CSCI E-82
Advanced Machine Learning, Data Mining & Artificial Intelligence
Lecture 5

Advanced Regression
Time Series Part II

Peter V. Henstock
Fall 2018

Last week on Regression

- Formulate a solution to fit y = f(x)
  - Predict house prices from properties
  - House price = f(sq ft, land, school, #bath, etc)
- Defined optimization of least-squares
  - Solve it using pseudo-inverse
  - Solve it using gradient descent
- Perform diagnostics
- Get information on fit and feature values

Formulating the model

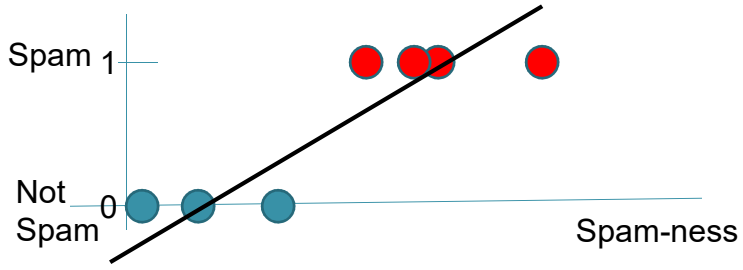
- Each input → probability of significance
  - Automatically computed from software
  - Includes all polynomial terms
  - Includes all interaction terms
- Insignificant terms are removed except:
  - Need to keep subsets (hierarchy principle)
- Simplest model → final model

Logistic Regression

Regression for Classification

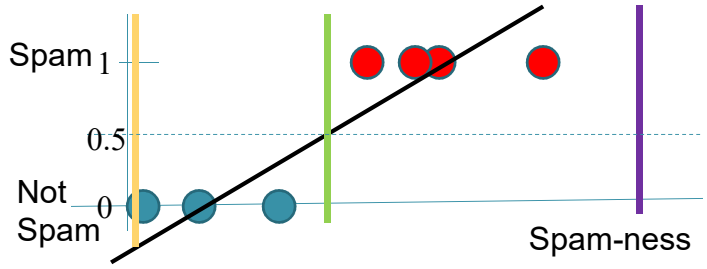
- Regression is a supervised method
- Previously studied fitting continuous values as in y = f(x)
- How can we convert this to a classifier?

Motivation for Logistic Regression



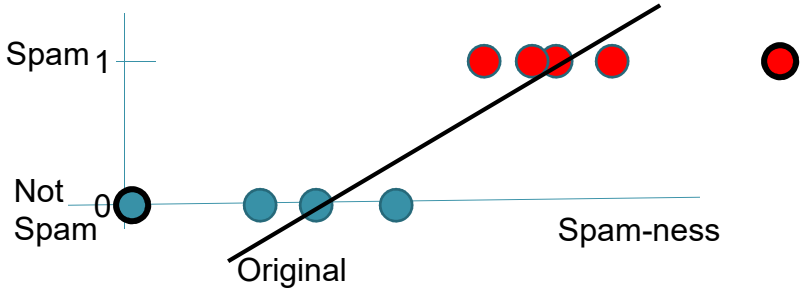
- Not the greatest fit
- Let's try to use it anyway

Motivation for Logistic Regression



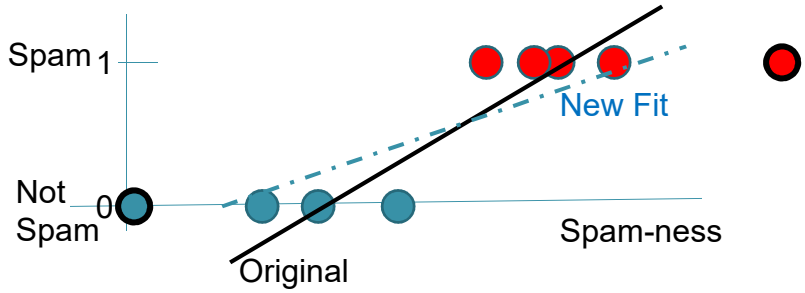
- Predictions for
  - Yellow?
  - Purple?
  - Green?

Motivation for Logistic Regression



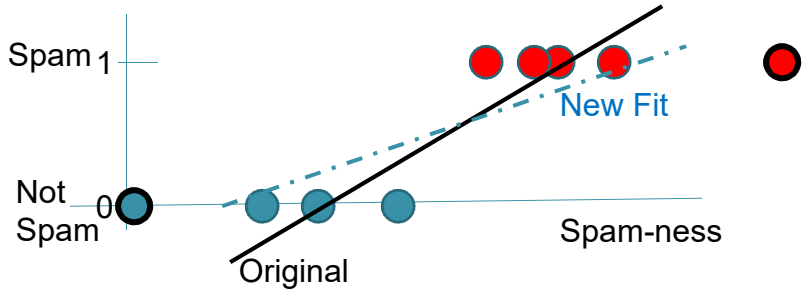
- What happens if we add extreme points far from the the boundary?

Motivation for Logistic Regression



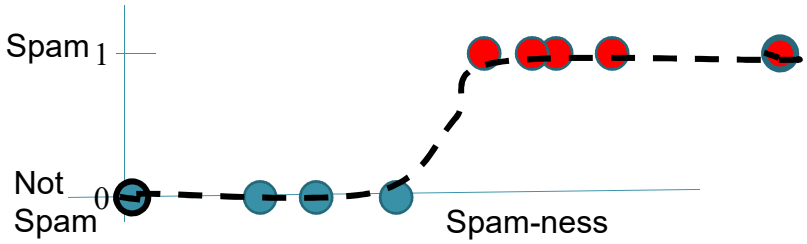
- Addition of new points far from the boundary shouldn't change predictions but it will using linear regression

Motivation for Logistic Regression



- Predictions for the new points are
  - -1 for the left blue
  - +2 for the right red
- Prefer more interpretable predictions that go with the [0,1] range near middle

Motivation for Logistic Regression



Solution is to fit a different shape curve that is used for logistic regression

What functions have this shape?

- Want something that is:
  - 0 at negative infinity
  - 1 at positive infinity
  - Smoothly changes between the two
- Any suggestions?

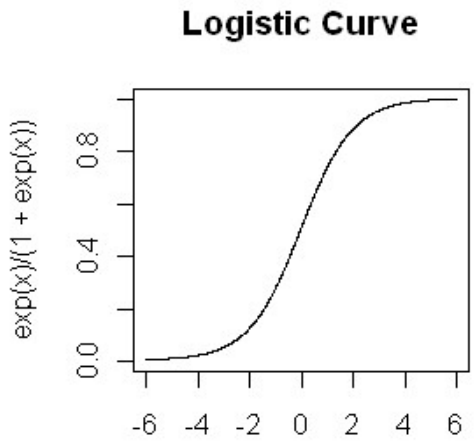
Logistic Equation

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k = \Sigma \beta_i x_i$
- Change the function to $f(Y) = \Sigma \beta_i x_i$
- $\log(\frac{p}{1-p}) = \Sigma \beta_i x_i$ = logit(p) = log-odds ratio
- Called the logit (hence logistic regression)
- What is p?
- p ~ probability that the value is true
- Previously we used 1 = true, 0 = false for Y
- But we might not want logit(p) but p

Logit conversion

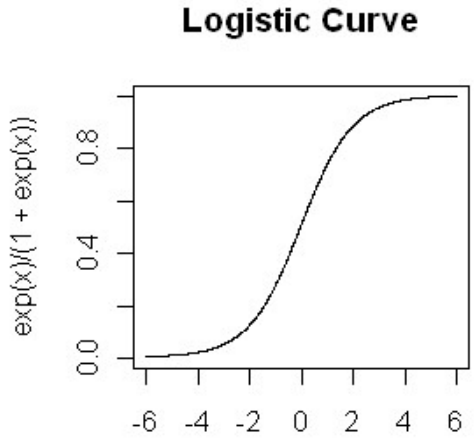
- $\log(\frac{p}{1-p}) = \Sigma \beta_i x_i$ = logit(p)
- $\exp[\log(\frac{p}{1-p})] = \exp[\Sigma \beta_i x_i]$
- $(\frac{p}{1-p}) = \exp[\Sigma \beta_i x_i]$
- $p = (1-p)\exp[\Sigma \beta_i x_i]$
- $p = \exp[\Sigma \beta_i x_i] - p \exp[\Sigma \beta_i x_i]$
- $p[1 + \exp[\Sigma \beta_i x_i]] = \exp[\Sigma \beta_i x_i]$
- $p = \frac{\exp[\Sigma \beta_i x_i]}{1+\exp[\Sigma \beta_i x_i]}$
- x = +∞? x = -∞? x = 0?

Logistic Curve



- Given our classification idea, for what values of x will it predict y = 1?

Logistic Curve



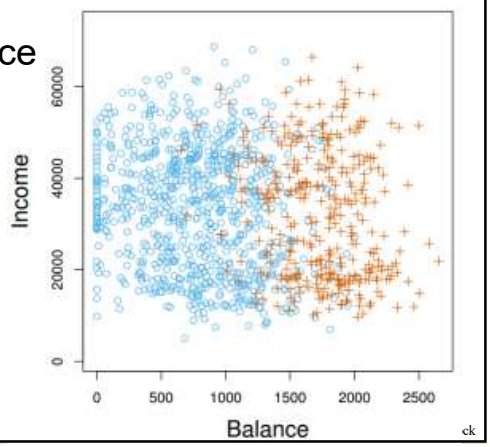
- Predict class 1 if x > 0 or logit > 0.5
- Predict class 0 if x < 0 of logit < 0.5

Cost Functions for Logistic Regression

- For linear regression:
  - $\arg \min_{m,b} \sum_{k=0}^{n} (\hat{y} - y)^2$
- Least squares optimization worked wel
- Logistic regression is not linear
- Nonlinear relationship results in a non-convex optimization
- Various alternatives can be used
- $J(\theta) = \frac{1}{N} \sum_{i=0}^{N} -y^{(i)} \log(h(x^{(i)}) - (1 - y^{(i)}) \log(1 - h(x^{(i)}))$

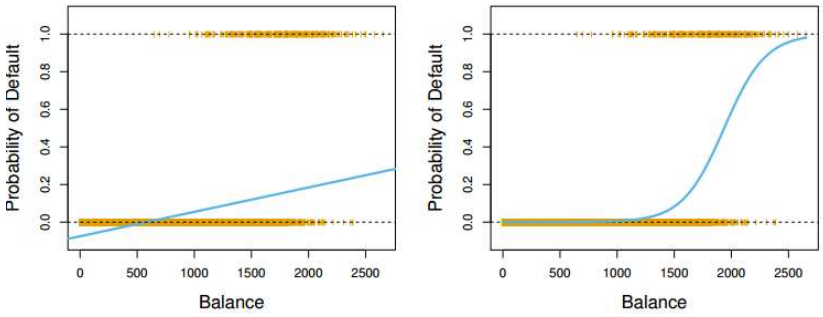
Example

- *Default* data set from "Introduction to Statistical Learning"
- Predict if will default
- 3 factors:
  - Credit card balance
  - Income
  - If student



Approaches for Default set

- Linear regression (left)
  - $p(X) = \beta_0 + \beta_1 X$
- Logistic regression (right)
  - $\log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 X$



Results of Logistic Regression

```
Coefficients:
             Estimate     Std. Error  z value  Pr(>|z|)
(Intercept)  -10.87       4.923e-01   -22.080  < 2e-16 ***
income         0.003033   8.203e-06     0.370  0.71152
balance        0.005737   2.319e-04    24.738  < 2e-16 ***
studentYes    -0.6468     2.363e-01    -2.738  0.00619 **
---
Signif codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5
```

- Is this fit any good?

Results of Logistic Regression

```
Coefficients:
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept) -10.75      3.692e-01  -29.116    < 2e-16 ***
balance       0.005738  2.318e-04   24.750    < 2e-16 ***
studentYes   -0.7149    1.475e-01   -4.846    1.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.7  on 9997  degrees of freedom
AIC: 1577.7
```

- What should we first do with this?
- How do we use this?

