A New Correlation Measure And A New Way To Analyze Skewness and Kurtosis

A Project Report submitted by

Amit Kumar
M22MA202

in partial fulfillment of the requirements for the award of the degree

Master of Science

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
Department of Mathematics

May 2024

Declaration

I hereby declare that the work presented in this Project Report titled **A New Correlation Measure And A New Way To Analyze Skewness and Kurtosis** - Master of Science, Mathematics, submitted to the **Indian Institute of Technology Jodhpur** in partial fulfillment of the requirements for the award of the degree of Master of Science, Mathematics, is a true and accurate record of the research work conducted under the supervision of **Dr.Vivek Vijay**. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree diploma or certificate.

Amit Kumar
M22MA202

Acknowledgements

I want to express my sincere gratitude to my supervisor **Dr. Vivek Vijay** , Department of Mathematics, Indian Institute of Technology Jodhpur, for giving me an opportunity to embark on this field of research. His constant support, vision, sincerity, and determination during my project have given me a new way to work. without his support, I would not be able to learn with so much depth.

I would also like to thank **Mr. Veeresh Chaudhary, Mr. Sunil Choudhary, Mr. Ankit Dalal, Mr. Utkarsh Sahai, Mr. Sahil Sharma** for their constant help and support, All my classmates and all my seniors for all those moments we experienced in IIT Jodhpur. I am also thankful for **my family** for giving me the opportunity to study and motivating me to chase my dreams.

Moreover, I would like to thank my **Department of Mathematics and Indian Institute of Technology Jodhpur** for providing such a good environment and facilities.

Amit Kumar
M22MA202
M.Sc. Mathematics

Certificate

This is to certify that the Project Report titled **A New Correlation Measure And A New Way To Analyze Skewness and Kurtosis**, submitted by **Amit Kumar (M22MA202)** to the Indian Institute of Technology, Jodhpur for the award of the degree of M.Sc., is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Vivek Vijay
Assosciate Professor
Department of Mathematics
Indian Institute of Technology, Jodhpur

Abstract

The first part explores the Information-based Correlation coefficient (ICC) for assessing information sharing between time series. ICC overcomes the limitations of traditional coefficients, proves effective in detecting non-linearity and spurious correlation, and performs well in both stationary and non-stationary time series. The ICC of order 1 is defined using the binary entropy function, showcasing its versatility and validated through simulations. While acknowledging computational complexity and biases in smaller samples, the ICC consistently outperforms traditional coefficients in detecting spurious correlations. Comparative Monte Carlo simulations highlight ICC's ability to yield correlations close to zero in uncorrelated data. Practical applications demonstrate ICC's superiority in detecting the absence of correlation in independent stationary time series and addressing spurious correlations. In conclusion, ICC emerges as a valuable and versatile tool for correlation measurement.

while in the second part, the paper talks about using a thing called kurtosis to see if data is similar to a Gaussian, or bell-shaped, distribution. We look at two types of kurtosis: the regular one and another version used in a different type of statistics called q-statistics. We found a relationship between two things, skewness and kurtosis, that some researchers thought existed. It actually only works for small sets of data, not big ones. And if the data's fourth moment (a special thing in statistics) is not finite, kurtosis doesn't have a fixed value when the dataset gets big.

For datasets where the fourth moment is finite, the size of the dataset needed for kurtosis to have a fixed value depends on how different the data is from a Gaussian. But even if the fourth moment is finite, relying only on kurtosis to decide how different the data is from a Gaussian can give wrong answers, especially with small datasets. When we switch to q-statistics, we see that while a type of kurtosis called q-kurtosis has finite values for certain values of q, it's not very useful for comparing different datasets with non-Gaussian distributions unless we pick the right q value for each dataset. Finally, we suggest a way to figure out the right q value and calculate q-kurtosis for datasets that follow q-Gaussian distributions.

Contents

Chapter 1

Introduction

In the first part of the paper, we will discuss a new type of correlation based on the entropy of a time series. Correlation is important in every field, but standard coefficients like the Pearson correlation coefficient are sensitive when dealing with outliers. When dealing with time series data, spurious correlation invalidates the relationship between two variables, resulting in false results. To deal with such crucial problems, we need a correlation that can handle these problems and provide better results. Spearman's correlation has some advantages over Pearson's, such as being less affected by outliers and can handle data that are not normally distributed. Moreover, Pearson and Spearman's rank correlation coefficient may show variability in detecting correlation for independent stationary processes, such as the white noise process. Therefore, there is a need for an alternative coefficient that can provide a good result when dealing with such problems. This coefficient, called the Information-based Correlation Coefficient (ICC), can address many of the disadvantages of Pearson, Spearman, and Kendall's tau. For example, it shows less variability when dealing with independent stationary data and is more reliable in the presence of outliers. This correlation has many more advantages, which we will discuss in sections 1.3, 1.4, 1.5, and 1.6 of the paper.

In the second part of the paper, we have attempted to find a method to compare two non-Gaussian distributions. Initially, it was assumed that each distribution held a Gaussian distribution. However, scientists discovered some datasets that didn't conform to a Gaussian shape, which raised a significant question. Was the deviation from Gaussian indicative of a data collection error, or was evolution truly yielding a non-Gaussian distribution? When Pearson examined this problem, he observed that the frequency distribution of certain datasets diverged from a Gaussian pattern. Recognizing the limitations of Gaussian theory, Pearson proposed a new theory that better captured the distributions, called kurtosis, to understand whether a distribution is Gaussian or not. These curves are now recognized as part of the Pearson family.

Pearson identified three conditions for a distribution to be classified as Gaussian: kurtosis, skewness, and the distance between median and mean. He emphasized the need to consider higher moments to measure differences between distributions and the Gaussian curve, introducing kurtosis and skewness. The primary aim of defining kurtosis was to decide if a distribution is Gaussian. Therefore, the standard definition of kurtosis references a Gaussian curve. However, Pearson did not attempt to compare two non-Gaussian distributions based on their kurtosis values. Our main purpose here is to find a way to compare two non-Gaussian distributions.

Chapter 2

An Entropy-based Correlation-Coefficient

2.1 What is Correlation

A correlation coefficient is a numerical value between -1 and +1 that represents the strength and direction of the relationship between two variables.. The coefficient's value determines whether the variables are positively related, negatively related, or have no correlation.

1. **Positive Correlation-**: This term Refers to a relationship between two variables where an increase or decrease in one variable leads to a corresponding increase or decrease in the other. The correlation ranges from 0 to +1.

2. **Negative Correlation-**: This term Refers to a relationship between two variables where an increase or decrease in one variable leads to a corresponding decrease or increase in the other. The correlation ranges from -1 and 0.

3. **Zero Correlation-**: The relationship with no defined increase or decrease between two variables. The correlation is around 0.

2.1.1 Types of Correlation Coefficient

We can choose many different correlation coefficients based on the linearity of the relationship, Type of measurement of data, Type of distribution, etc.

Type of Correlation Coefficient	Relationship	Level of Measurement	Data Distribution
Pearson	Linear	Quantitative data	Normal Distribution
Spearman rank	Non-linear	Ordinal data	Any Distribution
Kendall's tau	Non-linear	Ordinal data	Any Distribution

There are many other correlations defined on the basis of different data properties, such as DCCA, spurious, etc.

1. **Pearson Correlation Coefficient-**:
   It is the most common measure of linear correlation and is more appropriate when the given variables are quantitative and normally distributed.

   Let X and Y be two Quantitative Variables of n data points. Then, the Pearson correlation coefficient(r) of X and Y is given by

$$r(X,Y) = \frac{\sum_k (X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_k (X_k - \bar{X})^2}\sqrt{\sum_k (Y_k - \bar{Y})^2}} = \frac{Cov(X,Y)}{Sd(X) * Sd(Y)}$$

   Where, X_k is ith element of X, Y_k is kth element of Y, $\bar{X}$ is Mean of X, $\bar{Y}$ is Mean of Y, Cov is Covariance and Sd is the Standard deviation.

   - If r lies between 0 and 1 it is a positive correlation.
   - If r lies in -1 and 0 there is a negative correlation and
   - for r value 0 there is no correlation.

   Limitations of Pearson's Correlation

   - It can't determine the non-linearity between two variables.
   - It is more sensitive to outliers.
   - If the data is not normally distributed, the correlation coefficient may be less reliable, etc.

   EXAMPLE: Pearson correlation coefficient for the given Data X=$\{21, 15, 18, 12, 27\}$ and Y=$\{8, 4, 6, 2, 12\}$

   Solution:

$$\bar{X} = \frac{21 + 15 + 18 + 12 + 27}{5} = 18.6$$
$$\bar{Y} = \frac{8 + 4 + 6 + 2 + 12}{5} = 6.4$$

X	Y	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$
21	8	2.4	1.6
15	5	-3.6	-2.4
18	6	-0.6	-0.4
12	2	-6.6	-4.4
27	12	8.4	5.6

Substitute these values into the correlation coefficient formula:

$$
\begin{aligned}
r &= \frac{(-6.6)(-4.4) + (-3.6)(-2.4) + (-0.6)(-0.4) + (2.4)(1.6) + (8.4)(5.6)}{\sqrt{\sum (X_i - 18.6)^2 \sum (Y_i - 6.4)^2}} \\
&= \frac{29.04 + 8.64 + 0.24 + 3.84 + 47.04}{\sqrt{(43.56 + 12.96 + 0.36 + 5.76 + 70.56)(19.36 + 5.76 + 0.16 + 2.56 + 31.36)}} \\
&= \frac{88.8}{\sqrt{(133.2)(59.2)}} \\
&= \frac{88.8}{\sqrt{7885.44}} \\
&\approx \frac{88.8}{88.8} \\
&\approx 1
\end{aligned}
$$

2. **Spearman rank Correlation Coefficient-**:
It assesses the strength and direction of the monotonic relationship between the ranked (Rank the values of each variable separately, from lowest to highest) values of two variables without assuming a specific distribution for the data.

Let X and Y be two Variables of size n. Then, the spearman rank correlation coefficient(ρ) of X and Y is given by

$$\rho(X, Y) = 1 - \frac{6 \sum k d_i^2}{n(n^2 - 1)}$$

Where, d_i = difference between the ranks of ith values

- If ρ value lies in 0 and 1 it is positive monotonic.
- If ρ value lies in 0 and -1 it is negative monotonic.
- For ρ value 0, no monotonic relation.

Limitations of Spearman rank Correlation

- Converting the data to ranks involves a loss of information.
- Spearman correlation tends to be less stable with small sample sizes.
- Less sensitive to outliers

- Reliable only for ordinal data and etc.

Example : Calculate the Spearman rank correlation coefficient for the given Data X=$\{10, 6, 9, 12, 18\}$ and Y=$\{8, 7, 5, 6, 9\}$

Solution:

X	Y	R_X	R_Y	d_i	d_i^2
10	8	2	2	0	0
6	7	5	3	2	4
9	5	3	5	-2	4
12	6	1	4	-3	9
8	9	4	1	3	9

We have,

$$\sum_{i=1}^{5} d_i^2 = 0 + 4 + 4 + 9 + 9 = 26$$

$$\rho = 1 - \frac{26}{5^3 - 5} = 0.79$$

3. **kendall's tau:-**

It measures the relationship between columns of ranked data and referenced ranked data of the same data.

