Data Science in Financial Markets CSE4009

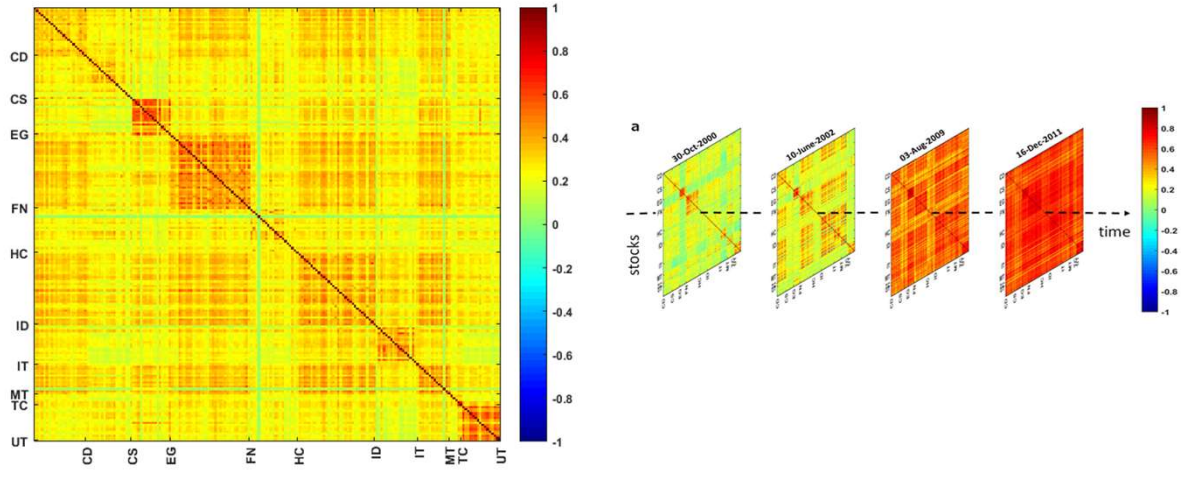
Dr. Hirdesh Kumar Pharasi
Associate Professor
BML Munjal University, Haryana, India

1

Long-term and Short-term Financial Investments

Correlation Matrix Analysis

S&P 500 (USA): 1985 to 2016 (T = 8068 days); Number of stocks N = 194



2

CORRELATION

3

Correlation

- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

4

More examples

- **Positive relationships**
- water consumption and temperature.
- study time and grades.

- **Negative relationships**:
- alcohol consumption and driving ability.
- Price & quantity demanded

5

TYPES OF CORRELATION

Positive and Negative Correlation

- **Positive Correlation** *(When two variables move in the same direction)*
- **Negative Correlation** *(When two variables move in opposite direction)*

Linear and Non-Linear Correlation

- **Linear Correlation** *(When change in one variable tends to bear a constant ratio to the amount of change in other variable)*
- **Non-Linear Correlation** *(When change in one variable does not bear a constant ratio to the amount of change in other related variable)*

Simple, Multiple and Partial Correlation

- **Simple Correlation** *(When only two variables are studied)*
- **Multiple Correlation** *(When relationship among three or more than three variables is studied)*
- **Partial Correlation** *(When relationship between two variables is examined keeping other variables as constant)*

6

Types of Correlation

- **Simple correlation**: Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation**: Under Multiple Correlation three or more than three variables are studied. Ex. $Q_d = f (P, P_C, P_S, t, y)$
- **Partial correlation**: analysis recognizes more than two variables but considers only two variables keeping the other constant.
- **Total correlation**: is based on all the relevant variables, which is normally not feasible.

7

Methods of Studying Correlation

- Scatter Diagram Method
- Graphic Method
- Karl Pearson's Coefficient of Correlation

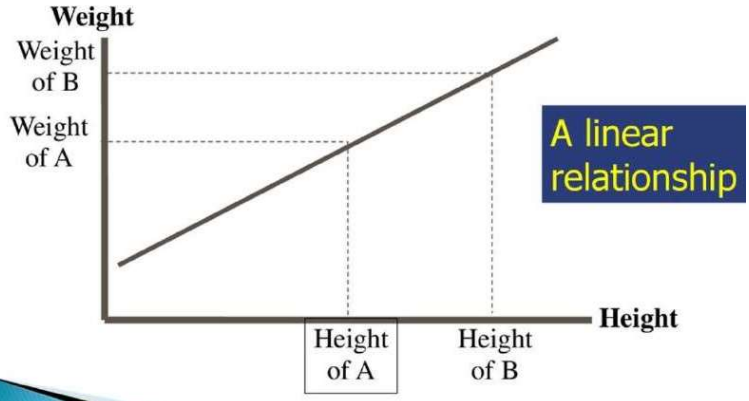
8

Scatter Diagram Method

- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate. It portrays the relationship between these two variables graphically.

