Correlation Coefficient Measures

Revanth Kota

2018-04-21

```r
# Libraries used to check relation between two variables before we start our analysis:
library(tidyverse) # To deal with tidy structured data.
library(acepack) # Library to check maximal correlation
library(minerva) # Library to check Maximal Information Coefficient.
```

In this presentation I would like to talk about various correlation coefficients used to quantify associations or relations between various variables.

Correlation in simple terms is "How much can you explain or predict behaviour of one variable based on another variable"

Looking for correlations among variables in data and quantifying them is important in any analysis as it helps us in identifying important variables and improve prediction accuracy without loss of information

In time series analysis as successive observations are not iid's and present values depend on past values, we should look for various correlation coefficients to identify and quantify association between lags and current values as successive observations carry mutual information.

In this presentation we see will see:

a. Pearson's correlation coefficient

b. Spearman's Correlation coefficint

c. Maximal Correlation

d. Mutual Information Coefficient's to quantify association between variable and how they are used in time series analysis:

Defining the above correlation coefficients:

Pearson's correlation: Pearson's correlation gives us the measure and direction of linear association between two variables. It values varies from -1 to +1. They do excellent job in identifying linear relation relationships between variables but fail in identifying non-linear association between two variables. Pearson's correlation does a great job in identifying and quantifying information transfer across lag in linear time series processes like : ARMA or ARIMA.

Spearman's correlation: In spearman's correlation we rank observations of both the variables in a descending order to calculate the difference in ranks and use the calculated difference in ranks to calculate the quantity of association between variables.

Maximal Correlation: Maximal correlation is an optimization problem that is trying to search for transformations of X and Y such hat Pearson's correlation between transformed X and Y is maximized. It is robust to noise unlike Maximal Information Coefficient.

Maximal Information Coefficient: Maximal Information Coefficient (MIC) is one of the most important measures of Independence.

MIC takes value between zero and one, and it has two main properties: Generality and equitability.

Generality means that with sufficiently large sam- ple size, the statistic should capture a wide range of asso- ciation such as linear, exponential, or periodic.

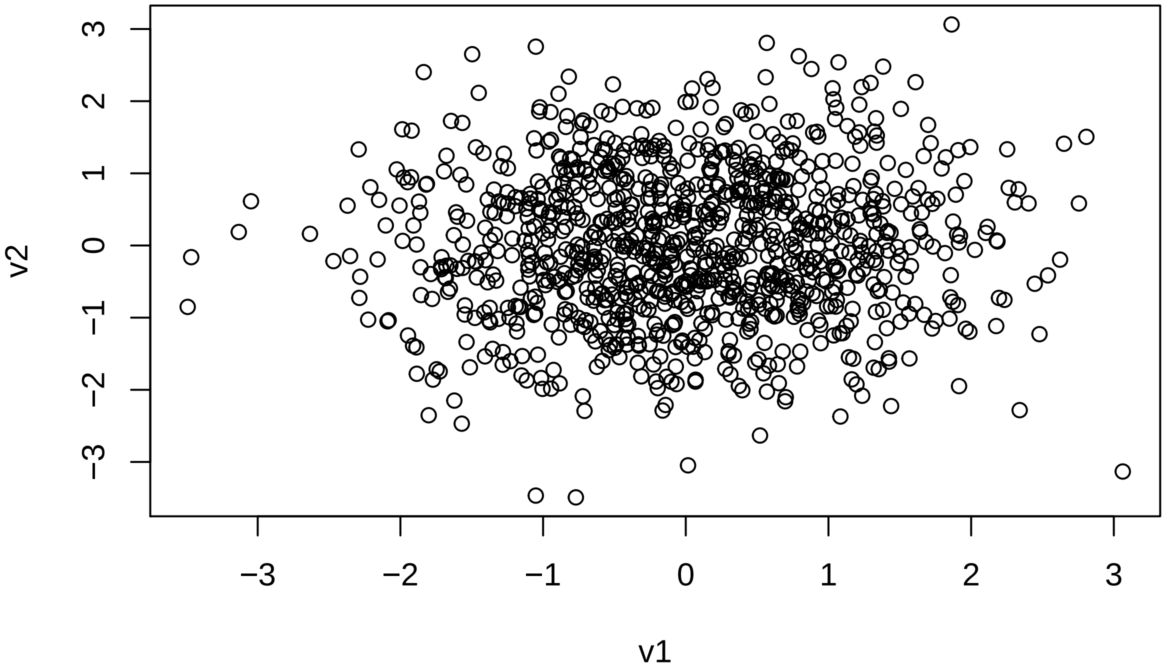
Equitability means that MIC gives similar scores to equally noisy rela- tionships regardless the type of relationships.

In addition, with probability approaching 1 as sample size grows, MIC gives scores of one to all noiseless functional relationships and gives scores that tend to 0 to statistically independent variables.