Let us have some ranked data of some length, and there is another rank for the same data, then the value of Kendall's tau is given by

$$\tau = \frac{C - D}{C + D}$$

Where, C is number of '+' and D is number of '-'.

First, arrange one of the ranked data from 1 to n. C and D are the values when comparing the first element of other ranked data to all and give value + if the element is greater than the first element and - if it is less than the first element. We will do the same with the second element and compare the second element with all the elements after the second element and continue the process till the last second element. Then count the number of + and - in the table to calculate τ.

Example: Six different batsmen of a team are ranked by two different coaches on the basis of their batting ability. $Coach1 = \{1, 2, 3, 4, 5, 6\}$ and $coach2 = \{4, 2, 3, 1, 5, 6\}$. Then, calculate how much the two rankings correlated using Kendall's tau.

Solution:

players	Coach1	Coach2	C	D
1	1	4	2	3
2	2	2	3	1
3	3	3	2	1
4	4	1	2	0
5	5	5	1	0
6	6	6		

$$\tau = \frac{(2+3+2+2+1)-(3+1+1)}{(2+3+2+2+1)+(3+1+1)} = \frac{5}{15} = 0.33$$

2.2 Information Based Correlation Coefficient(ICC)

The ICC provides an interpretation of how informative one series is in comparison to another. It retains many of the valuable qualities of traditional correlation coefficients. This measure is symmetric, has an asymptotically normal distribution under independence, and is bound between -1 and +1.

ICC potentially overcomes the limitations of other coefficients like:

- It can determine non-linearity between two Variables.
- Perform better in the case of spurious correlation.
- More reliable in the situation of stationary and non-stationary time series
- More reliable for independent white noise

2.2.1 ICC of Order 1

For given two quantitative time series X and Y. ICC of order 1 is defined by

$$ICC(X,Y) = \begin{cases} 1 - H[M(X,Y)] & \text{if } p \geq \frac{1}{2} \\ H[M(X,Y)] - 1 & \text{if } p < \frac{1}{2} \end{cases} \tag{2.2.1.1}$$

Where H is **Binary Entropy Function** whose value lies between 0 and 1. M(X, Y) is **Matching Series of X and Y**. p is the probability of matching of series X and Y.

Binary Entropy Function-:
Shannon's entropy for a binary random variable is called a Binary entropy function.

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p) \tag{2.2.1.2}$$

Where p is the probability of a series match,$1-p$ is the probability of a series mismatch.

To Estimate H[M(X, Y)]
For a time series X, We define a dichotomized series

$$d_k(X) = \begin{cases} + & \text{if } X_k > X_{k-1} \cup A_1 \cup A_2 \\ - & \text{if } X_k < X_{k-1} \cup A_3 \end{cases} \tag{2.2.1.3}$$

'+' is called runs 'UP', and '-' is called runs 'Down'.
If $X_k = X_{k-1}$, Then
A_1= $X_1 = X_2$, If there is no change at the start of the series, the run is considered as a UP. .
A_2= If there is no change after an up, the run is considered as a UP.
A_3= If there is no change after a down, the run is considered as a DOWN.

Now, define Matching Series for two time series X and Y using dichotomized series as,

$$M(X,Y) = \begin{cases} M^+ & \text{if } d_k(X) = d_k(Y) \\ M^- & \text{if } d_k(X) \neq d_k(Y) \end{cases} \tag{2.2.1.4}$$

Where M^+ indicates a match and M^- indicates a mismatch of the series X and Y. From the above equation, it is also clear that **M(X, Y)=M(Y, X)**

To Estimate H[M(X, Y)] we estimate p using matching series as,

$$p = \frac{\sum_k I(M_k(X,Y) = M^+)}{n} \tag{2.2.1.5}$$

and

$$1 - p = \frac{\sum_k I(M_k(X,Y) = M^-)}{n} \tag{2.2.1.6}$$

Where I is an indicator function that shows how many times the symbol appears in a matching series. Now it is clear that if $p > \frac{1}{2}$, then the number of matches is higher than the number of mismatches and hence $ICC > 0$ positive correlation. Similarly, if $p < \frac{1}{2}$, then the number of matches is higher than the number of mismatches in this case $ICC < 0$ and is a negative correlation. for the case $p = \frac{1}{2}$, ICC=0 hence the series are uncorrelated.

Theorem* ICC between X and Y has the following properties.

(i) $-1 \leq ICC \leq +1$
(ii) $ICC(X,Y) = 0$ for M(X,Y) $\sim Bernoulli(\frac{1}{2})$
(iii) $ICC(X,Y) = \pm 1$ iff $p = 1$ or $1 - p = 1$
(iv) $ICC(X,Y) = ICC(Y,X)$
(v) $X = \pm Y \Rightarrow ICC(X,Y) = \pm 1$

2.2.2 Properties of ICC Estimators

When estimating ICC, the estimation of p and of H[M(Y, Z)] is necessary. In this section, we will use independent series to demonstrate how the estimator of ICC shows the same properties as traditional correlation coefficients for independent time series..
To Estimate entropy using the Harris Approach, we Change log base 2 to natural log then $\hat{H} = H(\hat{p}) = -\hat{p_1}\ln(\hat{p_1}) - \hat{p_2}\ln(\hat{p_2})$ due to this change in log ICC range will update to -0.69 to 0.69 and the unit of entropy change from bits to nats and the ICC value becomes

$$\widehat{ICC}(X,Y) = \begin{cases} 0.69 - \hat{H}[M(X,Y)] & \text{if } \hat{p} \geq \frac{1}{2} \\ \hat{H}[M(X,Y)] - 0.69 & \text{if } \hat{p} < \frac{1}{2} \end{cases}$$

In general, $\hat{H}$ is a Biased, Consistent, and Asymptotic normal estimator of entropy H with mean and variance

$$\mathbb{E}(\hat{H}) = H - \frac{1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^{2}\frac{1}{p_i}\right) + \mathcal{O}\left(\frac{1}{n^3}\right) \tag{2.2.2.1}$$

also

$$\mathbb{E}(\hat{H} - H)^2 = \frac{1}{n}\left(\sum_{i=1}^{2} p_i \ln^2 p_i - H^2\right) + \frac{3}{4n^2} + \mathcal{O}\left(\frac{1}{n^3}\right) \tag{2.2.2.2}$$

Some Propositions and Theorems Based on Above Equations are

Proposition 1 $\hat{H}$ is consistent estimator of H.

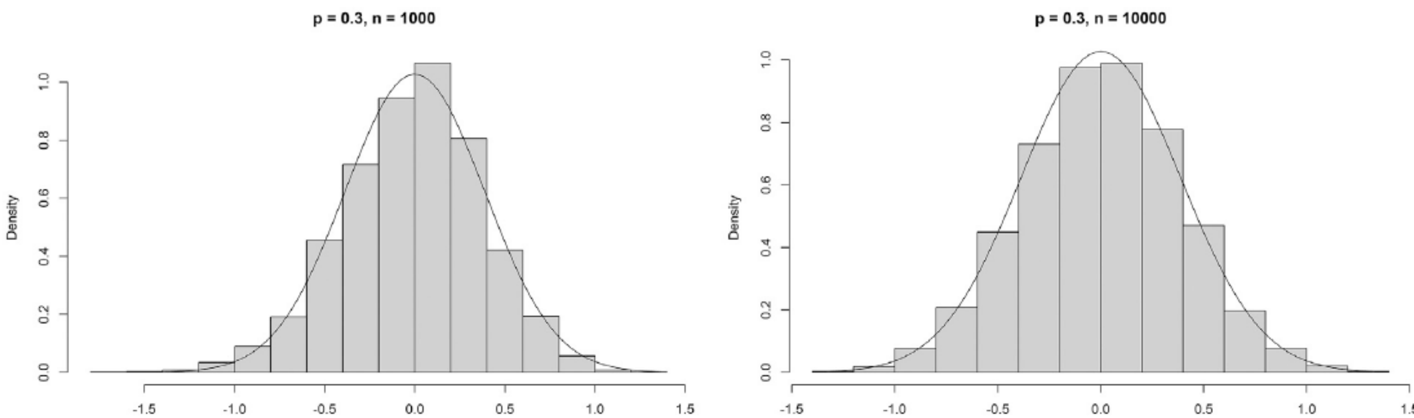
Theorem*1
(a)If $p \neq \frac{1}{2}$. Then, $\sqrt{n}(\hat{H} - H) \xrightarrow{D} \mathcal{N}\left(0, \Sigma_{i=1}^{2} p_i(\ln p_i + H)^2\right)$. Also, $\hat{H}$ is an asymptotically normal estimate of entropy.
(b) if $p = \frac{1}{2}$. Then, $2n(H - \hat{H}) \xrightarrow{D} \chi_1^2$

Theorem*2:
$\widehat{ICC}$ in general is not an unbiased estimator of ICC. But, $\widehat{ICC}$ is asymptotically unbiased.

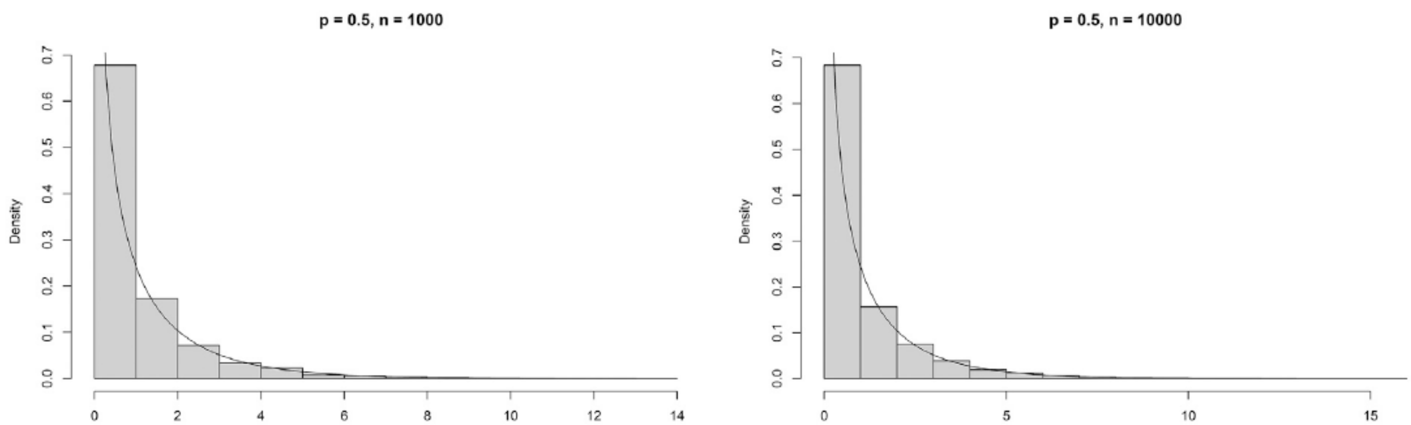
Theorem*3: $I\hat{C}C$ is consistent estimator of ICC.

Theorem*4:
If $p \neq \frac{1}{2}$. Then, $\sqrt{n}(\widehat{ICC} - ICC) \xrightarrow{D} \mathcal{N}\left(0, \Sigma_{i=1}^{2} p_i(\ln p_i + H)^2\right)$. Also, $\widehat{ICC}$ is an asymptotically normal estimate of ICC.

