MATH 7339 - Machine Learning and Statistical Learning Theory 2

Section Exploratory Data Analysis

1. Detrend
2. Difference operator
3. Frequency
4. Smoothing

Motivation:

In time series analysis, we need to account for the **dependence** between the values in the series. We frequently would prefer to analyze a **stationary** process.

Stationarity for a time series enables us to measure the dependence, since the dependence structure is regular and does not change over time. This allows us to better estimate autocorrelation and other quantities of interest.

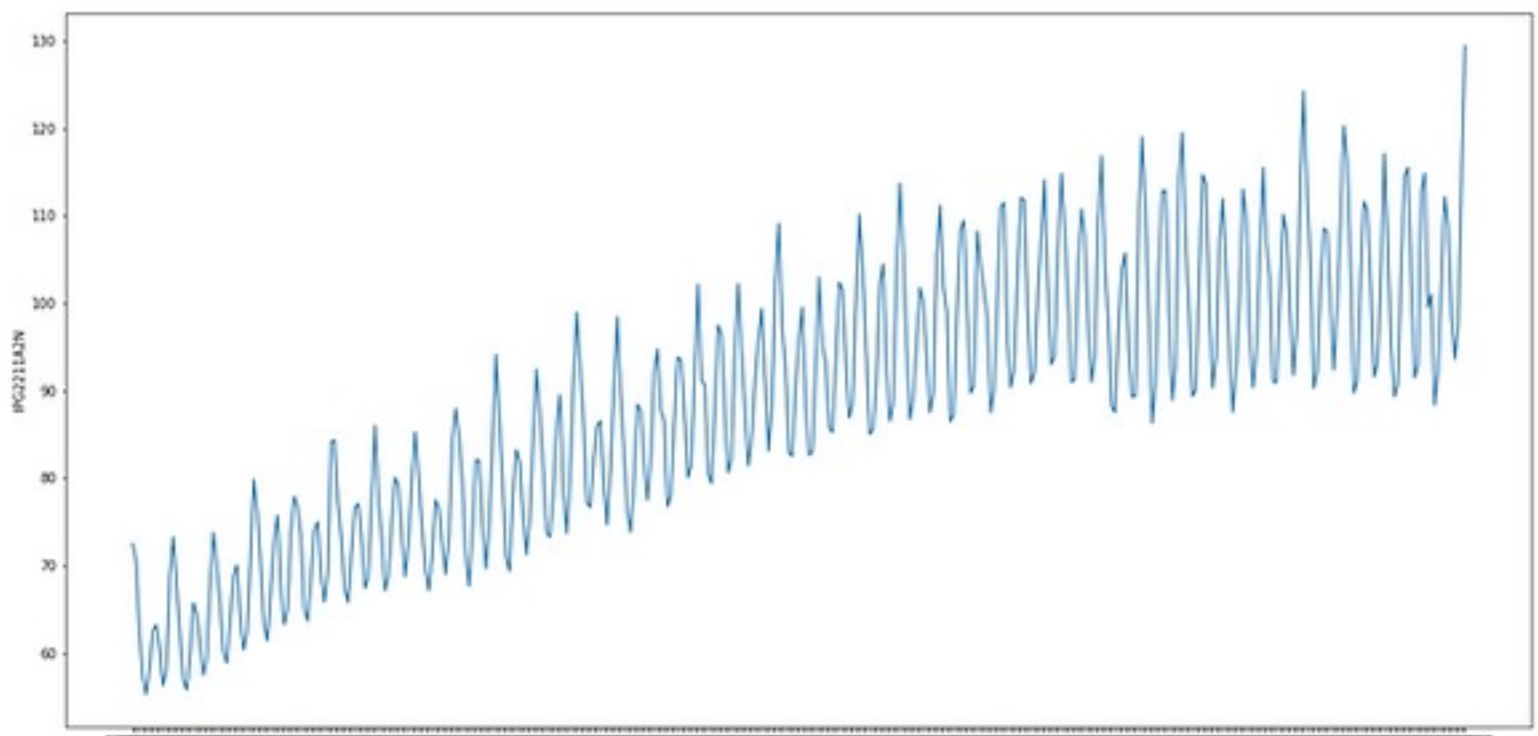
In addition, ARMA processes provide a rich framework for analyzing stationary processes.