An advantage of MIC is the ability to catch non- linear associations as well as linear associations

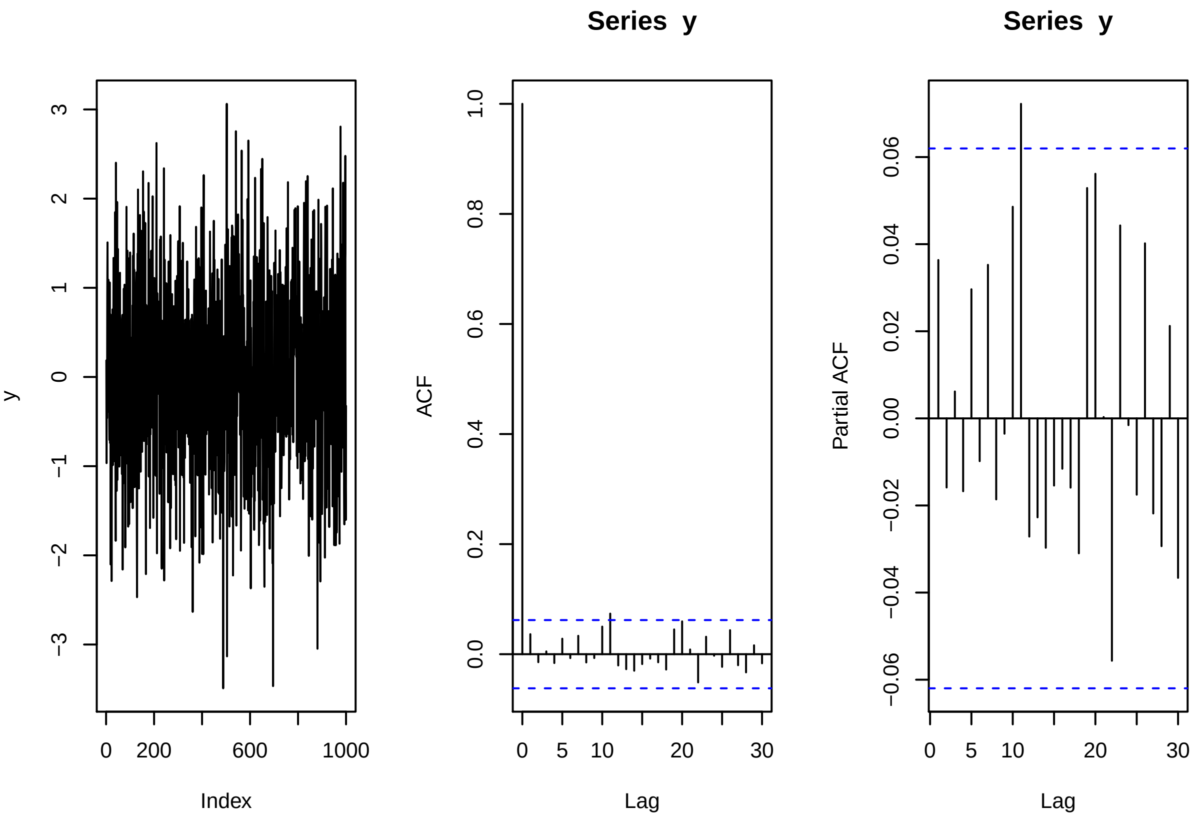
```r
y <- rnorm(1000) # iid's from normal distribution with 0 mean and variance = 1.
v1 <- y[1:999]
v2 <- y[2:1000]
```

```r
plot(v1,v2)
```



```r
# As observations are scattered around in scatterplot, we can say that succesive
# observations are iid's
```

```r
par(mfrow = c(1,3))
plot(y, type = "l")
acf(y)
pacf(y)
```



```r
cor(v1,v2) #pearson's correlation coefficient
```

```
## [1] 0.03636319
```

```r
cor(v1,v2, method = "kendall") # kendall's correlation coefficient
```

```
## [1] 0.02242924
```

```r
cor(v1,v2, method = "spearman") # spearman's correlation coefficient.
```

```
## [1] 0.03386122
```

```r
minerva::mine(v1,v2) # maximal information correlation
```

```
## $MIC
## [1] 0.1217723
##
## $MAS
## [1] 0.005787098
##
## $MEV
## [1] 0.1217723
##
## $MCN
## [1] 3
##
## $`MIC-R2`
## [1] 0.1204501
##
## $GMIC
## [1] 0.05267523
##
## $TIC
## [1] 6.900017
```

```
argmax = ace(v1,v2)
cor(argmax$tx, argmax$ty) # maximal corrrelation
```

```
##             [,1]
## [1,] 0.0437096
```