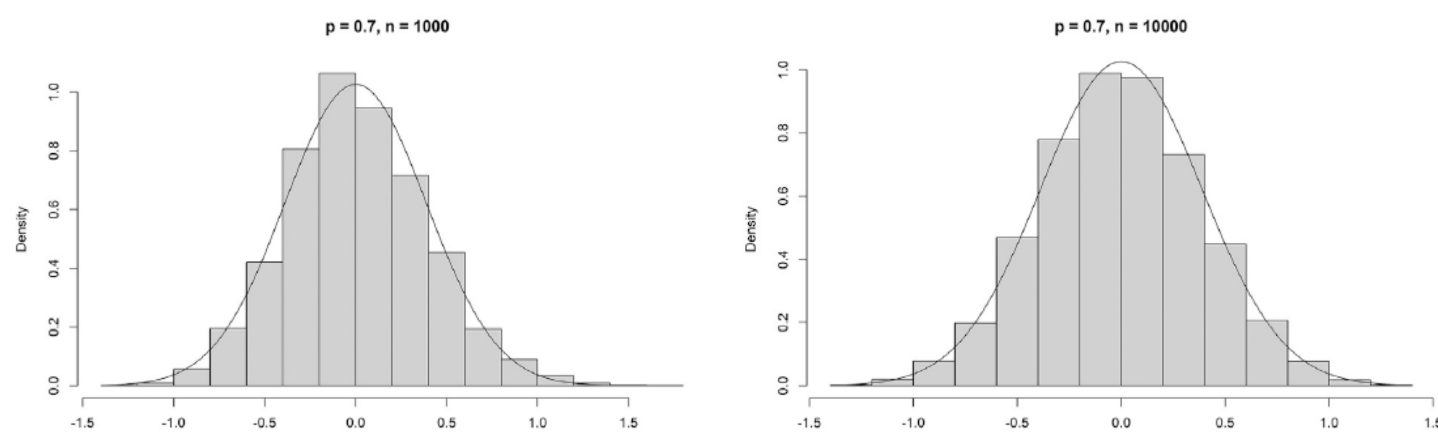
Validation of Asymptotic Theory



Histogram for 10000 simulations of $\sqrt{n}(\hat{H}-H)$ in the case $p \neq \frac{1}{2}$. From theorem 1 for $p \neq \frac{1}{2}$, $\mathcal{N}\left(0, V\right)$ to check the validity of the theorem.



Histogram for 10000 simulations of $2n(H - \hat{H})$ in the case $p = \frac{1}{2}$ from theorem 1 for $p = \frac{1}{2}$, χ^2_1 to check the validity of the theorem.

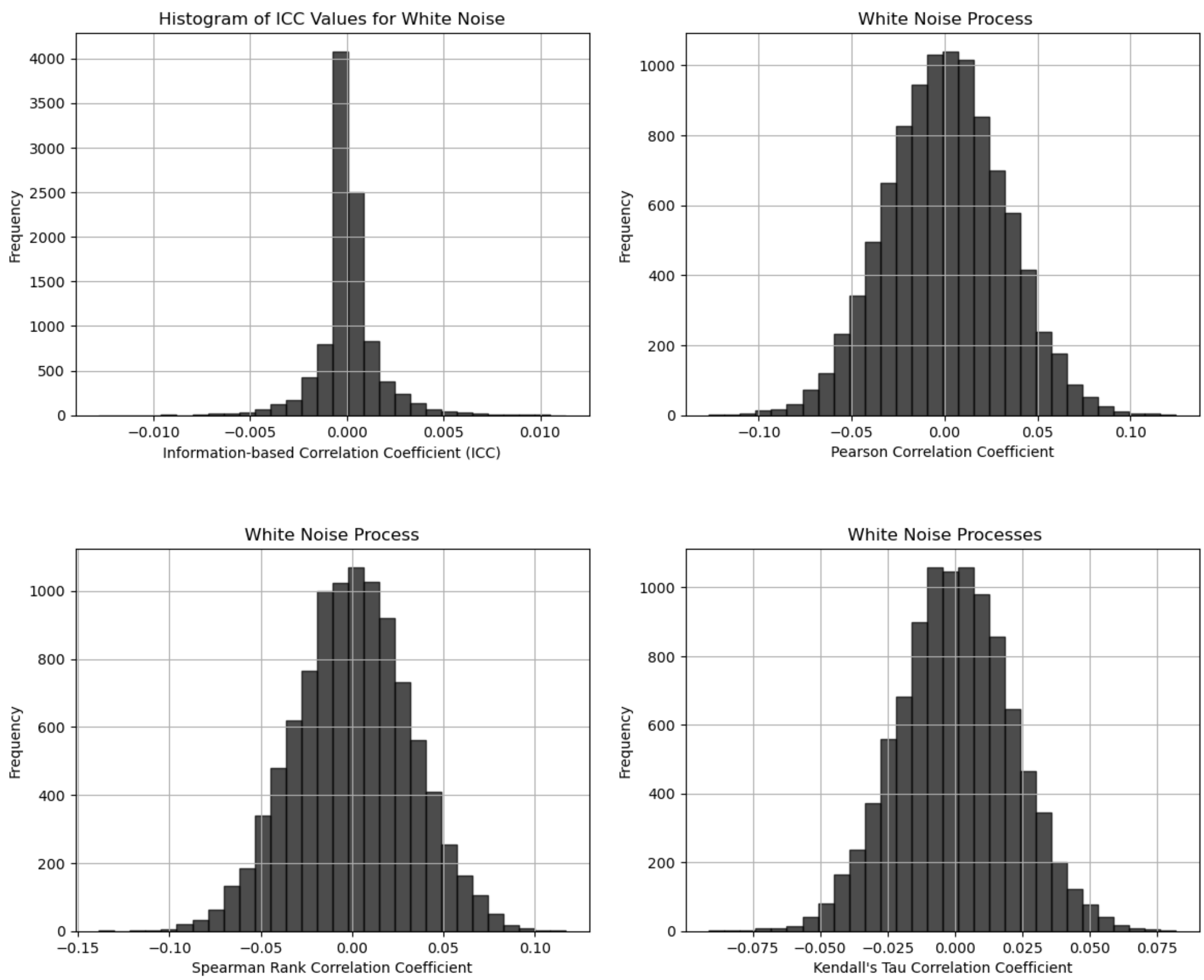


Histogram for 10000 simulations of $\sqrt{n}(\widehat{ICC} - ICC)$ in the case $p \neq \frac{1}{2}$. From theorem 4 for $p \neq \frac{1}{2}$, $\mathcal{N}\left(0, V\right)$ to check the validity of the theorem.

2.3 Comparision of ICC With other Correlation Coefficients

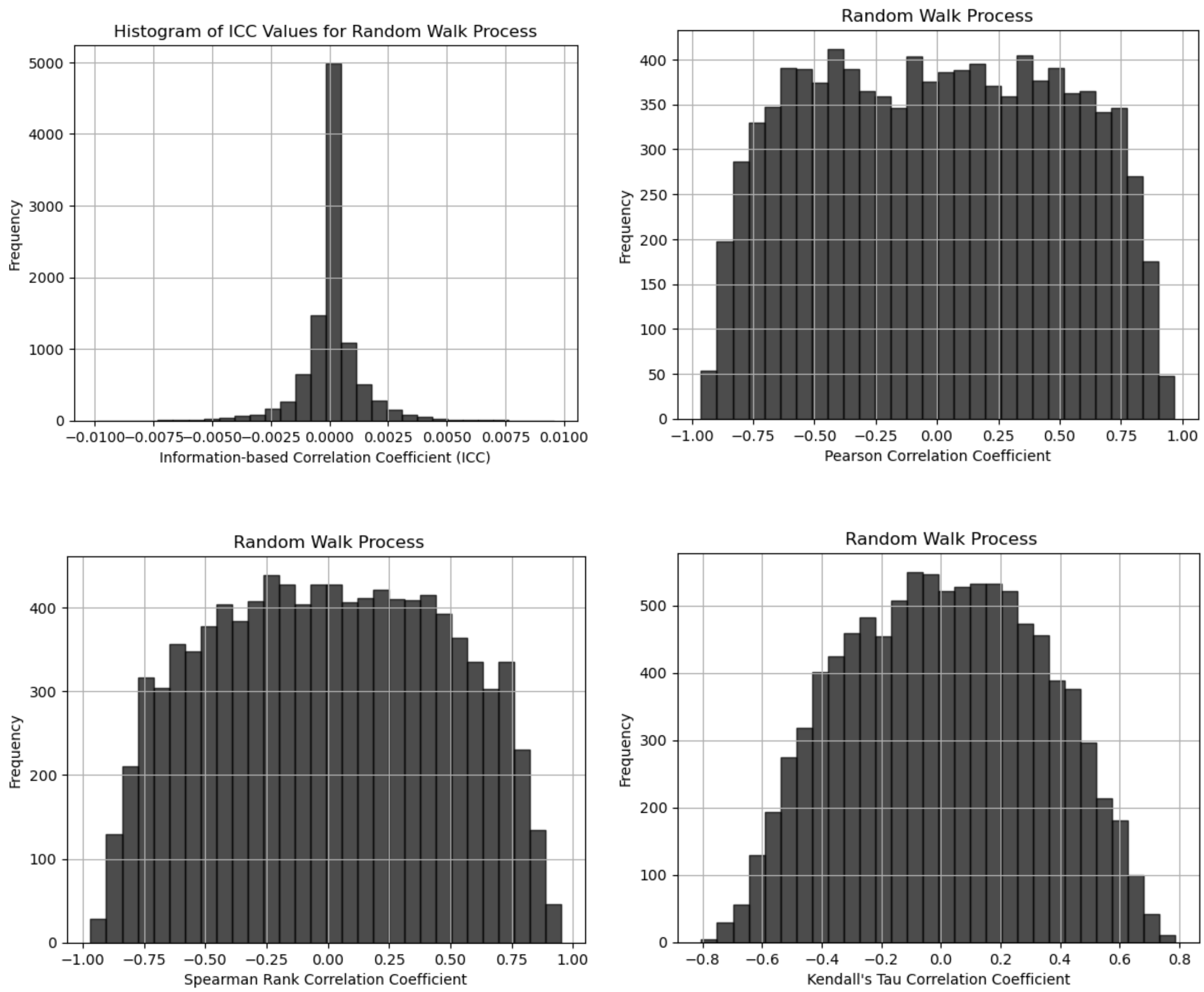
In this section, we will compare different correlation coefficients for the Independent white noise process, the Random walk process, the Student's t process, and the AR(1) process using Monte Carlo simulation for uncorrelated randomly generated data.

1. **For White Noise Process**:
   A white noise process is a sequence of random variables that are independent and identically distributed. With each variable having a constant mean and variance.
   A common example of a white noise process is a sequence of random variables X_t with mean Zero, i.e., $E(X_t) = 0$ and constant variance $Var(X_t) = \sigma^2$ and $E(X_t * X_h) = 0$ when $t \neq h$. Where t represents the index of the observation. A white noise process can represented by $X_t \sim IID(0, \sigma^2)$ and the sequence of white noise id generated by $X_t \sim N(0, \sigma^2)$.



