Introduction To Forecasting, Control And Time Series

Examples

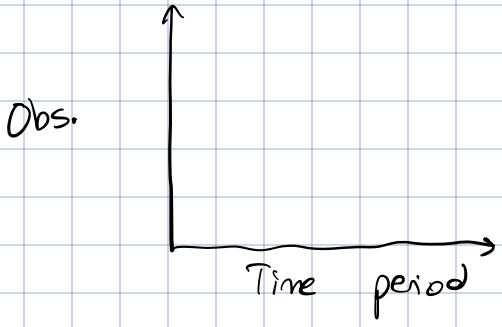
What makes this diff. than other statistics?

① Data is not indep.

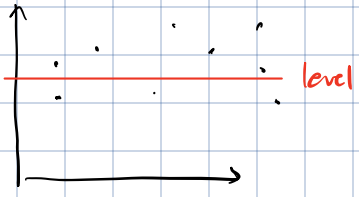
② Data is not i.i.d. ← can vary across time

③ Abrupt changes to data generation process

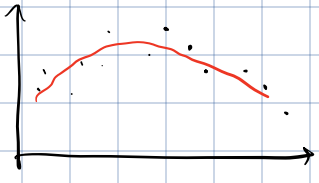
Time series plot:



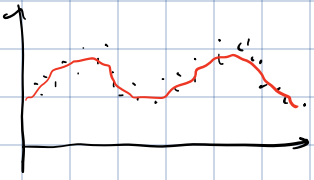
Level: local mean of observations



Trend: level varies w time



Seasonal effect: calendar-related effect which repeats in a given period



Change point: time where 1/more of following changes

① Data gen. process

② Method of data measurement

③ Obs. defn.

Additive-based Modelling

Modelling stationary data is easier than modelling changing data. We split data into stationary & non-stationary data

$$X_T = m_t + s_t + r_t$$

- m_t: trend
- s_t: seasonality
- deterministic func. of time
- r_t: Randomness. Hopefully 0-mean & uncorr. (stationary)

General process:

① Remove non-stationary elements: stabilize variance, remove seasonality & trend
↳ Similar to L.R.: remove trend to produce random residuals

② Model stationary data

③ Forecast on stationary data & add back non-stationary

Time Series Models

Stochastic process: family of random variables $\{X_t, t \in T\}$ defined on a prob. space
↳ T: index set, could be real-valued func.

Ideally, we can find the joint distr. of all $\{X_T\}$
↳ $F(x_1, \dots, x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$

We know that this joint distr. exists b/c of Kolmogorov's Existence Theorem:

Kolmogorov's Existence Theorem:

$F_T(\cdot)$ are distribution func. of some stoch. process $\Leftrightarrow \forall\, n \in \{1, 2, \dots\}$ & $t = (t_1, \dots, t_n) \in T$ and $1 \le i \le n$:

$$\lim_{x_i \to \infty} F_t(x) = F_{t(i)}(x(i))$$

} Like how joint → marginal!

$F_{t(i)}$: $(n-1)$-component vectors w/ i^{th} comp. deleted

Intuition: if we can't get CDF of $n-1$ R.Vs from n R.Vs, it's not a stoch. process

In reality, we don't need joint distr., we only need:

① $E(X_t)$ } $Var(X_t)$

② $E(X_t^2)$ }

③ $E(X_t, X_t^*)$ } $Cov(X_t, X_t^*)$

Since we can't find joint distr., we don't know if $\forall i,j$, X_i is indep. of X_j

↳ Recall: X_i is indep. of $X_j \iff F(X_i X_j) = F(X_i) F(X_j)$. We don't know any $F(\cdot)$!

Zero-Mean Time Series Models

If we follow our process listed above, our residuals ideally have $\mu = 0$. Such models that assume this are called zero-mean time series models, which include:

① i.i.d. Noise

Defn: $\{X_i, i = 1, \dots, \}$ where $E(X_i) = 0 \ \forall i$

Joint distr:

$$F(x_{i_1}, \dots, x_{i_n}) = P(X_{i_1} \le x_{i_1}, \dots, X_{i_n} \le x_{i_n})$$

$$= \prod_{k=1}^{n} P(X_{i_k} \le x_{i_k})$$

This implies:

$$P(X_{n+h} \le x \mid X_1 \le x_1, \dots, X_n \le x_n) = P(X_{n+h} \le x)$$

In other words, past does not help w/ predicting future

② Random Walk

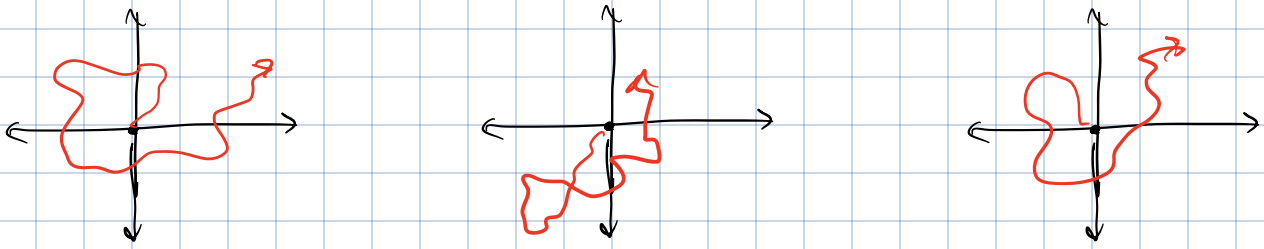
Defn: $\{S_T, t = 0, 1, \dots\}$ where $S_t = \sum_{i=1}^{t} X_i$, where $X_i \in \{X_t, t = 1, 2, \dots\}$ of i.i.d noise. Usually, $S_0 = 0$ (or some S_0 w/ $E(S_0) = 0$)

Note:

$$E(S_t) = E\left(\sum_{i=1}^{t} X_i\right) = \sum_{i=1}^{t} E(X_i) = 0 \quad \text{BUT not finite variance}$$

Ex:// Consider 2D discrete-time walk where $|dx| = |dy| = 1$

Sim. 1 Sim. 2 Sim. 3



③ White noise

Defn. $\{X_i,\ i = 1, 2, \ldots\}$ where each X_i is:

i) Not corr. w/ another (doesn't say anything about indep.)

ii) $E(X_i) = 0$

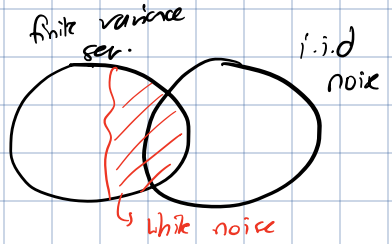
iii) $Var(X_i) = \sigma^2 < \infty$

Notation: $X_i \sim WN(0, \sigma^2)$

Random walks are a sum of white noise as well

If i.i.d noise has finite variance $\Rightarrow$ white noise.

If white noise is indep. $\Rightarrow$ i.i.d noise



WHY ZERO-MEAN RESIDUALS $\Rightarrow$ we will not be under/over predicting data

TREND, SEASONALITY, AND CLASSICAL DECOMPOSITION

Classical Decomposition

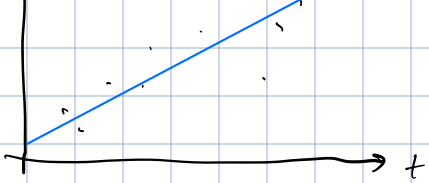
① Models w/ (non-periodic) trend

$$X_t = m_t + Y_t \leftarrow \text{random. var w/ } E(Y_t) = 0 \quad \forall t$$

m_t: deterministic, slowly changing

Note that $E(X_t) = m_t$

Ex:// $X_t = m_t + Y_t$, $m_t = 1 + 2t$, $Y_t \overset{i.i.d.}{\sim} N(0, 36)$



⇒ Similar to linear regression

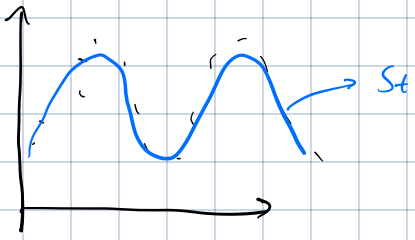
② Models w/ Seasonality (Periodic trends)

$X_t = S_t + Y_t$ ← random. var w/ $E(Y_t) = 0 \quad \forall t$

S_t: periodic deterministic function w/ period d

Note that $E(X_t) = S_t$

Ex:// $X_t = S_t + Y_t$, $S_t = 5 - 3\cos(0.2\pi t)$, $Y_t \overset{i.i.d.}{\sim} N(0, 1)$.

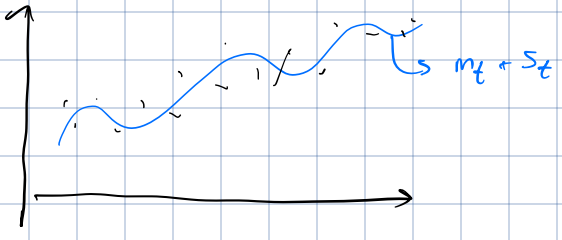


③ Models w/ both trend & seasonality

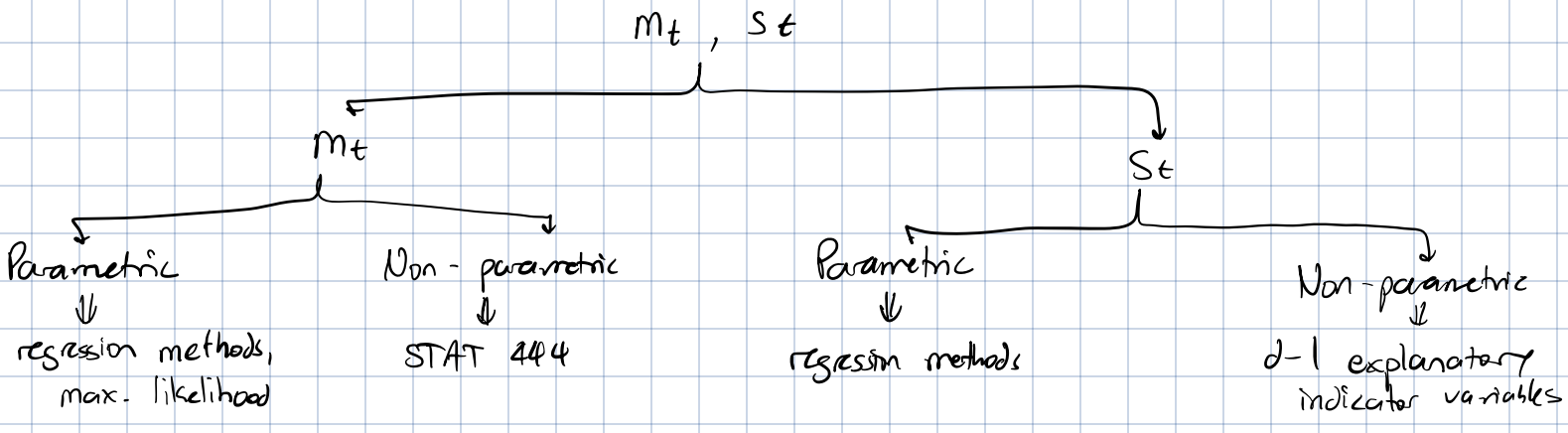
$X_t = m_t + S_t + Y_t$

$E(X_t) = S_t + m_t$

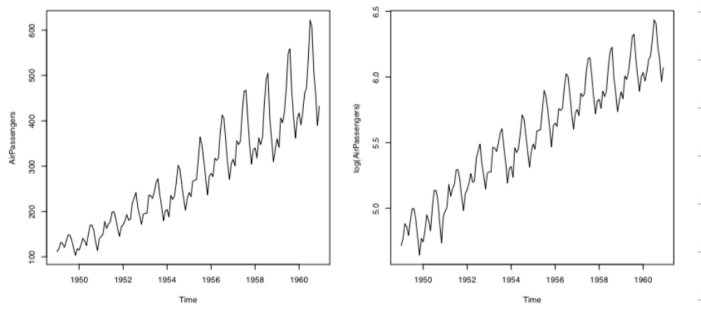
Ex:// $X_t = m_t + S_t + Y_t$, $m_t = 1 + 2t$, $S_t = 5\sin(0.3\pi t)$, $Y_t \overset{iid}{\sim} N(0, 7)$.