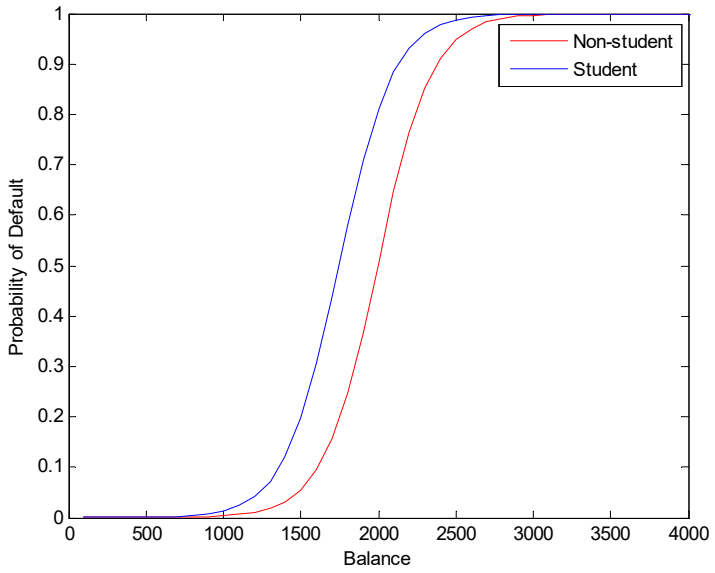
Predict results

- (Intercept) -10.75
- balance 0.005738
- studentYes -0.7149
- $\log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- $\log(\frac{p(x)}{1-p(x)})$ = -10.75 + 0.005738*balance – 0.7419*studentYes
- $p(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{(1+\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$
- $p(x) = \frac{\exp(-10.75 + 0.005738 \text{ balance} - 0.7419 \text{ studentYes})}{1+\exp(-10.75 + 0.005738 \text{ balance} - 0.7419 \text{ studentYes})}$

Prediction results

- $p(x) = \frac{\exp(-10.75 + 0.005738\ \text{balance} - 0.7419\ \text{studentYes})}{1+\exp(-10.75 + 0.005738\ \text{balance} - 0.7419\ \text{studentYes})}$
- If balance=$1000, studentYes=yes (1)
  - p(default) = 0.0032
- If balance=$2000, studentYes=yes (1)
  - p(default) = 0.4960
- If balance=$2000, studentYes=no (0)
  - p(default) = 0.6739

Graphical Result

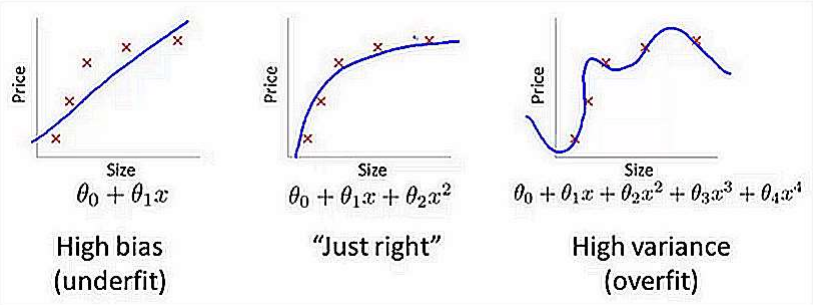


How to get around overfitting?

- Reduce features
- Use a simpler model
- Regularization method

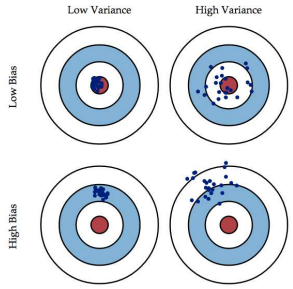
Bias vs. Variance for Regression

https://www.quora.com/How-would-you-explain-the-bias-variance-tradeoff-to-a-five-year-old



Bias vs. Variance

- $E([Y-f_{est}(x)]^2) = \text{noise} + \text{bias}^2 + \text{variance}$



- $E([Y-f_{est}(x)]^2) = \sigma_y^2 + \text{bias}(f_{est}(x))^2 + \text{var}(f_{est}(x))$
- Noise is what we cannot control: σ_y^2
- Bias = $E[f_{est}(x)] - f(x)]$
- Variance= Variance(f_{est})= $E[f_{est}(x)–E(f(x))^2]$

OLS Second Justification

- $Y = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4$
- $Y = 3 + 4.2x + -0.5x^2 + 142x^3 - 111_4x^4$

Regularization

- J = $\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} w_j^2$
  - Called ridge regression
- J = $\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |w_j|$
  - Called Lasso = "Least absolute shrinkage and selection operator"
  - Result is some parameters → 0 so it actually performs features selection automatically
- J = $\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} w_j^2 + \lambda_2 \sum_{j=1}^{P} |w_j|$
  - Called elastic net

What to do with regularization?

Fits into the gradient descent

- $w_j \leftarrow w_j - \alpha \sum_{i=1}^{N}(\hat{y}_i - y_i)x_i - \lambda w_j$
- $w_j \leftarrow w_j(1 - \alpha\lambda) - \alpha \sum_{i=1}^{N}(\hat{y}_i - y_i)x_i$

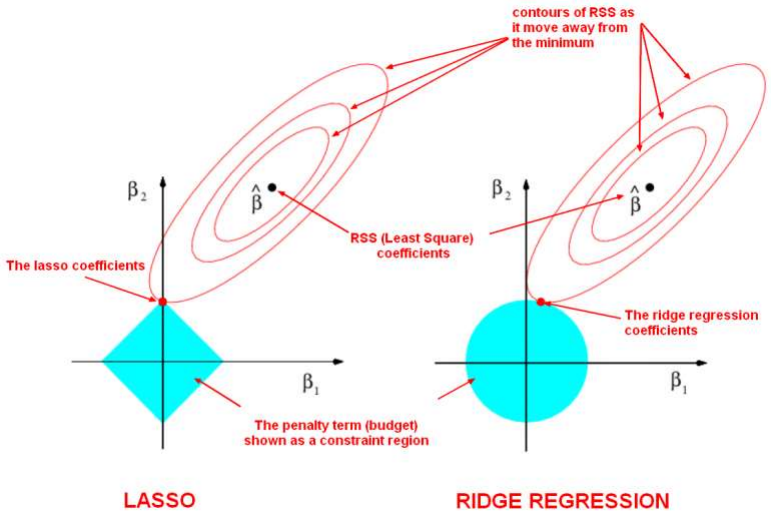
Fits into the matrix solution

- $W = (X^T X + \lambda I')^{-1} X^T y$
  - I' has size (#params+1) square matrix
  - I' is identity matrix with 0 for the top left term

How large should λ be?

- If λ is very large
  - Penalize large values of all terms
  - All weights go to 0 and underfit
- If λ is very small
  - No effect so prone to overfitting
- λ should be > 0
- λ often is often < #parameters being fit
- Usually try multiple values at factors of 2
- Cross-validate

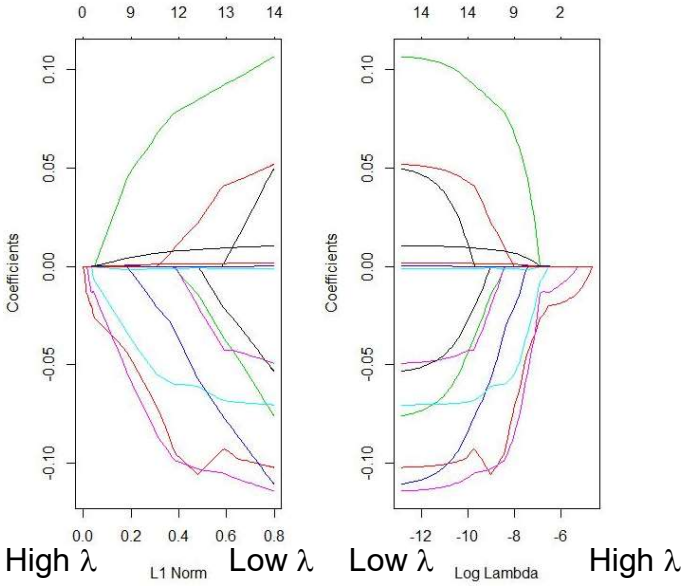
Classic Picture



http://gerardnico.com/wiki/data_mining/lasso

LASSO Variable Trace Plots

http://stats.stackexchange.com/questions/68431/interpretting-lasso-variable-trace-plots



High λ Low λ Low λ High λ

$$\beta_{lasso} = \text{argmin}\,[RSS(\beta) + \lambda * \text{L1-Norm}(\beta)]$$

Who cares?

- Trading variance and bias
- Avoid overfitting
- More effective when correlated features
  - Variance inflation is thereby avoided
  - Variance inflation often means large and offsetting coefficient values
- Standard approach for fitting many variables

Comparing Models

How to compare models?

- So far, talked only about 1 model
- What if we wanted to compare
  - $Y = b_0 + b_1X$
  - $Y = b_0 + b_1X + b_2X^2$
- Which is better?

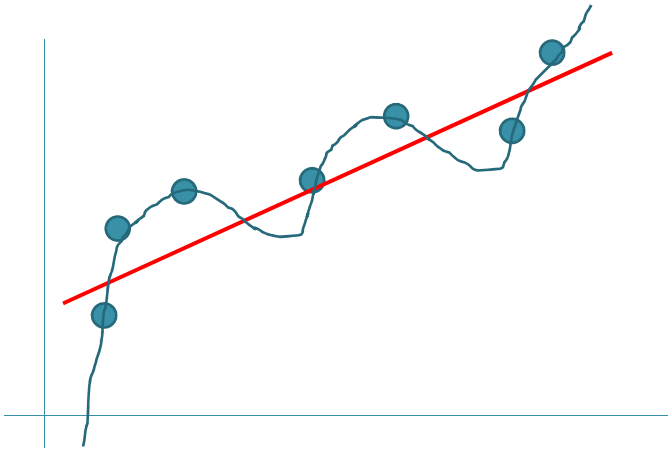
Model Comparisons

- Coefficient of determination
  - R^2 = SSreg/SST = 1-RSS/SST
  - If model includes intercept, R = corr coeff
  - SST = total variation = $\sum_{i=0}^{n}(y_i - \bar{\bar{y}})^2$
  - RSS = sumsq of residuals = $\sum_{i=0}^{n}(y_i - \hat{y})^2$
  - SSreg = regression sum of squares= $(\hat{y} - \bar{\bar{y}})^2$
    - Explained sum of squares through model
  - R^2 = explained variation / total variation
- Adjusted $R^2 = R^2 - (1- R^2)\frac{p}{n-p-1}$
  - n = #samples, p=#params

Model Comparisons

- Coefficient of determination
  - R^2 = SSreg/SST = 1-RSS/SST
  - If model includes intercept, R = corr coeff
  - SST = total variation = $\sum_{i=0}^{n}(y_i - \bar{\bar{y}})^2$
  - RSS = sumsq of residuals = $\sum_{i=0}^{n}(y_i - \hat{y})^2$
  - SSreg = regression sum of squares= $(\hat{y} - \bar{\bar{y}})^2$
    - Explained sum of squares through model
- Adjusted $R^2 = R^2 - (1- R^2)\frac{p}{n-p-1}$
  - n = #samples, p=#params
- AIC = Akaike Information Criterion
- BIC = Bayesian Information Criterion

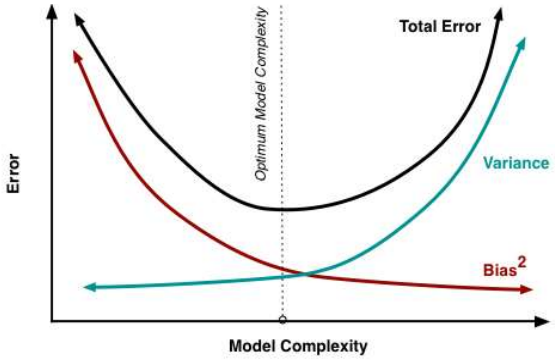
What is the best fit?



Bias vs. Variance Tradeoff

Goals of any model:

- Fit on unseen data
- Accurately fit trained data



Bias ~ Underfit
Variance ~ Overfit

http://scott.fortmann-roe.com/docs/BiasVariance.html

Regression with 40 features

- 1-40 features in combinations
  - 40 single features
  - ~40^2 pairs of features
  - ~40^3 triples of features …
- Don't forget about the interactions
- What is the overall strategy to try these?

Search Strategy

- Bottom-up strategy (forward)
  - Find best single feature
  - Find best 2nd feature that works with first
- Top-down strategy (backward)
  - Start with all the features
  - Remove the worst single feature
- Forward/Backward
  - Bottom-up with option to remove features between steps

Variance Inflation

www.racelies.com

One last issue

- Each parameter contains an estimate and a standard error
- What would happen if we included two copies of the same feature?
- $Y = 3.1 + 4X_1 + 6X_2$ where $X_1 = X_2$
- What's variance would you expect for the X_1 and X_2 estimates?

Variance Inflation Factor

- $Y = 3.1 + 4X_1 + 6X_2$ where $X_1 = X_2$
- $Y = 3.1 + 10X_1 + 0X_2$
- $Y = 3.1 + 0X_1 + 10X_2$

- Variance will be very high for these
- Same problem for correlated features

- Need to remove correlated features

How to get around the problem?

- What have we learned that:
  - Takes correlated column vectors
  - Converts them into orthogonal vectors?

How to get around the problem?

- What have we learned that:
  - Takes correlated column vectors
  - Converts them into orthogonal vectors?
- Take PCA of the X matrix → X'
- Apply regression on X' instead of X
- "Principal Component Regression"/PCR
- How many eigenvectors do we keep?