9

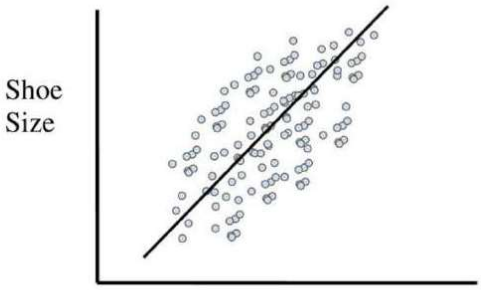
A perfect positive correlation



10

Degree of correlation

- **Moderate Positive Correlation**

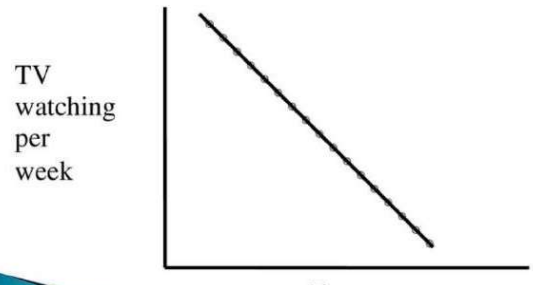


r = + 0.4

11

Degree of correlation

- **Perfect Negative Correlation**



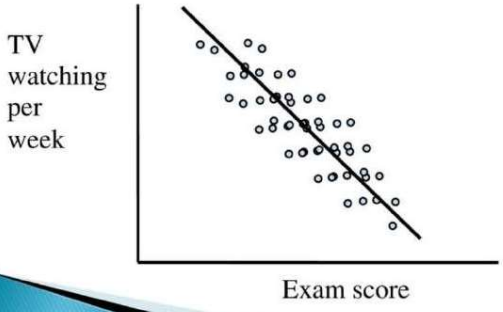
r = -1.0

12

Degree of correlation

- Moderate Negative Correlation

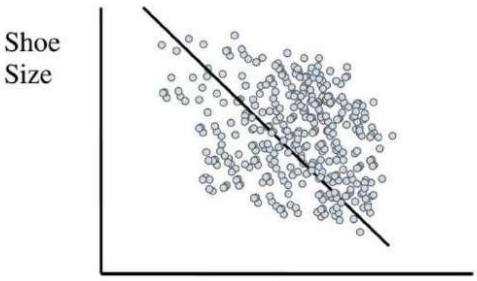
r = -.80



13

Degree of correlation

- Weak negative Correlation

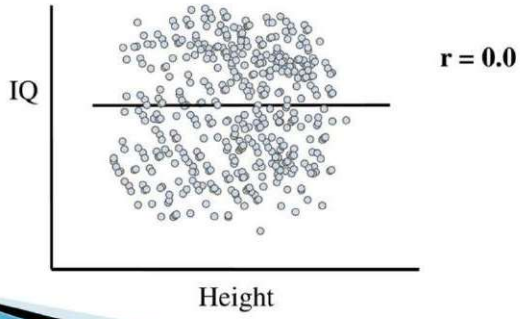


r = - 0.2

14

Degree of correlation

- No Correlation (horizontal line)



15

Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

16

Disadvantage of scatter diagram

Can not adopt an **exact** degree of correlation

17

Correlation r – basic assumptions

- No distinction between explanatory (x) and response (y) variable.
- The null hypothesis test that r is significantly different from zero (0).
- Requires both variables to be quantitative or continuous variables
- Both variables must be normally distributed. If one or both are not, either transform the variables to near normality or use an alternative non-parametric test of Spearman
- Use Spearman Correlation coefficient when the shape of the distribution is not assumed or variable is distribution-free.

18

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

19

Karl Pearson's Coefficient of Correlation

- Pearson's 'r' is the most common correlation coefficient.
- Karl Pearson's Coefficient of Correlation denoted by– 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y.

20

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

The coefficient of correlation:

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2} \cdot \sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

where r = coefficient of correlation, $-1 \le r \le +1$
n = number of data points

21

SN	Maths X	Science Y	Xi-Xmean	Yi-Ymean	(Xi-Xm)(Yi-Ym)	(Xi-Xm)^2	(Yi-Ym)^2
1	98	88	13.28571429	1.571428571	20.87755102	176.5102041	2.469387755
2	87	92	2.285714286	5.571428571	12.73469388	5.224489796	31.04081633
3	91	95	6.285714286	8.571428571	53.87755102	39.51020408	73.46938776
4	75	82	-9.714285714	-4.428571429	43.02040816	94.36734694	19.6122449
5	81	74	-3.714285714	-12.42857143	46.16326531	13.79591837	154.4693878
6	68	78	-16.71428571	-8.428571429	140.877551	279.3673469	71.04081633
7	93	96	8.285714286	9.571428571	79.30612245	68.65306122	91.6122449
Mean	**84.71428571**	**86.42857143**			**396.8571429**	**677.4285714**	**443.7142857**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Using formula	0.7238536037
Direct formula	0.7238536037