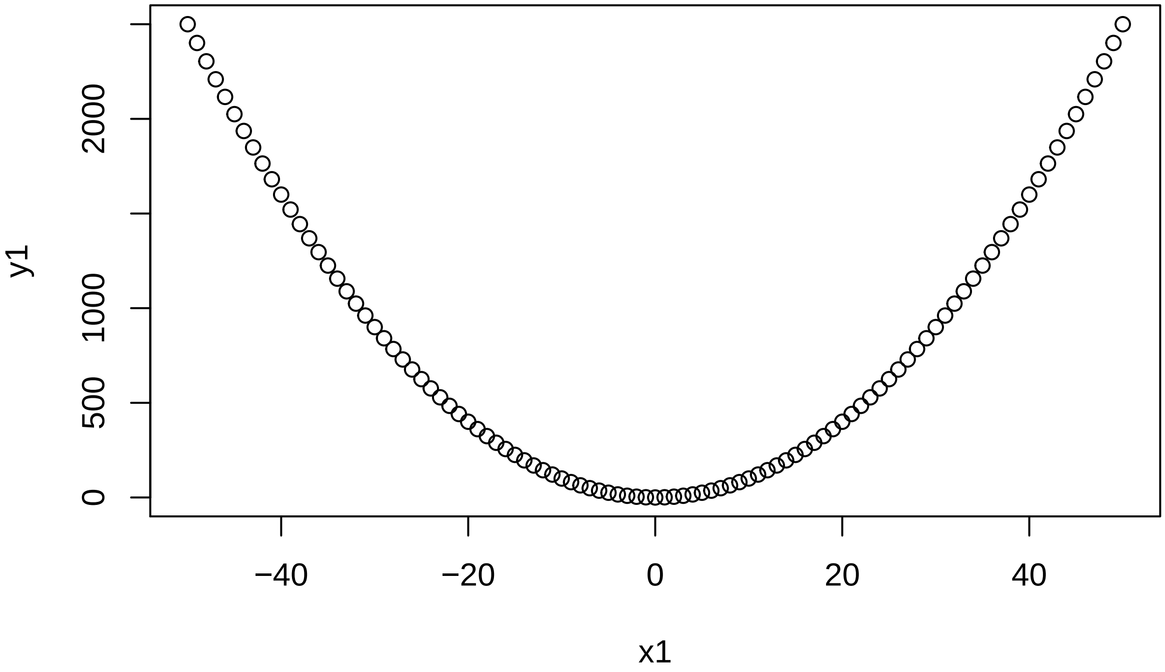
As all correlation measures are almost near to zero, we can say that all observation are iid's

Non linearity correlation measures:

Correlation Measures for parabolic relation between x and Y

```
x1 <- (-50:50)
y1 <- x1*x1 #Parabolic data generation.
```

```
plot(x1,y1) # Here the function is symmertric about Y axis.
```



```
cor(x1,y1) # Linear correlation comes to Zero using pearson's linear correlation measure
```

```
## [1] 0
```

```
 cor(x1,y1, method = "spearman")
```

```
## [1] 0
```

```
cor(x1,y1, method = "kendall")
```

```
## [1] 0
```

```
# even spearman's and kendall's correlation could not detect
# the non linear association between x and y as their correlation scores are zero
```

```
minerva::mine(x1,y1)
```

```
## $MIC
## [1] 0.9999293
##
## $MAS
## [1] 0.6876475
##
## $MEV
## [1] 0.9999293
##
## $MCN
## [1] 2.584963
##
## $`MIC-R2`
## [1] 0.9999293
##
## $GMIC
## [1] 0.8789617
##
## $TIC
## [1] 10.51554
```

```
argmax = ace(x1,y1)
cor(argmax$tx, argmax$ty)
```

```
##           [,1]
## [1,] 0.9999999
```

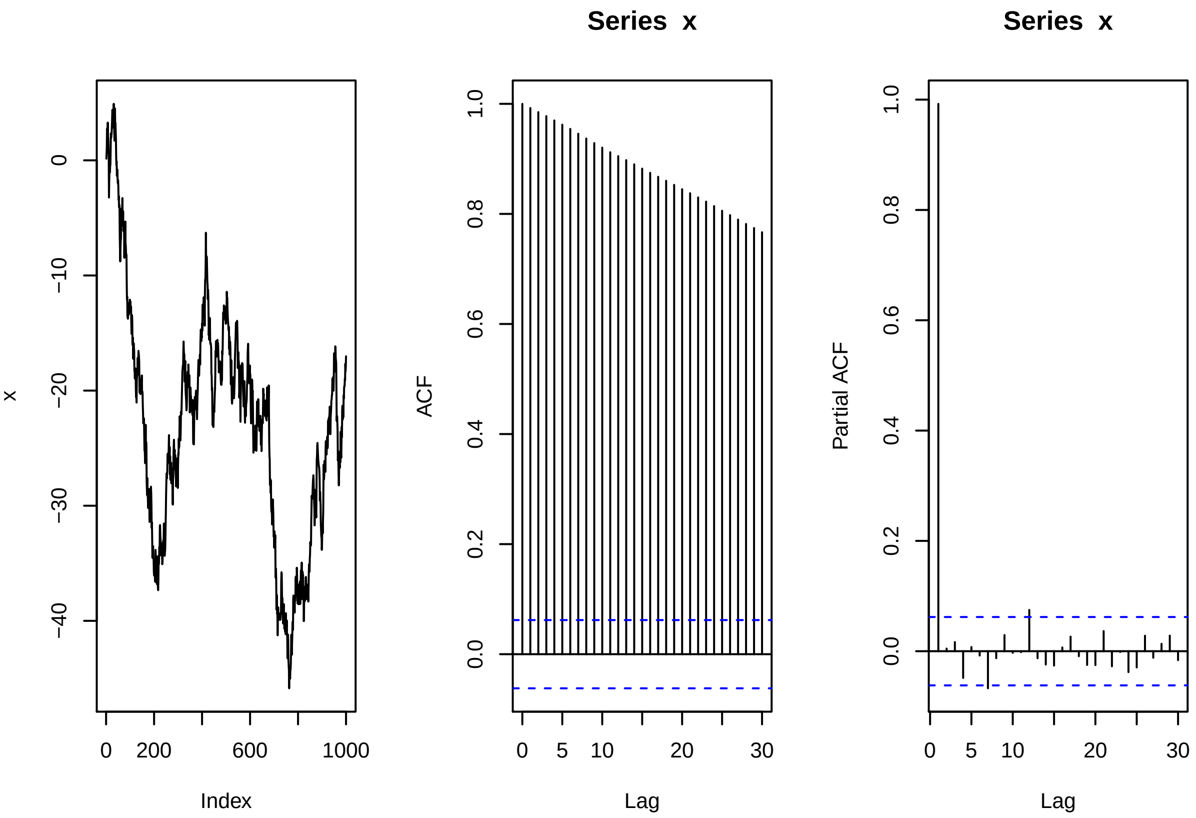
```
# Here, unlike pearson's, spearman's and kendall's correlation, we can see that
# Both maximal information coeffient and maximal correlation could detect and measure the
# Non linear association between x and y values as their correlation measures are 0.9999 respectively.
```