Estimating trend & seasonality via regression



Ex:// Fit models for the following pattern



It seems like there is a periodic, increasing trend. Log both sides to remove exponential behaviour when fitting.

$$\log(Y_t) = m_t + S_t + \beta_t$$
$$= \underbrace{\beta_0 + \beta_1 t}_{m_t} + \underbrace{\beta_2 x_1 + \dots + \beta_{12} x_{11}}_{S_t} + \beta_t$$

S_t composed of 11 indicator variables b/c period is 12 months

↳ $S_t = \beta_0 + \beta_1 x_1 + \dots + \beta_{11} x_{11}$ ← Expected change from Jan. baseline

↑ January baseline

Exponentiate back:

$$Y_t = e^{m_t + s_t + r_t}$$

Evaluating model, we see problems (add)

Add a quadratic component to trend:

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

Loss Functions & Forecasting

Introduction & Definition

Loss function: measures inconsistency b/w true & fitted values in model

↳ Notation: $L(Y, f(X))$ (true value, estimated values)

↳ Conditions:

① $L(Y, f(X)) \geq 0$

② $Y = f(X) \Rightarrow L(Y, f(X)) = 0$

Risk: $E_{X,Y}[L(Y, f(X))]$. If $X = x \Rightarrow E_{Y|X}[L(Y, f(X)) \mid X = x]$ ← Know how to take conditional expectations

Types of Loss Functions

① Quadratic loss / squared error

$$L_{SE}(Y, f(X)) = (Y - f(X))^2$$

Risk / expected value is mean squared error:

$$MSE(f) := E_{X,Y}((Y - f(X))^2)$$

Conditional expectation theorem:

X, Y are R.V., $E(Y) = \mu$, $Var(Y) < \infty \Rightarrow$ f that minimizes MSE is $f(X) = E_{Y|X}[Y|X]$ (R.V. depends on X, NOT #)

↳ Proof outline:

① Show that constant c that minimizes $E_Y[(Y-c)^2]$ is $c = \mu$ by solving following for c:

$$\frac{\partial}{\partial y} E_Y[(Y-c)^2] = 0$$

② From ①, $\hat{c}(x)$ that minimizes $E_Y[(Y - c(x))^2 \mid X = x]$ is $E[Y|X=x]$

③ We know $Z := z(x) \geq W := w(x) \Leftrightarrow \forall x,\ z(x) \geq w(x)$. From ②, for any $g(\cdot)$:

$$E_Y[(Y - g(x))^2 \mid X = x] \geq E[(Y - \hat{c}(x))^2 \mid X = x]$$

④ Take E_X on both sides & by law of iterated expectations:

$$E_{X,Y}[(Y - g(X))^2] \geq E[(Y - \hat{c}(x))^2]$$

⑤ Conclude minimality

Ex:// In least squares regression

$$Y = X\beta + \varepsilon$$

Since $E(Y) = X\beta$, $Var(Y) \geq \infty$, MSE function is $E[Y|X] = X\beta$

② Absolute error loss

$$L_{abs}(Y, f(x)) = |Y - f(x)|$$

Function that minimizes $E(L_{abs}(Y, f(x)))$:

$$\hat{f}(x) = \text{Median}(Y|X)$$

Know proof

③ Zero-one loss

If Y is discrete:

$$L_{01}(Y, f(x)) = \begin{cases} 0 & \text{if } Y = f(x) \\ 1 & \text{o.w.} \end{cases}$$

Expected value:

$$E(L_{01}(Y, f(x)) = P(Y \neq f(x))$$

Mostly for classification

Function that minimizes risk

$$\hat{f}(x) = \max_{y \in S} P(y|x) \Rightarrow \text{Bayes classifier}$$

Mostly interested in quadratic loss $\Rightarrow$ conditional expectations

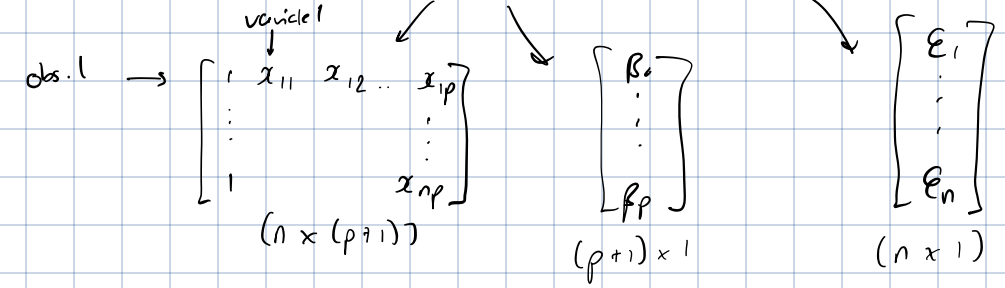
REGRESSION REVIEW

Main assumption of linear model: If $X = (x_1, \ldots, x_p)$, then:

$$f(x; \beta) = E[Y|X] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

In matrix notation:

$$Y = X\beta + \varepsilon$$



Parameter Estimation

Set loss function to residual sum of squares

$$RSS(\beta) := \sum_{i=1}^{n} \left(y_i - \underbrace{\beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}_{X\beta} \right)^2$$

To solve this, solve:

$$\hat{\beta} = \min_{\beta} \left(\underbrace{(Y - X\beta)^T}_{\varepsilon} \underbrace{(Y - X\beta)}_{\varepsilon} \right)$$

Taking derivative:

$$\frac{\partial}{\partial \beta} RSS(\beta) = -2X^T (Y - X\hat{\beta}) = 0$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Use second derivate test → positive definite → solution is minima

Hat matrix

Fitted values:

$$\hat{Y} = X\hat{\beta}$$

$$\hat{Y} = \underbrace{X(X^T X)^{-1} X^T}\ Y$$

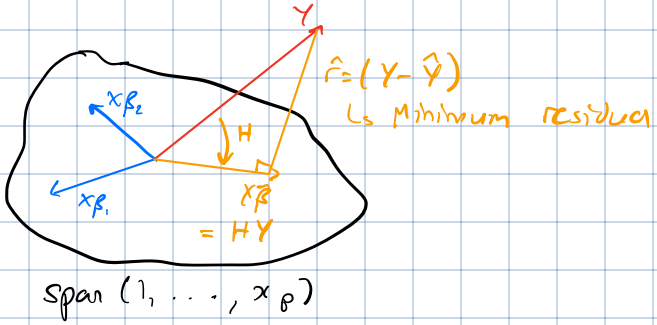
transforms $Y \to \hat{Y}$ ⇒ hat matrix $(n \times n)$ (H)

Properties:

① Idempotent

$$H^2 = H \rightarrow \text{orthogonal}$$

② Geometry:



How do we know $\hat{Y}$ & Y are orthogonal:

$$\hat{r} = Y - \hat{Y} = Y - HY = (I - H)Y$$

$$\hat{r} \cdot \hat{Y} = H(I - H) = H - H^2 = H - H = 0 \checkmark$$

③ Symmetric

$$H = H^T$$

④ Trace:

$$\text{trace}(H) = \sum_{i=1}^{n} h_{ii} = p + 1$$

⑤ Prediction of new value is just transformation of Y

$$\hat{Y}_i = \sum_{i=1}^{n} h(x_i)\, y_i$$

Statistical inference

Another assumption: $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$ → only for inference

Distn. of $\tilde{\beta} = (X^T X)^{-1} X^T \tilde{Y}$:

$$\tilde{\beta} \sim MVN(\beta, \sigma^2 (X^T X)^{-1})$$

Estimating σ^2:

$$\tilde{\sigma}^2 = \frac{SSE}{n - p - 1}$$

↳ Distn:

$$(n - p - 1) \cdot \frac{\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

Possible tests:

A: $H_0: \beta_j = 0$

$$z_j = \frac{\hat{\beta}_j}{\sigma \sqrt{v_j}} \sim t_{n-p-1}$$

$(j+1)^{th}$ diagonal elem $(X^T X)^{-1}$

B: Test if smaller model w/ p_0 variables is better than one w/ p_1 variables ($p_0 < p_1$)

$$\frac{(RSS_0 - RSS_1) \div (p_1 - p_0)}{RSS_1 \div (n - p_1 - 1)} \sim F_{p_1 - p_0,\, n - p - 1}$$

Further notes

① Sum of squares

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{Total sum of squares}} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSReg} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE}$$

total variability = model variability + error

Balance → bias-variance tradeoff (eg. decreasing SSE (variance) → increase $SSReg$ (bias))

② $\hat{\sigma}$:

residual standard error → $\hat{\sigma}$

③ Prediction interval

If predict on x is $\hat{y}$ w/ $[a, b]$ as 95% interval → $P(a < Y_0 < b) = 0.95$ (Y₀ is r.v.)

Why is this interpretation diff than C.I. ⇒ Y_0 is R.V., param in C.I. is fixed

④ Interpretation:

$\hat{\beta}_j = a$ ⇒ expected d.f. in Y w/ 1 unit change in x_j w/ all other variables the same yields a

RESIDUAL DIAGNOSTICS

Assuming $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ → residual checking for assumption correctness

① Normality

Graphical: plot residuals on Q-Q plot. If straight line → approx. Normal

Note: histogram ⇔ Q-Q plot

Formal test: Shapiro-Wilk test (non-parametric)

↳ H_0: $Y_1 \dots Y_n \sim N(\dots, \dots)$

↳ Low p-val → evidence of non-normality. We want high p-value

② Constant mean & variance

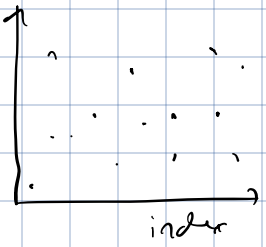
Graphical:

A: residual vs. fitted values

⇒ Ideally, no trend

If funnel-shape ⇒ variance not constant

B: Residuals vs. index / time (only if data order known)



⇒ Ideally, no trend

Formal test: non-parametric Fligner-Killeen test for non-constant variance

↳ $H_0: \sigma_1^2 = \ldots = \sigma_k^2$

↳ Test divide data into k segments & checks variance for each segment

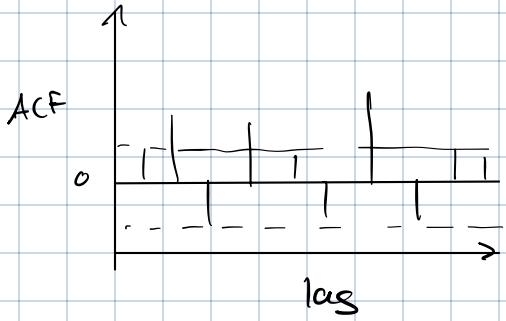
↳ Low-probe → evidence of non-constant variance

③ Uncorrelatedness

If correlated ⇒ dependent. Independence ⇒ uncorrelated, uncorrelated $\overset{\text{joint normality}}{\Rightarrow}$ indep

Test to check if residuals are correlated → not indep.

Residuals' sample ACF: 95% of spikes should be ∈ confidence bands



⇒ Spikes should also be in random lags

If trend ⇒ dependence

Spikes at lag = 1, 2 → concern

④ Randomness

A: Difference sign test

1. S = # of values s.t. $y_i - y_{i-1} > 0$
2. For large n, $S \sim N(\mu_S = \frac{n-1}{2}, \sigma_S^2 = \frac{n+1}{12})$. Thus

$$\frac{S - \mu_S}{\sqrt{\sigma_S^2}} \sim N(0, 1)$$

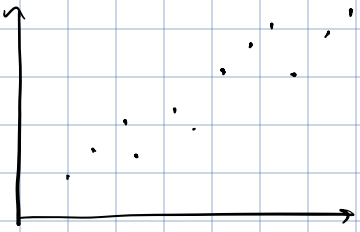
3. H_0: data is random → run a test.

Doesn't work if data has a lot of seasonality

$y_i = y_{i-1}$ ⇒ exclude

False negative problem: test doesn't consider <u>order</u> of sign increase / decrease.

↳ Ex: //



⇒ Test doesn't consider this random

B: Runs signs test

1. Estimate median m
2. $n_1 :=$ num. of obs $> m$, $n_2 :=$ num of obs. $< m$
3. $R :=$ # of consec. seq. of data where all $< m$
4. $\frac{R - \mu_R}{\sigma_R} \sim N(0,1) \rightarrow \mu_R = 1 + \frac{2n_1 n_2}{n_1 + n_2}$

$$\sigma_R^2 = \frac{(\mu_R - 1)(\mu_R - 2)}{n_1 + n_2 - 1}$$

5. Run test w/ H_0: data is random

False negative still a problem ⇒ clustered sequences not weighted highly

Run many tests → if any show problem, assumptions may be faulty.

