi ijt}{n}$ cancels $exp(-\frac{2\pi ij(n-t)}{n}$.
The above formula basically shows that you can subtract the mean of the data without changing b_j: $b_j = \sum_{t=0}^{n-1}(x_t - \bar{x})exp(-\frac{2\pi ijt}{n})$
$|b_j|^2 = b_j\bar{b_j} = \sum_{t=0}^{n-1}(x_t - \bar{x})exp(-\frac{2\pi ijt}{n})\sum_{s=0}^{n-1}(x_s - \bar{x})exp(\frac{2\pi ijs}{n})$
$= \sum_{t=0}^{n-1}\sum_{s=0}^{n-1}(x_t - \bar{x})(x_s - \bar{x})exp(-\frac{2\pi ijt}{n})exp(\frac{2\pi ijs}{n})$
$= \sum_{t=0}^{n-1}\sum_{s=0}^{n-1}(x_t - \bar{x})(x_s - \bar{x})exp(-\frac{2\pi ij(t-s)}{n})$
$= \sum_{h=-(n-1)}^{n-1}\sum_{t,s:t-s=h}(x_t - \bar{x})(x_s - \bar{x})exp(-\frac{2\pi ij(t-s)}{n})$
$= n\sum_{|h|<n}\hat{\gamma(h)}exp(-\frac{2\pi ijh}{n})$ The sample ACVF and the Periodogram give useful summaries of a sample that are linked. Move back and forth between frequency and covariance representation.

Process Representation

Goal: represent stationary process as a sum of sinusoids with random coefficients.

Simple Process

$X_t = Acos(2\pi\lambda t) + Bsin(2\pi\lambda t)$ where λ is fixed frequency and A and B uncorrelated random variables with mean 0 and variance σ^2. Stationary because $E[X_t] = 0$ and
$Var(X_t) = Var(A)cos^2(2\pi\lambda t) + Var(B)sin^2(2\pi\lambda t) =$
$\sigma^2(cos^2(2\pi\lambda t) + sin^2(2\pi\lambda t)) = \sigma^2$
Stationary covariance: $Cov(X_t, X_s) = Var(A)cos(2\pi\lambda t)cos(2\pi\lambda s) +$
$Cov(A,B)(cos(2\pi\lambda t)sin(2\pi\lambda s)) + Cov(A,B)(cos(2\pi\lambda s)sin(2\pi\lambda t)) +$
$Var(B)sin(2\pi\lambda t) + sin(2\pi\lambda t)sin(2\pi\lambda s))$
$= \sigma^2(cos(2\pi\lambda t)cos(2\pi\lambda s) + sin(2\pi\lambda t)sin(2\pi\lambda s))$
$= \sigma^2(cos(2\pi\lambda(t-s)))$

Complex Processes

$X_t = \sum_{j=1}^{m}(A_j cos(2\pi\lambda_j t) + B_j sin(2\pi\lambda_j t))$
Approximate any stationary process arbitrarily well if m is large enough and correct frequencies and variances are chosen. For white noise:
$\lambda_j = \frac{j}{2m}$ and $\sigma_j^2 = \frac{\sigma^2}{m}$
For more complicated process, we select a large N where N is the number of frequencies to consider and define the frequencies to be evenly spaced: $\lambda_j = \frac{j}{2N}$. We define the variance for each A_j and B_j to be as follows: $\sigma_j = \frac{f(\lambda)}{N}$. We generate random samples for A_i's and B_i's following distribution $N(0, \sigma_i^2)$. Then plug back into the equation for X_t.

Spectral Density

A sample from any stationary proces has a periodogram $I(j/n)$ that is a noisy verson of the spectral density $f(\lambda)$. We choose λ_j and σ_j^2 for processes with the spectral density.
$f(\lambda) = \sum_{h=-\infty}^{\infty}\gamma(h)exp(-2\pi i\lambda h)$ for $0 \leq \lambda \leq 1/2$

Properties

Symmetric about 0: $f(-\lambda) = f(\lambda)$
Symmetric about 1/2: $f(\lambda) = f(1/2 + (1/2 - \lambda))$

ACVF Relationship

$\gamma(h) = \int_{-1/2}^{1/2} exp(2\pi i\lambda h)f(\lambda)d\lambda$
$\gamma(0) = \sigma_X^2$

Process Representation

Choose λ_j, σ_j^2 where $\sigma_X^2 = \sum_{j=1}^{m}\sigma_j^2$. Fix $\lambda_j = \frac{j}{2m}$. Evenly spaced frequencies. $\sigma_j^2 = \frac{f(\lambda_j)}{m}$. As m approaches ∞, converges to a stationary process.