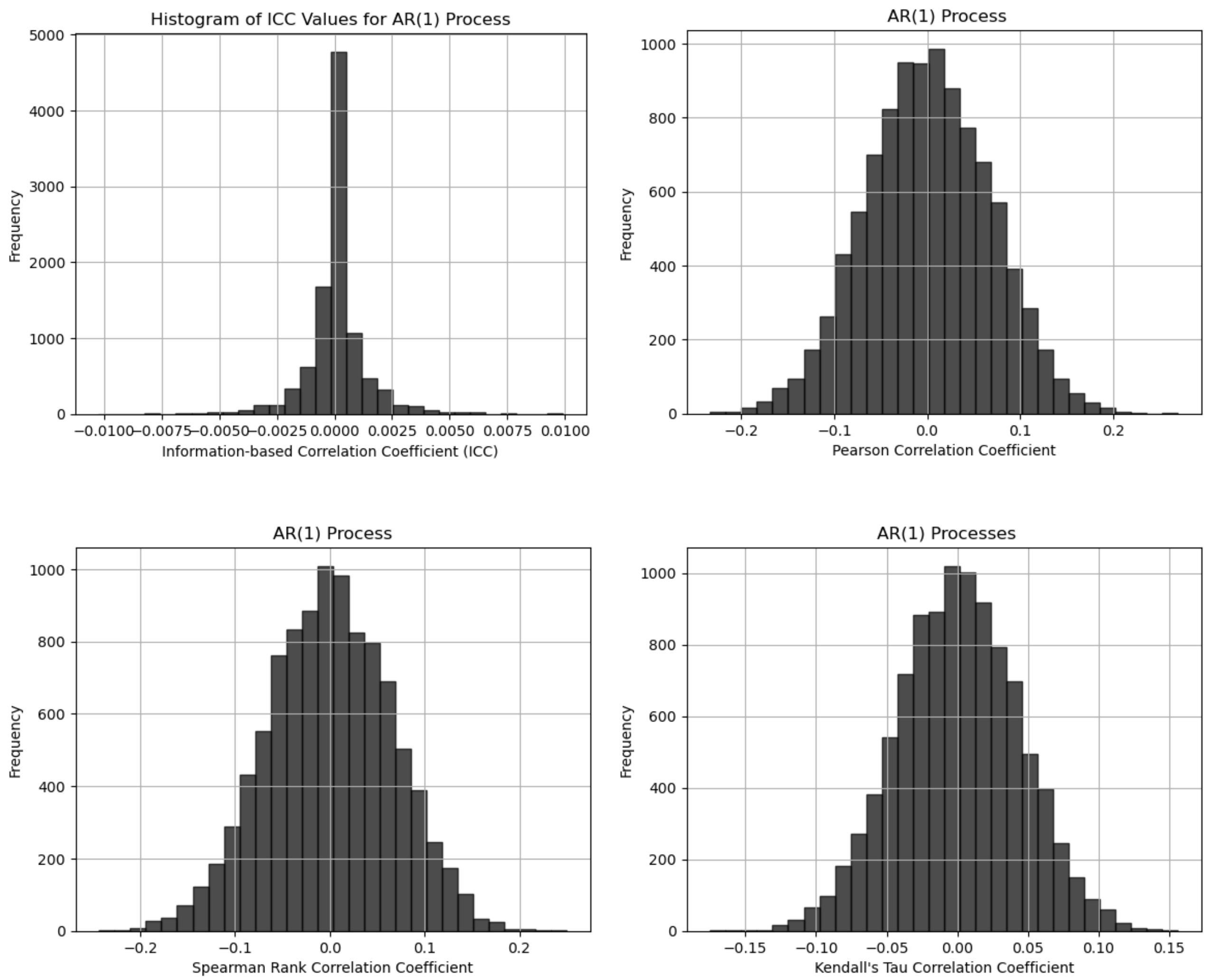
The figure shows an estimation of ICC, Spearman rank correlation, Pearson correlation, and Kendall's tau correlation for the independent white noise process via Monte Carlo simulation. For each simulated series n=1000 and iterations=10000 for $p \neq \frac{1}{2}$.

2. **Random Walk Process**
   It is a process to determine the probable location of a point subject to random motion. To define Random walk, Take independent random variables $\epsilon_1, \epsilon_2, \epsilon_3 ...$ where each variable is +1 or -1 with a probability of 50% of either value. Consider $S_0 = 0$ and $S_n = S_{n-1} + \epsilon_n$ where S_n is the value of random walk at time n. The series S_n is the net distance walked. If the n value increases then Expectation of S_n is 0 i.e. $E(S_n) = 0$
   We can write $S_t = \epsilon_1 + \epsilon_2 + \epsilon_3 + ... + \epsilon_t$ then $E(S_t) = E(\epsilon_1) + E(\epsilon_2) + E(\epsilon_3) + ...E(\epsilon_t) = 0$. Consequently, $\sum_{i=1}^{t} E(\epsilon_i) = 0$. For a one-dimensional continuous random walk distribution of the walker's position after t steps approached a normal distribution as t becomes large. The random walk-generation process of the form $S_n = S_{n-1} + \epsilon_n$ with $\epsilon \overset{iid}{\sim} N(0, 1)$



The figure shows an estimation of ICC, Spearman rank correlation, Pearson correlation, and Kendall's tau correlation for the Random walk process via Monte Carlo simulation. For each simulated series n=1000 and iterations=10000 for $p \neq \frac{1}{2}$.

3. **AR(1) Process**
In the AR(1) process, AR is for autoregressive, and '1' is for the first-order autoregressive process. Let $\epsilon_1, \epsilon_2, \epsilon_3, ...$ be white noises then at time t the value of AR(1) process is given by $X_t = \phi X_{t-1} + \epsilon_t$. Where ϕ is the autoregressive parameter that determines the impact of X_t on X_{t-1} and AR(1) model applied for uncorrelated stationary data only for which $|\phi| < 1$. This process is used for those time series data where the value at t is linearly dependent on the value at $t-1$.



The figure shows an estimation of ICC, Spearman rank correlation, Pearson correlation, and Kendall's tau correlation for the AR(1)process via Monte Carlo simulation. For each simulated series n=1000 and iterations=10000 for $p \neq \frac{1}{2}$ for $\phi = 0.8$.

From the above comparisons of ICC with other correlations, it is clear that the ICC estimator has a more frequent 0 for uncorrelated data than other correlations. Which shows that ICC is a better estimator than other correlations.

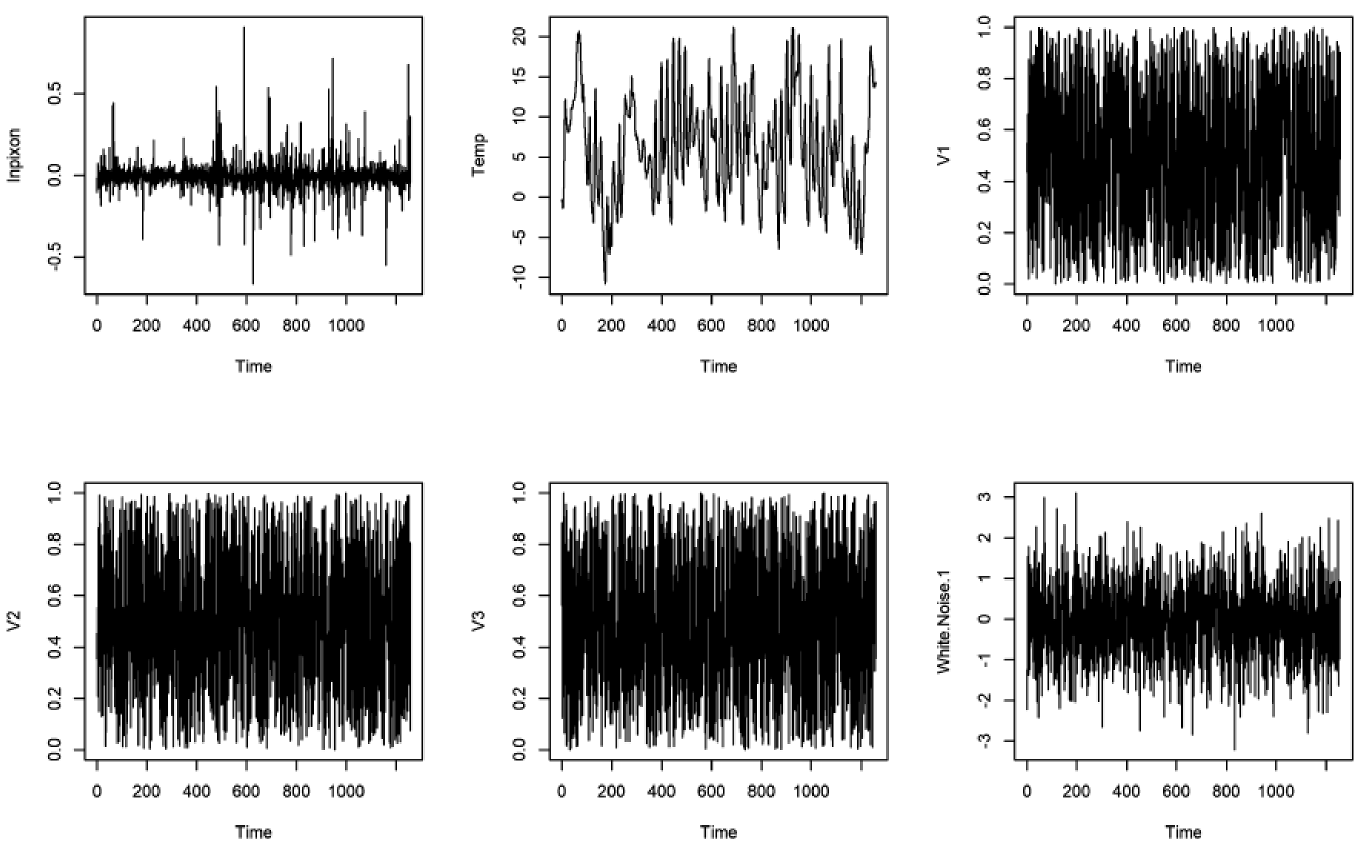
2.4 Limitations of ICC

Although ICC performs better than many other correlation coefficients but it also has some limitations like (1) Computational complexity ICC has a greater time complexity than other coefficients. Because it requires to construction and estimation of many factors like dichotomized series and matching series to estimate probability(p) and entropy(H) and (2) For large samples of data like finance, ICC estimation is unbiased but when it comes to small sample of data the biasness term is negligible. We can see this using the equation $\mathbb{E}(\hat{H}) = H - \frac{1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^{2} \frac{1}{p_i}\right) + \mathcal{O}\left(\frac{1}{n^3}\right)$. Moreover, ICC performs better when it comes to spurious correlation, stationary data, Chaotic time series...etc

2.5 Applications of ICC

In this section, we will see how ICC is better than Pearson's in detecting the absence of correlation between two-time series data.

1. **Independent Stationary Time Series**
   Stationary data are non-periodic time series data whose mean and variance don't change over time. Consider six variables as a daily return of INPIXON stock from Jan 2015 to Jan 2020, Random temperature in degrees Celsius such that data is stationary. Three random stationary processes V_1, V_2, V_3 and a white noise process(WN1) such that all six variables have the same length.



From the above time-series plot of six stationary variables, It is obvious that these six time series are completely uncorrelated. If we calculate ICC between any two, its value should be near ZERO, and the correlation that gives the nearest value to zero is the best correlation.