PREDICTION WITH LINEAR REGRESSION

Forecasting w Multiple L.R. Models

Divide dataset into 2:

① Training set: data for model fitting

② Testing set: evaluate model perf.

Prediction interval theorem: $(Y_0, X_0) \notin$ training set. Define point forecast of Y_0 at $X = x_0$ as $\hat{\mu} = x_0^T \hat{\beta}$

$$100(1-\alpha)\% \text{ prediction interval} = \hat{\mu} \pm c_\alpha \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

⇒ Std. error: estimated std. dev. of estimator

c_α: $(1 - \frac{\alpha}{2})$ quantile of t_{n-p-1}

$\hat{\sigma}^2$: MSE of model

Proof sketch:

$$Y_0 - x_0^T \hat{\beta} \sim N\left(0, \sigma^2 \left(1 + x_0^T (X^T X)^{-1} x_0\right)\right)$$

$$\therefore T := \frac{y_0 - x_0^T \hat{\beta}}{\hat{\sigma}\sqrt{(1 + \ldots)}} \sim t_{n-p-1} \quad \text{if } \hat{\sigma} \text{ is estimate}$$

Thus:

$$P(-c_\alpha \le T \le c_\alpha) = 1 - \frac{\alpha}{2}$$

Rearranging:

$$x_0^T \hat{\beta} \pm c_\alpha \hat{\sigma} \sqrt{\ldots}$$

Q: What can we do to reduce the width of prediction interval?

Components

c_α

$c_\alpha \downarrow \rightarrow$ less confidence in our interval

$\hat{\sigma}$

Inherent var. in data $\rightarrow$ can't control

$\sqrt{1 + x_0^T (X^T X)^{-1} x_0}$

$$(X^T X)^{-1} = \frac{1}{|X^T X|} \text{adj}(X^T X)$$

Increased $\det(X^T X) \rightarrow$ P.I. length $\downarrow$.

Multicollinearity $\rightarrow$ $\det(X^T X) \downarrow$

Multicollinearity

Defn: two columns of X are highly correlated

Impacts:

① $\text{Var}(\tilde{\beta})$ is high

$$\text{Var}(\tilde{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \cdot \frac{\text{adj}(X^T X)}{\det(X^T X)}$$

Since multicollinearity $\Rightarrow$ small $\det(..)$, $\text{Var}(\tilde{\beta})$ is high

$\hat{\beta}$ is not a stable estimate

② Prediction interval width

$$P.I. = \hat{\mu} \pm c_\alpha \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

High value $\rightarrow$ interval $\uparrow$

Identification

① Check corr. b/w explanatory variables

② $SE(\hat{\beta}_j)$ is high

③ Overall test of sig. does not match test of sig of $\hat{\beta}_j$

④ $\hat{\beta}_j$ sign is not expected

⑤ Variance inflation factor

For some variate j, regress j on all other variates & extract R^2.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j | x_{-j}}}, \quad j = 1, \dots, p$$

↳ If R^2 ↑, VIF ↑

$VIF(\hat{\beta}_j) \geq 5 \Rightarrow$ collinearity

Q: How to deal w this?

① Drop variables

② Regularization models

Interpolation & Extrapolation

Idea: find range of validity where prediction makes sense

$h_{max} := \max(H_{ii}) \Rightarrow$ convex hull of data

Assume predicting on point x.

$$x^T (X^T X)^{-1} x \leq h_{max} \Rightarrow \text{interpolation}$$

Why is extrapolation bad? Assumptions may not hold for data outside of range

BIAS-VARIANCE TRADEOFF

Bias-variance decomposition:

Assume we are estimating on $x = x_0$, $\hat{y}_0 = x_0' \hat{\beta}$:

$$MSE(x_0) = \underbrace{(y_0 - \mathbb{E}(\hat{y}_0))^2}_{Bias(\hat{y}_0)^2} + \underbrace{\mathbb{E}(\hat{y}_0 - \mathbb{E}(\hat{y}_0))^2}_{Var(y_0)}$$

Bias: error of $f(x)$ (model) not belonging in actual model space

Variance: variability of params ← new data

Expected prediction error:

Let $Y = f(x) + \varepsilon$, $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$. We choose a model $\hat{f}(x)$. We fit value x_0.

$$\text{Err}(x_0) = E\left[(Y - \hat{f}(x_0))^2\right]$$

$$= \sigma_\varepsilon^2 + \text{Bias}^2 + \text{Variance}$$

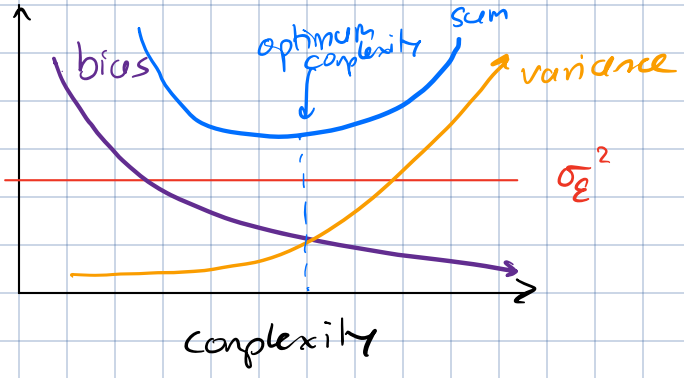
Choosing $\hat{f}(x)$ instead $f(x)$

Irreducible error: error inherent in data that we model (ε term)

Irreducible error not in control, others depend on model selection

The more complex the model, more variability but lower bias

" simple " , less " " higher bias



VARIABLE SELECTION

R^2 & Adjusted R^2

Defn of R^2:

$$R^2 = \frac{SS_{reg.}}{SS_{total}} = 1 - \frac{SSE}{SS_{Total}}$$

Issue: adding more variables will always increase R^2 → overfitting

Fix: adjusted R^2

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

- Penalizes large, complex models

Akaike's Information Criterion

Defn: $AIC = -2\ell(\hat{\beta}) + 2N_p$ → # of model params

The lower the AIC, the better

This measures model fit, not pred.

Issue: $n \downarrow \rightarrow$ AIC doesn't work well

- Corrected AIC: $AIC_c = AIC + \frac{2N_p(N_p+1)}{n - N_p + 1}$ → 0 as $n \uparrow$

Bayesian Information Criteria

Defn: $BIC = -2\ell(\hat{\beta}) + \log(n) \cdot N_p$ → Penalizes harsher than AIC

Same as AIC in interpretation & usage

Stepwise model selection

Either add more variables / subtract variables one by one → choose best AIC / BIC model

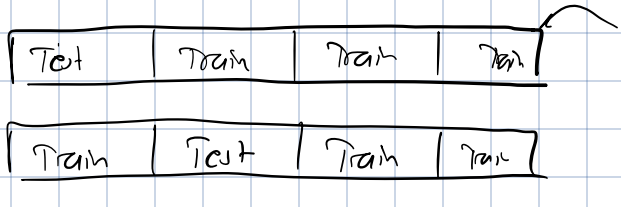
Average Prediction Squared Error (APSE)

Unlike others, this is a measure of predictive power

$$APSE = \frac{\sum_v (y - \hat{y})^2}{|v|} \quad (v \text{ is test set})$$

Smaller APSE → better forecasting power

K-fold cross validation



Partition data into k disjoint & ~ equal sets

⋮

After each fold:

$$sMSE_i = \frac{1}{n_k} \sum_{j \in T_i} (y_j - \hat{f}^{-i}(x_j))^2$$

This is just 1 example of a loss func, can be anything

$\hat{f}^{-i}(x_j)$: $\hat{y}_j$ fitting on T_{-i}

Then, calculate:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} sMSE_i$$

How to choose k? Large k is low bias & high variance + more comp. time

Randomness in creating k-fold ⇒ won't get same ans.

Model selection techniques

① Subset selection

Create subsets of variables (2^p models) → train → evaluate on metrics.

Comp. expensive

② Shrinkage

Penalize loss function on # of variables (known as regularization)

Effect: some $\hat{\beta} \to 0$, $Var(\hat{\beta}) \downarrow$

③ Dimensionality reduction

Project p variables → M-dimensional space ($M < p$) ⇒ dimensions incorporate all p variables in L.C.

REGULARIZATION METHODS

$$\tilde{\beta}_{regularized} = \arg\min_{\beta} \left\{ (Y - X\beta)'(Y - X\beta) + \lambda \underbrace{Pen(\beta)}_{\geq 0} \right\}$$

($\lambda \geq 0$)

Pros:
- + multicollinearity ↓
- + $n > p$ not an issue
- − variable selection
- + $Var(\tilde{\beta}_{regularized}) < Var(\tilde{\beta}_{ols})$

Cons:
- + Bias: $E[\tilde{\beta}_{reg.}] \neq \beta$ ← can bias correct
- − Comp. expensive
- + Optimality of β not achieved

Ridge Regression

Defn: $Pen(\beta) = l_2 = \sum_{j=1}^{p} \beta_j^2$ ← β_0 not penalized

Equivalent form:

$$(\beta_0, \hat{\beta}_{ridge}) := \arg\min_{\beta_0, \vec{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j \le t \quad (t \leftrightarrow \lambda)$$

Solving minimization:

1st col $\neq \vec{1}$

$$RSS(\lambda) = (y - \beta_0 \vec{1} - X\beta)'(y - \beta_0 \vec{1} - X\beta) + \lambda \beta'\beta$$

$$\vdots$$

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'Y, \quad \hat{\beta}_0 = \bar{y} - \frac{1}{n}\sum_{ij} \hat{\beta}_j x_{ij}$$

(λI: diagonal shifted by λ)

Since $(X'X + \lambda I)$ is invertible for some λ, multicollinearity not an issue

Why do we not penalize β_0? β_0 should absorb constant changes to $Y \to$ no change to $\hat{\beta}_{ridge}$ soln.

↳ Can drop β_0: Replace $\_{ij} \to x_{ij} - \bar{x}_j$, $j = 1 \ldots p$ $\Rightarrow \hat{\beta}_0 = \bar{y}$. $y \to y - \bar{y} \to \beta_0 = 0$

We need to standardize predictors s.t. all contrib. to penalty is equal.

↳ $\dfrac{x_{ij} - \bar{x}_j}{\hat{\sigma}_x}$

The larger λ, the more shrinkage. Does not set β to $0 \to$ no direct variable selectn.

LASSO Regression

Defn: $Pen(\beta) = \ell_1 = \sum_{j=1}^{n} |\beta_j|$

Minimization problem

$$(\hat{\beta}_0, \hat{\beta}_{lasso}) := \arg\min_{\beta_0, \beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Full form: least absolute selection and shrinkage operator (does variable sel.)

Compared to ridge:

① Harder to minimize b/c abs. val.

② No closed form soln for $\hat{\beta}_{lasso}$

④ Good if few predictors w/ outsized effects

⑤ Chooses collinear vars randomly → not

③ Variable selection — sparse for interpretation

$\lambda \uparrow \rightarrow$ more shrinkage

How to choose λ? C.V.