A strong **trend**, however, may **obscure** the behavior of the stationary process. It may, therefore, be necessary to **remove** a **trend**; one way to do that is via **regression**.

For example,

$$x_t = \mu_t + y_t$$

where y_t is a zero mean stationary process, e.g. MA(2), AR(1), white noise, etc., and μ_t is a deterministic trend, e.g. $\mu_{\mathrm{t}} = \beta_0 + \beta_1 t$.



Linear Regression Basics

The basic data type for regression consists of a list of pairs of numbers, $(x_1, \overrightarrow{z_1}), \ldots . (x_n, \overrightarrow{z_n})$, where the x_i are thought of as the **response** variables and z_i are thought of as the **predictor** variables.

The **linear regression model** would then be

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_{tq} z_{tq} + w_t$$

$$= \vec{\beta}^T \overrightarrow{z_t} + w_t$$

$$\overrightarrow{z_t} = \begin{bmatrix} 1 \\ z_{t1} \\ z_{t2} \\ \vdots \\ z_{tq} \end{bmatrix}$$

Here, w_t are iid normal random noise with mean zero and variance σ^2.

Estimating the parameter vector $\vec{\beta}$ is done by **minimizing** the sum of squares **error**

$$L = \sum_{t=1}^{n} \left(x_t - \vec{\beta}^T \overrightarrow{z_t}\right)^2$$

Solution is the ordinary least squares (OLS) estimator

$$\hat{\vec{\beta}} = (Z^T Z)^{-1} Z^T \vec{x}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Denote the *minimized error sum of squares*

$$SSE := \sum_{t=1}^{n} \left(x_t - \hat{\vec{\beta}}^T \vec{z_t} \right)^2$$

An unbiased estimator for the variance σ^2

$$s^2 = MSE = \frac{SSE}{n - (q+1)}$$

Fitted values $\widehat{x_t} := \hat{\vec{\beta}}^T \vec{z_t}$

Residuals: $e_t = x_t - \widehat{x_t}$

Inference

Assuming independent Gaussian errors, we can build **confidence intervals** using statistics such as

$$\frac{\hat{\beta}_i - \beta_i}{standard\ error\ (\hat{\beta}_i)}$$

which have a t-distribution with $n-(q+1)$ d.f, and s_w^2 is distributed proportionally to a $\chi^2_{n-(q+1)}$

Model Selection:

Subset selection using AIC, BIC,

Assumptions for Linear Regression

The assumptions for linear regression are:

- There is a **linear** relationship between the response and predictor variables.
- There is a random noise w_i
- $E(w_i) = 0$.
- $Var(w_i) = \sigma^2$ is constant and finite.
- w_i are **iid** normal.

Diagnostics of linear assumptions:

Check regression assumptions are satisfied:

- **Residual plot**: to check if right regression equation used, variance of errors is constant, mean of errors is zero.
- **ACF plot**: to determine correlation.
- **Normal probability plot**: to check for normality. (QQ plot)

❑ Detrending

If our process has a **linear trend**, we could use linear regression to remove the trend ("**detrend**").

Consider the model:

$$x_t = \mu_t + y_t$$

where y_t is a zero mean stationary process, e.g. MA(2), AR(1), white noise, etc., and μ_t is a deterministic trend, e.g. $\mu_\mathrm{t} = \beta_0 + \beta_1 t$.

We can view x_t as having stationary behavior around a trend.
A strong trend, μ_t, can obscure the behavior of the stationary process, y_t .

Remove the trend:

1. Obtain an estimate of the trend component, $\hat{\mu}_t$, e.g. via OLS.
2. Work with the residuals $e_t = x_t - \hat{\mu}_t$

❑ Differencing

The **first difference** of x_t is

$$\nabla x_t := x_t - x_{t-1}$$

Using backshift operator, $\nabla = 1 - \mathrm{B}$

In general, **the $\boldsymbol{d}$-th difference operator** is $\nabla^{\mathrm{d}} = (1 - B)^d$