Random Walk Data generation

```
x <- rnorm(1)
for (i in 2:1000) {
  x[i] <- x[i-1] +rnorm(1)
}
```

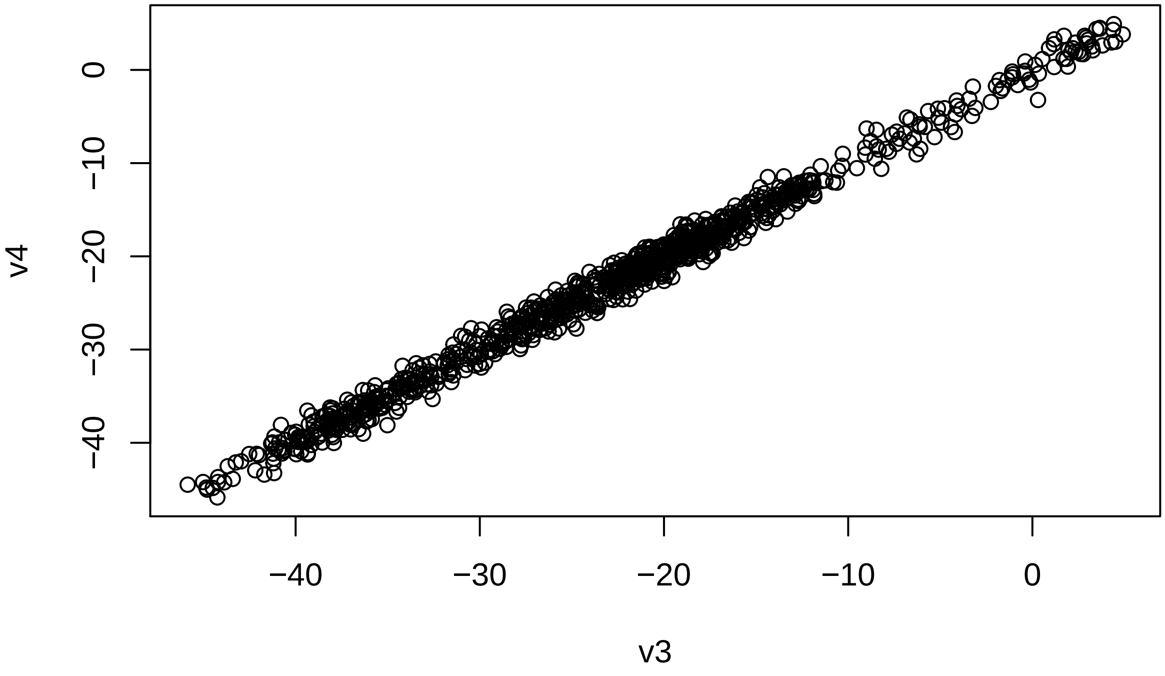
Random Noise Process generation and properties:

```
par(mfrow = c(1,3))
plot(x, type = "l")
acf(x)
pacf(x)
```



```
v3 <- x[1:998]
v4 <- x[2:999]
```

```
plot(v3,v4)
```



```
# As, we know that random walk is an AR(1) Process, That is, current value depends on past lags,
# we can see that there is a strong correlation between past lag value and current values of y
# And this can be identified below.
```