Elastic net regression

$$\hat{\beta}_{en} := \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left[\underbrace{(1-\alpha)\beta_j^2}_{\text{ridge}} + \underbrace{\alpha|\beta_j|}_{\text{lasso}} \right] \right\}, \quad \lambda > 0, \ \alpha \in [0,1]$$

How to choose α & λ? Grid search (coarse increm for α, fine for λ

STATIONARY PROCESS

Ideal time series model is finite dimensional distributions of process $\{X_t, t \in T\}$

- Finite dimensional distn.: joint distn. of finite subset of time series

Stationarity: finite dimensional distn. unchanged under time shift

Strong stationarity

Defn: $\forall$ finite subsets $\{t_1, \ldots, t_n\} \subseteq T$ & $\forall h$, $F(X_{t_1}, \ldots, X_{t_n}) = F(X_{t_1+h}, \ldots, X_{t_n+h})$

Ex:// Show any seq. of IID R.V. is strictly stationary

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \Big\} \ x_i \overset{D}{=} x_j$$

$$= \prod_{i=1}^{n} F_{X_{i+h}}(x_i)$$

$$= F_{t_1+h,\ldots,t_n+h}(x_1,\ldots,x_n)$$

Practically, strict stationarity is v. hard to prove & rarely true

Weak stationarity

Mean function: $\mu_X(t) = E[X_t]$, $t \in T$

Auto-covariance function (ACVF): $\gamma(r,s) = \text{Cov}(X_r, X_s)$ (assumes all $\text{Var}[X_t] < \infty$)

$= E(X_r X_s) - E(X_r)E(X_s)$

Ex:// Derive mean & ACVF for $X_t = 2 + 3t + Z_t - 0.5 Z_{t-1}$, $Z_j \overset{iid}{\sim} N(0, \sigma^2)$

Mean function:

$$\mu(t) = E[X_t] = 2 + 3t \quad \forall\, t = 1, 2, \ldots$$

ACVF:

$$\gamma(r,s) = Cov(2 + 3r + Z_r - 0.5Z_{r-1},\ 2 + 3s + Z_s - 0.5Z_{s-1})$$

↓ Constants don't matter

$$= Cov(Z_r - 0.5Z_{r-1},\ Z_s - 0.5Z_{s-1})$$

↘ Linearity props.

$$= Cov(Z_r, Z_s) - 0.5\,Cov(Z_r, Z_{s-1}) - 0.5\,Cov(Z_{r-1}, Z_s) + 0.25\,Cov(Z_{r-1}, Z_{s-1})$$

Case analysis: For each term, check when time index is equal

① $r = s$

$$\gamma(r,s) = \sigma^2 + 0 + 0 + 0.25\sigma^2 = 1.25\sigma^2$$

② $r = s-1$

$$\gamma(r,s) = -0.5\sigma^2$$

③ $s = r-1$

$$\gamma(r,s) = -0.5\sigma^2$$

Thus:

$$\gamma(r,s) = \begin{cases} 1.25\sigma^2 & |r-s| = 0 \\ -0.5\sigma^2 & |r-s| = 1 \\ 0 & |r-s| > 0 \end{cases}$$

Weak stationarity defn: $\{X_T\}$ is weakly stationary if:

① $E(X_t^2) < \infty \quad \forall\, t \in T$ (or $Var(X_t) < \infty$)

② $E(X_t) = \mu \quad \forall\, t \in T \Rightarrow$ indep. of t

③ $\forall\, r, s, r+t, s+t \in T$, $\gamma(r,s) = \gamma(r+t, s+t) \Rightarrow$ $\gamma(r,s)$ & $\gamma(r+t, s-t)$ only function of $|r-s|$

If ③ is true, then ACVF at lag h is:

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(t, t+h) = \gamma_X(t+h, t)$$

Corollary: Strictly stationary process ⇒ weakly stationary, converse not true!

- Proof: Assume $\{X_t\}$ is strictly stationary w $E(X_t^2) < \infty$. Proving weak:

① $E(X_t^2) < \infty$ by defn.

② If $\{X_t\}$ is strictly stationary, then $F_{X_1}(x) = \ldots = F_{X_n}(x) \quad \forall x$.

$$\Rightarrow \quad F_{X_1}(x) = F_{X_t}(x) \quad \forall x, t$$

$$\Rightarrow \quad E[X_t] = E[X_1]$$

$$\Rightarrow \quad E[X_t] = \mu$$

③ $\gamma(h) = \mathrm{Cov}(X_t, X_{t+h})$

$= \vdots$

$= \mathrm{Cov}(X_1, X_{h+1})$

Prove. Show $X_t, X_{t+h} \overset{D}{=} X_1, X_{h+1}$

Auto-correlation function (ACF):

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Cor}(X_{t+h}, X_t)$$

$\gamma_X(0) = \mathrm{Var}(X_t)$

Used b/c ACVF can be scaled up/down.

Ex:// Is $\{Z_t\} \sim WN(0, \sigma^2)$ stationary?

① $\mathrm{Var}(Z_t) = \sigma^2 < \infty$

② $E(Z_t) = 0 \Rightarrow$ not a func. of time

③ $\gamma(h) = \begin{cases} 0 & h \neq 0 \\ \sigma^2 & h = 0 \end{cases} \Rightarrow$ not a func. of time.

Concl. Yup! $\{Z_t\}$ is stationary

Ex:// Is random walk stationary?

Recall that $\{S_t\}$ is random walk if $S_t = \sum_{i=1}^{t} X_i$, X_i is iid noise.

Consider 3rd cond.:

$$\gamma(t, t+h) = \mathrm{Cov}(S_{t+h}, S_t)$$

$$= \mathrm{Cov}\left(\sum_{i=1}^{t} X_i + \sum_{i=t+1}^{t+h} X_i, \sum_{i=1}^{t} X_i\right)$$

$$= \mathrm{Cov}\left(\sum_{i=1}^{t} X_i, \sum_{i=1}^{t} X_i\right) + \mathrm{Cov}\left(\sum_{i=t+1}^{t+h} X_i, \sum_{i=1}^{t} X_i\right)$$

$= 0$ b/c no overlap

$$= \sum_{i=1}^{t} \mathrm{Cov}(X_i, X_i)$$

$$= t\sigma^2$$

Since $\gamma(h) \propto t$, 3rd cond. is violated

Ex:// Show that $X_t = 5 + 4t + Z_t$, Z_t is white noise is not stationary.

From cond. 2 $\Rightarrow$ $\mathbb{E}[X_t] = 5 + 4t$ $\rightarrow$ dependent on time.

Exercise: show cond. 1 & 3 are met

Ex:// Example of strict stationary time series which is not weakly stationary

Intuition: strictly stationary $\xrightarrow{\text{Var}(X_t) < \infty}$ weakly stationary

$\{X_t\}$ is iid. Cauchy. Strictly stationary b/c finite distn. are identical

Not weakly stationary b/c variance is infinite.

Ex:// Example of weak stationary process that is not strictly stationary

Intuition: we want $\{X_t\}$ whose distn. changes under time shift.

Let $\{X_t : t \geq 1\}$ be:

$$X_t = \begin{cases} Z_t, & t \text{ is even} \\ \frac{1}{\sqrt{2}}(Z_t^2 - 1), & t \text{ is odd} \end{cases}, \quad Z_t \overset{iid}{\sim} N(0, \sigma^2)$$

Proving weak stationarity:

① $$\text{Var}(X_t) = \begin{cases} \sigma^2 \\ \frac{1}{2}\text{Var}(Z_t^2) \end{cases} = \begin{cases} \sigma^2 & t \text{ even} \\ \frac{1}{2} \cdot 2 = 1 & t \text{ odd} \end{cases}$$

↳ $\sim \chi_1^2$, Var = 2 · d.f.

② $$\mathbb{E}[X_t] = \begin{cases} 0 \\ \frac{1}{\sqrt{2}}(\mathbb{E}(X_t^2) - 1) \end{cases} = 0$$

($\mathbb{E}$ of χ_1^2 is 1)

③ $\gamma_X(h) = 0 \quad \forall\, h \neq 0$ b/c no overlap

To prove this is not strongly stationary, consider prob. func. of X_2 & X_3.

$t = 2$

$$P(X_2 \leq 0) = P(Z_t \leq 0) = 1/2$$

$t = 3$

$$P(X_3 \leq 0) = P\left(\tfrac{1}{\sqrt{2}}(Z_t^2 - 1) \leq 0\right)$$
$$= P(Z_t^2 \leq 1)$$
$$\neq 1/2$$

Univariate distn. are diff. under time shift $\rightarrow$ not strictly stationary.

Ex:// $X_t = Z_t + \theta Z_{t-1}$, $Z_t \sim WN(0, \sigma^2)$. Show that $\{X_t\}$ is weakly stationary.

① $Var(X_t) = Var(Z_t + \theta Z_{t-1})$

$= Var(Z_t) + \theta^2 Var(Z_{t-1}) + 2\theta \cancel{Cov(Z_t, Z_{t-1})}$ → 0 b/c no overlap

$= \sigma^2 + \theta^2 \sigma^2 < \infty$

② $E(X_t) = E(Z_t) + \theta E[Z_{t-1}] = 0$, doesn't depend on t

③ $Cov(X_t, X_{t+h}) = Cov(Z_t + \theta Z_{t-1}, Z_{t+h} + \theta Z_{t+h-1})$

$= Cov(Z_t, Z_{t+h}) + \theta Cov(Z_t, Z_{t+h-1})$

$+ \theta Cov(Z_{t-1}, Z_{t+h}) + \theta^2 Cov(Z_{t-1}, Z_{t+h-1})$

Case analysis:

$$\gamma(h) = \begin{cases} \sigma^2 + \theta^2\sigma^2, & h = 0 \\ \theta\sigma^2, & h = 1 \\ \theta\sigma^2, & h = -1 \\ 0, & o.w. \end{cases} \Rightarrow \text{Doesn't depend on } t$$

①, ②, ③ $\Rightarrow \{X_t\}$ is weakly stationary.

Properties of ACVF & ACF

Assume $\gamma(r, s)$ is ACVF of stationary process $\{X_t\}$. Then:

① $\forall r$, $\gamma(h) := \gamma(r+h, r)$ → r doesn't matter

② $\gamma(0) \geq 0$ → $Cov(X_r, X_r) = Var(X_r) \geq 0$

③ $|\gamma(h)| \leq \gamma(0)$ → $|\rho(h)| \leq 1 \Rightarrow \frac{|\gamma(h)|}{\gamma(0)} \leq 1 \Rightarrow |\gamma(h)| \leq \gamma(0)$

④ ACVF is even: $\gamma(h) = \gamma(-h)$ $\Rightarrow \gamma(h) = Cov(X_t, X_{t+h}) = Cov(X_t, X_{t-h})$

Q: How do we go about modelling stationary process?

We want to keep model simple yet useful.

Eg:// If we tried to model w MVN under stationarity, variance-covariance matrix $\in O(n^2)$ params. → more data, more params. Too complicated.

Examples: MA(1), AR(1) → will see later.

Estimating ACF

In real data, we don't have model to construct ACF; need to use actual data.

Sample ACVF:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x})$$

$\hat{\gamma}(h)$: $\widehat{Cov}(X_{t+|h|}, X_t)$

Sample ACF:

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

If $\{X_t\}$ is i.i.d. noise w/ finite variance

$$\hat{\rho}(h) \sim N(0, 1/n)$$

so.

$$C.I._{95\%} = \hat{\rho}(h) \pm 1.96 \cdot 1/\sqrt{n}$$

If data has trend, $|\hat{\rho}(h)|$ has slow decay.

If data has periodicity, $|\hat{\rho}(h)|$ will also have periodicity w/ same period.

Note: it is fine for ACF plot to have 5% of values $\geq$ C.I., but must be in random order!

SMOOTHING METHODS

Defn: estimating patterns (deterministic) while leaving out noise

Non-stationary data $\xrightarrow[\text{variance stabilization}]{\text{smoothing}}$ stationary data $\longrightarrow$ model

$\Downarrow$

Want this to be WN/iid noise

Trend estimation

Estimate m_t & s_t $\rightarrow$ analyze & model residuals

① Finite moving avg. filter:

Assume $X_t = m_t + Y_t$ (Y_t is random component), q is non-neg. #

$$W_t := \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t-j}$$

$$= \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t-j}$$

$W_t \approx m_t$

Law of large #s: sample mean → pop. mean.
Avg. of $Y_{t-j} \to 0$.

Intuition: avg. of q points b/f & afte each X_t datapoint

Effect of q: $q \uparrow \longrightarrow$ bias ↑, variance ↓

Problem: can't use in forecasting b/c needs future data

↳ Can't use C.V. to find optimal q

② Exponential smoothing

Defn:

$$\begin{cases} \hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1} \\ \hat{m}_1 = X_1 \end{cases}$$

$$\Downarrow$$

$$\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1$$

Intuition: estim. trend is avg. of current data point & previous estim. trends exponentially weighted

Effect: $\alpha \uparrow \longrightarrow$ bias ↓, variance ↑

Finding α: 1. $\hat{\alpha} = \min_{\alpha} \left(\sum (X_t - \hat{x}_t)^2 \right)$

2. leave-one-out cross-validation

③ Polynomial regression

1. Make model assumption
2. Fit model
3. Model selection

} None of this requires $\varepsilon \sim N(0, \sigma^2)$ assumption!