For example, if $x_t = \beta_0 + \beta_1 t + y_t$

$$\nabla x_t = \beta_1 + y_t - y_{t-1}$$

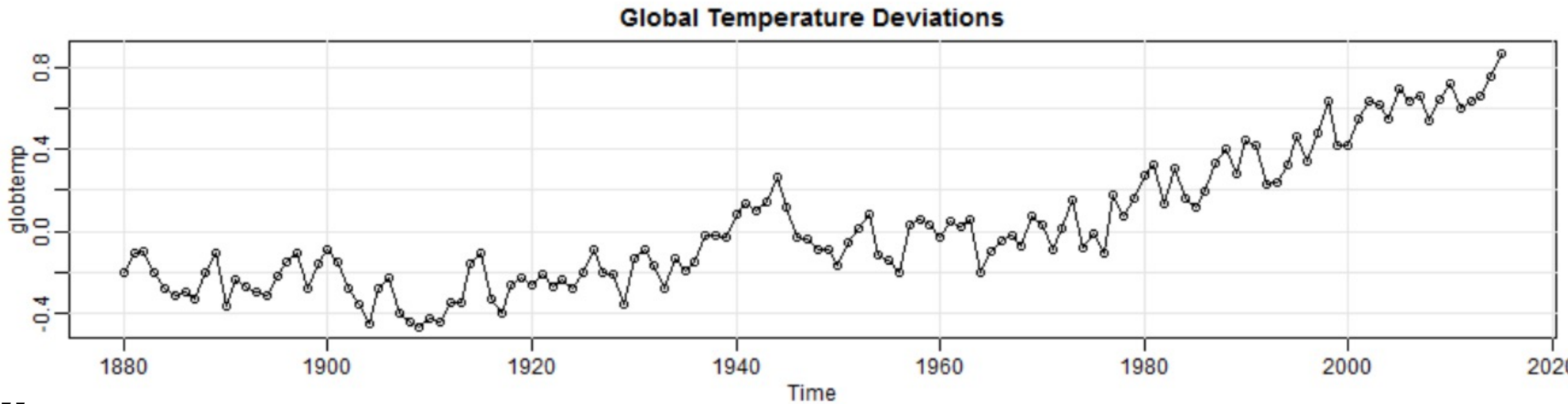
The detrending may give us a more accurate representation, whereas differencing completely removes β_0 and turns β_1 in to the mean of the series $\{\nabla x_t\}$.

In addition, second difference eliminates a quadratic trend.

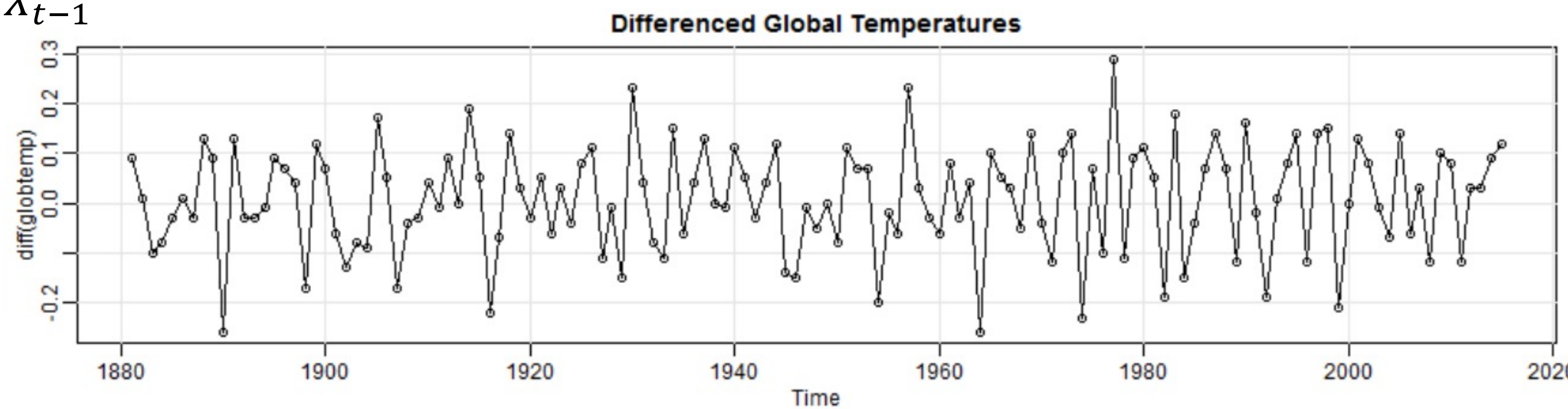
Random Walk Trend

Not stationary, but differenced data are stationary

X_t



$X_t - X_{t-1}$



Differencing Vs Detrending

- An advantage of differencing over detrending is that fewer parameters are estimated after the differencing operation.
- A disadvantage of differencing is that it often makes an estimate of the stationary process y_t more difficult.