(a) Matrix for Pearson Correlation for above stationary variables.

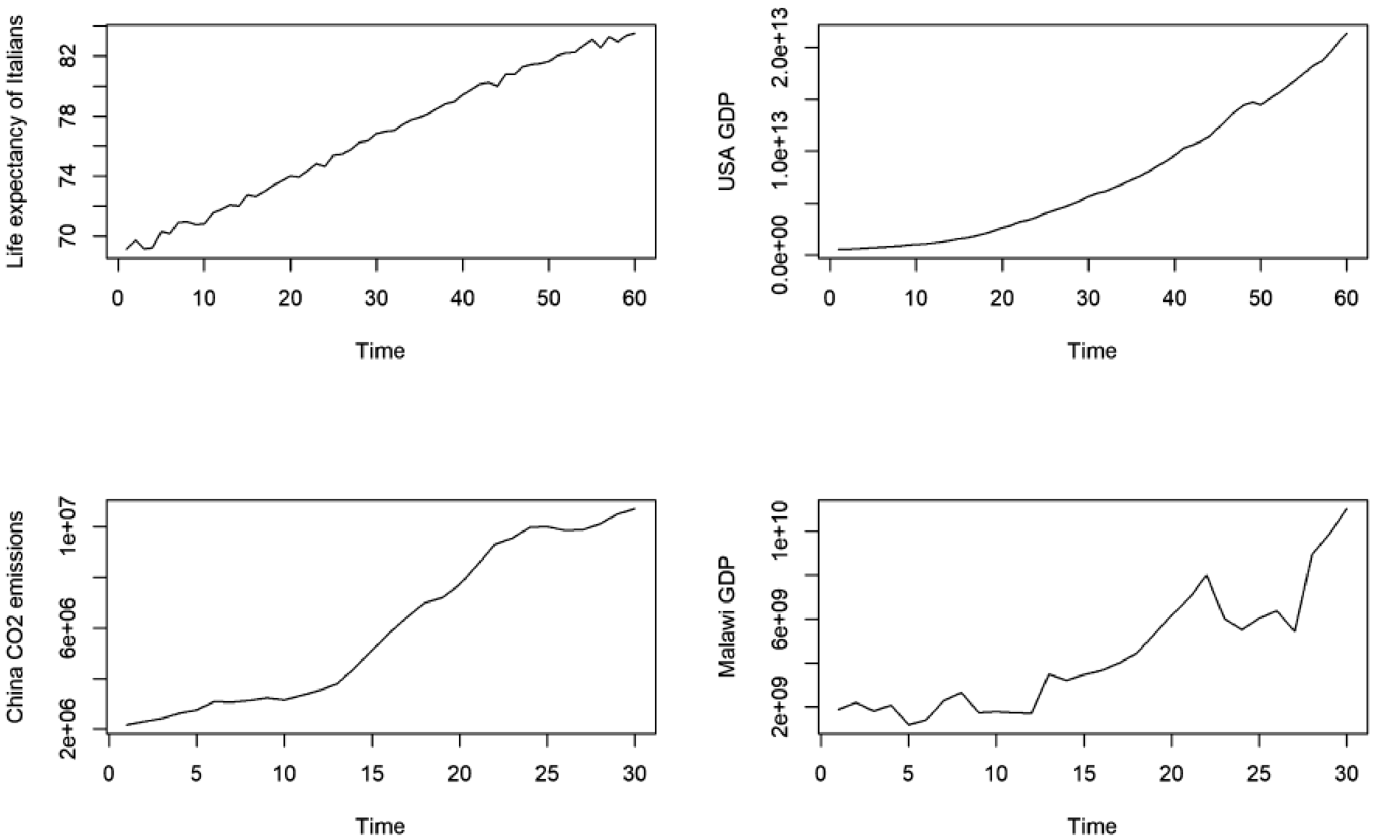
Pearson	INPIXON	Temp.	V_1	V_2	V_3	WN(1)
INPIXON	1					
Temp.	0.0912	1				
V_1	0.0527	0.0095	1			
V_2	0.0127	0.0759	-0.0115	1		
V_3	0.0431	-0.0221	-0.0254	0.0049	1	
WN(1)	-0.0139	0.0252	-0.0379	-0.0092	-0.0075	1

(b) Matrix for ICC value for above stationary variables.

ICC	INPIXON	Temp.	V_1	V_2	V_3	WN(1)
INPIXON	1					
Temp.	0.0000	1				
V_1	0.0009	-0.0001	1			
V_2	-0.0001	0.0001	-0.0001	1		
V_3	0.0003	0	-0.0001	0.0004	1	
WN(1)	0.0016	0.0029	-0.0028	0.0013	0	1

From matrix (a) and (b), it is clear that the ICC value is closer to zero than the Pearson correlation coefficient for each data. This shows ICC is more reliable than the Pearson correlation coefficient. If we compare ICC with other coefficients, then again, we will get the same result that ICC performs better when it comes to stationary data.

2. **Spuriously Correlated Time Series**
A spurious correlation is a situation where two non-stationary variables appear to be correlated because of their same trend over time but are not correlated in real.
Consider four non-stationary time series variables. Life expectancy of Italians at the time of birth from 1960 to 2019, the USA's GDP between the years 1960 and 2019, CO2 emission during the years 1990 and 2019, and Malawi's GDP between the years 1990 and 2019. It is clear that when we find different correlations between the life expectancy of Italians and the GDP of the USA, their correlation value should be around zero as these variables are uncorrelated. Also, when we find the correlation between the CO_2 emission of China and the GDP of Malawi, their value should be around zero as these variables are completely uncorrelated.



Considered four spurious correlated variables

Correlations values for the life expectancy of Italians and USA GDP are

- Pearson= 0.9580
- Spearman= 0.9985
- ICC= 0.3060

Correlation values for China's CO2 emission and Malawi's GDP are

- Pearson= 0.9164
- Spearman=0.8798
- ICC=0.0216

Since all four spurious variables are equivalent in nature(Increase with time), correlations like Pearson and Spearman show value near one. But in reality, these variables are completely independent. Here, the ICC value is closer to zero. Hence, ICC performs better in the case of spurious correlation.

2.6 Conclusions

In many practical applications, ICC can be used as an alternative to traditional coefficients. Pearson and Spurious show a high correlation in the case of spurious correlation, whereas ICC gives a value closer to Zero than spurious and Pearson.
We can list some of the benefits of ICC as follows:

- ICC Provides a simple calculation with respect to other correlations.
- It offers many properties traditional correlation coefficients offer, such as symmetry and boundedness. between -1 and 1.
- Analyzing the presented simulations as the figure, it performs better than other correlation coefficients for white noise processes, random walks, and AR(1)processes.
- ICC is less sensitive to outliers.
- "ICC is less influenced by Spurious correlation than other coefficients.".

ICC can be applied practically in weather forecasting, signal processing and audio recognition, chaotic time series, and many others.

2.7 Future Work

1. **Extension to multivariate time series**
   This involves considering more than two time series simultaneously and developing a measure that can measure the information shared in these series.
2. **Applications in Signal Processing**
   The applications of ICC in signal processing, such as detecting patterns or correlations in signals with a focus on information content. This can be particularly relevant in fields like communication systems.
3. **Generalization of correlation for a generalized entropy**

2.8 References

- Guiseppe Pernagallio, An entropy-based measure of correlation for time series.
- B. Harris, The statistical estimation of entropy in the non-parametric case.

Chapter 3

Analysis on Skewness and Kurtosis for non-Gaussian distributions

3.1 Skewness of a distribution

Skewness measures the asymmetry of a distribution. A distribution could be left-skewed, right-skewed, or zero-skewed. Skewness is a crucial measure in understanding the shape and characteristics of a dataset, providing insights into its underlying distribution.

Skewness is defined using the third moment of a distribution as:

$$S = \frac{1}{\sigma^3}\left[\frac{1}{N}\sum_k (x_k - \mu)^3\right] \tag{3.1.0.1}$$

Where μ shows the mean of the distribution, σ shows the standard deviation, and N shows the number of data points.

Types of Skewness:

- **Positive Skewness** ($S > 0$)**:** In a Positive-skewed distribution, the peak is longer on the right side. It indicates that outliers typically occur on the right side of the distribution. Mean is greater than the median in this case of distribution.
- **Negative Skewness** ($S < 0$)**:** In a Negative-skewed distribution, the peak is longer on the left side. It indicates that outliers typically occur at the left side of the distribution. The mean is less than the median in this case of distribution.
- **Zero Skewness** ($S = 0$)**:** When a distribution has zero skews, it is symmetrical. The mean and median have the same value. eg. normal distribution

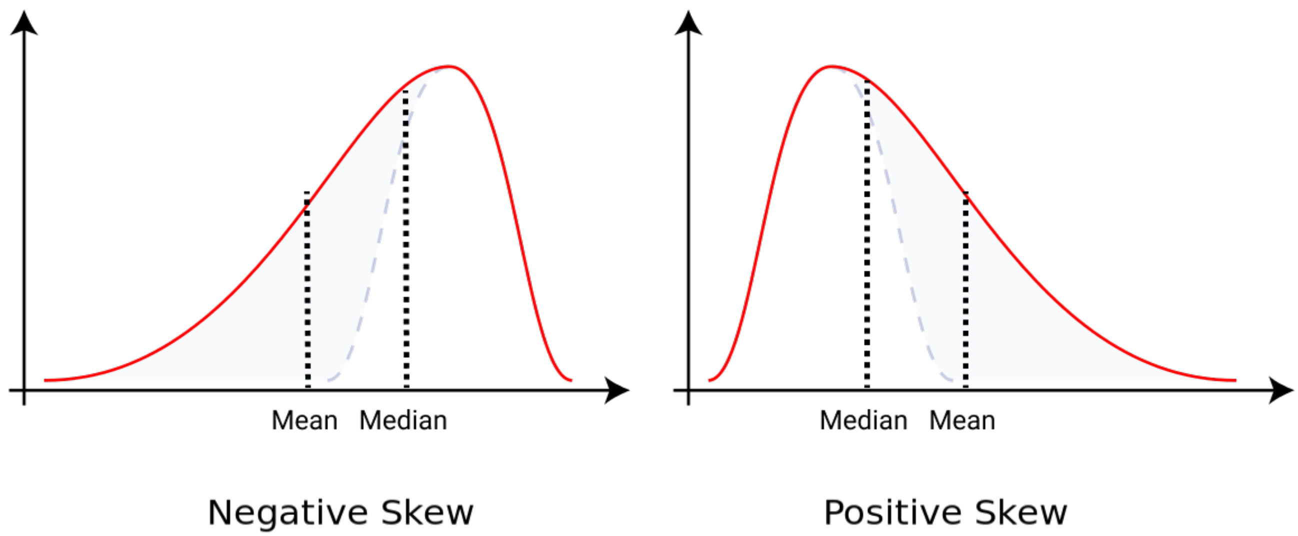


Fig.1. The figure shows how an outlier can affect the symmetry of a distribution. Fig (a) for negative skew shows outliers on the left tail. Fig (b) for positive skew shows outliers on the right tail.

3.2 Kurtosis of a distribution:

Kurtosis measures the degree of the distribution's peakedness and tail thickness. It determines whether a distribution is Gaussian or not by measuring its deviation from a normal distribution.