Correlation measures between successive lags using various correlation coefficients:

```
cor(v3,v4) # Pearson's correlation coefficient
```

```
## [1] 0.9950181
```

```
cor(v3,v4, method = "spearman") # Spearman's correlation coefficient
```

```
## [1] 0.9919197
```

```
cor(v3,v4,method = "kendall") # Kendall's Tau correlation coefficient
```

```
## [1] 0.926517
```

```
minerva::mine(v3,v4)
```

```
## $MIC
## [1] 0.9472033
##
## $MAS
## [1] 0.02473052
##
## $MEV
## [1] 0.9472033
##
## $MCN
## [1] 4.70044
##
## $`MIC-R2`
## [1] -0.04285767
##
## $GMIC
## [1] 0.9135525
##
## $TIC
## [1] 128.0989
```

```
argmax = ace(v3,v4)
cor(argmax$tx, argmax$ty)
```

```
##           [,1]
## [1,] 0.9950356
```

based on all the above correlation measures

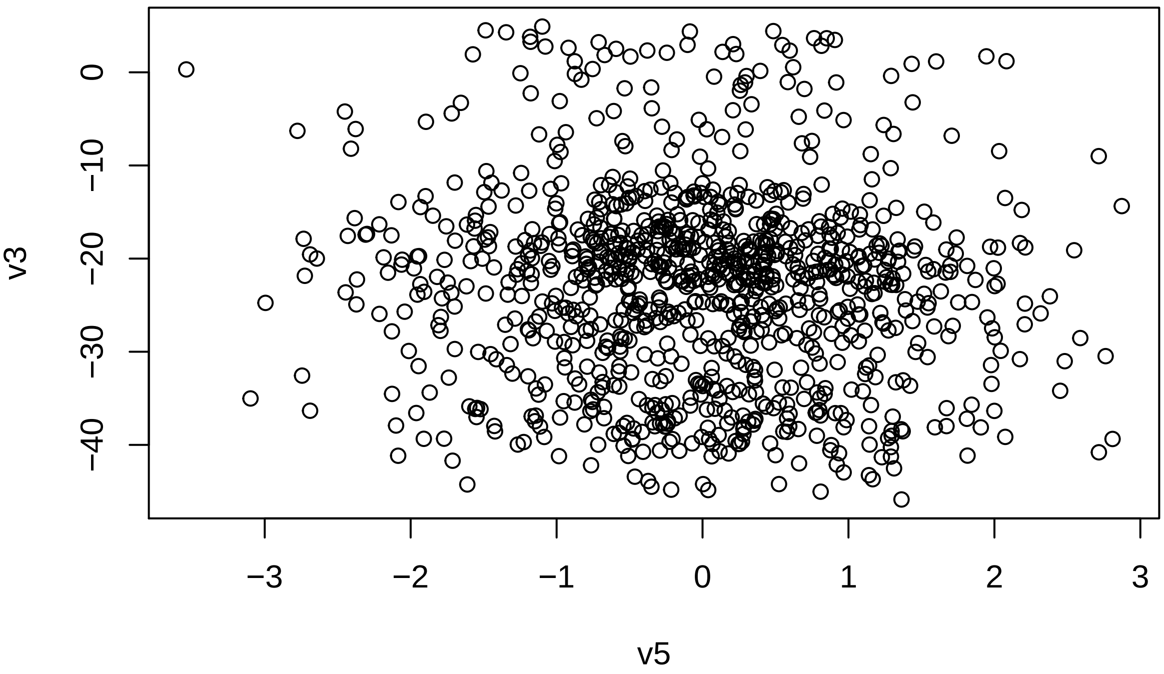
we can say that lag1 has high correlation with current value and

current value can be explained using past values.

Correlation coefficients calculated after substracting lag1 values from current values in random walk data generation process:

```
v5 <- (v4-v3) # substracting lag1 we are left with noise.
```

```
plot(v5,v3)
```



```
# After removing first lag y(t-1) from current values y(t),
# we can see that the residuals are iid's
```

```
cor(v3,v5)
```

```
## [1] -0.07356573
```

```
cor(v3,v5, method = "spearman")
```

```
## [1] -0.07750572
```

```
cor(v3,v5,method = "kendall")
```

```
## [1] -0.05176451
```

```
minerva::mine(v3,v5)
```

```
## $MIC
## [1] 0.1161012
##
## $MAS
## [1] 0.004228412
```

```
##
## $MEV
## [1] 0.1161012
##
## $MCN
## [1] 3
##
## $`MIC-R2`
## [1] 0.1106892
##
## $GMIC
## [1] 0.05401321
##
## $TIC
## [1] 6.555386
```

```
argmax = ace(v3,v5)
cor(argmax$tx, argmax$ty)
```

```
##            [,1]
## [1,] 0.08486721
```