Generalizing Regression

Linear Basis Function Regression

- $y = \sum_{j=0}^{M-1} w_j \varphi_j(x) = w^T \varphi(x)$
- ϕ_j are basis functions
- Typical property is $\phi_0(x) = 1$
  - Mathematical convenience that lets w_0= bias
- Multiple linear regression:
  - $\phi_0(x) = 1$
  - $\phi_i(x) = x_i$

Types of basis functions

- Polynomial $\phi_i(x) = x^i$
- Gaussian $\phi_i(x) = \exp\left\{\frac{(x-\mu_i)^2}{2\sigma^2}\right\}$
- Sigmoidal $\phi_i(x) = \frac{1}{1+\exp\left\{-\frac{(x-\mu_i)^2}{\sigma}\right\}}$

Optimizing for least squares

- $W = (\Phi^T\Phi)^{-1}\,\Phi^T\,Y$

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

A single example

N x M

A basis function

- Comparison to the matrix form from multiple linear regression
  - $B = (X^TX)^{-1}X^TY$

Feature Engineering

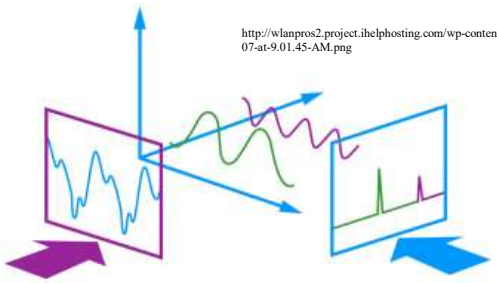
- Take any function of the inputs
- Convert to a linear regression problem

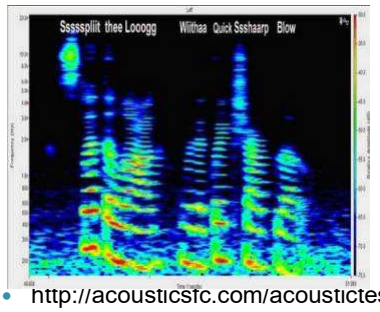
Feature Engineering Approaches

- At a later date…

Time vs. Frequency Domain

http://wlanpros2.project.ihelphosting.com/wp-content/uploads/2014/01/Screen-Shot-2012-11-07-at-9.01.45-AM.png

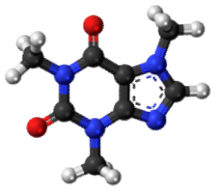


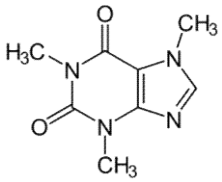


- http://acousticsfc.com/acoustictest.htm

Chemical Modeling

- Caffeine





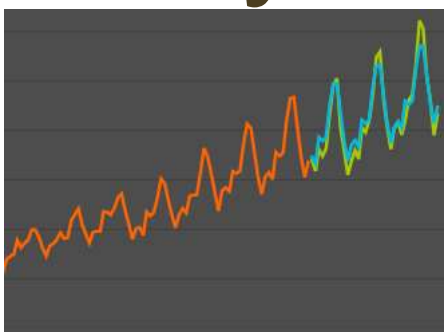
- Topological descriptors
- 2D connectivity
- 3D spatial characterization
- Charge modeling

Text mining

- Counts of words
- Frequencies of words
- Normalized frequencies TFIDF
- Adjacency

www.turnafrownaround.org

Time Series Analysis

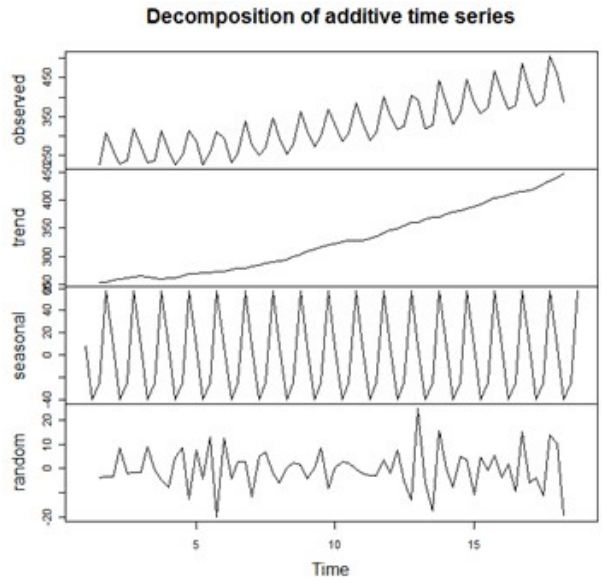


https://www.wolfram.com/mathematica/new-in-10/expanded-time-series-processes/

Applications of Time Series

- Forecast sale of apple watch 4
- Predict blood glucose levels in diabetics
- Model the gait in Parkinson's patients
- Assess global warming temperatures

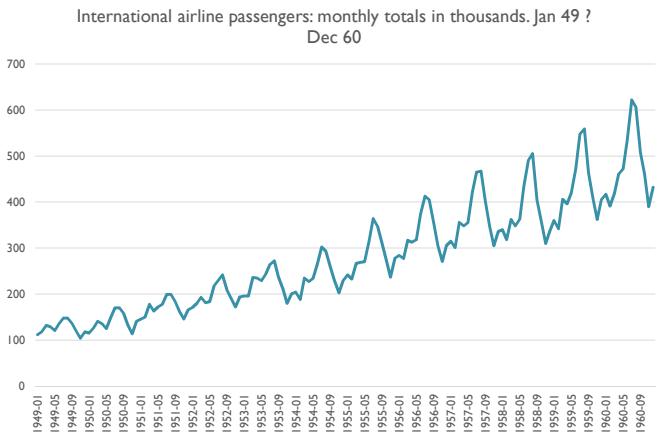
Signal Decomposition



https://onlinecourses.science.psu.edu/stat510/node/69

Decomposition: not just additive

- International Airline Passengers Data
- https://datamarket.com/data/set/22u3/international-airline-passengers-monthly-totals-in-thousands-jan-49-dec-60#!ds=22u3&display=line



Decomposition Types

- Additive
  - Full = noise + seasonal + trend
- Multiplicative
  - Full = noise * seasonal * trend

Time Series Lecture

- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic
- Seasonality Extension
- Example

Key Properties

- Stationary: invariance w.r.t. time
  - Every value should have same distribution
  - Real process are rarely strictly stationary
- Weak Stationary
  - Mean doesn't change over time: no trends
  - Covariance is function of Δt or lag
- Ergodic:
  - Use part of sequence to estimate statistics of the full process

WSS or Weak Stationarity

- Mean is time invariant
- Mean and variance are finite
- Auto-covariance is a function of only the time-difference and not k (i.e. time alone)
  - $\text{cov}(X_{k1}, X_{k2}) = E[(X_{k1}-\mu_1)(X_{k2}-\mu_2)]$
  - Covariance(1, 3) = Cov(10, 12)

Mathematical idea of stationary

x(k) = Acos(wk + u) where u is U(0,π)

Is this stationary?

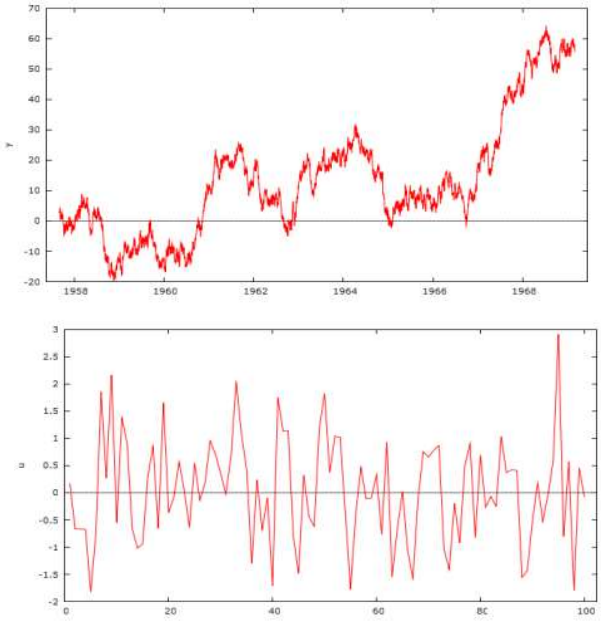
How would we determine the answer?

E(x|k) = $\int_0^{\pi}$ Acos(wk + u)$f(u)du$

E(x|k) = $\frac{A}{\pi}\int_0^{\pi}$ cos(wk + u)$1du = \frac{-2}{\pi}\sin(wk)$

E(x|k) depends on k so not stationary

Random Walk (top), iid(0,1) (bottom)



Random Walks

Random walk

- x(0) = w(0)
- x(k) = x(k-1) + w(k)
- x(k) = $\sum_{n=0}^{k} w(n)$

where E(w)= 0, var(w) = σ^2 and w are i.i.d.

Let's characterize x:

- Mean(x) = ?
- Variance(x): kσ^2

- Is x stationary [weak]?

Random Walks & Differencing

- x(0) = w(0)
- x(k) = x(k-1) + w(k)
- x(k) = $\sum_{n=0}^{k} w(n)$

- X random process is not stationary
  - Mean(x) = 0
  - Variance(x): kσ^2

- Is difference x(k+1) - x(k) stationary?

Random Walks & Differencing

Random walk

- x(0) = w(0)
- x(k) = x(k-1) + w(k)
- x(k) = $\sum_{n=0}^{k} w(n)$

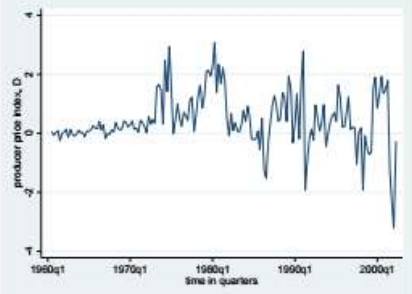
where E(w)= 0, var(w) = σ^2 and w are i.i.d.

- Is x stationary [weak]?
  - Mean(x) = ?
  - Variance(x): $k\sigma^2$
    - Individual terms are uncorrelated
- Is difference x(k+1)-x(k) stationary?
  - x(k+1)-x(k) = w(n) so yes

Apply Time Differencing

- Y(t) = x(t)-x(t-1)

Differenced variable (Δppi)



Is it stationary?

Why care about stationarity?

- Good models exist for stationary data
- Frequently transform data to become stationary using differencing
- Can also use log(), sqrt(), etc. but differencing is the most common due to trends

Time Series Lecture

- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic
- Seasonality Extension

Time Series Review

- Random Process
  - w(k) = 0, 2, 0, -1, 2, -1, 2, 1, 0, 0, 1, 2, -1, -2
- AR Model
  - $x(k) = \phi_1 x(k-1) + \phi_2 x(k-2) + \ldots \phi_p x(k-p) + w(k)$
  - Let's use a 1st order model $\phi_1 = 0.9$
  - x(0) = 0
  - x(1) = 0.9*0 + 2 = 2
  - x(2) = 0.9*2 + 0 = 1.8
  - x(3) = 0.9*1.8 + -1 = 0.62

Time Series Review

- Random Process
  - w() = 0, 2, 0, -1, -2, -1, 2, 1, 0, 0, 1, 2, -1, -2
- MA Model (clearer notation)
  - $x(k) = w(k) + \theta_1 w(k-1) + \theta_2 w(k-2) + \ldots \theta_p w(k-p)$
  - Let's use a 2nd order model $\theta_1 = 0.7$, $\theta_2 = 0.3$
  - Assume x(-1) = 0
  - x(0) = 0
  - x(1) = 2 + 0.7(0) = 2
  - x(2) = 0 + 0.7(2) + 0.3(0) = 1.4
  - x(3) = -1 + 0.7(0) + 0.3(2) = -0.4

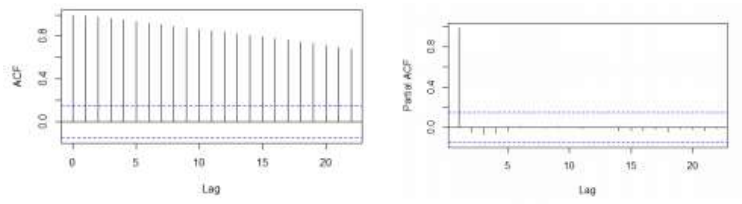
Key Diagnostics

- Autocorrelation Function (ACF)
- Partial Autocorrelation (PACF)

- Examine ACF and PACF to figure out what model to apply to stationary data

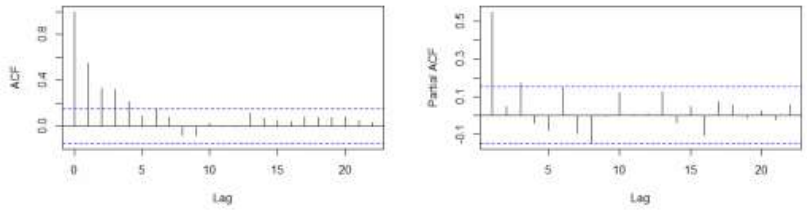
- Verify model:
  - ACF to ensure residuals zero
  - Usual models: no residual pattern, etc.
  - Ljung-Box statistic