22

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient 'r' ranges from −1 to +1
- If r = +1, then the correlation between the two variables is said to be perfect and positive
- If r = −1, then the correlation between the two variables is said to be perfect and negative
- If r = 0, then there exists no correlation between the variables

23

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

24

Limitation of Pearson's Coefficient

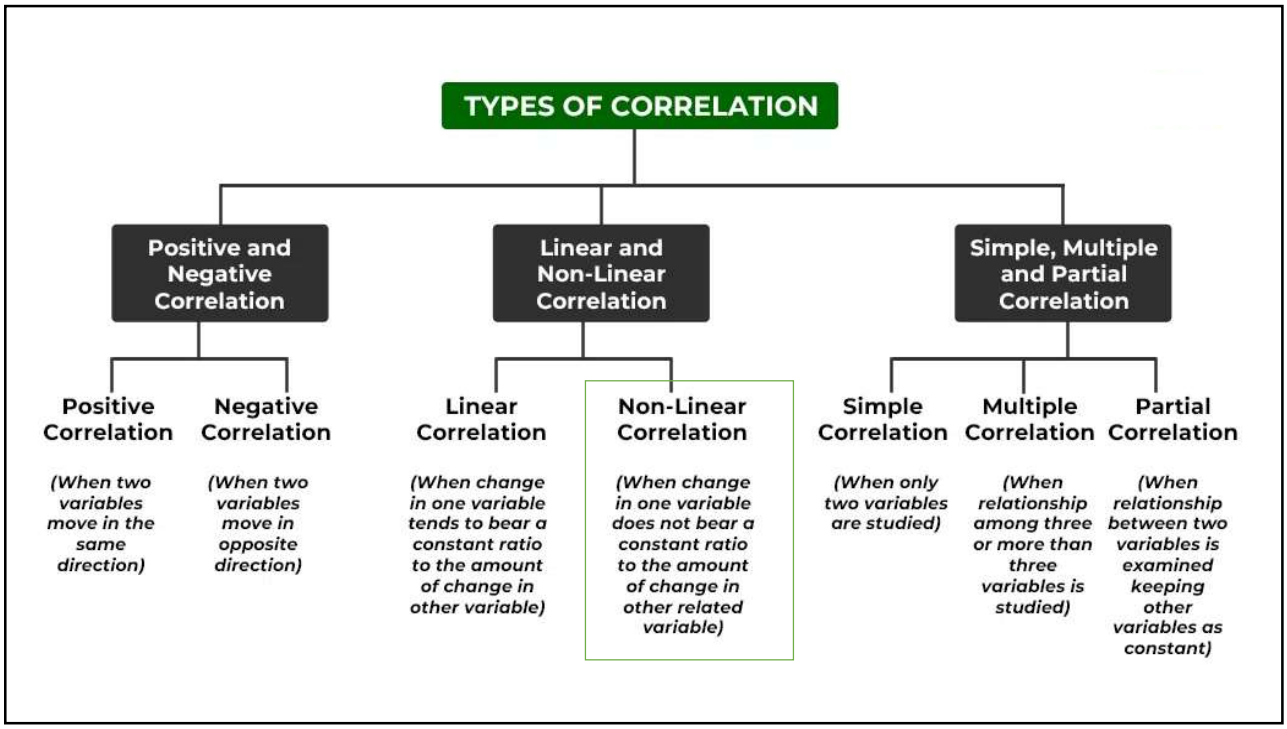
- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