Kurtosis is defined using the fourth moment of a distribution as:

$$K = \frac{1}{\sigma^4}\left[\frac{1}{N}\sum_k (x_k - \mu)^4\right] \tag{3.2.0.1}$$

Where μ shows the mean of the distribution, σ shows the standard deviation, and N shows the number of data points.

Types of Kurtosis:

- **Mesokurtic:** Mesokurtic is a medium-tailed distribution that shows outliers are neither highly frequent nor less frequent in the distribution. Kurtosis value K =3 in this case. e.g., Normal distribution.
- **Platykurtic:** Platykurtic is a small-tailed distribution that shows outliers are less frequent in the distribution. Kurtosis value K <3 in this case. e.g., Uniform distribution
- **lepokurtic:** lepokurtic is a long-tailed distribution that shows outliers more frequent in the distribution. Kurtosis value K >3 in this case. e.g., Laplace distribution

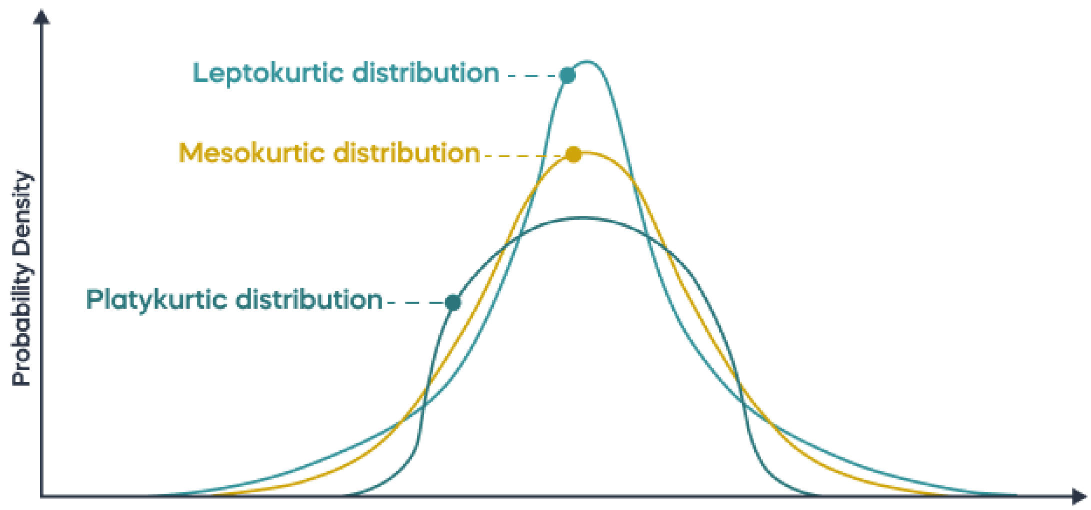


Fig.2. The figure shows how outliers affect the tails of a distribution.

3.3 A Skewness-Kurtosis relation

In the previous section, we have seen the main purposes of Skewness and Kurtosis. In this section, we will discuss a $\frac{4}{3}$ power law relation between Skewness and Kurtosis.

$$K = \frac{1}{\sigma^4}\left[\frac{1}{N}\sum_k (x_k - \mu)^4\right]$$
$$S = \frac{1}{\sigma^3}\left[\frac{1}{N}\sum_k (x_k - \mu)^3\right]$$

The Skewness-Kurtosis relation follows a 4/3 power-law. To explain this behavior, Consider there is an outlier in the data set; this outlier will dominate the summation of other data points, and the other data points can be considered negligible. Then, the Skewness and Kurtosis values turned out to be

$$S \approx \frac{\frac{(x-\mu)^3}{N}}{\sigma^3}$$

and

$$K \approx \frac{\frac{(x-\mu)^4}{N}}{\sigma^4}$$

following these two equations we have

$$\frac{(x-\mu)}{\sigma} \approx (NS)^{1/3}$$

and

$$\frac{(x-\mu)}{\sigma} \approx (NK)^{1/4}$$

If we solve the above two equations, we can get the 4/3 power law Skewness-Kurtosis relation as:

$$K \approx (N)^{1/3}(S)^{4/3} \tag{3.3.0.1}$$

The purpose is to find answers to the following questions:

1. Is a high value of Kurtosis always indicative of a significant deviation from the normal distribution?

2. If not, How can we compare two distributions that are not Gaussian?

3. Does changing the definition of kurtosis provide an advantage in comparing non-Gaussian distributed data sets?

We are going to discuss these questions in upcoming sections for a particular type of non-Gaussian distribution(q-Gaussian distribution). In the next section, we describe how to Synthesize q-Gaussian distributions. Then, we try to find answers to the three questions raised above.

3.4 Generating q-Gaussian distributions

One of the most important methods to generate Gaussian distribution is the Box-Muller method. However, not all data systems exhibit Gaussian distribution. Many distributions exhibit q-Gaussian distribution. These q-Gaussian distributions are optimized by non-additive entropy $S_q = \frac{1-\sum_{i=1}^{N}(p_i)^q}{q-1}$. For the limiting case $q \longrightarrow 1$, it holds the Shannon's entropy.

If the value of q ranges between 1 and 3, then the distributions that follow q-Gaussian are long-tailed non-Gaussian distributions. These distributions have a finite second moment for $1 < q < \frac{7}{5}$ and have infinite second moment for $\frac{7}{5} < q < 3$. In this section, we will generalize the Box-Muller method to generate a q-Gaussian distributed data set. Suppose U_1 and U_2 are independent random variables from the uniform distribution defined on (0, 1). Then two random variables Z_1 and Z_2 are defined as

$$Z_1 = \sqrt{-2\ln_{q'} U_1}\cos(2\pi U_2) \tag{3.4.0.1}$$

$$Z_2 = \sqrt{-2\ln_{q'} U_1}\sin(2\pi U_2) \tag{3.4.0.2}$$

Then Z_1 and Z_2 are q-Gaussian distributed. Where, $q' = \frac{1+q}{3-q}$ and $ln_q(x)$ is defined as

$$\ln_q(x) = \frac{x^{1-q}-1}{1-q}; x > 0 \tag{3.4.0.3}$$

Now, the q-Gaussian distribution is:

$$p(x;\mu_q;\sigma_q) = B_q\sqrt{C_q}[1+(q-1)C_q(x-\mu_q)^2]^{\frac{1}{1-q}} \tag{3.4.0.4}$$

Where B_q is a normalization factor and C_q is the width of distribution and are defined as:

$$B_q = \begin{cases} \frac{\Gamma[\frac{5-3q}{2(1-q)}]}{\Gamma[\frac{2-1}{1-q}]}\sqrt{\frac{1-q}{\pi}} & \text{if } q < \ 1 \\ \frac{1}{\sqrt{\pi}} & \text{if } q = \ 1 \\ \frac{\Gamma[\frac{1}{q-1}]}{\Gamma[\frac{3-q}{2(q-1)}]}\sqrt{\frac{q-1}{\pi}} & \text{if } 1 \ < q < \ 3 \end{cases} \tag{3.4.0.5}$$

and

$$C_q = [(3-q)(\sigma_q)^2]^{-1} \qquad \text{if } q \in (-\infty, 3) \tag{3.4.0.6}$$

Where μ_q is q-mean, and σ_q is q-variance are the first and second q-moments respectively, and the general definition of q-moment is given by

$$\langle x^n \rangle_{q_n} = \frac{\sum_{i=1}^{N} (x_i - \mu_q)^n \left[p(x_i)\right]^{q_n}}{\sum_{i=1}^{N} \left[p(x_i)\right]^{q_n}} \tag{3.4.0.7}$$

where $p(x)$ is the probability distribution of x and

$$q_n = 1 + \frac{n}{2}(q-1) \tag{3.4.0.8}$$

Generating q-Gaussian distributed data sets can be achieved by adjusting the q-value.. q=1 shows a normal distribution.

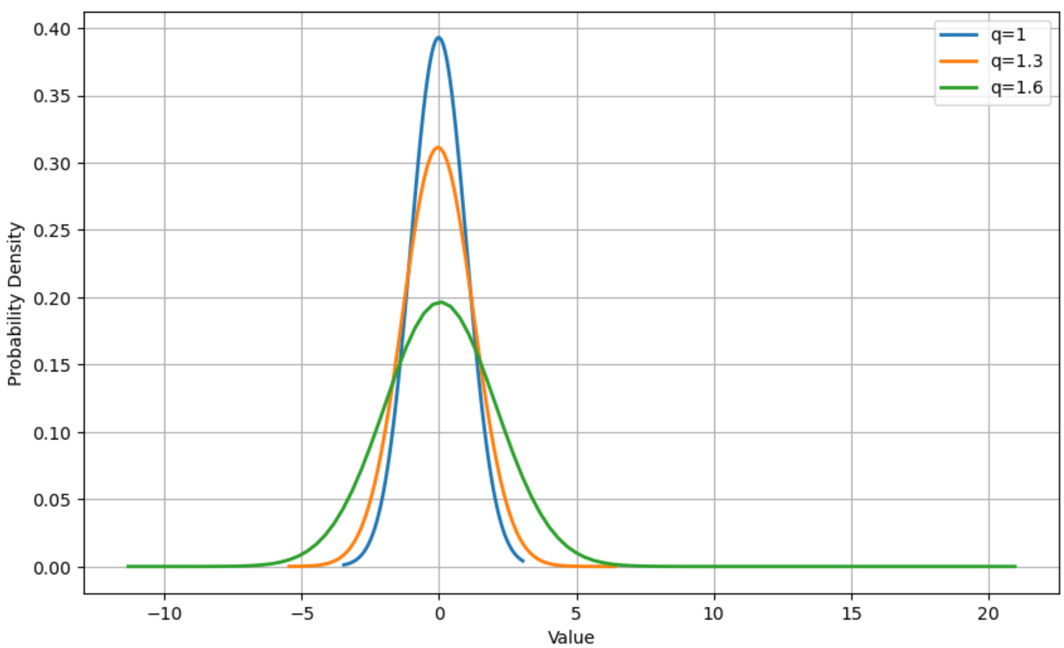


Fig.3. q-Gaussian distribution for different q values. The figure shows that the distribution becomes longer-tailed as we increase the q value.

3.5 Standard Skewness and Kurtosis for q-Gaussian distributions

In this section, we will discuss how changing the number of data points and q value affects the Kurtosis of a q-Gaussian distribution.
For this, we will generate q-Gaussian distributed data sets for different numbers of data points(N) for different q values. Then, we simulate it 1000 times. The data generated in each simulation can be referred to as a window. Now, for each data set, we calculate standard Skewness and Kurtosis. Now, we test the Skewness and kurtosis 4/3 exponent power law relation as mentioned above.