ACF and PACF



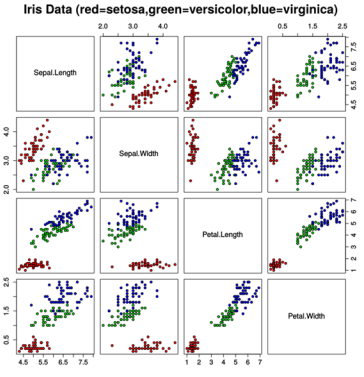
- ACF has a slow decay on raw data
- PACF has a single peak at lag=1

ACF and PACF Differenced Data

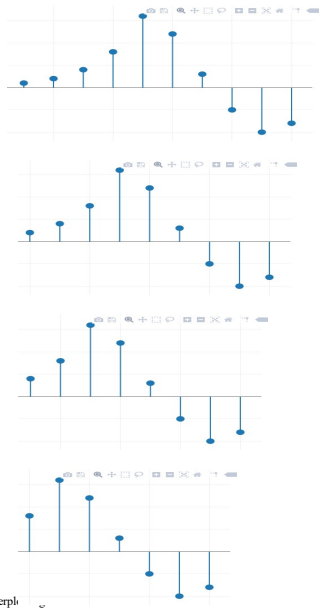


- Took x(t)-x(t-1) and ran ACF and PACF
- ACF has slow decay
- PACF has 1 significant peak

Correlation vs. Time Series ACF



https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_dataset_scatterpl



X(k)

X(k-1)

X(k-2)

X(k-3)

Covariance

- $x_1, x_2, \ldots x_n$
- Compute covariance between all pairs
- For weak stationary process:
  - Mean is invariant
  - Distribution is function of lag k_2-k_1
- Autocorrelation Variance Function for weak stationary:
  - Let $\mu = E[x(k)]$
  - Cov(lag) = E[(x(k) - μ) (x(k-lag) - μ)]

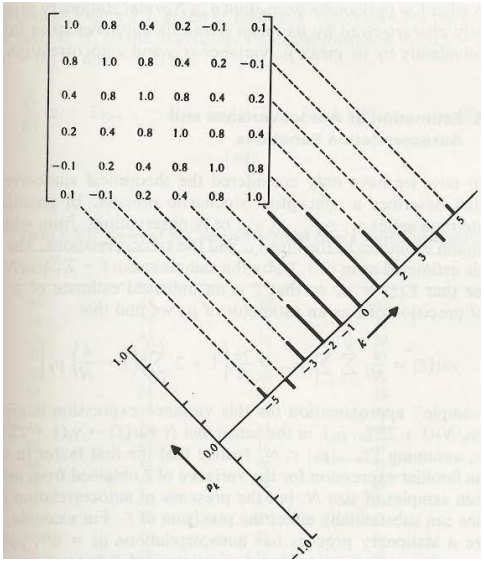
Math concept

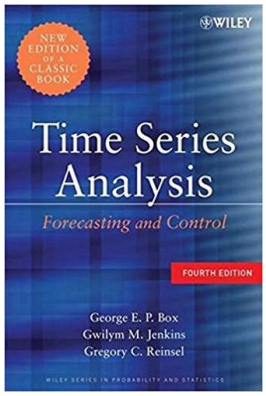
- How would you describe the format of the covariance matrix for weak stationary process?

$$A = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix}$$

- Called a Toeplitz matrix
- Also symmetric and positive semi-definite

Autocorrelation Matrix Idea



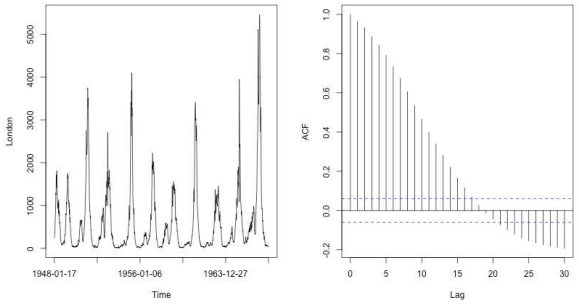


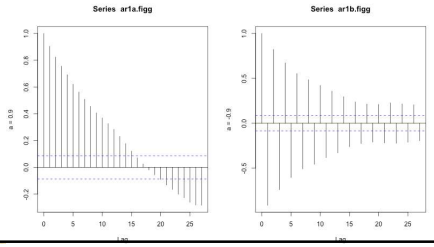
Correlation coefficient

- r = CorrCoef(X,Y) = cov(X,Y) / ($s_X s_Y$)
  - Technically, the Pearson corr. coef.
- $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$
- Range of values for correlation is [-1, 1]
- No units

Autocorrelation Function

- Autocorrelation(l) = $\frac{var(l)}{var(x_k)var(x_{k-l})}$



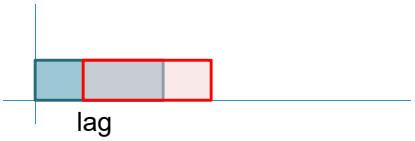


https://people.maths.bris.ac.uk/~magpn/Research/LSTS/STSIntro.html

ock

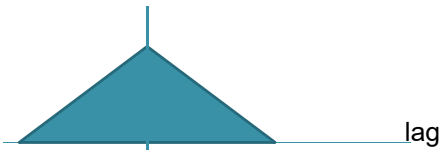
ACF (Autocorrelation Function)

- Use ACF to build a better model
- Shows Pearson correlations between all lags as function of the lag
- General form:
  - $R(s,t) = E[X_t-\mu_t)(X_s-\mu_s)] / [\sigma_t\sigma_s]$
  - R(s-t) = cov(s-t) / cov(0)
- For WSS with μ and σ
  - $R(lag) = E[X_t-\mu)(X_{t+lag}-\mu)] / \sigma^2$



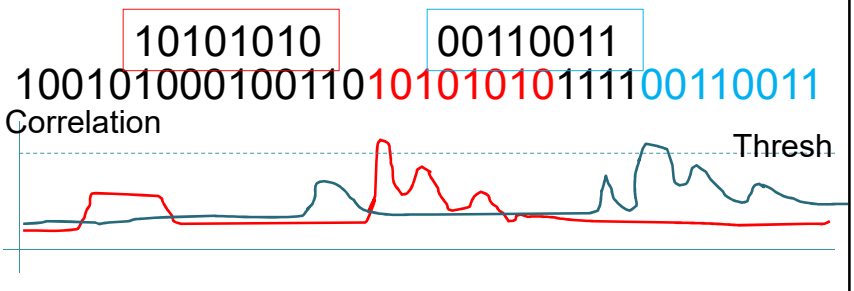
Sliding window f(lag)

- Sliding window → function(lag)
- Have you seen this before?



Correlation in Signal Processing

- Trying to send you 1's and 0s
- Too much noise to send you a direct sequence so I send a longer code
- 10101010 → 1
- 00110011 → 0



Autocorrelation

- Want to have a clear "match" for your key
- Correlate signal against itself

- 10101010
  10101010
  10101010
  10101010
  10101010
  10101010
- Would have slow taper in even ACF values

Autocorrelation Properties = ACF

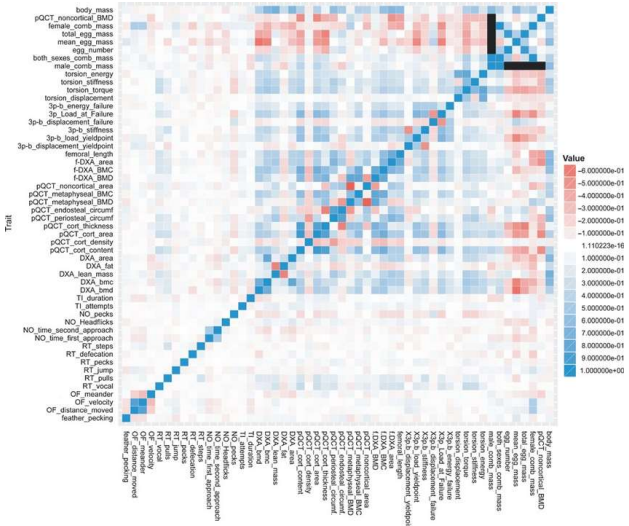
- For WSS with μ and σ
  - $R(lag) = E[X_t-\mu)(X_{t+lag}-\mu)]/\sigma^2 = Cov(X_t,X_{t+lag})/\sigma^2$
- R(lag) = R(-lag)
- R(lag=0) = Variance
- Correlation is func(lag) but same for all t

- For time series modeling, usually the lag=0 is excluded (depends on software)
- Test positive lags for non-significant values of the residuals

Partial Autocorrelation Function

- Conditional correlation measuring strength and direction of 2 terms while controlling effects of others
- PACF = $$\frac{Cov(y,x_i|x_{j\neq i})}{\sqrt{Var\left(y\middle|x_{j\neq i}\right)Var(x_i|x_{j\neq i})}}$$
- Describing how y is related to x_i given the other independent variables x_j
- Showing it in terms of how y relates to the x_j and how x_i relates to the x_j

Aside: Partial Correlation is useful



Genetic architecture of domestication in the chicken: Wright et al. 2010

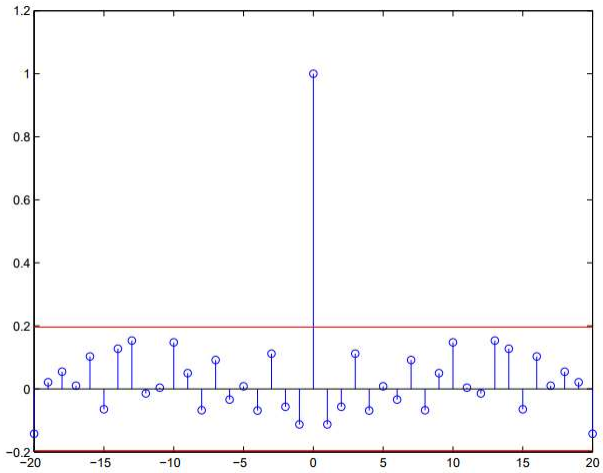
Partial Autocorrelation Function

- Time series version
- PACF = $\frac{Cov(x_t, x_{t-lag} | x_{j \neq t, lag})}{\sqrt{Var\left(x_t \middle| x_{j \neq t, lag}\right) Var(x_{t-lag} | x_{j \neq t, lag})}}$
- PACF(lag=1): same as autocorrelation
  - Only include t-1 term so nothing left
- PACF(lag=2): $x_{j \neq t, lag} \rightarrow x_{j=1}$
- PACF(lag=3): $x_{j \neq t, lag} \rightarrow x_{j=1,2}$

White Noise Process

- Random process
- Terms are:
  - Uncorrelated
  - Mean 0
  - Finite variance
- ACF(lag) =
  - 1 if lag=0
  - 0 otherwise

ACF for white noise



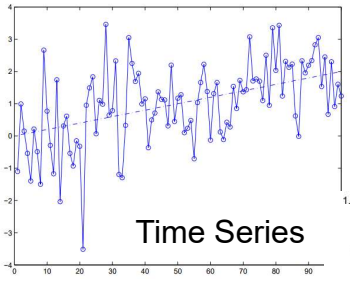
- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

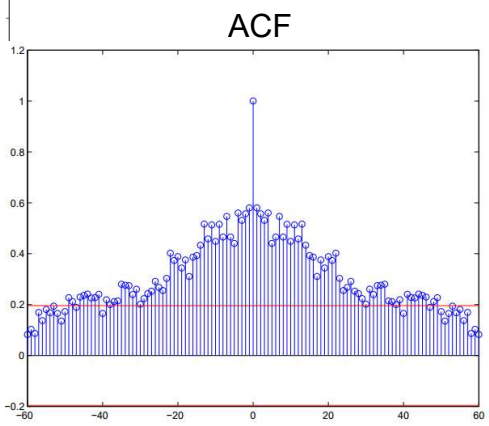
Why care about white noise?