Based on the above correlation coefficient values

Before and after substracting lag1 values from current values,

we can say that the above synthetic data is an AR process of first order

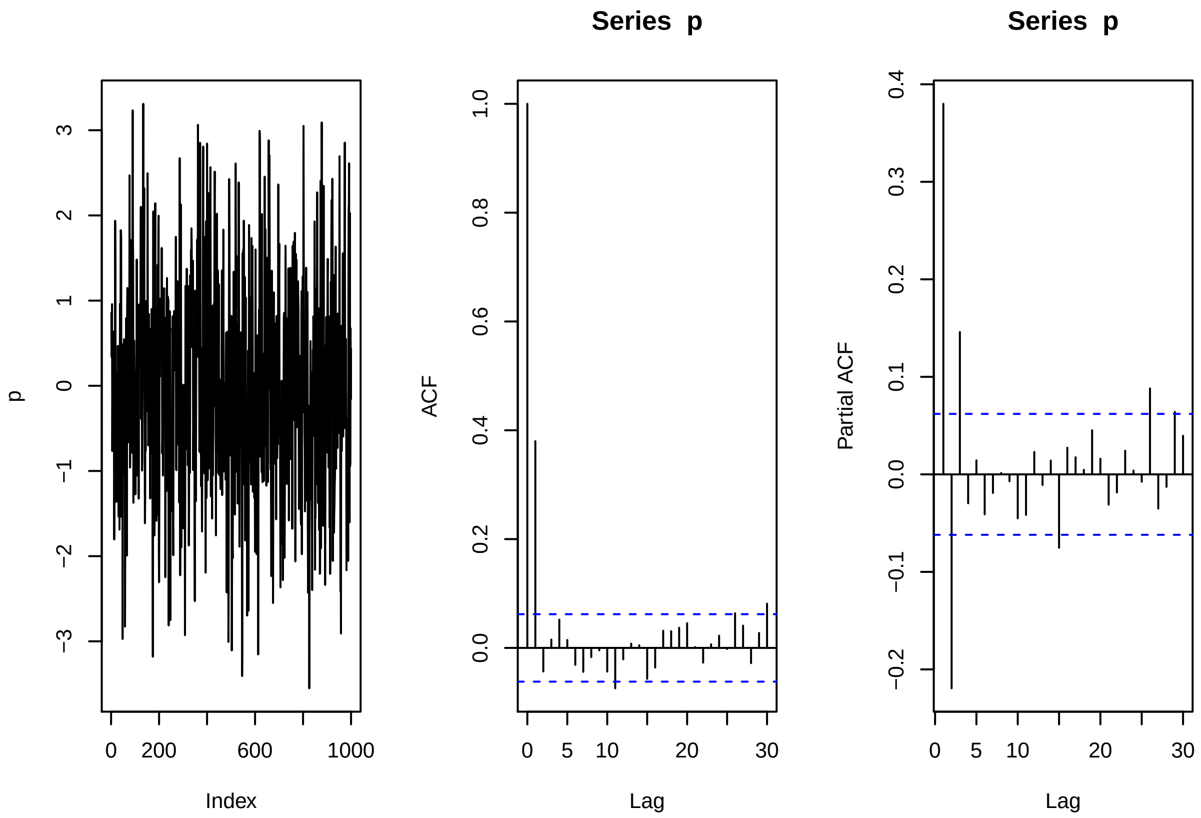
and it is Random walk based on ACF and PACF plots.

MA(1) Process generation:

```
e <- rnorm(1000)
p <- e[1]
for (i in 2:1000) {
p[i] <- e[i] + 0.5*e[i-1]
}
```

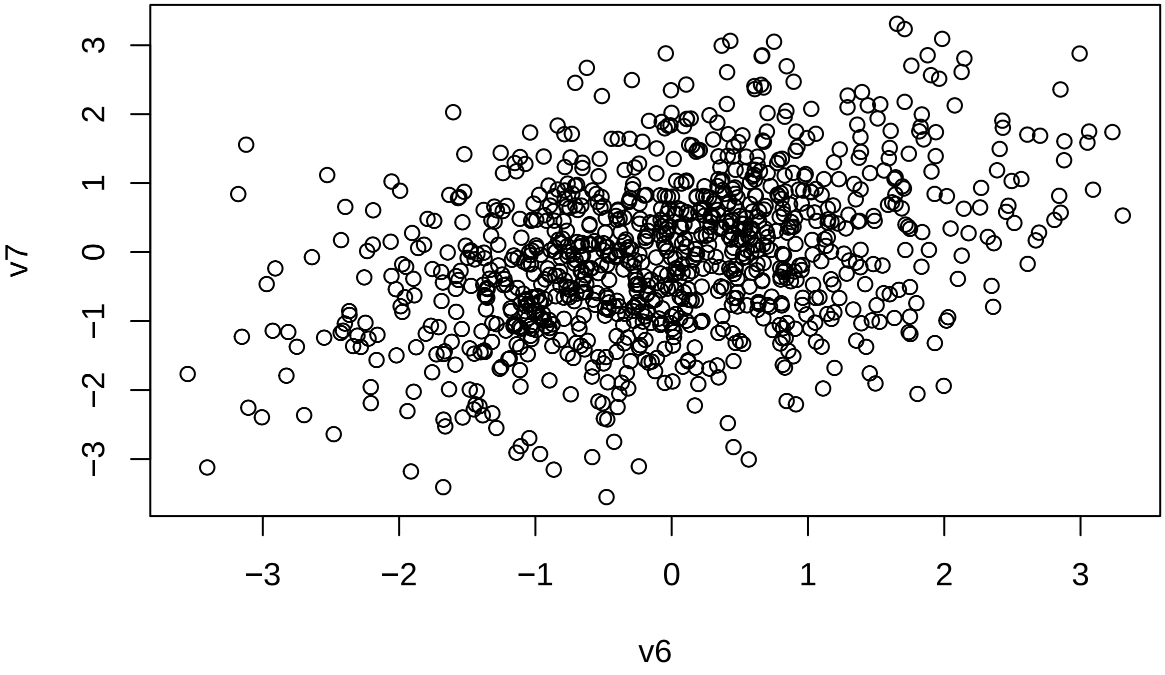
MA(1) Process generation and properties:

```
par(mfrow = c(1,3))
plot(p, type = "l")
acf(p)
pacf(p)
```



```r
v6 <- p[1:999]
v7 <- p[2:1000]
```

```r
plot(v6, v7)
```