25

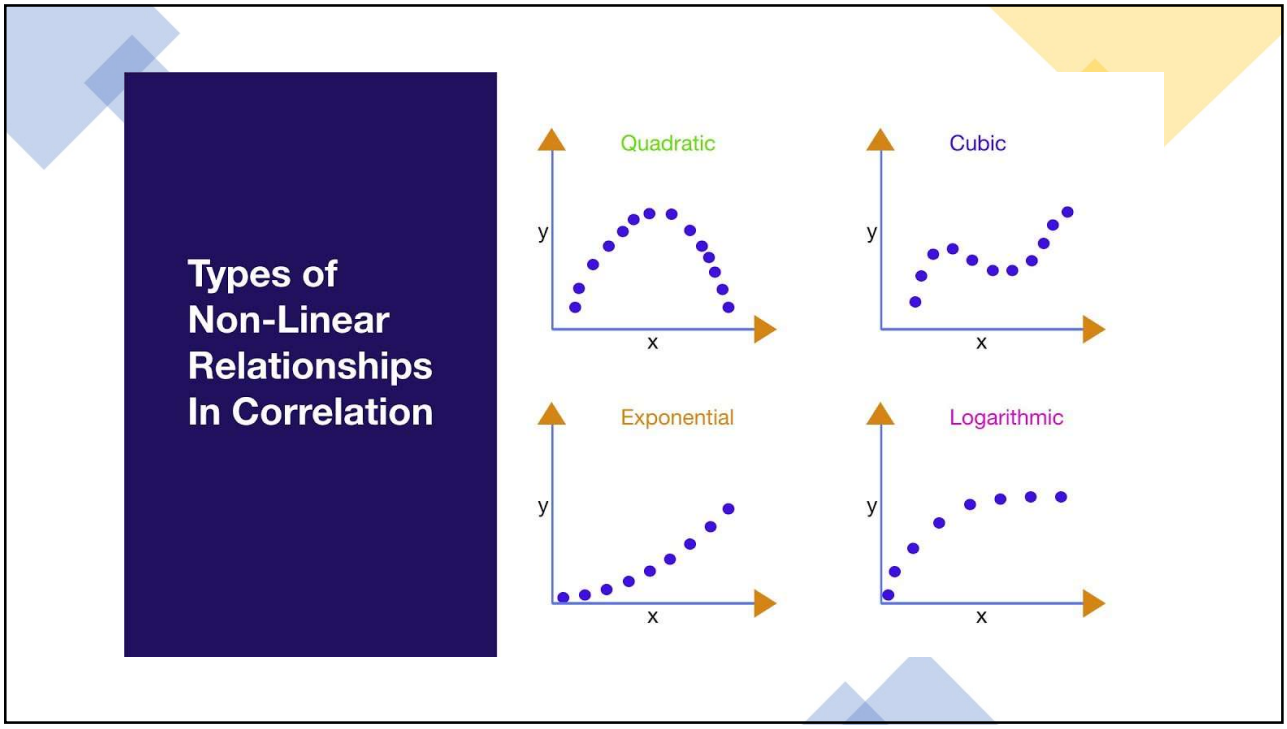
Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

26


TYPES OF CORRELATION
Positive and Negative Correlation
Linear and Non-Linear Correlation
Simple, Multiple and Partial Correlation
Positive Correlation
(When two variables move in the same direction)
Negative Correlation
(When two variables move in opposite direction)
Linear Correlation
(When change in one variable tends to bear a constant ratio to the amount of change in other variable)
Non-Linear Correlation
(When change in one variable does not bear a constant ratio to the amount of change in other related variable)
Simple Correlation
(When only two variables are studied)
Multiple Correlation
(When relationship among three or more than three variables is studied)
Partial Correlation
(When relationship between two variables is examined keeping other variables as constant)

27


Types of Non-Linear Relationships In Correlation
Quadratic
Cubic
Exponential
Logarithmic
y
x

28

Spearman's Rank Coefficient of Correlation

- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson's correlation coefficient can not be used in such case Spearman Rank correlation can be used.
- $R = 1 - (6\ \Sigma D^2) / N (N^2 - 1)$
- R = Rank correlation coefficient
- D = Difference of rank between paired item in two series.
- N = Total number of observation.

29

Spearman's Rank Correlation (ρ) is a **non-parametric measure** of the strength and direction of the **monotonic relationship** between two variables.

The **Spearman correlation** between two variables is equal to the **Pearson correlation** between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

Spearman's correlation is excellent for **nonlinear monotonic relationships**, but not suitable for **non-monotonic patterns**.

Pearson vs Spearman

- Pearson's correlation (r): measures the *strength of a linear relationship*. If the data follow a curved (nonlinear) but monotonic pattern, Pearson may give a misleadingly low correlation.
- Spearman's correlation (ρ): measures *monotonic association* (whether values consistently increase or decrease together), regardless of whether the relationship is straight-line or curved.

When to Use Spearman?

- Data are ordinal or contain outliers.
- Suspected curvilinear but monotonic relationship.
- Variables don't meet Pearson's assumptions (normality, linearity, homoscedasticity).

30

Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from −1 to +1
- If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction
- If R = −1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction
- If R = 0, then there is no correlation

31

SN	Maths X	Rank X	Science Y	Rank Y	D=RankX -RankY	D^2
1	98	1	88	4	-3	9
2	87	4	92	3	1	1
3	91	3	95	2	1	1
4	75	6	82	5	1	1
5	81	5	74	7	-2	4
6	68	7	78	6	1	1
7	93	2	96	1	1	1
Mean	**84.71428571**		**86.42857143**			**18**

$$\hat{\rho} = 1 - 6\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Formula	0.6785714286
Correlation	0.7238536037

32

Example Data

X: 1, 2, 3, 4, 5, 6, 8

Y: 8, 6, 5, 4, 3, 2, 1

Table:

R_x	R_y	d	d^2
1	7	-6	36
2	6	-4	16
3	5	-2	4
4	4	0	0
5	3	2	4
6	2	4	16
7	1	6	36

$\Sigma d^2 = 112$

- $\rho = 1 - (6 \times 112) / [7(49 - 1)]$
- $\rho = 1 - (672 / 336)$
- $\rho = -1$

- Result: Perfect negative correlation.
- As one variable increases, the other decreases monotonically.