- If we put white noise into a system, then you produce the random process
- Standard approach for signal processing modeling

- Input → System → Output
- Noise →System → “Transfer_function”
- Input * Transfer_function → Output

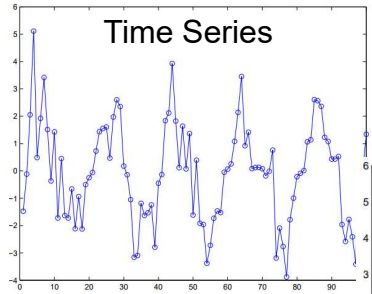
ACF of a "Trend" signal

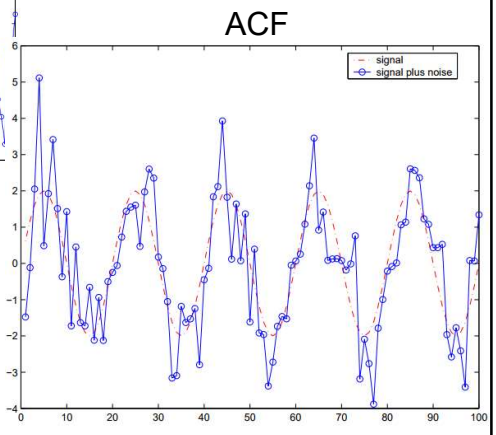




http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

ACF of Periodic Signal





- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

Time Series Lecture

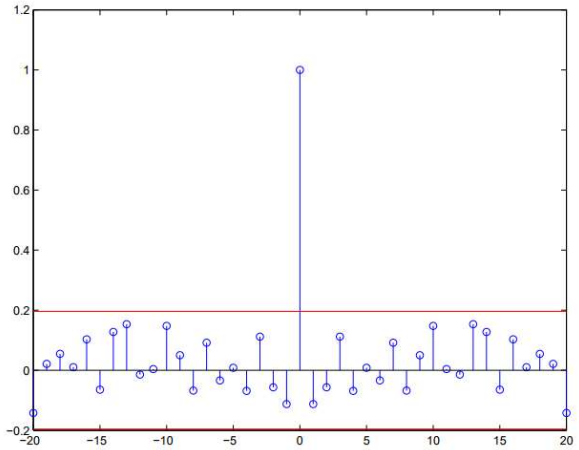
- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic
- Seasonality Extension
- Example

How do we find the model?

- Want to have a minimal model that meets the criteria:
  - Regression fit so normal residuals
  - Residuals are white noise (ACF, PACF)
- Two strategies:
  1) Examine ACF & PACF from raw data to determine approximate model from pattern
  2) "Whack-a-mole"

Checking ACF(residuals) are ~ 0

- White noise ACF(0) = 1.0 , ACF(k≠0) = 0
- Residuals should be like white noise



- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

AR(1) Model

- AR(1) = first order model
- $x_t = c + \phi_1 x_{t-1} + w_t$

- c is a constant to be estimated
- ϕ_t is a weight to be estimated
- w_t is N(0, σ_w^2) i.i.d. random variable

- This is a regression model

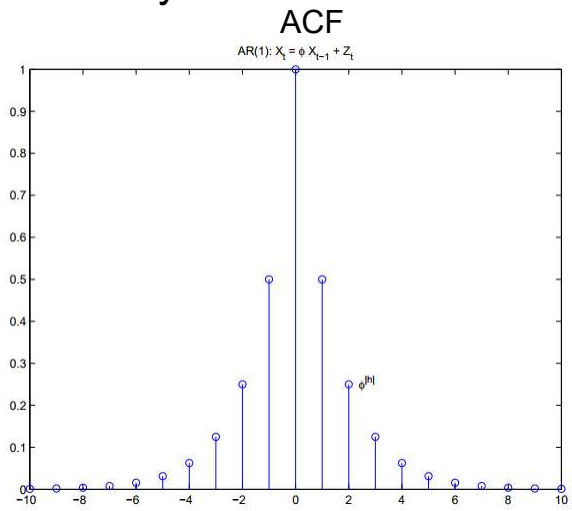
AR(1) Model

- AR(1) = first order model (since only 1 ϕ_1)
- $x_t = c + \phi_1 x_{t-1} + w_t$
- Mean: $c / (1-\phi_1)$
- Var: $\sigma_w^2 / (1-\phi_1^2)$
- ACF values as f(lag) = ϕ_1^{lag}

https://datascience.stackexchange.com/questions/18268/acf-function-shows-error-while-fiiting-time-series/18327

ACF of AR(p=1)

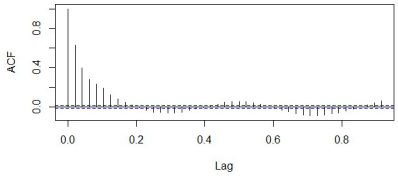
Exponential decay



http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

AR Model

- ACF values as f(lag) = ϕ_1^{lag}



What would it look like if ϕ_1 were negative?

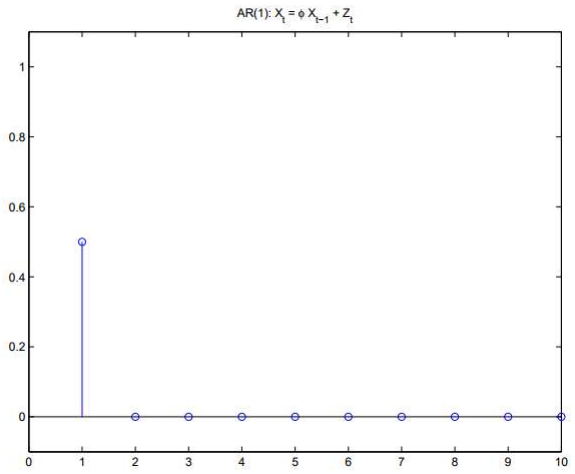
- PACF ~ 0 after the order of the model

https://datascience.stackexchange.com/questions/18268/acf-function-shows-error-while-fiiting-time-series/18327

Generalization of AR(1) to AR(p)

- AR(1)
  - $x_t = c + \phi_1 x_{t-1} + w_t$
- AR(2)
  - $x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$
- AR(p)
  - $x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + w_t$

PACF of AR(p=1) model: 0 for h>p



How does the ACF help?

- Look at the raw data in time series form
- Might take differences i.e. y=x(t)-x(t-1)
- Look at the ACF of raw signal y to figure out the kind of model

- Fit the model
- Look at the residuals
- Check that ACF of residuals are ~ 0

Moving Average Models

- w_t is N(0, σ_w^2) i.i.d. random variable

- MA(1): $x_t = \mu + w_t + \theta_1 w_{t-1}$
- MA(2): $x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$
- MA(q): $x_t = \mu + w_t + \Sigma_{i=[1..q]} \theta_i w_{t-I}$

- For MA(1)
  - Mean(x_t) = μ
  - Var(x_t) = $\sigma_w^2(1+\theta_1^2)$
  - ACF(1) = $\theta_1 / (1+\theta_1^2)$ but ACF(x) = 0 for x>=2

Properties for higher order MA

- For MA(2)
  - Mean(x_t) = μ
  - Var(x_t) = $\sigma_w^2(1+\theta_1^2+\theta_2^2)$
  - ACF(1) = $(\theta_1 + \theta_1\theta_2)/(1+\theta_1^2+\theta_2^2)$
  - ACF(2) = $\theta_2 / (1+\theta_1^2+\theta_2^2)$
  - ACF(>=3) = 0

- In general MA(q)
  - Mean = μ
  - ACF values are non-zero for <= q

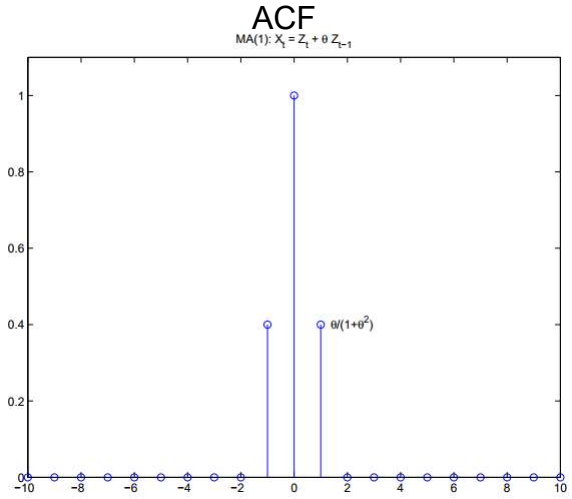
MA Model

- ACF = 0 for lags > model order
- PACF tapers to 0 but not a clear pattern

https://datascience.stackexchange.com/questions/18268/acf-function-shows-error-while-fiiting-time-series/18327

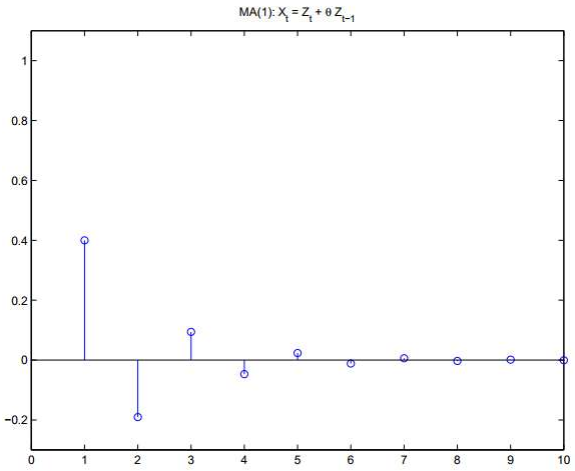
ACF of MA(q=1)

- Zero for |h| > q where q = order of MA



- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

PACF of MA(1) model decays



Using ACF & PACF

	ACF	PACF
White noise	All 0	All 0
Non-stationary	Stay significant	Stay significant
AR(1)	Nonzero at lag=1 only	0 for lag > 1
AR(2)	Sinusoid → 0 after lag 2	0 for lag > 2
MA(1)	Goes to 0	Decays
MA(2)	0 for 3rd and higher lags	Decays
AR(p)	Decays exponentially	0 for lag > p
MA(q)	0 for lag > q	Decays
ARMA	Goes to 0	Goes to 0

From earlier (as it relates to the above):

- The lag=0 is excluded usually
- Focus on positive lags for non-significant values of the residuals

Non-Seasonal Models

- Autoregressive model = AR model
  - Only includes AR terms
- Moving Average model = MA model
  - Only includes MA terms
- ARMA: contains both but no differencing
- ARIMA = Box-Jenkins
  - Contains AR + MA with differencing

Examine Your Zipper

- Pretty Darn Quick

- https://www.keepcalm-o-matic.co.uk/p/keep-calm-and-xyz-pdq/

ARIMA(p,d,q)

- Notation: ARIMA(p,d,q)
  - p: order of the AR model
  - d: differencing order
  - q: order of the MA model
- Differencing:
  - x_t-x_{t-1} is 1st order d=1
  - (x_t-x_{t-1}) - (x_{t-1}-x_{t-2}) is 2st order d=2 (quadratic)
    - Generally try not to go beyond d=2
  - Use differencing to produce a wide-sense stationary model
- ARIMA(2,1,0):
  - AR(2), no MA applied to x_t-x_{t-1}

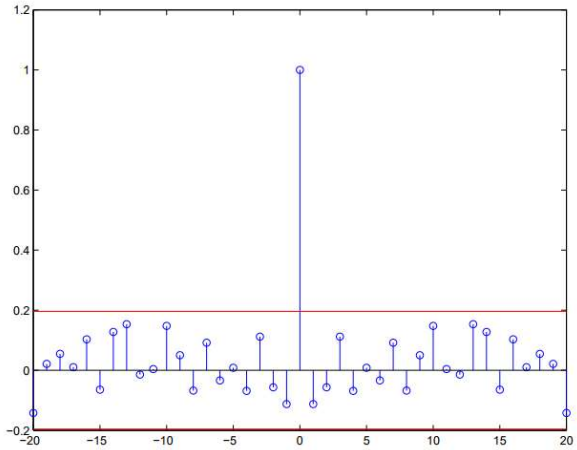
Time Series Modeling is like Whack-a-mole

No moles were harmed in the making of this slide.