###
As error is added there is high density at the middle and ### from the plot we can see that there is randomness.

```r
cor(v6,v7)
```

```
## [1] 0.3800958
```

```r
cor(v6,v7, method = "spearman")
```

```
## [1] 0.362241
```

```r
cor(v6,v7, method = "kendall")
```

```
## [1] 0.248162
```

```r
minerva::mine(v6,v7)
```

```
## $MIC
## [1] 0.215981
## 
## $MAS
## [1] 0.0162463
## 
## $MEV
## [1] 0.215981
## 
## $MCN
## [1] 2
## 
## $`MIC-R2`
## [1] 0.07150825
## 
## $GMIC
## [1] 0.1584332
## 
## $TIC
## [1] 16.04425
```

```r
argmax = ace(v6,v7)
cor(argmax$tx, argmax$ty)
```

```
##           [,1]
## [1,] 0.3810668
```

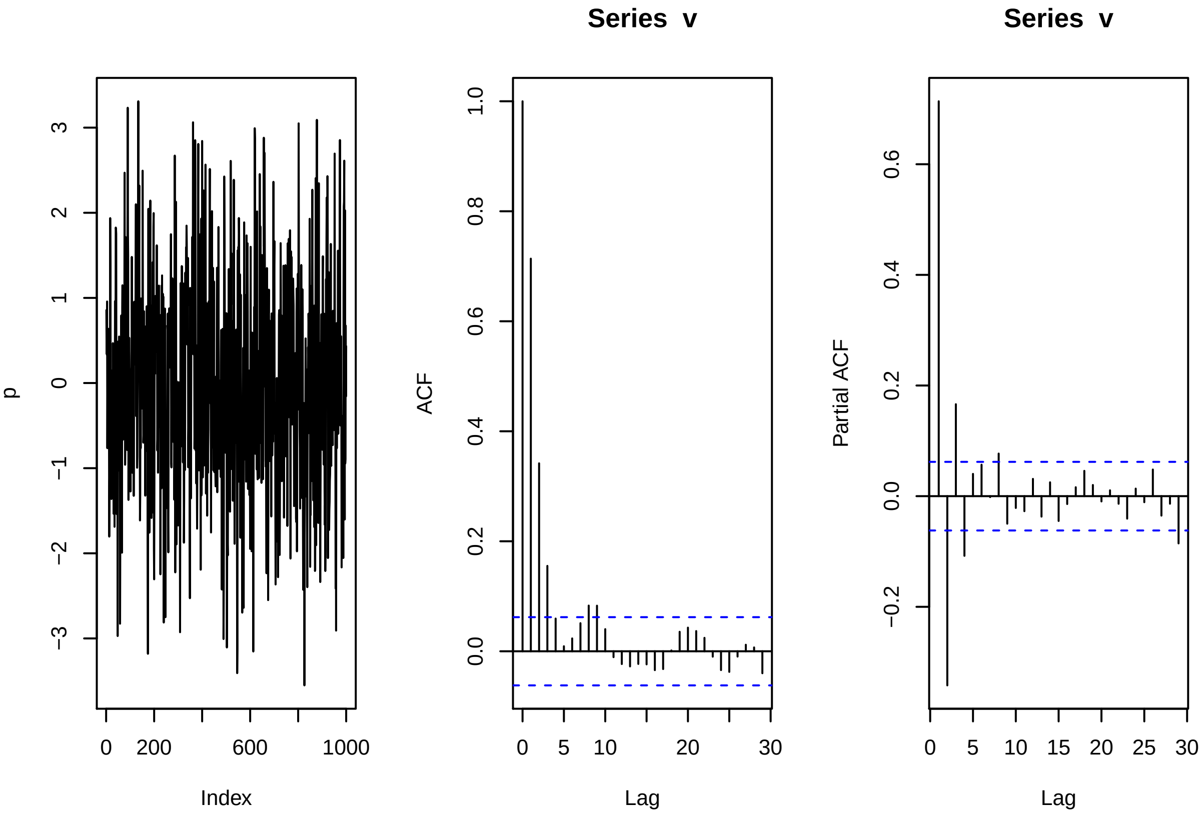
Based on the above values, we can say that as error is propagating along the data, correlation measures between succesive observations is not strong.