White Noise

$f(\lambda) = \sum_{h=-\infty}^{\infty}\gamma(h)\exp(-2\pi i\lambda h) = \gamma(0) = \sigma_X^2$. All $\gamma(k)$ are zero except the variance of X_t.

MA(1)

Consider an MA(1) process with white noise variance σ_W^2. Thus,
$\gamma(0) = \sigma_W(1+\theta^2)$ and $\gamma(\pm 1) = \theta\sigma_W^2$
$f(\lambda) = \sum_{h=-\infty}^{\infty}\gamma(h)\exp(-2\pi i\lambda h) =$
$\gamma(-1)\exp(2\pi i\lambda) + \gamma(0)\exp(0) + \gamma(1)\exp(-2\pi i\lambda)$
$= \gamma(0) + \gamma(1)(\exp(2\pi i\lambda) + \exp(-2\pi i\lambda))$
$= \gamma(0) + 2\gamma(1)cos(2\pi\lambda)$
$= \sigma_W^2(1 + \theta^2 + 2\theta\cos(2\pi\lambda))$ for $-\frac{1}{2} \leq \lambda \leq \frac{1}{2}$

Linear Time-Invariant Filter

LTI set of coefficients a_k to transform input X_t to output Y_t according to $Y_t = \sum_{k=-\infty}^{\infty} a_k X_{t-k}$.

Impulse Response Function

Set of coefficients a_k defined as a function of k. $X_t = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{o.w.} \end{cases}$.

Other examples are moving average filter: $a_k = \begin{cases} \frac{1}{2q+1} & \text{for } |k| \leq q \\ 0 & \text{o.w.} \end{cases}$

and differencing: $a_k = \begin{cases} -1 & \text{for } k = 1 \\ 1 & \text{for } k = 0 \\ 0 & \text{o.w.} \end{cases}$

Power Transfer Function Derivation

ACVF Modification

Input X_t with ACVF $\gamma_X(h)$. Output Y_t with ACVF $\gamma(h)$.
$\gamma_Y(h) = Cov(\sum_{k=-\infty}^{\infty} a_k X_{t-k}, \sum_{l=-\infty}^{\infty} a_l X_{t+h-l})$
$= \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_k a_l Cov(X_{t-k}X_{t+h-l})$
$= \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_k a_l \gamma_X(h - l + k)$

Spectral Density Modification

Using the result above, let us derive $f_Y(\lambda)$ where $f_X(\lambda)$ is the spectral density of X_t. Recall that $\gamma_X(h) = \int_{-1/2}^{1/2}\exp(2\pi ih\lambda)f_X(\lambda)d\lambda$.
$\gamma_Y(h) = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_k a_l \gamma_X(h - l + k)$
$= \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_k a_l \int_{-1/2}^{1/2}\exp(2\pi i(h - l + k)\lambda)f_X(\lambda)d\lambda$
$=$
$\int_{-1/2}^{1/2} f_X(\lambda)\exp(2\pi ih\lambda)\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} a_k a_l \exp(-2\pi il\lambda)\exp(2\pi ik\lambda)d\lambda$
Let us define $A(\lambda) = \sum_{k=-\infty}^{\infty} a_k \exp(-2\pi ik\lambda)$.
$= \int_{-1/2}^{1/2} f_X(\lambda)\exp(2\pi ih\lambda)A(\lambda)\bar{A(\lambda)}d\lambda$
$= \int_{-1/2}^{1/2} f_X(\lambda)\exp(2\pi ih\lambda)|A(\lambda)|^2 d\lambda$
That means that $f_Y(\lambda) = f_X(\lambda)|A(\lambda)|^2$. $A(\lambda)$ is the transfer function or frequency response function. $|A(\lambda)|^2$ is the power transfer function, specifying which frequencies get amplified and damped by the filter.

ARMA Spectral Density

$U_t = \phi(B)X_t = \theta(B)W_t$ where U_t is in terms of two linear filters.
$f_U(\lambda) = |A_\phi(\lambda)|^2 f_X(\lambda)$ and $f_U(\lambda) = |A_\theta(\lambda)|^2 f_W(\lambda) = \sigma_W^2|A_\theta(\lambda)|^2$.
Impulse response functions are just the coefficients $-\phi_k$ and θ_k.
$f_X(\lambda) = \frac{|A_\theta(\lambda)|^2}{|A_\phi(\lambda)|^2}\sigma_W^2$ $A(\lambda) = \sum_{k=-\infty}^{\infty} a_k \exp(-2\pi ik\lambda)$
$A_\phi(\lambda) = 1 - \phi_1\exp(-2\pi i\lambda) - \phi_2\exp(-2\pi i(2\lambda)) - \cdots -$
$\phi_p\exp(-2\pi i(p\lambda)) = \phi(\exp(-2\pi i\lambda)$. $A_\theta(\lambda) = \theta(\exp(-2\pi i\lambda)$.
$f_X(\lambda) = \sigma_W^2\frac{|\theta(\exp(-2\pi i\lambda)|^2}{|\phi(\exp(-2\pi i\lambda)|^2}$