- http://news4wide.livedoor.biz/archives/2144242.html

Checking ACF(residuals) are ~ 0

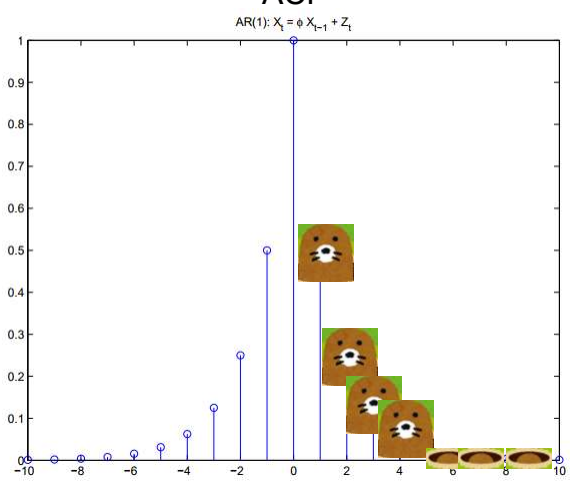
- White noise ACF(0) = 1.0 , ACF(k≠0) = 0
- Residuals should be like white noise



- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

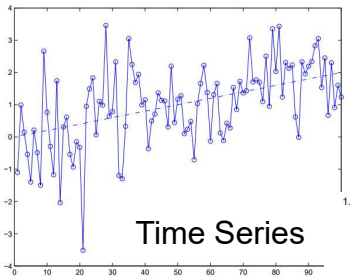
ACF of AR(p=1)

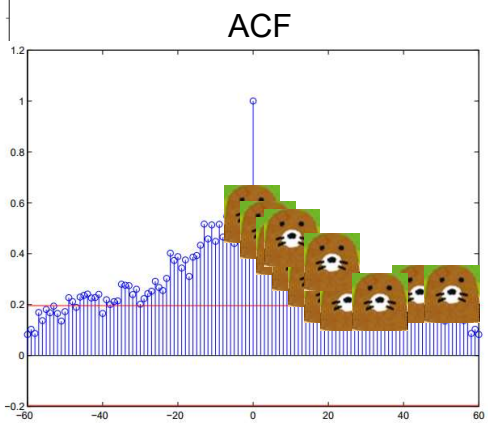
- Ignore k = 0 (most ACFs show 1…N)
- Need to find model to whack down [P]ACF



- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

ACF of a "Trend" signal





- http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/3.pdf

Brute Force approach

ARIMA Models

	ARIMA (1,0,0)	ARIMA (2,0,0)	ARIMA (0,0,1)	ARIMA (1,0,1)	ARIMA (1,1,0)	ARIMA (0,1,1)	ARIMA (1,1,1)	ARIMA (1,1,3)	ARIMA (2,1,3)
Const	64.37	64.18	64.69*	64.67*	0.46*	0.47*	0.43*	0.43*	0.44*
L1.ar	0.999*	1.64*		0.99*	0.55*		0.72*	0.73*	1.51*
L2.ar		-0.64*							-0.71*
L1.ma			1.00*	0.53*		0.48*	-0.25*	-0.24	-1.05*
L2.ma								-0.10	0.21
L3.ma								0.12	0.32*
AIC	502	424	1401	441	393	405	393	392	390
BIC	511	426	1420	543	402	414	406	411	412

* These are the Stata results. R has very similar coefficients. In the SAS output, the MA components have reverse signs than what is reported in this table and some coefficients have different magnitudes.

- We know that the variable is not stationary so we need to use differenced variable ARIMA (p,1,q). But here we also include models with the original variable ARIMA (p,0,q).
- The coefficient on the lagged dependent variable is close to 1 indicating non-stationarity.
- To select a model to use, look at the significance of the coefficients and also low AIC or BIC. Usually, there are a few models that perform similarly, so it is up to the researcher to try a few models and decide which one to use. The recommendation is to go with the simplest model.
- ARIMA(1,1,1) is a good choice based on low AIC and BIC.
- ARIMA(2,1,3) is also a good choice based on the significance of the lags.

Time Series Lecture

- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic & diagnostics
- Seasonality Extension

Choosing the Right Model

- We want to whack down the significant terms of the ACF and PACF
- Tools for “whacking” are:
  - Differencing x(i)-x(i-1) for all i
  - AR models
  - MA models
  - Combination of AR+MA = ARMA or ARIMA
- Use as few terms as possible
- But how do we know which?

Linear vs. Time Series Model

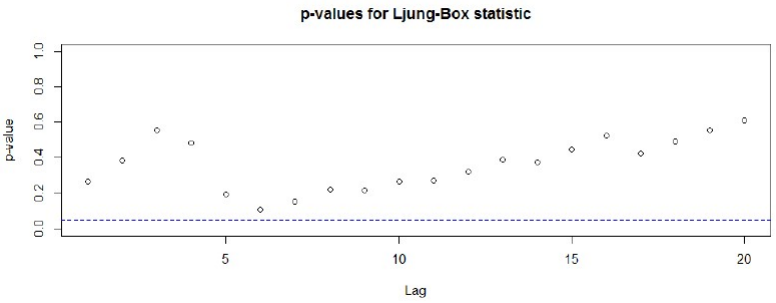
- Linear Regression:
  - Modeling Y as function of columns X1..XN
  - Minimizing sum-squared of residuals
  - Assuming (and checking) residual properties
- Time Series Modeling
  - Modeling stationary x(k) as function of
    - x(k-1),x(k-2), and other past sequences (AR)
    - w(k), w(k-1), … and other past noise (MA)
  - Apply linear regression to data
  - Assuming (and checking) residual properties
  - Checking ACF(residuals) are ~ 0

Linear vs. Time Series Model

- Y = WX + e
- Same equation for both models
- In the time series:
  - Y are now X(k)
  - X matrix: now past values: X(k-*)
- Residuals have similar assumptions
- Under the hood: same regression model
- Tailored diagnostics:
  - Ljung-Box & ACF plots

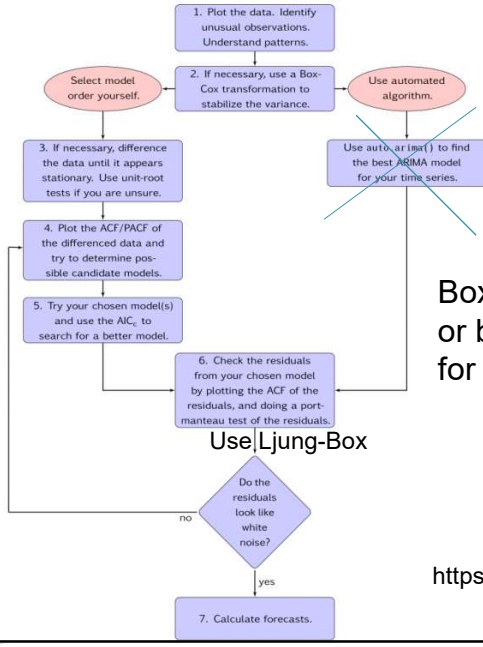
Ljung-Box Statistic

- Goal is that the residuals are zero
- Ljung-Box calculates their accumulated autocorrelation from a lag of 1 on up
- Want values to be non-significant (> 0.05)



https://www.researchgate.net/figure/Plot-of-Ljung-Box-p-values-of-fitted-ARIMA2-1-0_fig9_263505554

Model to Prediction Flowchart



R not python

Box-Cox finds sqrt or best Y^u power for the data

Use Ljung-Box

https://www.otexts.org/fpp/8/7

Making Predictions from Time Series

Predictions of Time Series

- Making predictions is almost easy
- AR model example:
  - $x_{t+1} = c + \phi_1 x_t + \phi_2 x_{t-1} + w_{t+1}$
  - You estimate your parameters c, ϕ_1, ϕ_2
  - Then what?

Two problems of prediction

- $x_{t+1} = c + \phi_1 x_t + \phi_2 x_{t-1} + w_{t+1}$
- Data goes from time 1 to 100.
- [2.3 2.6 1.1 …. 9.2]
- How do we predict time 101?
  - t
  - x_t
  - x_{t-1}
  - w_{t+1}

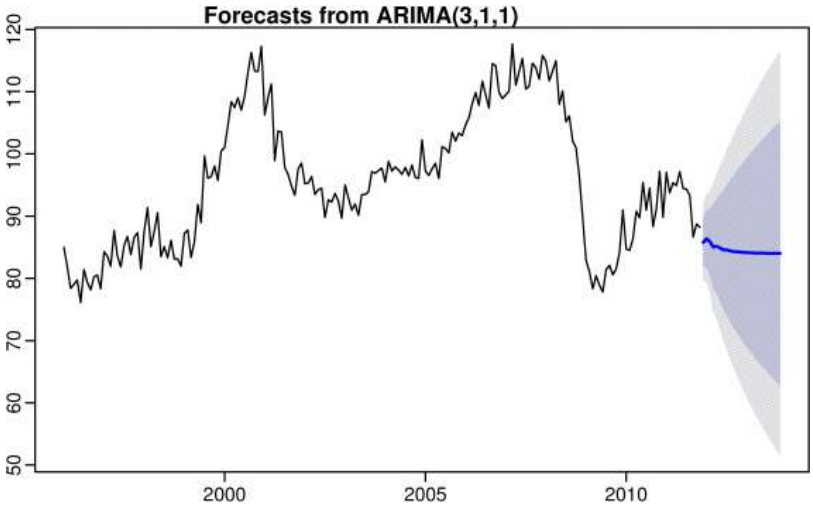
Two problems of prediction

- $x_{t+1} = c + \phi_1 x_t + \phi_2 x_{t-1} + w_{t+1}$
- Data goes from time 1 to 100.
- How do we predict time 101?
  - x_t
  - x_{t-1}
  - w_{t+1}
- How do we predict time 102?
  - x_t
  - x_{t-1}
  - w_{t+1}

Prediction Generalization

- Generalization:
  - Use known w_k when known else 0
  - Use known x_k when known else use estimate
  - "Known" defined as k < t (your last samples)
- What happens to the error bars on the estimate as your predictions continue?
- Variance at m points into the future:
  - $= \sigma_w^2 \sum_{i=0}^{m-1} \Psi_i^2$
  - Any ARIMA models ←→infinite MA model
  - $x_t - \mu = w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots$ for $\sum_{i=0}^{\infty} |\Psi_i| < \infty$

Typical Forecast Pattern – why?



- https://www.otexts.org/fpp/8/7

Time Series Lecture

- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic & diagnostics
- Seasonality Extension
- Example

Seasonality

- Many time series have a cyclic seasonal pattern such as consumer products
- Seasonality really means seasons rather than other cyclic patterns
- How do you detect seasonality?
  - Expert knowledge
  - Eye-balling
  - Plot average value across all months
  - Frequency domain techniques
    - FFT
    - Periodogram

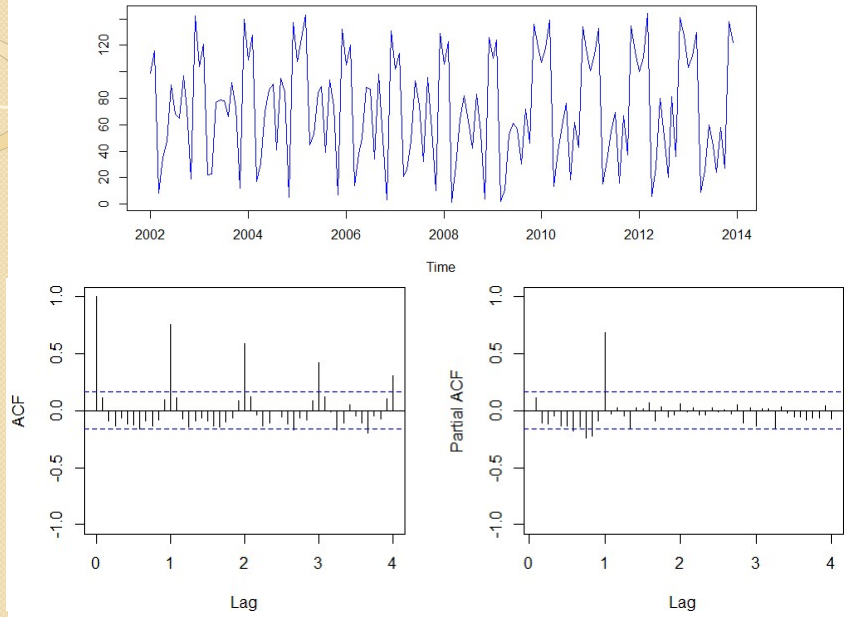
Seasonality ←→ stationary?

- Seasonal pattern will not be stationary
- Need a stationary pattern to model
- Usually when we have a trend, we could "difference" it away using
  y(t) = x(t) – x(t-1) and modeling y
- What could we do about a seasonal pattern to make it stationary?

Seasonality ←→ stationary?