} Try your best to make residuals have constant mean of 0

Trend Elimination (Differencing)

Important note: cannot help w/ non-constant variance, only removes trend & seasonality

Backwards shift operator: $B X_t := X_{t-1}$

$B^k X_t := X_{t-k}$

Differencing operator:

$$\nabla X_t = (1 - B) X_t = X_t - X_{t-1}$$

$$\nabla^k X_t = (1 - B)^k X_t \Rightarrow \text{Differencing } k \text{ times}$$

How many times should we difference (k?)

↳ More you difference → the more variability you introduce, model more complicated

↳ Only want to difference until trend/seasonality removed.

Ex:// $Y_t = 1 + t + Z_t$, $Z_t \sim WN(0, \sigma^2)$, $\sigma^2 < \infty$.

1) Is Y_t stationary

$E[Y_t] = 1 + t \rightarrow$ dependent on $t \rightarrow$ non-stationary

2) Is ∇Y_t stationary

$$W_t = \nabla Y_t = (1 + t + Z_t) - (1 + t - 1 + Z_{t-1})$$

$$= Z_t - Z_{t-1} + 1$$

1. $Var(W_t) = Var(Z_t) + Var(Z_{t-1}) - 2Cov(Z_t, Z_{t-1})$ (0 b/c WN (no corr.))

$$= \sigma^2 + \sigma^2$$

$$= 2\sigma^2 < \infty$$

2. $E(W_t) = 1$ → not dependent on t

3. $Cov(W_t, W_{t+h}) = Cov(Z_t - Z_{t-1}, Z_{t+h} - Z_{t+h-1})$

$\vdots$

Not a function of t

1, 2, 3 → non-stationary

Theoretical result: if trend is polynomial of degree p, $\nabla^p X_t$ will eliminate trend

Can also eliminate seasonality by differencing in order of seasonal period.

↳ Eg:// If X_t has period of $d = 6$ → $X_t - X_{t-6}$ will eliminate seasonality
$(1 - B^6) X_t$

↳ Notation: $\nabla^k X_t$ → differencing k times, lag of 1 ⇒ $(X_t - X_{t-1})^k$

$\nabla_k X_t$ → difference 1 time, lag of k ⇒ $(X_t - X_{t-k})$
$= (1 - B^k) X_t$

Ex:// $X_t = \sin\left(\frac{\pi t}{10}\right) + Z_t$, $Z_t \sim WN(0, \sigma^2)$.

1. Is X_t stationary?

$E[X_t] = \sin\left(\frac{\pi t}{10}\right)$ → dependent on t, non-stationary

2. $\nabla_{20} X_t$ stationary?

Aside: what is period of X_t?

$$\sin\left(\frac{\pi(t+d)}{10}\right) = \sin\left(\frac{\pi t}{10}\right)$$

$$\sin\left(\frac{\pi t}{10} + \frac{\pi d}{10}\right) = \sin\left(\frac{\pi t}{10}\right)$$

$$\frac{\pi d}{10} = 2\pi$$

$$d = 20$$

$$\nabla_{20} X_t = \underbrace{\sin\left(\frac{\pi t}{10}\right) - \sin\left(\frac{\pi(t-20)}{10}\right)}_{\text{Equiv.}} + Z_t - Z_{t-20}$$

$$= Z_t - Z_{t-20}$$

Show that this is stationary

If you have both trend & seasonality, try regular & seasonal diff. seperately, and only combine if neither works.

Decompose

We introduce a method to remove trend & seasonality ⇒ only for EDA

Decompose func. in R does following:

① Estimate trend: Apply moving avg. filter w $q = d/2$ (d is period of time series we model)

$$\hat{m}_t = \begin{cases} \dfrac{0.5\,x_{t-q} + x_{t-q+1} + \dots + x_{t+q-1} + 0.5\,x_{t+q}}{d}, & d = 2q,\ q < t \le n-q \\ \dfrac{1}{2q+1}\sum_{i=-q}^{q} x_{t-i}, & d = 2q+1 \end{cases}$$

Same period, so sum x_{t+}

} Even period

} odd period

② Detrendize data: $x_t - \hat{m}_t$

③ Estimate seasonality from detrendized data

Assumptions for seasonality:

1. $S_t = S_{t+d}$ ⇒ d is period
2. $\sum_{j=1}^{d} S_j = 0$

Process:

1. Take avg. of every period value in detrdized data ⇒ $\{w_k\}$, $k = 1, \dots, d$

   Ex:// Avg. for Jan., Feb., ..., Dec.

2. Centralize each to overall avg.

$$\hat{s}_k = w_k - \frac{\sum w_j}{d}$$ ⇒ Guarantees assump. 2 is fulfilled

④ Model $x_t - \hat{m}_t - \hat{s}_k$ (residuals)

Special note: decompose will always estimate seasons, it's not always relevant

↳ Signature seasonality not important: scale of seasonal plot is ≈ scale of residuals.

Decompox model can also be used for multiplicative models ($X_t = m_t s_t R_t$)

↳ Outcomes: avg. of seasonal is 1, not 0

Another note: freq. in time series model may not be correct. Look @ ACF, det. correct period & change time series for proper decomposition.

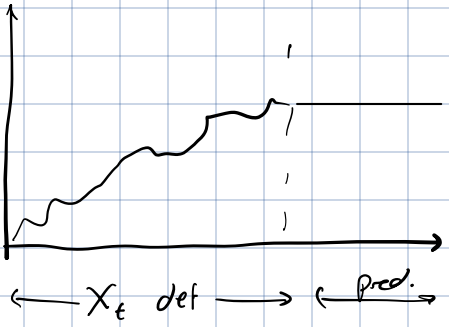
Holt - Winters

Problems w/ exp. smoothing

① Cannot accomodate seasonality

② Prediction results in constant value prediction

} Obj.: generalize exp. smoothing & solve this



Also (additive):

level → $L_T = \alpha(X_t - T_{t-p}) + (1-\alpha)(L_{t-1} + T_{t-1})$

trend → $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

seasonal → $I_t = \gamma(X_t - L_t) + (1-\gamma)I_{t-p}$

} Exp. smoothing

where L_0, T_0 & I_0 are defined.

↳ No need to memorize

Forecast of h periods ahead:

$$\hat{X}_{t+h} = L_t + hT_t + I_{\text{period of } t+h}$$

Also (multiplicative):

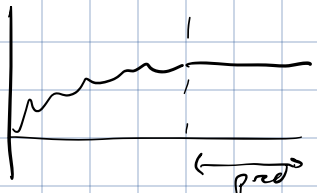
$$\left\{ \begin{matrix} L_t \\ T_t \\ I_t \end{matrix} \right. \Rightarrow \text{Forecast: } \hat{X}_{t+h} = (L_t + hT_t) I_{\text{period of } t+h}$$

In R: need starting values & $\{\alpha, \beta, \gamma\}$ → R does this automatically

Special cases:

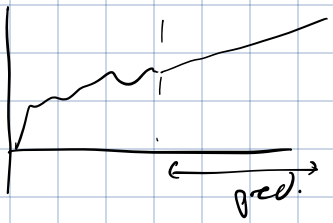
① $\beta = \gamma = 0$ (regular exp. smoothing).

↳ $\hat{x}_{t+h} = L_t$



② $\gamma = 0$ (no season, double exp. smoothing)

↳ $\hat{x}_{t+h} = L_t + hT_t$



Try out both additive & multiplicative to see which is better

IMPORTANT NOTE: No stat. assumptions needed!

↳ Needed only for prediction interval:

$$\hat{x}_{t+h} \pm z_{\alpha/2}\, SE(x_{t+h} - \hat{x}_{t+h})$$

STATIONARITY & LINEAR PROCESSES

We will use stationarity as an assumption when developing models

Moving Average Process: MA(q)

Defn: $Z_t \sim WN(0, \sigma^2)$, θ_i are non-zero constants. Then:

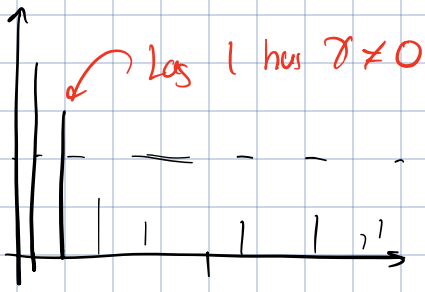
$$x_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$$

⇒ Not the same as MA filter!

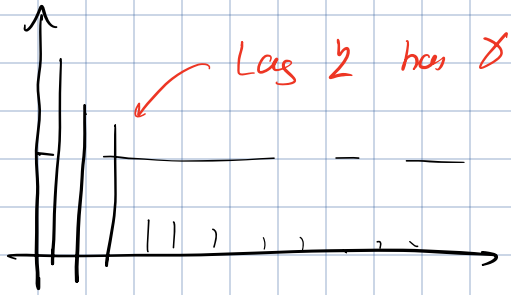
Q: How do you know what q should be?

A: Check ACF! q^{th} lag will be last non-zero lag

Ex:



ACF of MA(1)



ACF of MA(2)

Theorem: MA(q) is stationary

① $\text{Var}[X_t] = \text{Var}(Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q})$

$= \sigma^2 + \theta_1^2 \sigma^2 + \dots + \theta_q^2 \sigma^2$

$= \sigma^2 (1 + \theta_1^2 + \dots + \theta_q^2) < \infty$ bc θ_i & σ are finite

② $E[X_t] = E[Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}]$

$= 0 \Rightarrow$ indep. of t & finite

③ $\text{Cov}(X_t, X_{t+h}) = \text{Cov}(\underbrace{Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}}_{i}, \underbrace{Z_{t+h} + \theta_1 Z_{t+h-1} + \dots + \theta_q Z_{t+h-q}}_{ii})$

↳ Case analysis:

Func. of h, not t

A: $h = 0 \Rightarrow$ i & ii perfectly match

$\therefore \text{Cov}(X_t, X_t) = \text{Var}(X_t) = \sigma^2 (1 + \theta_1^2 + \dots + \theta_q^2)$

B: $h \leq q \Rightarrow$ overlap

$\therefore \text{Cov}(X_t, X_{t+h}) = \theta_w \sigma^2 + \theta_1 \theta_{1+w} \sigma^2 + \dots + \theta_{q-w} \theta_q \sigma^2$

C: $h > q \Rightarrow 0$, no overlap

①, ②, ③ $\Rightarrow \{X_t\}$ is stationary

ACVF of MA(q):

$$\gamma(h) = \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}$$

ACF of MA(q):

$$\rho(h) = \frac{\sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}}{\sum_{j=0}^{q} \theta_j^2}$$

Note that for $h > 0$, $\rho(h) = 0$ ⇓ Signature of MA(q) proc.

q-dependence: Process $\{X_T\}$ is q-dep. if X_t & X_s are dep. $\Leftrightarrow |t-s| = q$. and indep. if $|t-s| > q$

↳ Pretty strict, need joint prob distrns.

q - correlated: Process $\{x_t\}$ is q-corr. if

$$\gamma(h) = \begin{cases} \text{non-zero} & h = q \\ 0 & h > q \end{cases}$$

Ex:// Prove/disprove that iid sq. of R.V. is 0-dependent.

True:

Case A: $|t-s| = 0 \Rightarrow X_t = X_s$. This is trivially dependent.

Case B: $|t-s| > 0 \Rightarrow$ B/c iid, $X_t \perp X_s$.

Ex:// Prove/disprove that white noise is 0-corr & 0-dep.

0-corr:

Suppose $Z_t \sim WN(0, \sigma^2)$. At lag 0, there is corr. ($\text{Var} = \sigma^2$) b/c it's same variable

At lag $h > 0$, $\text{Corr}(Z_t, Z_{t+h}) = 0$ by defn. of WN

0-dep:

Uncorrelation of $Z_t \not\Rightarrow Z_t \perp Z_{t+h}$

$\Rightarrow$ jointly normal

Ex:// Show all MA(q) processes are q-corr.

Suppose X_t is MA(q) process:

$$\gamma(h) = \begin{cases} \sigma^2(1+\theta_1^2+\dots+\theta_q^2), & h = 0 \\ \sigma^2 \sum_{j=0}^{q+h} \theta_j \theta_{j+h}, & h \le q \\ 0 & h > q \end{cases}$$

At $h = q$, $\gamma(h) \neq 0$. At $h > q$, $\gamma(h) = 0 \Rightarrow$ q-corr.