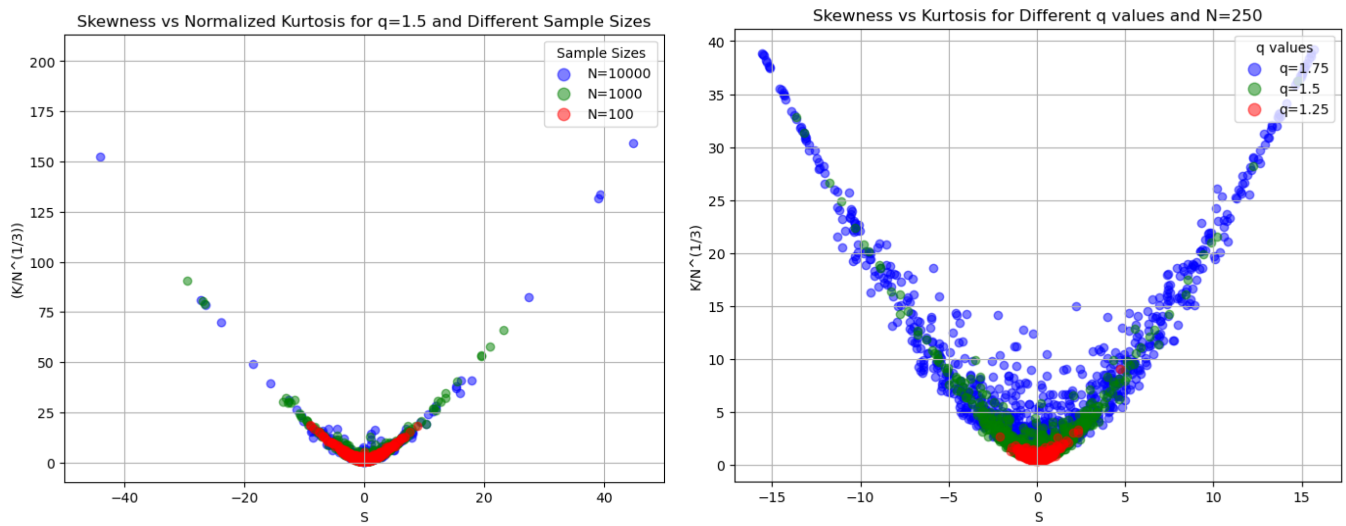


Fig.4. Kurtosis v/s Skewness for q=1.5 and different values of N and q.

By analyzing the figure, we can observe how the behavior of the S-K changes with changing N and q. In the above plot of Skewness v/s Kurtosis, it is clear that as the value of N or q increases, Kurtosis reaches a higher value.
From Fig.3, it is clear that as the q-value increases, the distribution becomes more long-tailed, and the possibility of discovering extreme events will be expected to rise. The same is true for increasing N. If we increase N. From Fig.4(a), although the data from each window are from the same distribution, the window with more frequent data from the tail has a larger kurtosis. It is important to note that a higher kurtosis value does not necessarily indicate a more long-tailed distribution. It shows only that the contribution of the tail in that window is higher than the other window.
The analysis in Fig.4(a) is enough to answer our first question that a relatively large value of kurtosis doesn't mean a higher deviation from Gaussian. It only shows that the contribution on the tail is higher for that distribution.
Let's take a look at Figure 4(b). We are dealing with two distributions that have different q-values. The distribution with a higher q value has a shape with a more elongated tail, which means it is expected to have a higher kurtosis. However, it's important to note that extreme events are randomly distributed throughout the entire data set, so it's not guaranteed that any given window will accurately represent the distribution's tails. This is especially true for smaller data sets where most data may come from the central part of the distribution.

Kurtosis saturated to a finite value for $1 < q < 7/5$ for any finite N. The result can be seen in Fig.5 and Fig.6.

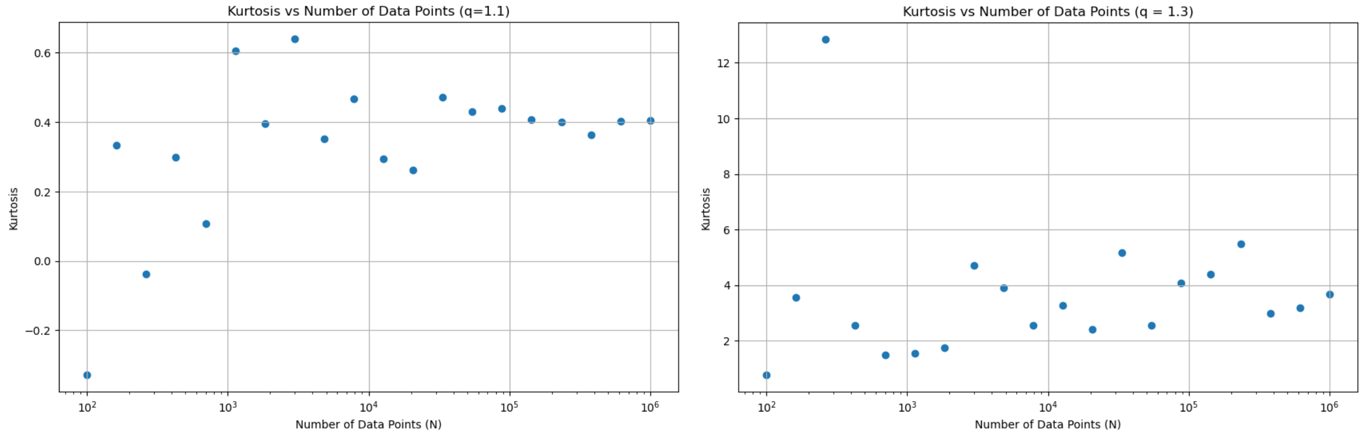


Fig.5 Plolt for Kurtosis v/s N for $q < 7/5$

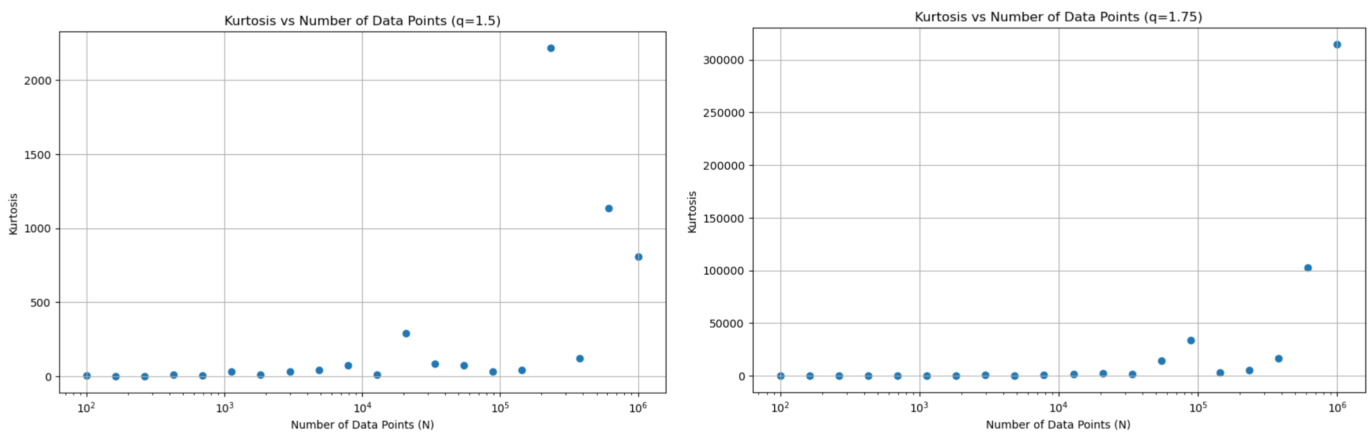


Fig.6 Plot for Kurtosis v/s N for $q > 7/5$

From the above figures, it is clear that Kurtosis finds a fixed value for $q < 7/5$ as N is large, and Kurtosis doesn't saturate for even a large value of N in the case of $q > 7/5$. The value of saturation increases with increasing q, and The distribution is moving further away from the normal or Gaussian distribution.
Now the part concludes that for sufficiently large N and $q < 7/5$, Kurtosis will no longer be dependent on N and become independent, and the Kurtosis(Fourth moment) will be finite in the case, and for the case, $q > 7/5$ the distribution has an infinite fourth moment, i.e., K doesn't saturate but increases with N.
The analysis in Fig.5 and Fig.6 is enough to answer our second question that we can compare two non-Gaussian distributions by using their Kurtosis value with sufficient large N and $1 < q < 7/5$.
In the next sections, we will try to find a way to calculate how much is the deviation from Gaussian.

3.6 q-Skewness and q-Kurtosis for q-Gaussian distribution

We can define q-Skewness and q-Kurtosis for a q-Gaussian distribution as follows:

$$S_q = \frac{\langle x^3 \rangle_{q3}}{\langle x^2 \rangle_{q2}^{3/2}} = \frac{\frac{\sum_{i=1}^{N}(x_i-\mu_q)^3[p(x_i)]^{q3}}{\sum_{i=1}^{N}[p(x_i)]^{q3}}}{\left(\frac{\sum_{i=1}^{N}(x_i-\mu_q)^2[p(x_i)]^{q2}}{\sum_{i=1}^{N}[p(x_i)]^{q2}}\right)^{3/2}} \tag{3.6.0.1}$$

$$K_q = \frac{\langle x^4 \rangle_{q4}}{\langle x^2 \rangle_{q2}^{2}} = \frac{\frac{\sum_{i=1}^{N}(x_i-\mu_q)^4[p(x_i)]^{q4}}{\sum_{i=1}^{N}[p(x_i)]^{q4}}}{3\left(\frac{\sum_{i=1}^{N}(x_i-\mu_q)^2[p(x_i)]^{q2}}{\sum_{i=1}^{N}[p(x_i)]^{q2}}\right)^{2}} \tag{3.6.0.2}$$

Where N is the number of data points,μ_q is the q-mean of the sample calculated using the definition of q-moments and $q_n = 1 + \frac{n}{2}(q-1)$. For $q \longrightarrow 1$, the definition goes over a standard definition of Skewness and Kurtosis.
Using the definition of q-Kurtosis, the analytical expression can be obtained as:

$$K_q = \frac{3-q}{1+q} \tag{3.6.0.3}$$

It was previously mentioned that kurtosis can be misleading for insufficient data sets when $1 < q < 7/5$ and for any N when $7/5 < q < 3$. Now, we need to determine whether q-kurtosis will yield more satisfactory results.
To do this, we are generating q-Gaussian distributed data with different q values, which we will refer to as q_{dist}. In cases where we don't know the prior q_{dist} of the distribution, we will use an arbitrary q-value denoted as q_{ref} to calculate q-skewness and q-kurtosis.
We generate the data points using the q_{dist} value and then calculate q-Skewness and q-Kurtosis using the q_{ref} value then we plot the q-skewness(S_q) v/s q-Kurtosis(K_q) for different q_{dist} values and a q_{ref} value using the power law relation for 1000 simulations.