Periodogram Distribution

We can estimate the power transfer function using estimates to the spectral density functions: $|A(\hat{\lambda})|^2 = \frac{\hat{f_y}(\lambda)}{\hat{f_x}(\lambda)}$
To estimate $f(\lambda)$ we can calculate the sample ACVF and plug into the periodogram: $I(\lambda) = \sum_{h:|h|<n}\hat{\gamma(h)}\exp(-2\pi i\lambda h)$ for $-1/2 \leq \lambda \leq 1/2$
When $\lambda \in (0, 1/2]$: $I(j/n) = \frac{|b_j|^2}{n}$ where $b_j = \sum_t x_t \exp\Big(-\frac{2\pi ijt}{n}\Big)$.
Distribution for General ARMA: $\frac{2I(j/n)}{f(j/n)}$ follows χ_2^2 for $0 < j < n/2$.
$E(\chi_2^2) = 2$.

Spectral Estimation By Smoothing

$f(\hat{j/n}) = \sum_{k=-m}^{m} W_m(k)I(\frac{j+k}{n})$ where $W_m(k)$ is called the kernel or spectral window and is an impulse response function.
Daniell Spectral Window/Kernel: $W_m(k) = \frac{1}{2m+1}$ for $-m \leq k \leq m$.
Confidence Interval
Given a Daniell Spectral Kernel with hyperparameter m, we can approximate the confidence of the kernel.
Find probability distribution is within those bounds. We let $\chi_2^2(\alpha/2)$ and $\chi_2^2(1 - \alpha/2)$ be quantiles.
$P\Big(\chi_{2(2m+1)}(\alpha/2) \leq \chi_{2(2m+1)}(\alpha/2) \leq \chi_{2(2m+1)}(1 - \alpha/2)\Big) = 1 - \alpha$
$C(\alpha) = \Bigg(2(2m+1)\frac{\hat{f}(j/n)}{\chi^2_{2(2m+1)}(1-\alpha/2)},\ 2(2m+1)\frac{\hat{f}(j/n)}{\chi^2_{2(2m+1)}(\alpha/2)}\Bigg)$

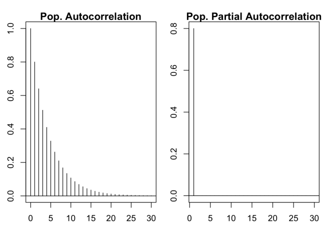
ARMA Summaries

AR(p)

PACF:
$PACF(p) = \phi_p$
$PACF(h) = 0$ for $h > p$

AR(1)

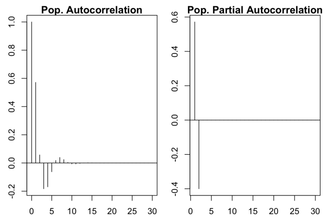
ACF and PACF Graph: ($\phi = 0.8$)



ACVF:
$\gamma(0) = \frac{\sigma^2}{1-\phi_1^2}$
$\gamma(k) = \frac{\sigma^2}{1-\phi_1^2}\phi_1^k$
Population ACF:
$\rho(i) = \rho^i$ when $|\phi| < 1$
$\frac{\rho(k+1)}{\rho(k)} = \phi_1$
Sample ACF Distribution:
$Var(R_1) \approx \frac{1-\phi^2}{n}$
$Var(R_i) \approx \frac{1}{n}\frac{1+\phi^2}{1-\phi^2}$
Spectral Density:
$f_X(\lambda) = \frac{\sigma_W^2}{1+\phi^2-2\phi\cos 2\pi\lambda}$ for $-\frac{1}{2} \leq \lambda \leq \frac{1}{2}$
Estimation:
$\Gamma_p = \Gamma_1 = \gamma(0) = \frac{\sigma^2}{(1-\phi_1^2)}$
$\hat{\phi_1} \sim N\Big(\phi, Var(\hat{\phi})\Big)$
$Var(\hat{\phi}) = \frac{\sigma^2}{n}\Gamma_1^{-1} = \frac{\sigma^2}{n}\frac{(1-\phi_1^2)}{\sigma^2} = \frac{1-\phi_1^2}{n}$