33

Rank Correlation Coefficient (R)

a) Problems where actual rank are given.

1) Calculate the difference 'D' of two Ranks i.e. (R1 – R2).
2) Square the difference & calculate the sum of the difference i.e. ΣD^2
3) Substitute the values obtained in the formula.

34

Rank Correlation Coefficient

b) Problems where Ranks are not given :If the ranks are not given, then we need to assign ranks to the data series. The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1. We need to follow the same scheme of ranking for the other series.

Then calculate the rank correlation coefficient in similar way as we do when the ranks are given.

35

Merits Spearman's Rank Correlation

- This method is simpler to understand and easier to apply compared to karl pearson's correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Limitation Spearman's Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.
- This method should be applied where N exceeds 30.

36

Partial Correlation

37

Partial correlation

- Partial correlation measures the correlation between *X* and *Y*, controlling for *Z*
- Comparing the bivariate (zero-order) correlation to the partial (first-order) correlation
  - Allows us to determine if the relationship between *X* and *Y* is direct, spurious, or intervening
  - Interaction cannot be determined with partial correlations

- Example:
- Exercise, Cholesterol, and Age
- Education, Income, and Work Experience
- Partial correlation is useful in social sciences, medicine, economics, and finance

38

Formula for partial correlation

- Formula for partial correlation coefficient for *X* and *Y*, controlling for *Z*

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

- We must first calculate the zero-order coefficients between all possible pairs of variables (*Y* and *X*, *Y* and *Z*, *X* and *Z*) before solving this formula

39

Index	Study Hours (X)	Exam Score (Y)	Sleep Hours (Z)
1	2	50	8.0
2	4	70	7.0
3	6	80	6.2
4	8	95	5.1

- Study vs Score (X,Y): Strong positive ≈ 0.99
- Study vs Sleep (X,Z): Strong negative ≈ -0.99
- Score vs Sleep (Y,Z): Strong negative ≈ -0.99

40

Partial Correlation (X & Y controlling Z)

Partial correlation ≈ -0.05 → almost zero

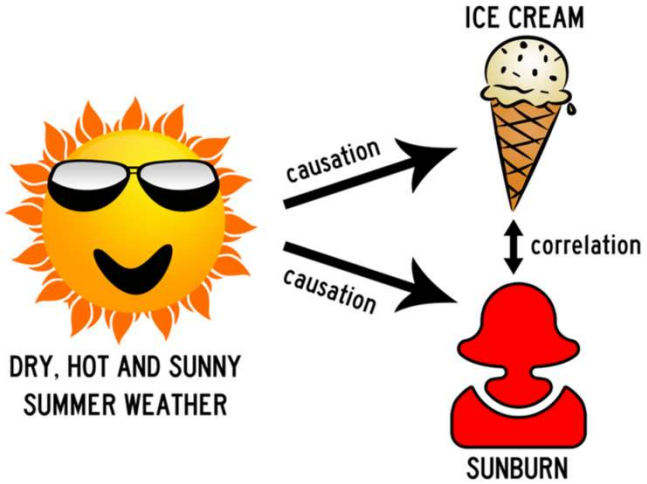
Interpretation:

- At first, Study hours and Exam score looked strongly related.
- But after removing the effect of Sleep, the relation disappears.
- Sleep was the **hidden factor** and driving the correlation.

Note: the strong correlation was **mainly due to sleep being linked with both** study and score.