Proposition: q-correlated w/ mean 0 & stationary $\iff$ MA(q) process

AR(P) Process

AR(1): $X_t = \Phi X_{t-1} + Z_t$, $Z_t \sim WN(0, \sigma^2)$.

$\Rightarrow (1 - \Phi B) X_t = Z_t$

- $|\Phi| < 1 \Rightarrow X_t$ is stationary
- ACF of AR processes have exp. decay $\to 0$

Another way of representing AR(1) is via MA(∞)

- Defn: X_t is MA(∞) if
  1. $X_t := \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, $Z_t \sim WN(0, \sigma^2)$
  2. $\sum_{j=0}^{\infty} |\psi_j| < \infty$

- Wold decomposition theorem: any stationary process = MA(∞) process + indep. det. process

$\Rightarrow X_t = \sum_{j=0}^{\infty} \psi Z_{t-j} + \mu_t$

  - Conditions:
    1. $\psi_0 = 1$, $\sum |\psi_j| < \infty$
    2. $\{\psi_j\}$, $\{\mu_t\}$ are unique
    3. $\{\mu_t, t \in \mathbb{Z}\}$ is a linear combo of past & det.
    4. $Cov(\mu_t, Z_s) = 0 \ \forall t, s$

- AR(1) $\to$ MA(∞)

$$\begin{aligned} X_t &= \Phi X_{t-1} + Z_t \\ &= \Phi(\Phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \Phi^2 X_{t-2} + \Phi Z_{t-1} + Z_t \\ &= \vdots \\ &= \sum_{j=0}^{\infty} \Phi^j Z_{t-j} \end{aligned}$$

$\Rightarrow$ We also know converges if $|\Phi| < 1$

$\therefore \sum_{j=0}^{\infty} |\Phi^j| = \frac{1}{1 - |\Phi|} < \infty$

AR(p) process defn:

$$\Phi(B) := 1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p$$

↳ $\Phi(B) X_t = Z_t$ (AR(p) process)

$\Rightarrow X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \dots + \Phi_p X_{t-p} + Z_t$

- Signature: exp. decay behaviour on ACF
- By Wold decomposition, we know:

$$\{X_T\} = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \quad \Rightarrow \text{exists}, \ \psi_j \text{ is some func. of } \Phi_i$$

- ACVF:

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$$

$$\rho(h) = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+h}}{\sum_{j=0}^{\infty} \psi_j^2}$$

} No q s.t. $\rho(h) = 0 \ \forall h > q$
↳ Not q-corr.

Linearity

Defn: $\{X_T\}$ is linear $\Leftrightarrow$

① $X_T = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $Z_t \sim WN(0, \sigma^2)$

② $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$

To show a process is linear:

① Show it can be rep. as $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, define ψ_j

② Show that ψ_j converges

Ex:// Show that AR(1) process is linear

① Under $|\Phi| < 1$, $X_T = \sum_{j=0}^{\infty} \Phi^j Z_{t-j}$

$$\Rightarrow X_T = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad \psi_j = \begin{cases} \Phi^j & \text{if } j \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

(v) $\sum_{j=-\infty}^{\infty} |\psi_j| = \sum_{j=-\infty}^{0} 0 + \underbrace{\sum_{j=0}^{\infty} |\phi^j|}_{<\infty} \Rightarrow$ converges.

Causality: $X_T = f(Z_s)$, $s \le t$ (doesn't need future info)

- If X_T is linear, X_T is causal $\Leftrightarrow \psi_j = 0 \ \forall j < 0$
- Ex// AR(1) & MA(q) are causal processes

Linear predictors

Linear processes are easy to work w/ → best predictors will usually be linear

Let's create a best linear predictor for a Gaussian process

Gaussian process: $\{X_T\}$ where all finite dimensional distn. are MVN

Note:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]$$

$$\Rightarrow X_1 | X_2 = x_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 + \frac{\sigma_{12}^2}{\sigma_2^2}\right)$$

$$= N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 (1 - \rho^2)\right)$$

↳ If X_1 & X_2 are corr., $\rho^2 \uparrow \Rightarrow \mathrm{Var}(X_1 | X_2 = x_2) \downarrow$ so pred. is easy!

Want to forecast X_{n+h} $(h > 0)$ using a func. of X_n, $m(X_n)$. $\{X_n\}$ is stat.

$$MSE = E\left[(X_{n+h} - m(X_n))^2\right]$$

What is best $m(X_n)$ given X_n?

We already know from Loss Functions that

$$m(X_n) = E(X_{n+h} | X_n) \Rightarrow \text{minimizes MSE}$$

Finding $E(X_{n+h} | X_n)$:

$$\begin{pmatrix} X_{n+h} \\ X_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \gamma(h) \\ \gamma(h) & \sigma^2 \end{pmatrix} \right)$$

(b/c stat.)

$$\Rightarrow X_{n+h} \mid X_n = x_n \sim N\left(\mu + \frac{\gamma(h)}{\gamma(0)}(x_n - \mu),\ \gamma(0)(1-\rho(h)^2) \right)$$ ↓ from aside

$$= N\left(\mu + \rho(h)(x_n - \mu),\ \gamma(0)(1-\rho(h)^2) \right)$$

$$\therefore E[X_{n+h} \mid X_n] = \mu + \rho(h)(x_n - \mu)$$

$$= \rho(h) X_n + \mu(1-\rho(h))$$

This is best predictor of X_{n+h}. Note that this is linear in X_n. b/c Gaussian.

W/b if process is not Gaussian?

↳ Best predictor of X_{n+h} may not be linear, but we don't care! Focus on best linear predictor b/c easy

↳ For any process $\{X_n\}$, we derive $m(X_n) = aX_n + b$ via

$$\min_{a,b} E[(X_{n+h} - aX_n - b)^2]$$

$$\frac{\partial M}{\partial a} = -2E[X_n(X_{n+h} - \hat{a}X_n - \hat{b})] = 0 \quad -①$$

$$\frac{\partial M}{\partial b} = -2E[X_{n+h} - \hat{a}X_n - \hat{b}] = 0$$

$$E[X_{n+h}] - \hat{a}E[X_n] - \hat{b} = 0$$ ↓ $E[X_{n+h}] = E[X_n] = \mu$ b/c stationary

$$\mu - \hat{a}\mu - \hat{b} = 0$$

$$\hat{b} = \mu(1-\hat{\alpha})$$

①: $E[X_n X_{n+h}] - \hat{a}E[X_n^2] - \hat{b}\cancel{E[X_n]}^{\mu} = 0$

$$E[X_n X_{n+h}] - \hat{a}E[X_n^2] - \mu^2(1-\hat{\alpha}) = 0$$

$$E[X_n X_{n+h}] - \mu^2 - \hat{a}(E[X_n^2] - \mu^2) = 0$$

$E[X_n X_{n+h}] - \mu^2 = \gamma(h)$, $E[X_n^2] - \mu^2 = \sigma^2$

$$\gamma(h) - \hat{a}\sigma^2 = 0$$

$$\hat{a} = \frac{\gamma(h)}{\sigma^2} = \rho(h)$$

$$\therefore \hat{m}(X_n) = \rho(h) X_n + \mu(1-\rho(h)) \Rightarrow$$ Best linear predictor for any stationary process

Reminder: Best linear predictor may not be good for non-linear process predictions.

Q: Forecast using all historical data → best linear predictor?

A: Theorem, best linear predictor of X_{n+h} given $X_1 .. X_n$ is

$$\hat{X}_{n+h} = P_n X_{n+h} = \mu + \sum_{j=1}^{n} a_j (X_{n+1-j} - \mu)$$

predict X_{n+h} w/ n history

$\mu = E[X_t]$. $a_0 = \mu(1 - \sum_{i=1}^{n} a_i)$. $\Gamma_n a_n = \gamma_n(h) \Rightarrow a_n = \Gamma_n^{-1} \gamma_n(h)$

$$\Gamma_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ & \gamma(0) & & \gamma(n-2) \\ & & \ddots & \vdots \\ & & & \gamma(0) \end{bmatrix}, \quad \gamma_n(h) = \begin{bmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(n+h-1) \end{bmatrix}$$

Properties:

① Only need $\vec{X}, \mu, \gamma(h)$

② $MSE = E([X_{n+h} - P_n X_{n+h}]^2) = \gamma(0) - a_n' \gamma_n(h)$

③ Unbiased: $E(X_{n+h} - P_n X_{n+h}) = 0$

④ $E([X_{n+h} - P_n X_{n+h}] X_j) = 0$, $j = 1, 2, \ldots$

↳ Proof. $Cov(Err, X_j) = E[Err\, X_j] - \cancel{E[Err]}^{0} E[X_j]$

$= E[Err\, X_j]$

Error should be uncorr. w/ past values, so $E[Err\, X_j] = 0$

⑤ $P_n[a_1 X_{n+h_1} + a_2 X_{n+h_2} + \beta] = a_1 P_n X_{n+h_1} + a_2 P_n X_{n+h_2} + \beta$

↳ Predictions are linearly decomposable

⑥ $Pred(\sum_{i=1}^{n} a_i X_i + \beta \mid X_1 .. X_n) = \sum a_i X_i + \beta$

↳ Prediction on fitted vals will be fitted values themselves

⑦ $P_n X_{n+h} = \mu \Leftrightarrow \gamma_n(h) = Cov(X_{n+h}, \vec{X}) = 0$

Ex:// Given $X_1 \dots X_n$, derive $P_n X_{n+1}$ for an AR(1) process & calculate MSE

① AR(1) process representation:

$$X_t = \Phi X_{t-1} + Z_t, \quad |\Phi| < 1, \quad Z_t \sim WN(0, \sigma^2)$$

② Solve for a_n:

$$a_n = \Gamma_n^{-1} \gamma_n(h)$$
$$\vdots \text{ in } R$$

③ One-step prediction $(h = 1)$

$$\Gamma_n a_n = \gamma_n(h)$$

$$\begin{bmatrix} \gamma(0) & \Phi\gamma(0) & \cdots & \Phi^{n-1}\gamma(0) \\ & \ddots & \ddots & \Phi^{n-2}\gamma(0) \\ & & \ddots & \vdots \\ & & & \gamma(0) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \Phi\gamma(0) \\ \Phi^2\gamma(0) \\ \vdots \\ \Phi^n\gamma(0) \end{bmatrix}$$

$\div \gamma(0)$

$$\begin{bmatrix} 1 & \Phi & \cdots & \Phi^{n-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \Phi \\ & & & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \Phi \\ \vdots \\ \Phi^n \end{bmatrix}$$

$$\therefore \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \Phi \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

④ Write in form:

$$\hat{X}_{n+1} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu)$$

$\downarrow \mu = 0$

$$= \sum_{i=1}^{n} a_i X_{n+1-i}$$

$\downarrow a_i = \begin{cases} \Phi, & i = 1 \\ 0, & \text{o.w.} \end{cases}$

$$= \Phi X_n$$

⑤ MSE:

$$MSE = E\left[(\hat{X}_{n+1} - X_{n+1})^2\right] = E\left[(\Phi X_n - X_{n+1})^2\right]$$

$\downarrow$ from AR(1) defn.

$$= E\left[\Phi X_n - \Phi X_n - Z_t^2\right]$$

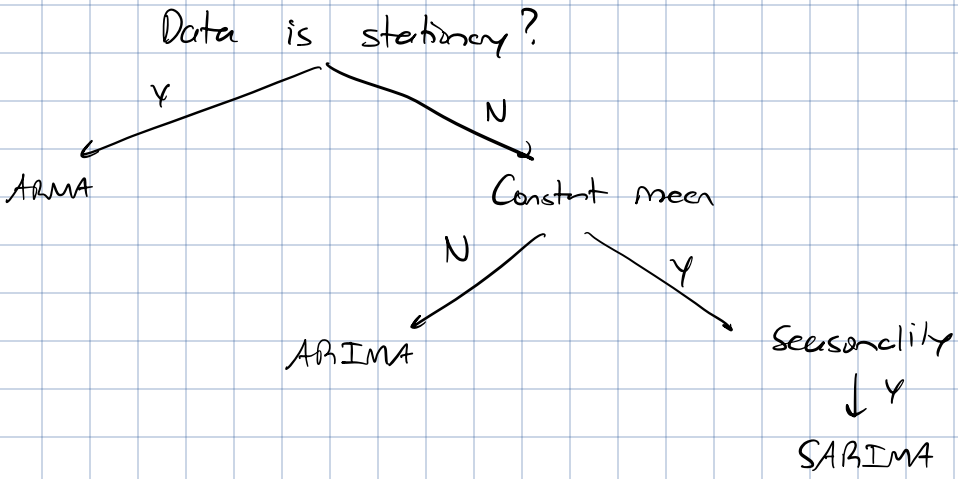
$$= E\left[Z_t^2\right]$$

$\downarrow \quad E[X^2] = \cancel{\mathrm{Var}(X)}^{\sigma^2} + \cancel{E[X]^2}^{0}$

$$= \sigma^2$$

Box Jenkins Models

ARMA: autoregressive moving average process



ARMA (p, q)

Defn: $\{X_t\}$ is ARMA (p, q) if

1. $X_t - \Phi_1 X_{t-1} - \dots - \Phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$, $Z_t \sim WN$

   AR (p)-like | MA (q)-like

2. $\{X_t\}$ is stationary (constant mean is fix, we can adjust $\{X_t - \mu_t\}$)

3. $\left.\begin{aligned} \Phi(z) &= 1 - \Phi_1 z - \dots - \Phi_p z^p \\ \theta(z) &= 1 + \theta_1 z + \dots + \theta_q z^q \end{aligned}\right\}$ no common factors.