- Solution is to use seasonal differencing
- For annual effect with monthly data
  - Model $x_t - x_{t-12}$ instead of x_t
- Can combine the trend differencing and the seasonal differencing together with two differ:
  - $(x_t - x_{t-1}) - (x_{t-12} - x_{t-13})$ or
  - $(x_t - x_{t-12}) - (x_{t-1} - x_{t-13})$
- Example:
  - https://people.duke.edu/~rnau/411sdif.htm

ACF and PACF of Seasonality

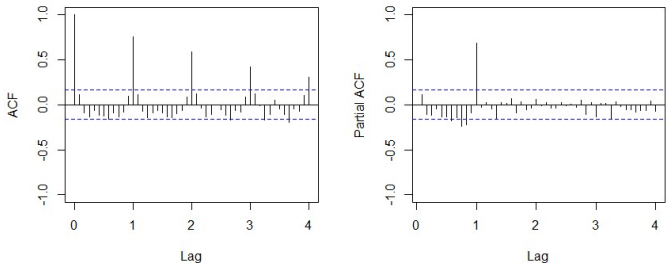


Seasonality Strategy

Exactly the same methods with the AR and MA models except…

- Lags are not 1,2,3… but 12,24,36
  - (Assuming months)
- Difference for stationary is x(t)-x(t-12)
- Pattern of residuals is at 12, 24, 36, …

To difference or not to difference



- Non-seasonal AR(1)
  - Gradual decay in ACF; Single spike in PACF
- Seasonal AR(1): $(0,0,0)(1,0,0)_{12}$
  - Shows gradual decay in ACF at lags = 12i
  - Shows single spike in PACF at lag=12
- Stationary may be solvable by seasonal model w/o differencing

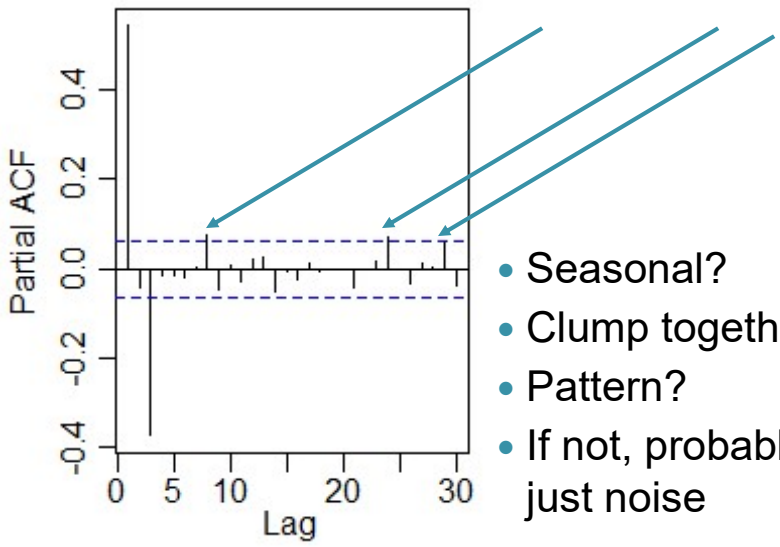
Seasonal ARIMA model

- ARIMA(p,d,q) x $(P,D,Q)_S$
- ARIMA(0,0,1) x $(1,0,0)_{12}$
- Apply a difference of 12 for months
  - D=0 or D=1, i.e. not 12 in notation
  - Could apply differencing to the series to remove annual trending
- Apply non-seasonal MA(1) model
- Apply seasonal AR(1) to season
  - Don't look at the ACF and PACF at lag 1,2
  - Look at ACF and PACF at multiples of 12 or 12, 24 in this case

Seasonal ARIMA model

- ARIMA(p,d,q) x (P,D,Q)S
- Concept is to apply differencing as needed for the season and/or trend removal
- View the ACF and PACF on differenced
- Perform usual ARIMA(p,d,q)
- Also look for ARIMA(P,D,S) and look for ACF and PACF patterns in multiples of the season
  - Decay for AR(2) after 2 lags→lag12, lag24

'bout them barely significant spikes



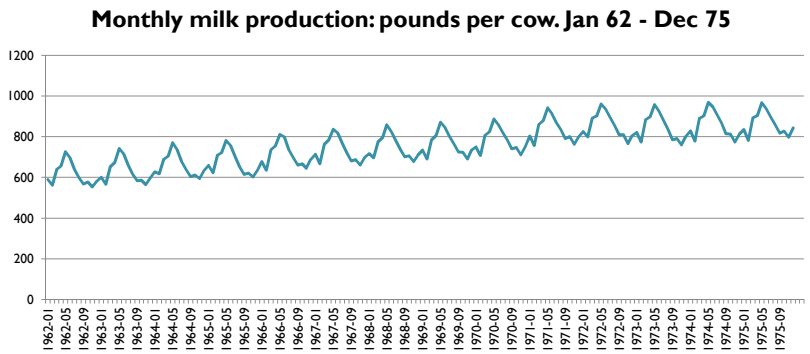
- Seasonal?
- Clump together?
- Pattern?
- If not, probably just noise

Time Series Lecture

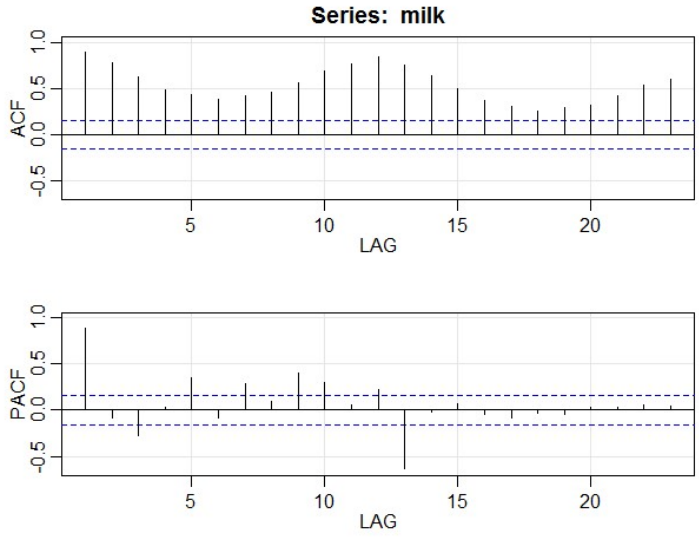
- Stationarity: WSS and how to get there
- ACF & PACF to understand shape
- Constructing AR, MA & ARIMA models
- Ljung-Box statistic & diagnostics
- Seasonality Extension
- Example

Seasonal Example

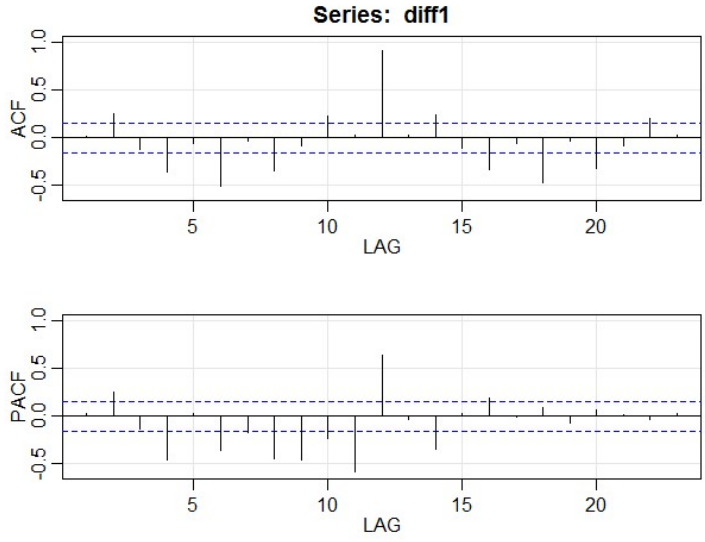
- Milk!



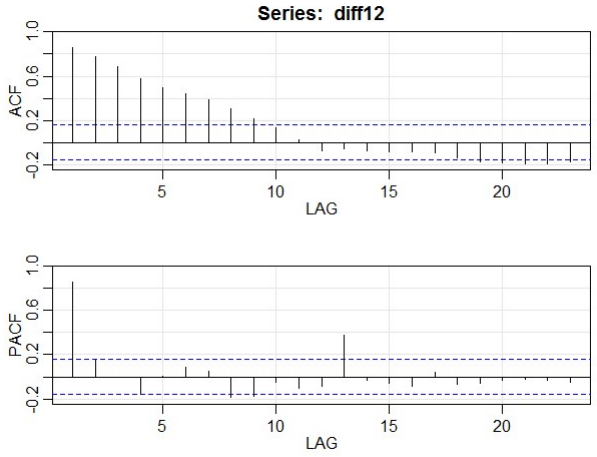
Raw Data



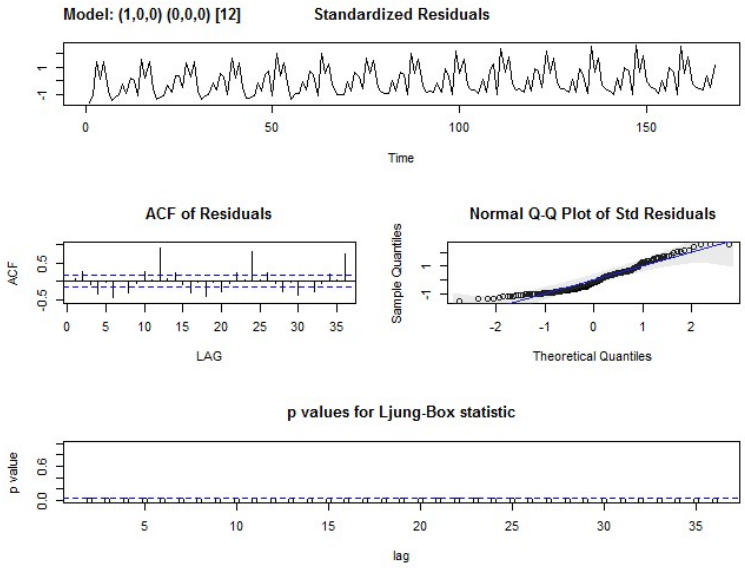
x(t)-x(t-1): Note the spike



Seasonal Differencing x(t)-x(x-12)

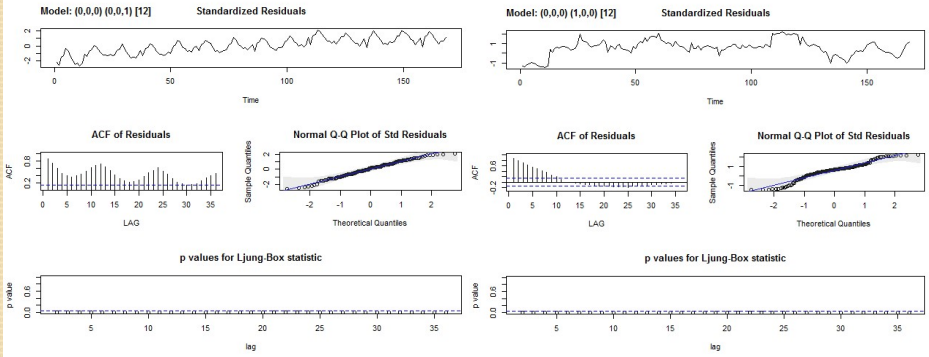


ARIMA(1,0,0) model (no seasonality)

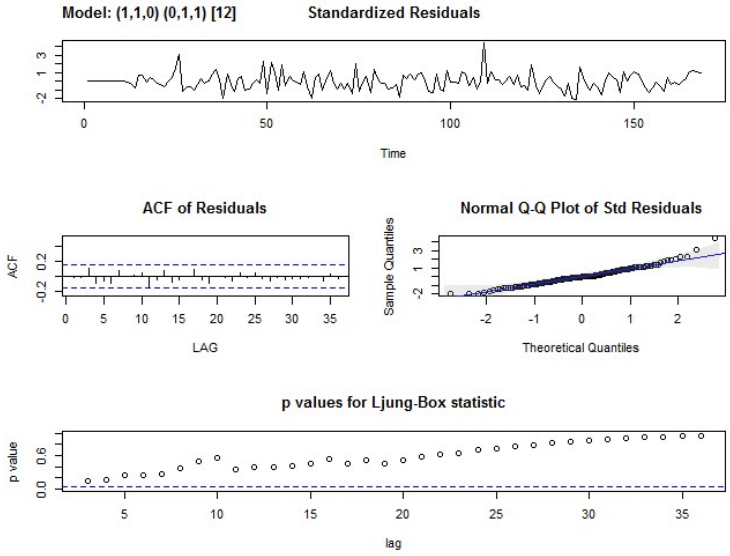


Seasonal (0,0,1) vs. Seasonal (1,0,0)

- Ignore non-seasonal ARIMA
- Try seasonal MA vs. AR model



Reasonable Model
Trend+Seasonal Diff AR(1)+seasonal MA(1)



Variations on ARIMA

- ARMA = AR + MA
  - Assumes no differencing
- ARIMA = AR + Integrating + MA
  - Generalization to allow for integer differencing to remove stationarity
- SARIMA = ARIMA with seasonal factors

How to approach time series

- Look for trends
  - Use differencing x(n)-x(n-1) → WSS model
  - Use seasonal differencing if appropriate
- Look at ACF & PACF
  - Figure out AR & MA components
  - Figure out seasonal AR & MA components
  - Build a model (ACF/PACF pattern) or 'whack'
- Diagnose residuals
  - No significant ACF or PACF *residuals*
  - Ljung-Box test → nothing significant
  - Standard regression diagnostics QQ, etc.

ARCH & ARIMAX

But Wait there's more

- ARIMAX in economics
  - ARIMA model with extra X variables
  - Model your data with context of GDP, oil prices, etc.
- VARIMA and VAR(p)
  - Extension to modeling multiple time series models together
- FARIMA or ARFIMA
  - Instead of ARIMA with integer d, can have a fractional d that converts it into a series

ARCH

- Autoregressive Conditionally Heteroscedastic model
- What does heteroscedastic mean?