AR(2)

ACF and PACF Graph: ($\phi_1 = 0.8$, $\phi_2 = -0.4$)



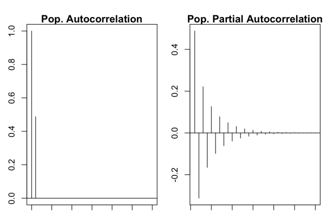
ACVF:
$\gamma(0) = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma^2}{(1-\phi_2)^2-\phi_1^2}$
$\gamma(k) = \frac{\phi_1^k \sigma^2}{1-\phi_1^2}$
Population ACF:
$\rho(1) = \frac{\phi_1}{1-\phi_2}$
Spectral Density:
$f_X(\lambda) = \frac{\sigma_W^2}{1+\phi_1^2+\phi_2^2-2\phi_1(1-\phi_2)\cos 2\pi\lambda-2\phi_2 \cos 4\pi\lambda}$ for $-\frac{1}{2} \leq \lambda \leq \frac{1}{2}$
Estimation:
$\gamma(0) = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma^2}{(1-\phi_2)^2-\phi_1^2}$
$\rho(1) = \frac{\phi_1}{1-\phi_2}$
$(\hat{\phi_1}, \hat{\phi_2}) \sim N\left(\begin{bmatrix}\hat{\phi_1}\\ \hat{\phi_2}\end{bmatrix}, \frac{1}{n}\begin{bmatrix}1-\phi_2^2 & -\phi_1(1+\phi_2)\\ -\phi_1(1+\phi_2) & 1-\phi_2^2\end{bmatrix}\right)$

MA(q)

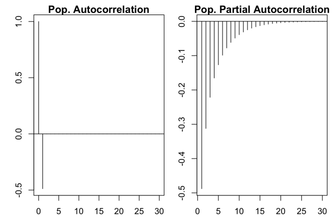
$X_t = \sum_{j=0}^{q}\theta_j W_{t-j}$
$Cov(X_t, X_s) = \begin{cases}\sigma^2\sum_{j=0}^{q-h}\theta_j\theta_{j+h} & for\ h = |t-s| \leq q\\ 0 & o.w.\end{cases}$ All lags k s.t. $k > q$ have ACF value $\rho(k) = 0$.

MA(1)

ACF and PACF Graph: ($\theta = 0.8$)



ACF and PACF Graph: ($\theta = -0.8$)



ACVF: $\gamma(0) = \sigma_W^2(1+\theta^2)$
$\gamma(\pm 1) = \theta\sigma_W^2$
$\gamma(k) = 0$ for $k > 1$
Population ACF:
$\rho(1) = \frac{\theta_1}{1+\theta_1^2}$
Spectral Density:
$f(\lambda) = \sigma_W^2(1+\theta^2+2\theta\cos(2\pi\lambda))$ for $-\frac{1}{2} \leq \lambda \leq \frac{1}{2}$
Estimation:
$E[\hat{\theta_1}] = \theta_1$
$Var(\hat{\theta_1}) = \frac{1-\phi_1^2}{n}$

MA(2)

ACVF:
$\gamma(0) = (1+\theta_1^2+\theta_2^2)\sigma^2$
$\gamma(1) = (\theta_1+\theta_1\theta_2)\sigma^2$
$\gamma(2) = \theta_2\sigma_w^2$
$\gamma(k) = 0$
Estimation:
$\Sigma = \frac{1}{n}\begin{bmatrix}1-\theta_2^2 & -\theta_1(1+\theta_2)\\ -\theta_1(1+\theta_2) & 1-\theta_2^2\end{bmatrix}$

MA(∞)

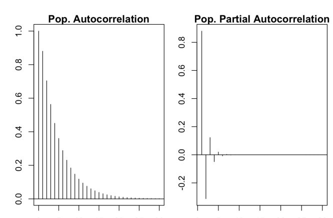
$X_t = \sum_{j=0}^{\infty}\theta_j W_{t-j}$
If absolute sum is finite, well-defined: $\sum_{j=0}^{\infty}|\theta_j| < \infty$.
$\gamma(h) = \sigma^2\sum_{j=0}^{\infty}\theta_j\theta_{j+h}$ Let us now assume that $\theta_j = \phi^j$ where $|\phi| < 1$:
$X_t = \sum_{j=0}^{\infty}\phi^j W_{t-j}$
$\gamma(h) = \sigma^2\sum_{j=0}^{\infty}\phi^j\phi^{j+h} = \sigma^2\phi^h\sum_{j=0}^{\infty}\phi^{2j} = \phi^2\phi^h\frac{1}{1-\phi^2}$
$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h$