Alt. way of writing: $\Phi(B) X_t = \theta(B) Z_t$

$$\begin{cases} \Phi(B) = 1 - \Phi_1 B - \dots - \Phi_p B^p \\ \theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q \end{cases}$$

Theorem: $\Phi(z)$ has no roots of unit circle (i.e. $\Phi(z) \neq 0 \ \forall z \in \mathbb{C}: |z| = 1$) then unique stationary sdn. exists for X_t.

↳ Intuition: Solving for X_t:

$$X_t = \frac{1}{\Phi(B)} \theta(B) Z_t$$

From Laurent series, we know that if no roots are on unit circle, we can write:

$$\frac{1}{\Phi(B)} = \sum_{j=-\infty}^{\infty} x_j \zeta^j \quad \text{for } 1 - \delta < |\zeta| < 1 + \delta \text{ where } \delta \text{ is really small.}$$

Let $\frac{1}{\Phi(B)} = \chi(B)$

$\therefore \quad X_t = \chi(B)\,\theta(B)\,Z_t$

$$= \sum_{j=-\infty}^{\infty} \psi_j B^j Z_t$$

$$= \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \& \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

We just showed that X_t can be written as a linear process if X_t is stationary!

Special cases for ARMA (p, q)

1) ARMA $(0, q) \Rightarrow \Phi(B) = 1 \Rightarrow X_t = \theta(B) Z_t$, MA (q)

2) ARMA $(p, 0) \Rightarrow \theta(B) = 1 \Rightarrow \Phi(B) X_t = Z_t$, AR (p)

3) ARMA $(0, 0) \Rightarrow X_t = Z_t$, WN

Causality & Invertibility

① Causal:

ARMA (p, q) causal $\Rightarrow$ can be represented as

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \quad \forall t, \quad \sum_{j=0}^{\infty} |\psi_j| < \infty$$

more restrictive than stationary cond.

Equivalent to showing no roots of $\Phi(z)$ in / on unit circle

Basically asking if we can write X_t as an MA process

② Invertibility:

ARMA (p, q) invertible $\Rightarrow$ can be represented as

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad \sum_{j=0}^{\infty} |\pi_j| < \infty$$

Equivalent to showing $\theta(z)$ has no roots in/on unit circle.

Basically asking if X_t can be written as AR process

Ex:// $X_t - 0.5X_{t-1} = Z_t + 0.4Z_{t-1}$, $Z_t \sim WN(0, \sigma^2)$.

a) Is X_t causal? If so, show causal soln.

① Check roots of $\Phi(z)$

$\Phi(z) = 1 - 0.5z \rightarrow z = 2$, so not on unit circle.
This is causal

② Find ψ_j

Note that: $\frac{\theta(z)}{\Phi(z)} = \psi(z) \Rightarrow \theta(z) = \Phi(z)\psi(z)$

Thus

$$1 + 0.4z = (1 - 0.5z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \dots) \quad \forall z$$

Find pattern by equating coefficients of same power of z:

z^0: $1 = \psi_0$

z^1: $0.4 = \psi_1 - 0.5\psi_0 \Rightarrow \psi_1 = 0.9$

z^2: $0 = \psi_2 - 0.5\psi_1 \Rightarrow \psi_2 = 0.5 \cdot 0.9$

z^3: $0 = \psi_3 - 0.5\psi_2 \Rightarrow \psi_3 = 0.5^2 \cdot 0.9$

To prove general form of ψ_j, let's use induction

Base case: $\psi_3 = 0.5^2 \cdot 0.9$

Assume: $\psi_k = 0.5^{k-1} \cdot 0.9$

Inductive step: prove $\psi_{k+1} = 0.5^k \cdot 0.9$

For z^{k+1}: $0 = \psi_{k+1} - 0.5\psi_k$

$\psi_{k+1} = 0.5\psi_k$

$= 0.5^k \cdot 0.9$ ☒

Thus:

$$\psi_j = \begin{cases} 0.5^{j-1} \cdot 0.9, & j \geq 1 \\ 1, & j = 0 \end{cases}$$

③ Soln:

Causal soln:

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = Z_t + 0.9 \sum_{j=1}^{\infty} 0.5^{j-1} Z_{t-j}$$

Also note that $\sum |\psi_j| < \infty$!

b) Is X_t invertible? If so show invertible solution

① Check if invertible:

$\theta(z) = 1 + 0.4z \Rightarrow z = -2.5$ not on unit circle so invertible

② Find ψ_j

$\Psi(z) = \theta(z)\pi(z)$

$1 - 0.5z = (1 + 0.4z)(\pi_0 + \pi_1 z + \pi_2 z^2 + \dots)$

$1 = \pi_0$

$-0.5 = \pi_1 + 0.4\pi_0 \Rightarrow \pi_1 = -0.9$

$0 = \pi_2 + 0.4\pi_1 \Rightarrow \pi_2 = 0.4(0.9)$

$$\pi_j = \begin{cases} -0.9(-0.4)^{j-1}, & j \geq 1 \\ 1, & j = 0 \end{cases}$$

② Soln:

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \Rightarrow Z_t = X_t - 0.9 \sum_{j=1}^{\infty} (-0.4)^{j-1} X_{t-j}$$

Also, $\sum_{j=0}^{\infty} |\pi_j| < \infty$

Ex:// $X_t = 0.7X_{t-1} - 0.1X_{t-2} + Z_t$ (AR(2)) process!

a) Is this invertible?

Check $\theta(z) = 1$ → no roots on unit circle

$\Psi(z)$ ↷ $1 - 0.7z - 0.1z^2 = \pi_0 + \pi_1 z + \pi_2 z^2 + \dots$

$\pi_0 = 1, \ \pi_1 = -0.7, \ \pi_2 = -0.1, \ \pi_3 = \pi_4 = \dots = 0$

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

b) Show MA(∞) rep.

To show X_t as MA, we need to get causal soln.

Can show: $\psi_j = \begin{cases} 0.7\psi_{j-1} - 0.1\psi_{j-2}, & j \geq 2 \\ 0.7, & j = 1 \\ 1, & j = 0 \end{cases}$ → no closed form

$$\Rightarrow X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

ACF of ARMA(p, q)

Consider a causal & stationary ARMA(p, q) process: $\Phi(B) X_t = \Theta(B) Z_t$

This will have an MA(∞) rep: $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$

We have:

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$$

Ex:// Derive ACVF & ACF of ARMA(1, 1).

$$X_t - \Phi X_{t-1} = Z_t + \theta Z_{t-1}$$

① Get MA(∞) if causal

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad \psi_j = \begin{cases} 1, & j = 0 \\ \Phi^{j-1}(\theta + \Phi), & j \geq 1 \end{cases}$$

② $\gamma(0)$:

$$\gamma(0) = \sigma^2 \sum \psi_j^2$$

$$= \sigma^2 \left(1 + \sum_{j=1}^{\infty} (\theta + \Phi)^2 (\Phi^{j-1})^2\right)$$

$$= \sigma^2 \left(1 + \frac{(\theta+\Phi)^2}{\Phi^2} \sum_{j=1}^{\infty} \Phi^{2j}\right)$$

$$= \sigma^2 \left(1 + \frac{(\theta+\Phi)^2}{\Phi^2} \cdot \frac{\Phi^2}{1-\Phi^2}\right)$$

$$= \sigma^2 \left(1 + \frac{(\theta+\Phi)^2}{(1-\Phi^2)}\right)$$

③ $\gamma(h)$:

$$\gamma(h) = \sigma^2 \left[\Phi^{h-1}(\theta + \Phi) + \sum_{j=1}^{\infty} \Phi^{j-1} (\theta+\Phi)^2 \Phi^{j+h-1}\right]$$

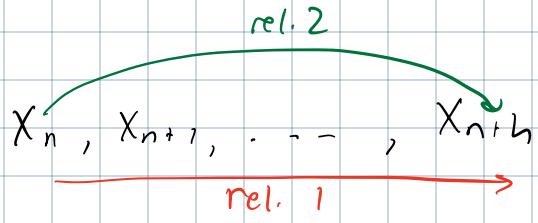
$$= \Phi^{h-1} \cdot \sigma^2 \left[(\theta+\Phi) + (\theta+\Phi)^2 \cdot \frac{\Phi}{1-\Phi^2}\right]$$

PARTIAL AUTOCORRELATION FUNCTION

Idea: If X is related to Y but Z is a confounding variable, then we should regress X on Z & Y on Z. Correlation b/w residuals

is partial corr.

In forecasting:



→ ACF measures corr. but don't know which are!

↳ This is a problem w/ AR processes b/c AR is a func. of many X_j

PACF looks at direct corr. b/w X_n & X_{n+h} by conditioning/regressing on X_{n+1} + X_{n+h-1}.

Definition: PACF of process $\{X_t\}$ is defined as

$$\alpha(h) = \begin{cases} 1 & h=0 \\ Cor(X_{n+1}, X_n), & |h|=1 \\ \rho_{X_{n+h}, X_n \cdot \{X_{n+1}, \ldots, X_{n+h-1}\}}, & |h|>1 \end{cases}$$

Corr b/w X_n & X_{n+h} conditioning on $X_{n+1} \ldots X_{n+h-1}$

Note that we can write ρ_{X_{n+h}, X_n} in another way

$$\rho_{X_{n+h}, X_n \cdot \{X_{n+1}, \ldots, X_{n+h-1}\}} = Cor(X_n, X_{n+h} \mid X_{n+1} \ldots X_{n+h-1})$$

$$= Cor(①, ②)$$

Corr. of residuals!