ARCH

- Autoregressive Conditionally Heteroscedastic model
- What does heteroscedastic mean?
  - Variance changes
- What does the ARIMA model assume in terms of variance?

ARCH

- $Y(t) = [x(t) - x(t-1)] / x(t-1)$

ARCH(m) model

- $Var(y_t \mid y_{t-1} \ldots y_{t-m}) = \sigma_t^2 + \alpha_0 + \alpha_1 y_{t-1}^2 + \ldots + \alpha_m y_{t-m}^2$
- Characterizing short bursts of variance in the time series data

Finding Cyclic Patterns

Cyclic Patterns & Seasonality

- Many time series have a cyclic seasonal pattern such as consumer products
- How do you detect periodicity?
  - Expert knowledge
  - Eye-balling
  - Plot average value across all months
  - Frequency domain techniques
    - FFT
    - Periodogram

Cyclic Series

- For cyclic series, the period may not always be obvious
- Period T = duration of one cycle
- Frequency $\omega = 1/T$
- Generally assume period constant
- $x_t = A\cos(2\pi\omega t + \phi)$
  - $\omega = 1/T$ = frequency
  - ϕ = phase or where it starts
  - A = amplitude

Spectral Analysis

- Examining the dominant frequencies of the observed time series
- Property:
  - Periodic signals can be decomposed into a weighted sum of sine and cosine functions at different frequencies
  - cos(a±b) = cos(a)cos(b) ∓ sin(a)sin(b)

Suppose x_n, from $n = 0$ to $N - 1$ is a time series (discrete time) with zero mean.

$$x_n = \sum_k A_k \cdot \sin(2\pi\nu_k n + \phi_k)$$

$$= \sum_k \left(\overbrace{a_k}^{A_k \sin(\phi_k)} \cos(2\pi\nu_k n) + \overbrace{b_k}^{A_k \cos(\phi_k)} \sin(2\pi\nu_k n) \right)$$

https://en.wikipedia.org/wiki/Spectral_density_estimation

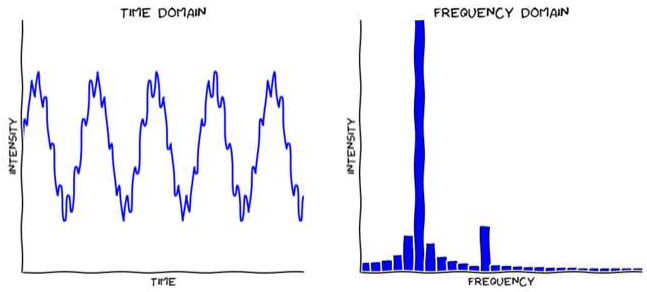
  - ak & bk become regression weights

Fast Fourier Transform (FFT)

- Mapping of periodic signal from time domain into the frequency domain
- FFT is a computationally efficient algorithm used by periodogram
- Typically generate size as 2^u for efficiency
- Longer the filter, the finer the detail

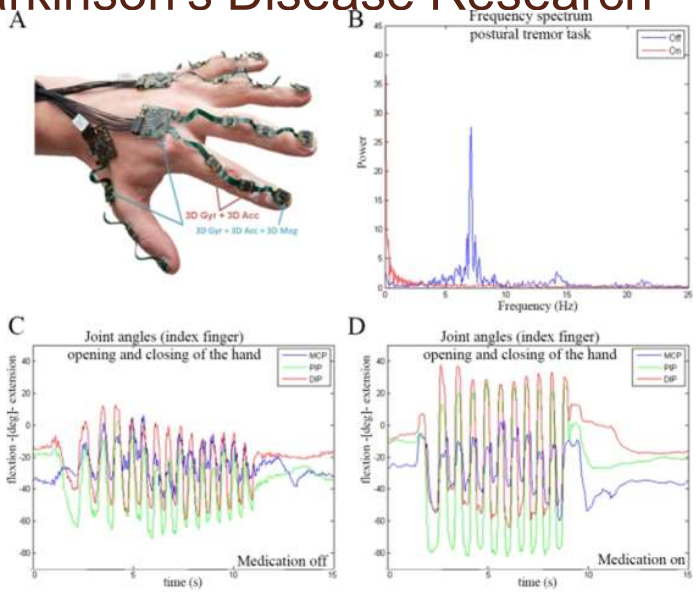
Example of FFT output

- https://learn.adafruit.com/assets/11430

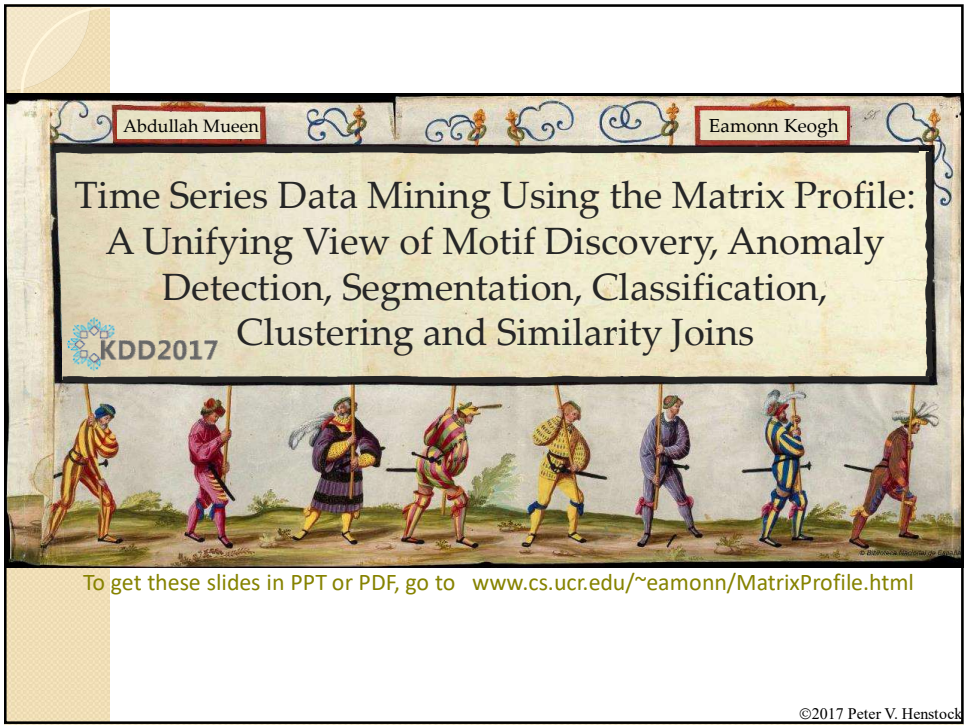


- Use is when you don't know what the period should be
- Can see major and minor effects

Parkinson's Disease Research

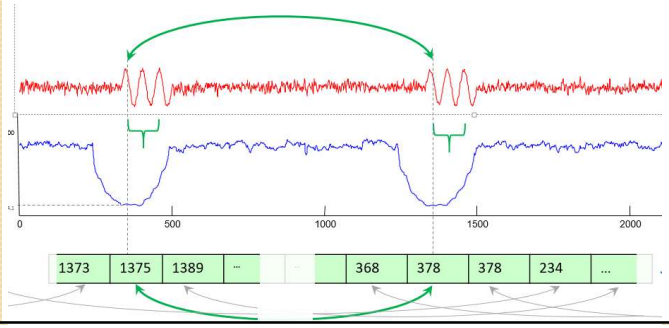


- http://www.mdsabstracts.com/abstract.asp?MeetingID=802&id=113637



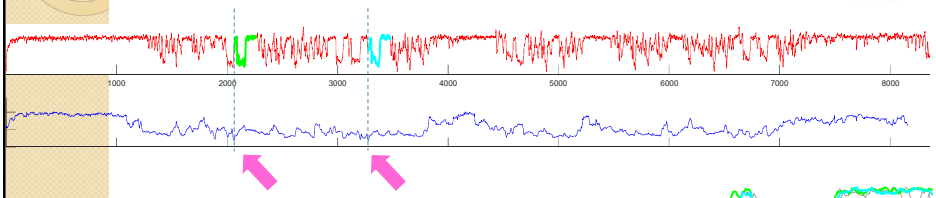
Matrix Profile

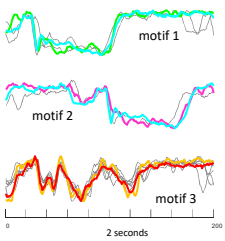
- Choose a window length ~100?
- Slide window through the sequence
- At each position:
  - Find the nearest "match" to sequence
    - Comparison of normalized sequences: $(x-\mu)/\sigma$
  - Record location of the match



Zebra Finch

(Zebra Finch Vocalizations in MFCC, 100 day old male)

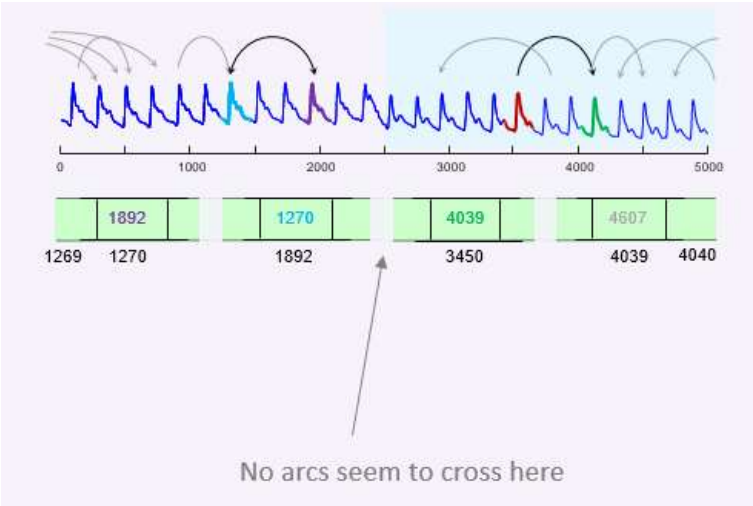




Motif discovery can often surprise you.

While it is clear that this time series is not random, we did not expect the motifs to be so well conserved or repeated so many times. There is evidence of a *vocabulary*, and maybe even a *grammar*…

Change Detection



This New Research is Awesome

- Matching everything against everything
  - Found some faster implementations
  - Takes hours not weeks
- Solution has 1 parameter: window

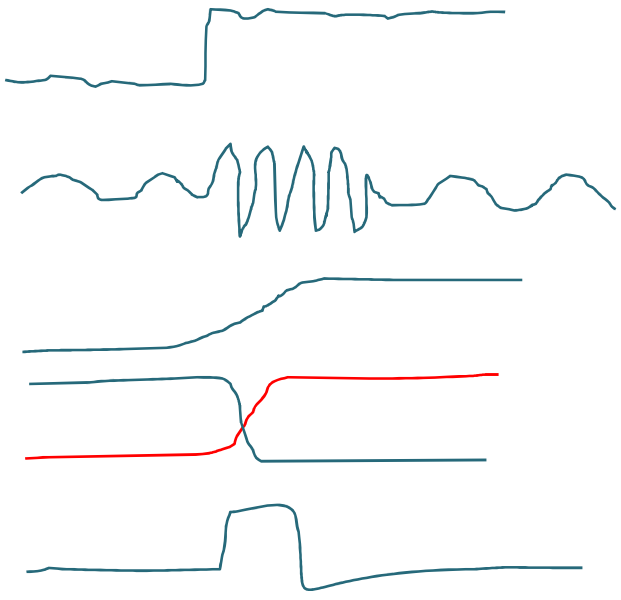
- Applications:
  - Seismology
  - Cardiac events
  - Bird mating calls
  - DNA motifs
  - Music patterns

Change Detection

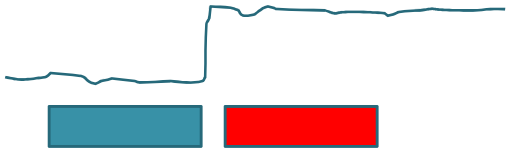
Why change detection?

- Useful information
- May need to switch out training/testing
- Many applications
  - Speaker detection (conversation)
  - Context changes
  - Normal to threat
  - Customization to a new user
  - Accurate models depending on contexts

Types of Changes



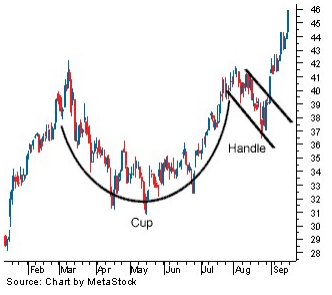
Types of Approaches



- Compute mean, stdev for blue and red
  - Check for significant differences
- Fuzzy or statistical modeling of “steady”
- Train decision trees on last K samples
  - If tree starts to rebalance, data changed

Other Problems

- Finding motifs = common patterns
  - Walking vs. Stirring vs. Standing
  - Cup-Handle motif
- Anomaly detection
- Clustering
- Segmentation



http://www.investopedia.com/university/charts/charts3.asp