41

Causation vs Correlation

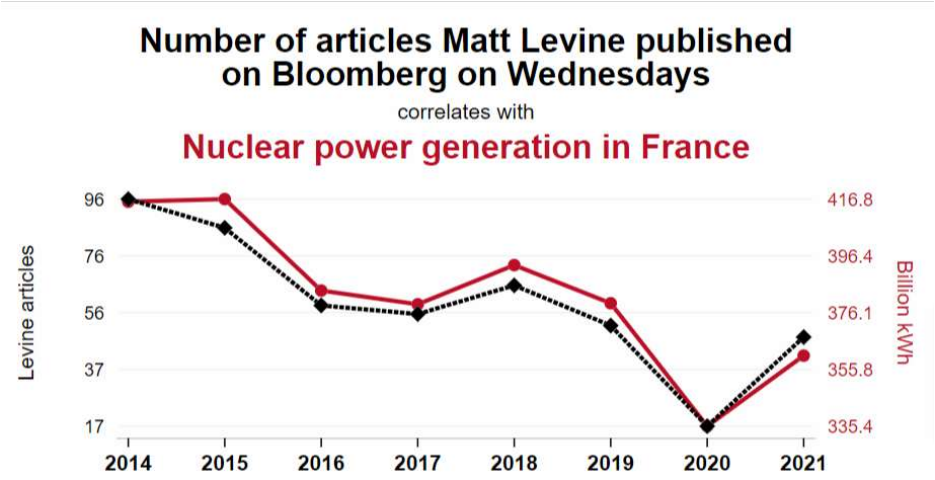


42

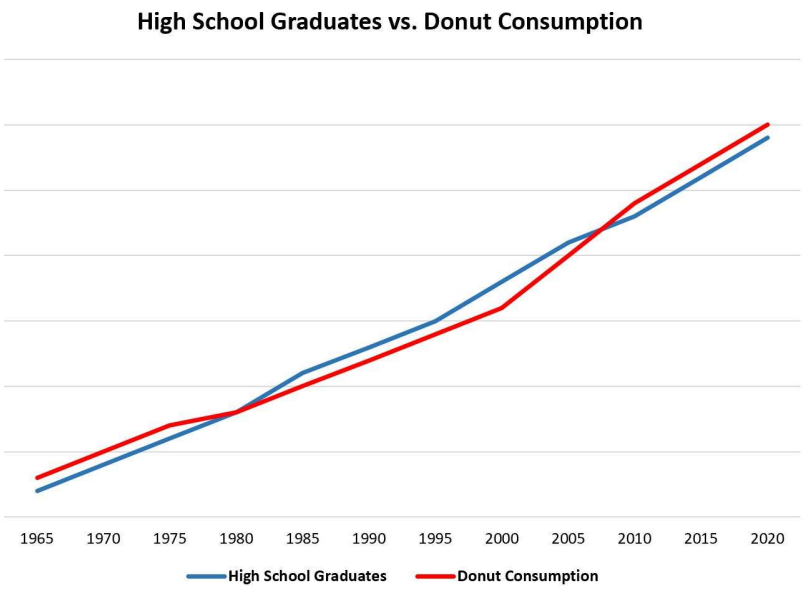
Spurious correlations: correlation is not causation

Sitelink: https://www.tylervigen.com/spurious-correlations

43



44



45

Granger causality

Causality
(Philosophy/Science)

- True cause–effect relationship
- Requires theory/experiments
- Example: Smoking → Cancer

Granger Causality
(Statistics)

- Predictive relationship in time series
- Based on statistical testing
- Stock prices of Company A predict Company B

- **Causality** = actual cause → effect.
- **Granger causality** = "X improves prediction of Y" (but doesn't prove true causation).

46

Granger causality

- **Granger causality is about predictive ability:**
  - A time series Xt **Granger-causes** another time series Yt if **past values of X** provide information that improves prediction of Yt, compared to using only past values of Yt.
  - It **does not mean true cause-and-effect**—it's only about whether one variable has useful predictive information.

Mathematical Formulation

Let's test if X Granger-causes Y.

- Restricted Model (AR model of Y):

$$Y_t = a_0 + a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \epsilon_t$$

- Unrestricted Model (add lags of X):

$$Y_t = a_0 + a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + b_1 X_{t-1} + b_2 X_{t-2} + \cdots + b_q X_{t-q} + \epsilon_t$$

If coefficients $b_1, b_2, \ldots, b_q$ are jointly **significant**, then **X Granger-causes Y.**

47

Hypothesis Testing

- **Null hypothesis (H0):**

$$b_1 = b_2 = \cdots = b_q = 0$$

(X does not Granger-cause Y)

- **Alternative hypothesis (H1):**
  At least one $b_i \neq 0$.
- **Test statistic (F-test):**

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(T-(p+q+1))}$$

where:

- SSR_r = sum of squared residuals (restricted model)
- SSR_{ur} = sum of squared residuals (unrestricted model)
- T = number of observations
- p = lags of Y
- q = lags of X

If F exceeds critical value (or p-value < significance level), reject H_0.

Assumptions

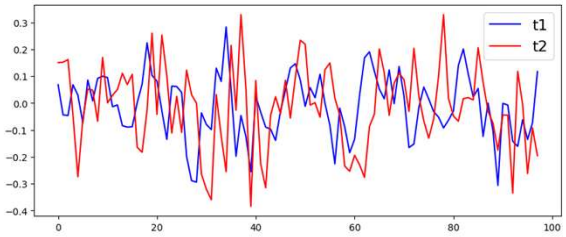
- **Stationarity**: Time series must be stationary (mean and variance constant over time).
  - Usually tested with ADF (Augmented Dickey-Fuller) test.
  - If not stationary → difference the series.
- **Lag length selection**: Number of lags (p, q) must be chosen carefully (using AIC, BIC, HQC criteria).
- **Linear relationship**: Granger causality assumes linear predictive power. (Nonlinear variants exist).