① $X_n - \hat{X}_n$ where $\hat{X}_n = Pred(X_n \mid X_{n+1}, \ldots X_{n+h-1})$

② $X_{n+h} - \hat{X}_{n+h}$ where $\hat{X}_{n+h} = Pred(X_{n+h} \mid X_{n+1}, \ldots, X_{n+h-1})$

Ex:// PACF of AR(1) process

① AR(1) defn:

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

② Find predictor

For AR(p) $\Rightarrow$ $\hat{X}_{n+1} = \sum_{j=1}^{p} \phi_j X_{n+1-j}$

For AR(1) $\Rightarrow$ $\hat{X}_{n+1} = \phi X_n$

③ Cases:

$h=0 \Rightarrow 1$

$h=1 \Rightarrow Cor(X_{n+1}, X_n) = \Phi$

$h=2 \Rightarrow Cor(X_t - \hat{X}_t, X_{t+2} - \hat{X}_{t+2})$

$= Cor(X_t - \hat{X}_t, X_{t+2} - \Phi X_{t+1})$

$= Cor(\underbrace{X_t - \hat{X}_t}_{\text{L.C. of } Z_i}, Z_{t+2})$

$= 0$

$\therefore \quad \alpha(h) = \begin{cases} 1, & h=0 \\ \Phi, & h=1 \\ 0, & h \geq 2 \end{cases}$

For ARMA processes:

$\alpha(0) = 1$

$\alpha(h) = \Phi(hh) \rightarrow$ last comp. of $\Gamma_h^{-1}\gamma_h$

Ex:// Calculate $\alpha(2)$ for MA(1)

① MA(1) process

$X_t = Z_t + \theta Z_{t-1}$

② $\gamma(h)$:

We know from ~~before~~ that:

$\gamma(h) = \begin{cases} (1+\theta)^2\sigma^2, & h=0 \\ \theta\sigma^2, & h=1 \\ 0, & h>1 \end{cases}$

③ Calculate:

$\alpha(0) = 1$

$\alpha(1) = \rho(1) = \frac{\theta}{(1+\theta)^2}$

$\alpha(2) \Rightarrow \Phi_2 = \Gamma_2^{-1}\gamma_2 = \begin{bmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma(1) \\ \gamma(2) \end{bmatrix}$

$= \frac{1}{(1+\theta^2)^2\sigma^4 - \theta^2\sigma^4} \begin{bmatrix} \theta((1+\theta^2)\sigma^4 \\ -\theta^2\sigma^4 \end{bmatrix}$

$\alpha(2) = \Phi_{22} = \frac{-\theta^2}{1+\theta^2+\theta^4}$

Recursive formula:

$\alpha(h) = \hat{\Phi}_{hh} = \frac{\rho(h) - \sum_{j=1}^{h-1} \hat{\Phi}_{h-1,j} \cdot \rho(h-j)}{1 - \sum_{j=1}^{h-1} \hat{\Phi}_{h-1,j} \cdot \rho(j)}$

Ex:// Calculate $\alpha(2)$ for MA(1) recursively

$$\alpha(2) = \frac{\rho(2) - \Phi_{1,1} \cdot \rho(1)}{1 - \Phi_{1,1} \cdot \rho(1)}$$

$$= \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}$$

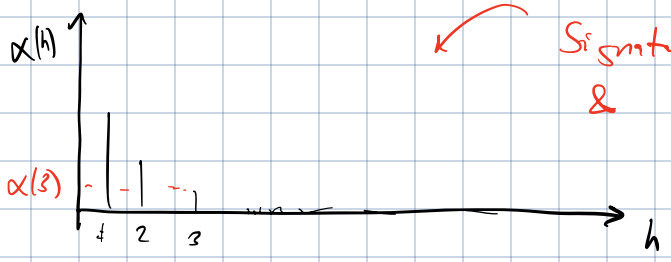
$$= \frac{-\theta^2}{1+\theta^2+\theta^4}$$

} Much simpler! Use this

Why is PACF important? Theorem:

$\{X_t\}$ is a causal AR(p) process $\Leftrightarrow$ $\alpha(p) \neq 0$, $\alpha(h) = 0 \ \forall \ h > p$

$\Leftrightarrow \alpha(p) = \Phi_p$



Signature of AR(3) process & $\Phi_3 = \alpha(3)$!

General ACF & PACF shapes of ARMA processes:

	ACF	PACF
MA(q)	cuts off after lag q	exp. decay / damped sinusoid
AR(p)	exp. decay / damped sinusoid	cut off after lag p
ARMA	"	exp. decay / damped sinusoid

To propose models:

① Construct ACF & PACF plots

② Look @ cutoffs & decays. Map to table above.

NOTE: PACF not useful if data is non-stationary

ARIMA & SARIMA Models

Up to this point:

Data ⟶ Remove trend & seasonality ⟶ Apply ARMA on data

Can we do this all at once? Yup!

ARIMA: autoregressive integrated moving avg.

↳ $\{X_t\}$ is ARIMA (p, d, q) is $Y_t := (1-B)^d X_t$ is ARMA (p, q) process

↳
$$\underbrace{\Phi(B)(1-B)^d}_{\Phi^*(B)} X_t = \Theta(B) Z_t$$

This type of model can only handle trend non-stationarity, not seasonality

Ex:// Process is $X_t = 0.8 X_{t-1} + 2t + Z_t$, $Z_t \overset{iid}{\sim} N(0, 35^2)$. Write process in ARIMA format.

① Rewrite:
$$(1 - 0.8B) X_t = Z_t + 2t \quad \rightarrow \text{Like AR(1)}$$

② Diff. to get rid of $2t$

$$\nabla X_t = X_t - X_{t-1}$$
$$= 0.8(X_{t-1} - X_{t-2}) + Z_t - Z_{t-1} + 2$$

Rewrite in terms of $Y_t := \nabla X_t$

$$Y_t = 0.8 Y_{t-1} + Z_t - Z_{t-1} + 2$$

Clearly, this process has a mean (+2). Finding mean.

$$(Y_t - \mu) - 0.8(Y_{t-1} - \mu) = Z_t - Z_{t-1}$$

E ↳ $$-\mu + 0.8\mu = 2$$
$$\mu = 10$$

$$\therefore \quad (Y_t - 10) - 0.8(Y_{t-1} - 10) = Z_t - Z_{t-1}$$

③ Model:

ARIMA (1, 1, 1)

Can we incorporate seasonality? Yes! → SARIMA

↳ Remember: ACF shows seasonality via periodicity! Cov at seasn. is NOT seasonality.

SARIMA defn: $\{X_t\} \in SARIMA(p,d,q) \times (P,D,Q)_s$ if

$$Y_t := \nabla^d \nabla_s^D X_t$$

↳ How many times seasonal diffing need to happen.

is causal ARMA process

$$\therefore \quad \Phi(B)\,\Phi^*(B^s)\,Y_t = \theta(B)\,\theta^*(B^s)\,Z_t$$

↑ Seasonal polynomials, only has z^q where q is multiple of s

Confusing! Just think of SARIMA as 2 ARMA models

1. ARMA on regular lags (w/ out seasonal lags)
2. ARMA on just seasonal lags

$\Phi(z) \wedge \Phi(z^s)$ cannot have roots on unit circle if causal.

If any params. missing → 0

If $d = D = 0$ → X_t is stationary

How to find params of SARIMA model?

① Plot ACF to check for patterns

② Difference seasonal & regular until stationary → d, D, s

③ ACF & PACF on differenced data → p, q, P, Q

④ Fit

NOTE: ACF of X_t^2 has trends → X_t has non-constant variance.

↳ Why? $\mathrm{Var}[X_t] = E[X_t^2] - E[X_t]^2$ ← 0 if trend removed

↳ Trend if $E[X_t^2]$ has trend.

SARIMA output on R gives us 4 plots:

① Residual plot ⇒ should look like white noise

② Residual ACF ⇒ no corr. except for lag 0

③ Normal Q-Q ⇒ normality for prediction intervals

④ p-val for Ljung-Box ⇒ $H_0: \rho(k) = 0 \ \forall k < h$

Ideally large p-value, no corr.

ESTIMATING ARMA(p, q) PARAMS

We have $p + q + 1$ params to estimate:

① $\Phi_1, \ldots, \Phi_p$

② $\theta_1, \ldots, \theta_q$

③ σ

Two methods of estimating:

① Likelihood.

Assume $(X_1, \ldots, X_n) \sim MVN$. Then:

$$Lik(\Phi_1, \ldots, \Phi_p, \theta_1, \ldots, \theta_q, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Gamma_n|}} \exp\left(\frac{-1}{2} X^T \Gamma_n^{-1} X\right)$$

Optimize Likelihood via params. This is comp. expensive!

② Yule-Walker Method (only AR(p))

$$\hat{\Phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p, \quad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\Phi}^T \hat{\gamma}_p$$

$\div \hat{\gamma}(0)$ ↓

$$\hat{\Phi} = \left(\frac{\hat{\Gamma}_p}{\hat{\gamma}(0)}\right)^{-1} \cdot \hat{\rho}_p, \quad \hat{\sigma}^2 = \hat{\gamma}(0)\left(1 - \hat{\Phi}^T \hat{\rho}_p\right)$$

Where $\hat{\gamma}_p = \begin{bmatrix} \hat{\gamma}(1) \\ \vdots \\ \hat{\gamma}(p) \end{bmatrix}$, $\hat{\Gamma}_p = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(p-1) \\ & \ddots & \ddots & \vdots \\ & & \ddots & \hat{\gamma}(1) \\ & & & \hat{\gamma}(0) \end{bmatrix}$

For large sample size:

$$\tilde{\Phi} \sim MVN\left(\Phi, \frac{\sigma^2}{n}\Gamma_p^{-1}\right)$$

FORECASTING IN ARMA MODELS

AR(p) forecasting

$$\hat{X}_{n+h} = \begin{cases} \sum_{j=1}^{p} \Phi_j X_{n+h-j}, & h = 1 \\ \sum_{j=1}^{h-1} \Phi_j \hat{X}_{n+h-j} + \sum_{j=h}^{p} \Phi_j X_{n+h-j}, & h = 2 \ldots p \\ \sum_{j=1}^{p} \Phi_j \hat{X}_{n+h-j}, & h > p \end{cases}$$

(All in history; use pred. values; history; All pred. values.)

h-step predictor is linear function of either historical data points or predicted values.

Ex:// Assume annual sales of store is modelled as AR(2) w/ $\hat{\Phi}_1 = 1$, $\hat{\Phi}_2 = -0.21$. Sales of 2021, 2020 & 2019 are 9, 11 & 10 million.

Forecast 2022:

$$\hat{X}_{2022} = \sum_{j=1}^{p} \hat{\Phi}_j \hat{X}_{2022-j} = \hat{\Phi}_1 X_{2021} + \hat{\Phi}_2 X_{2020}$$

$$\hat{X}_{2023} = \hat{\Phi}_1 \hat{X}_{2022} + \hat{\Phi}_2 X_{2021}$$

$$\hat{X}_{2024} = \hat{\Phi}_1 \hat{X}_{2023} + \hat{\Phi}_2 X_{2022}$$

MA(q) forecasting

$$\hat{X}_{n+h} = \begin{cases} \sum_{j=h}^{q} \theta_j Z_{n+h-j}, & 1 \le h \le q \\ 0, & h > q \end{cases}$$

$$Z_j = 0 \quad \forall j \le 0, \qquad Z_n = X_n - \sum_{j=1}^{q} \theta_j Z_{n-j}, \quad n \in 1, 2, \ldots$$

Ex:// Let $X_t = Z_t + 0.5 Z_{t-1}$. $x_1 = 0.3$, $x_2 = -0.1$, $x_3 = 0.1$. Forecast X_4 & X_5

① Classify X_t as MA

$X_t \sim MA(1)$ & invertible

② Let $Z_0 \dots Z_3$

$Z_0 = 0$

$Z_1 = x_1 - 0.5 Z_0 = 0.3$

$Z_2 = x_2 - 0.5 Z_1 = -0.25$

$Z_3 = x_3 - 0.5 Z_2 = 0.225$

③ Forecast

$\hat{X}_4 = \theta Z_3$

$= 0.5 \cdot 0.225$

$= 0.1125$

$\hat{X}_5 = \dots = 0$

ARMA (p, q) forecast

We have $X_n = \Phi_1 X_{n-1} + \dots + \Phi_p X_{n-p} + Z_n + \theta_1 Z_{n-1} + \dots + \theta_q Z_{n-q}$

It is just a combination of AR(p) & MA(q) combined for forecast

$$\hat{X}_{n+h} = \begin{cases} \sum_{j=1}^{p} \Phi_j X_{n+h-j} + \sum_{i=1}^{q} \theta_i Z_{n+h-i}, & h = 1 \\ \sum_{j=1}^{h-1} \Phi_j \hat{X}_{n+h-j} + \sum_{j=h}^{p} \Phi_j X_{n+h-j} + \sum_{i=h}^{q} \theta_i Z_{n+h-i}, & h > 1 \end{cases}$$

Now:

①
$$Z_n = \begin{cases} X_n - \sum_{j=1}^{p} \Phi_j X_{n-j} - \sum_{i=1}^{q} \theta_i Z_{n-i}, & n > p \\ Z_n = 0, & n \le p \end{cases}$$

② $\theta_j = 0, \quad \forall j > p$

③ $\sum_{i=h}^{q} \theta_i Z_{n+h-i} = 0, \quad h > q$ → obv.

(4) $\sum_{j=h}^{p} \Phi_j x_{n+h-j} = 0, \quad h > p$ → obv.