ARMA(p, q)

Spectral Density:
$f_X(\lambda) = \sigma_W^2\frac{|\theta(\exp(-2\pi i\lambda)|^2}{|\phi(\exp(-2\pi i\lambda)|^2}$

ARMA(1, 1)

ACF and PACF Graph:



ACVF:
$\gamma(0) = \sigma^2\frac{1+\theta^2+2\phi\theta}{1-\phi^2}$
$\gamma(k) = \sigma^2\phi^{k-1}\frac{(\theta+\phi)(1+\theta\phi)}{1-\phi^2}$

Population ACF
$\frac{\rho(k+1)}{\rho(k)} = \phi_1$ for lags $k \geq 1$
$\rho(x) = \frac{(\theta+\phi)(1+\theta\phi)}{1+\theta^2+2\phi\theta}\phi^{h-1}$
Estimation
$\Gamma_{\phi\theta} = Cov(A_1, B_1) = Cov(\phi A_0 + W_1, -\theta B_0 + W_1) = -\phi\theta\Gamma_{\phi\theta} + \sigma^2$
$\Gamma_{\phi\theta} = \frac{\sigma^2}{1+\phi\theta}$ $\Sigma = \frac{1}{n}\begin{bmatrix}(1-\phi^2)^{-1} & (1+\phi\theta)^{-1}\\ (1+\phi\theta)^{-1} & (1-\theta^2)^{-1}\end{bmatrix}^{-1}$

ARMA(1, q)

Geometric decay based on the solution of order 1 difference equations that starts after lag q.

Fundamentals

Covariance Properties

$Cov(aX, bY) = abCov(X, Y)$
$Cov(aX + bY, cW + dV) = acCov(X, W) + adCov(X, V) + bcCov(Y, W)$

Trigonometry

$\theta = \frac{\pi}{6}$: $\sin\theta = \frac{1}{2}$, $\cos\theta = \frac{\sqrt{3}}{2}$, $\tan\theta = \frac{1}{\sqrt{3}}$
$\theta = \frac{\pi}{4}$: $\sin\theta = \frac{\sqrt{2}}{2}$, $\cos\theta = \frac{\sqrt{2}}{2}$, $\tan\theta = 1$
$\theta = \frac{\pi}{3}$: $\sin\theta = \frac{\sqrt{3}}{2}$, $\cos\theta = \frac{1}{2}$, $\tan\theta = \sqrt{3}$
$\sin(-x) = -\sin(x)$
$\cos(x) = \cos(-x)$

Matrices

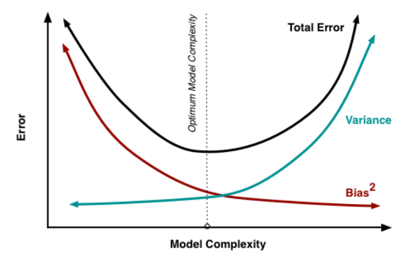
Inverse of a 2×2 matrix:
$A = \begin{bmatrix}a & b\\ c & d\end{bmatrix}$, $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$

Confidence Intervals

For x that follows normal distribution with mean μ and standard deviation σ, the confidence interval is defined as $\pm z\frac{\sigma}{\sqrt{n}}$ where z is the z-score for a given confidence interval. Common z-scores are 1.645 for 90 percent, 1.96 for 95 percent, 2.576 for 99 percent.

Bias-Variance Tradeoff

$MSE(\hat{(}\mu)) = E(\hat{\mu}-\mu)^2 = (E\hat{\mu}-\mu)^2 + E(\hat{\mu}-E\hat{\mu})^2 = Bias(\hat{\mu})^2 + Var(\hat{\mu})$



Bias - more you overfit the model the less bias you have because you are fitting to the noise.
Variance - the more you underfit the data, the less variance you have since outliers and noise do not affect the data.

Complex Numbers

$e^{i\theta t} = cos\theta t + isin\theta t$
$\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}$
$\sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}$
$1 - e^{i\theta} = -2i\sin(\theta/2)e^{i\theta/2}$
$|e^{i\theta}| = |\cos\theta + i\sin\theta| = \sqrt{\cos^2\theta + \sin^2\theta} = 1$

Chi-Squared Distribution

The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. Let $Y \sim \chi_k^2$. $E[Y] = k$ and $Var(Y) = 2k$