48

```
[8] #build the time series, just a simple AR(1)
    t1 = [0.1*np.random.normal()]
    for _ in range(100):
        t1.append(0.5*t1[-1] + 0.1*np.random.normal())

[9] #build the time series that is granger caused by t1
    t2 = [item + 0.1*np.random.normal() for item in t1]

    #adjust t1 and t2
    t1 = t1[3:]
    t2 = t2[:-3]
```




```
gc_res = grangercausalitytests(ts_df, 3)
```

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=4.0677  , p=0.0466  , df_denom=94, df_num=1
ssr based chi2 test:   chi2=4.1975  , p=0.0405  , df=1
likelihood ratio test: chi2=4.1092  , p=0.0426  , df=1
parameter F test:         F=4.0677  , p=0.0466  , df_denom=94, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=7.7292  , p=0.0008  , df_denom=91, df_num=2
ssr based chi2 test:   chi2=16.3078 , p=0.0003  , df=2
likelihood ratio test: chi2=15.0619 , p=0.0005  , df=2
parameter F test:         F=7.7292  , p=0.0008  , df_denom=91, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=42.3999 , p=0.0000  , df_denom=88, df_num=3
ssr based chi2 test:   chi2=137.3180, p=0.0000  , df=3
likelihood ratio test: chi2=84.9519 , p=0.0000  , df=3
parameter F test:         F=42.3999 , p=0.0000  , df_denom=88, df_num=3
```

49

Rolling Metrics in Time Series

Understanding dynamic, window-based measures in time series analysis

50

Concept & How it Works

- Rolling metrics are statistics computed over a sliding window of data.
- The window "**rolls**" forward step by step, recalculating the metric.
- Produces a dynamic series that tracks evolving behavior.

Example: 30-day rolling average in stock prices shows local trends.

51

Common Rolling Metrics & Uses

- Rolling Mean: Smooths short-term fluctuations.
- Rolling Variance/Std Dev: Captures volatility.
- Rolling Correlation: Tracks dynamic relationships.