ARMA(1) Process data generation

```r
v <- rnorm(1)
er <- rnorm(1000)
for (i in 2:999) {
v[i] = 0.5*v[i-1] + er[i] + 0.5*er[i-1]
}
```

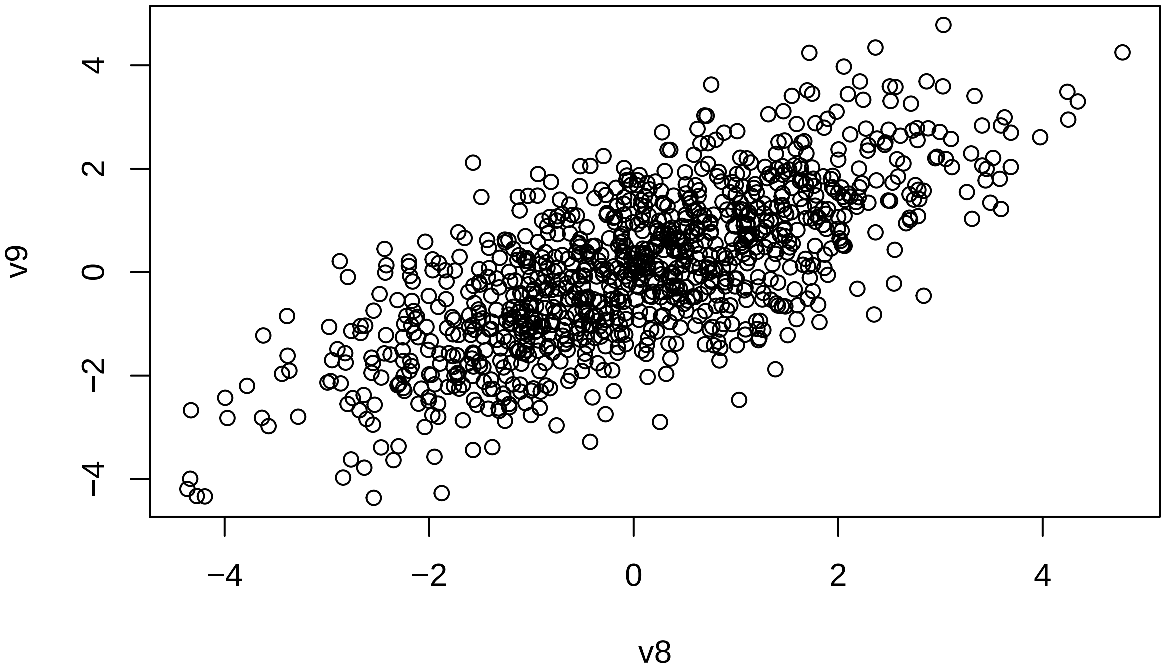
ARMA(1) Process and properties:

```
par(mfrow = c(1,3))
plot(p, type = "l")
acf(v)
pacf(v)
```



```
v8 <- v[1:998]
v9 <- v[2:999]
```

```
plot(v8, v9)
```



###
Here, we can see that there is a strong correlation between successive observations.

```
cor(v8,v9)
```

```
## [1] 0.7143257
```

```
cor(v8, v9, method = "spearman")
```

```
## [1] 0.694354
```

```
cor(v8, v9, method = "kendall")
```

```
## [1] 0.505812
```

```
minerva::mine(v8, v9)
```

```
## $MIC
## [1] 0.4046567
##
## $MAS
## [1] 0.02536183
##
## $MEV
## [1] 0.4046567
##
## $MCN
## [1] 2
##
## $`MIC-R2`
## [1] -0.1056044
##
## $GMIC
## [1] 0.352549
##
## $TIC
## [1] 41.90198
```

```
argmax = ace(v8,v9)
cor(argmax$tx, argmax$ty)
```

```
##           [,1]
## [1,] 0.7146145
```

Here as the underlying data generation process is ARMA(1) process with an error propagating along we can see that the correlation between successive observations is not that strong unlike AR(1).

Conclusions:

a. Unlike the above mentions scenarios where we know underlying data geneartion processes, in real world as we do not know the underlying data generation process it is wise to plot a scatter plot to see how data is distributed? That is, we will see if there is any pattern in the data spread or the spread is random to apply suitable association measures to quantify the association between variables.

b. If you see a linear or monotonic trends (like y = xˆ3 or y = eˆ(x)) then applying Spearman's or Pearson's correlation coefficients can quantify the association between variables.

c. if there isn't linear or monotonic trend but non-linear trends then, we may use maximal correlation or Mutual information coefficients to identify and quantify associations between variables.

d. Though Mutual Information coefficient can be used to identify and quantify different kinds of association between variables we should be careful with it as it performs poorly if the data corrupted by noise.

e. Unlike, mutual Information coefficient, mutual correlation can be used to identify associations or correlations between variables as it is robust to noise.