Differencing changes y_t and often introduces additional dependency.

For example, consider the MA(1) process $y_t = w_t + \theta_1 w_{t-1}$

Suppose $x_t = \beta_0 + \beta_1 t + w_t + \theta_1 w_{t-1}$

$$\begin{aligned}\nabla x_t &= \beta_1 + y_t - y_{t-1} \\ &= \beta_1 + w_t + \theta_1 w_{t-1} - w_{t-1} - \theta_1 w_{t-2} \\ &= \beta_1 + MA(2)\end{aligned}$$

∇x_t is stationary.

❑ Frequency and Periodic Functions

We've already seen how we can use differencing to obtain stationary processes. We are assuming that our observations can be written in the form

$$x_t = \mu_t + y_t$$

where y_t is a zero mean stationary process, and μ_t is a **trend**.

We have considered that μ_t to be a linear or polynomial functions.

Now, let us consider μ_t as a **periodic function.** For example,

$$\mu_t = A\cos(2\pi\omega t + \phi)$$

where

- A: amplitude
- ω: frequency
- $\frac{1}{\omega}$: Period
- ϕ: phase

Example:

Assume that y_t in model is white noise.

$$x_t = \mu_t + y_t$$

$$= A\cos(2\pi\omega t + \phi) + w_t$$

We could try to use non-linear least squares to fit A, ω, ϕ

In many settings, certain frequencies are natural.

For example, in monthly data a frequency ω=1/12 (corresponding to a period of 12) is quite natural. We may want to remove a periodic signal by fitting

$$x_t = \beta_1 \cos\left(\frac{2\pi}{12}t\right) + \beta_2 \sin\left(\frac{2\pi}{12}t\right) + w_t$$

Here, we have rewrite the model using $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$

We can use OLS estimate $\vec{\beta}$

➢ Smoothing

Sometimes, the time series data we have can be too noisy to be able to detect long term trends. Smoothing is used to smooth out short term random fluctuations so that longer term trends can be emphasized.

Assume the models of the form

$$x_t = \mu_t + y_t$$

where y_t is a zero mean stationary process, and μ_t is a trend or frequency.

One way to approximate μ_t is to take a moving average of the time series. Averaging, in general, reduces variability. It can also reduce "seasonal" fluctuations. Averaging can help in viewing longer term trends, because the seasonal variations will be dampened.

In general we may write a **moving average** as

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

where $a_j \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$

It is also called *centered moving average.* The smoothed value for a particular time is calculated as a linear combination of observations for surrounding times.

Averaging has the advantage of being adaptable to slow changes in μ_t across time. The disadvantage is that there may still be a substantial amount of variability in our estimate μ_t , and we may not know a priori what the **window size** k should be.

Question: What is an appropriate window size, k, to **smooth away** seasonality in **monthly** data, in order to identify **yearly** trends?

Variance Reduction with Averaging

It was mentioned earlier that averaging reduces variation, in general.

For example, assume that the original series x_t is stationary, such that $Var(x_t) = \sigma^2$. Let's create another time series

$$y_t = \frac{1}{3}x_{t-1} + \frac{1}{3}x_t + \frac{1}{3}x_{t+1}$$

Question: Derive the variance of y_t .

Kernel Smoothing

The idea with kernel smoothing is similar to the moving average; however, the contribution to the estimate of the smooth function at a point t from local points declines as a function of distance from the current point. The smooth function is estimated by

$$\hat{\mu}_t = \sum_{i=1}^{n} w_t(i)\, x_i$$

where $$w_t(i) = \frac{K\left(\frac{t-i}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{t-j}{b}\right)}$$

Here, $K(\,)$ is the **kernel function**, and b is the **bandwidth**.

For example, $K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$

Smoothing Splines

Textbook[Shumway-Stoffer]: Chapter 2