- Rolling Causality: Extend statistical tests (like Granger) into rolling windows.

Applications:

– Detect structural breaks

– Monitor financial risk

– Understand non-stationary relationships

52

Visualization & Insights

- Rolling Mean → Smoothed curve overlaid on original series.
- Rolling Correlation → Line graph of changing correlation.
- Heatmaps → Show dynamic dependencies across many variables.

Rolling metrics provide local, time-varying insights that static measures miss.

53

Rolling Mean (Moving Average)

- Averages data points within a sliding window.
- Reduces short-term noise, highlights trends.
- Formula: (1/k) * Σ (X_{t-i}), i=0..k-1
- Types: SMA (equal weights), WMA (weighted), EMA (exponential).

- Example:
  - Daily temperatures (7 days): [30, 32, 31, 29, 35, 36, 34]
  - 3-day rolling mean at day 5: (31 + 29 + 35) / 3 = 31.67
- Rolling mean produces a smoother series than raw data.

54

Applications & Insights

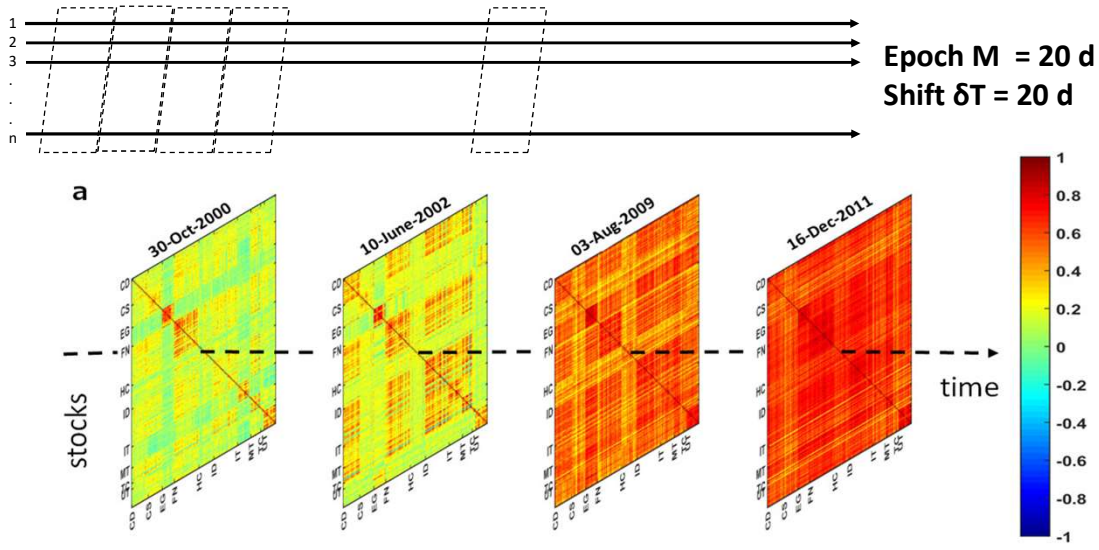
- Finance: Stock price trends
- Weather: Smooth daily fluctuations
- Economics: GDP growth (remove seasonal noise)
- Signal Processing: Noise reduction

Simple but powerful tool for smoothing & detecting trends.

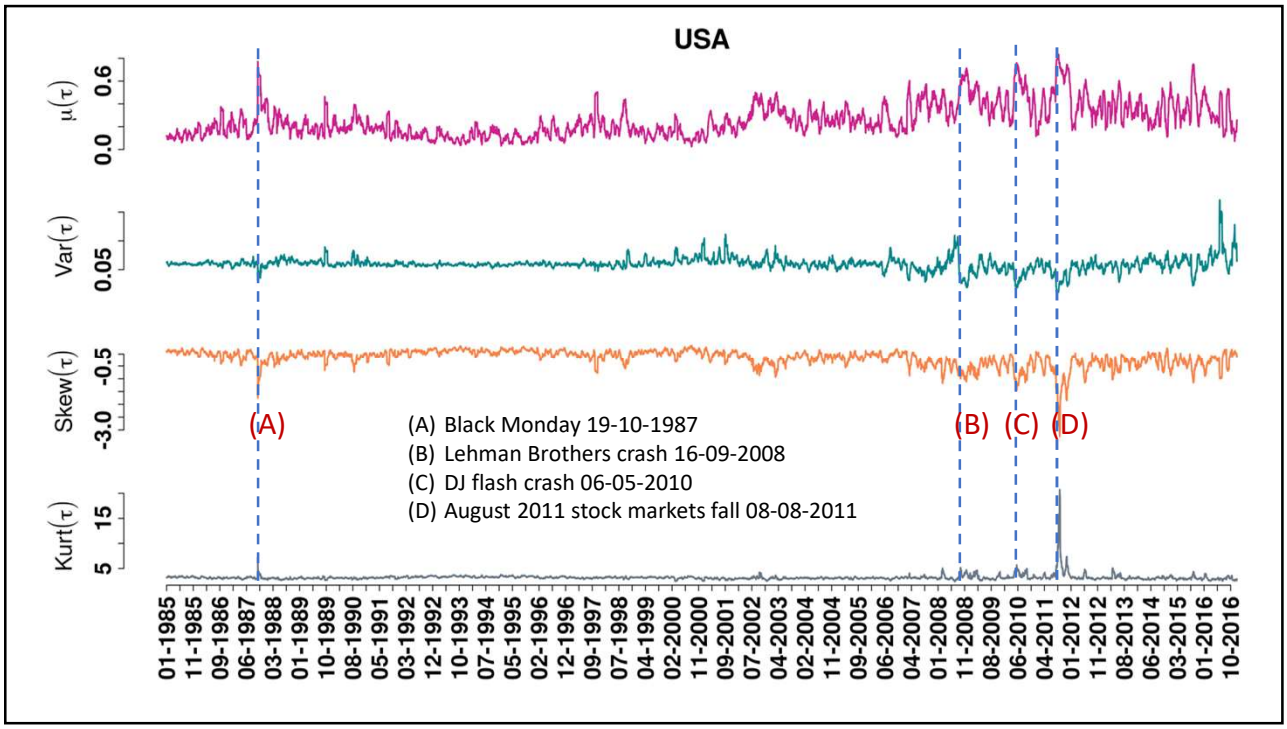
55

Rolling Cross-Correlation Matrices

Short or Medium Term strategy



56


USA
μ(τ)
0.6
0.0
Var(τ)
0.05
Skew(τ)
-0.5
-3.0
Kurt(τ)
15
5
(A)
(A) Black Monday 19-10-1987
(B) Lehman Brothers crash 16-09-2008
(C) DJ flash crash 06-05-2010
(D) August 2011 stock markets fall 08-08-2011
(B) (C) (D)
01-1985
11-1985
09-1986
06-1987
03-1988
01-1989
10-1989
08-1990
05-1991
03-1992
12-1992
10-1993
07-1994
05-1995
02-1996
12-1996
09-1997
07-1998
04-1999
02-2000
11-2000
09-2001
07-2002
04-2003
02-2004
11-2004
09-2005
06-2006
04-2007
01-2008
11-2008
08-2009
06-2010
04-2011
01-2012
11-2012
08-2013
06-2014
03-2015
01-2016
10-2016

57


Thank you

58