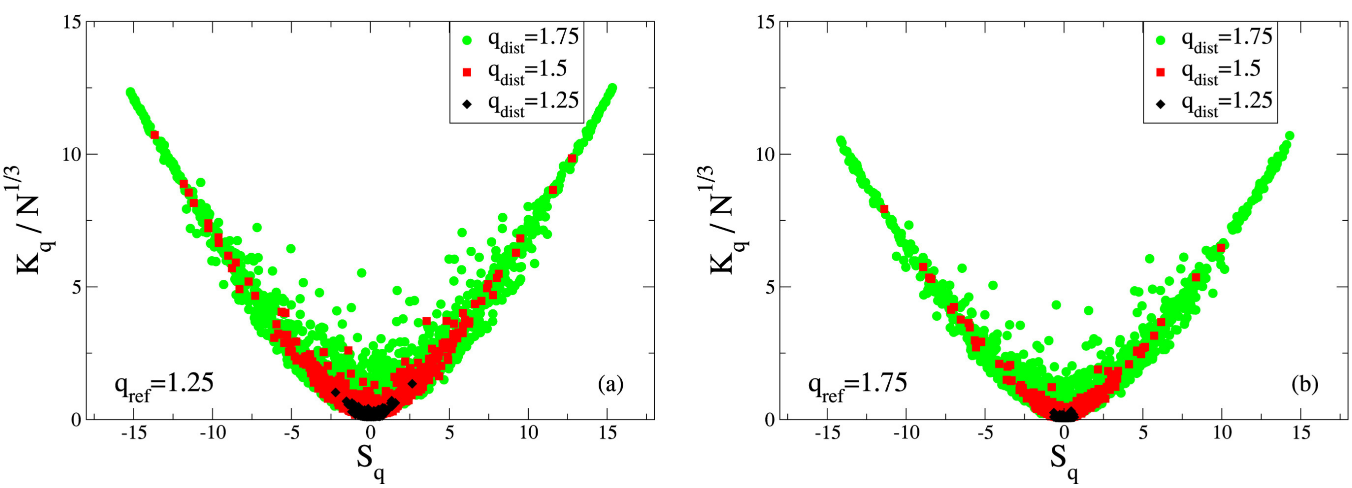
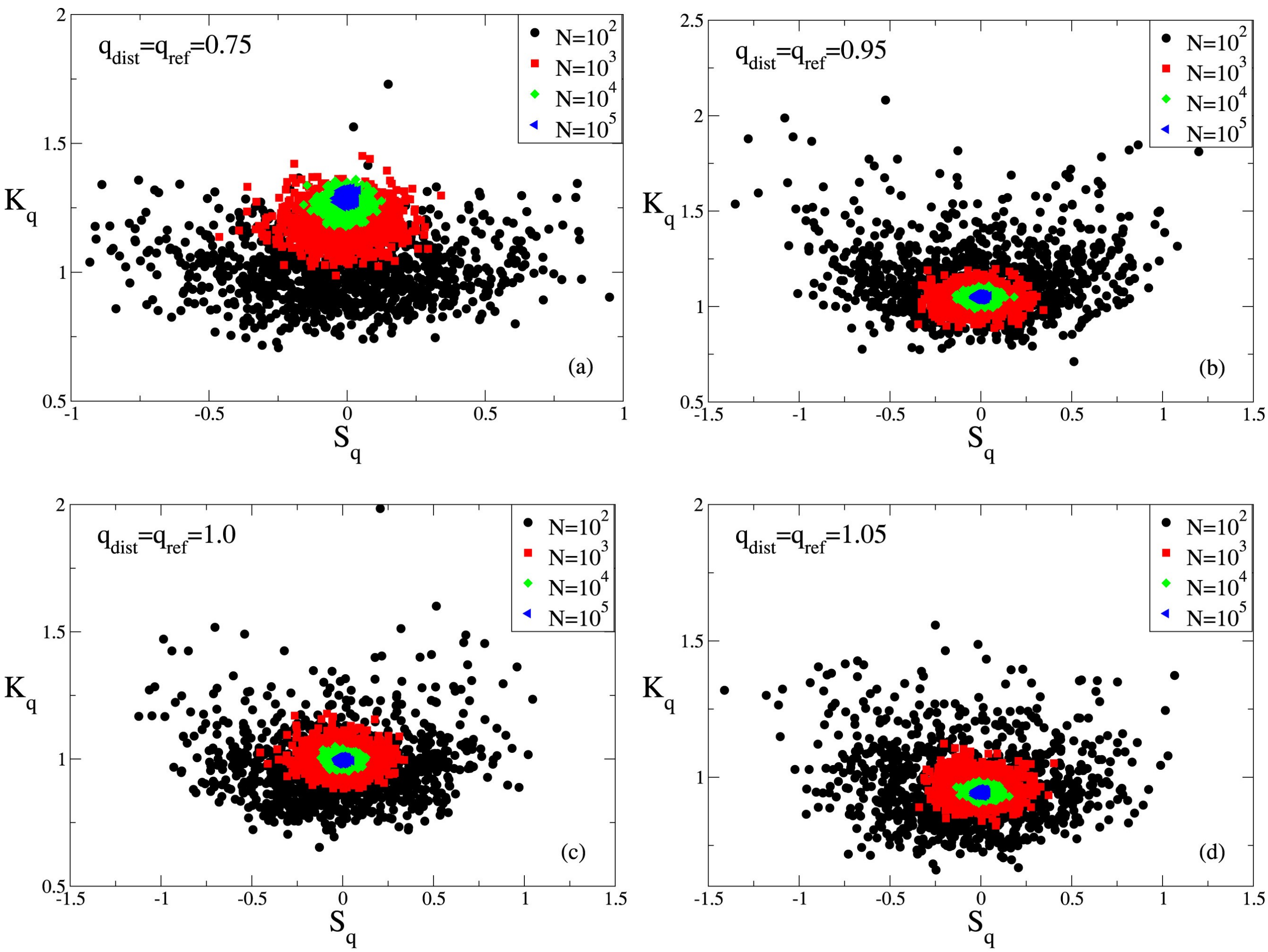


Fig.7 q-Kurtosis as a function of q-Skewness for N=250, and (a) $q_{ref} = 1.25$ and (b) $q_{ref} = 1.75$

Now, we have two different cases that arise from here, $q_{ref} \leq q_{dist}$ and $q_{ref} \geq q_{dist}$. Fig.(7a) and (7b) are the plottings of the first and second cases, respectively. In Fig.7, each point refers to a window of size N=250. Fig.7a is observed for $q_{ref} = 1.25$ and $q_{dist} = 1.25$ distribution accumulates at about a point close to $S_q = 0$ on the $S_q - K_q$ plane. Whereas for $q_{dist} = 1.5$ and $q_{dist} = 1.75$, there is no clustering, and it also shows that q-Kurtosis increases as q_{dist} increases. But in Fig.7b, when q-Skewness and q-Kurtosis are calculated using $q_{ref} = 1.75$, the maximum value of q-Kurtosis decreases. One can see that the clustering increases for a higher q_{ref} value. For a small q_{ref}, clustering needs a small N. For e.g., for a Gaussian distribution $q_{dist} = q_{ref} = 1$, N=100 is sufficient for clustering. Whereas, for a larger $q_{ref} = q_{dist}$, we need a large value of N for clustering.

From these observations, one can say that q-moments are finite in the range of $1 < q < 3$ for large N values and points on $S_q - K_q$ plane collapses at a point. The location of collapsing the points and for size N for that collapsing occurs differs with q_{ref} or q_{dist}. From Fig.8, one can observe that if we increase $q_{dist} = q_{ref}$ Clustering shifts to a lower value in the plane as $q_{dist} = q_{ref}$ increases. It is also observed from the Fig that the clustering happens for even small N values when $q_{dist} = q_{ref}$ is around 1, and For a larger value of $q_{dist} = q_{ref}$ clustering happens for a very large value of N.



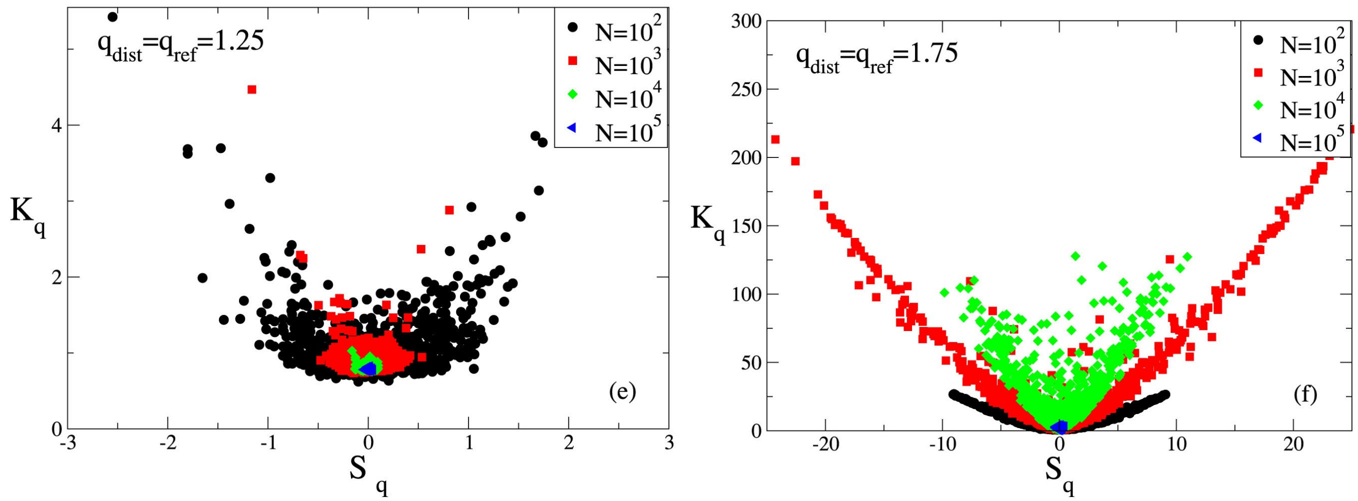


Fig.8 S_q v/s K_q planes for differnet $q_{dist} = q_{ref}$ and for different N.

If $q_{dist} = q_{ref}$, then this single point(cluster) on the plane corresponds to the value given by the analytical equation of kurtosis as mentioned in equation(3.6.0.3). But, if $q_{dist} \neq q_{ref}$ K_q converges to a finite value which is different from the analytical value for large N. Now, from these k_q values, we can try to find a way to decide which distribution is closer to Gaussian. for this, we use the plot between q-Kurtosis and the number of data points for two different q_{ref} values. We can see in Fig.8 that for $q_{ref} = 1.5$, the saturated K_q increases with increasing N. whereas, in Fig.9b, $q_{ref} = 2.0$ the saturated k_q values decrease with increasing N. The q-Kurtosis measure is only meaningful when the sample size (N) is large enough to saturate it. Otherwise, the measure may not provide an accurate representation of the data. Since there is no unique tendency of increase or decrease in different cases, it cannot be evidence to compare two different data sets. It may mislead the result if we still compare two data sets using q-Kurtosis for different q_{ref} values.

From eq.(3.6.0.3), it is clear that as we increase $0 < K_q < 1$, when $1 < q < 3$. K_q approaches to Zero(Numerically and Analytically) as q_ref value increases. Hence, there is no large distance between the K_q value(calculated when $q_{ref} = q_{dist}$) of the distribution and the point where K_q converges. This closeness can lead to mistakes. To overcome such problems, i.e., to compare two q-Gaussian distributions, one should calculate K_q numerically by using eq.(3.6.0.2), which we clear in the next section.

3.7 Localizing q-value of a distribution

In this section, we present a method to estimate the q-value of a q-Gaussian distribution. To find the q-value, we define it as:

$$\Delta K_q = \frac{K^{num}_{q_{ref}} - K^{an}_{q_{ref}}}{K^{an}_{q_{ref}}} \tag{3.7.0.1}$$

Where $K^{num}_{q_{ref}}$ and $K^{an}_{q_{ref}}$ are analytical and numerical value of q-Kurtosis respectively.

We have plotted a graph between ΔK_q and q_{ref} as shown in Fig.10. All the data sets exhibit two minima. The question is, which one of these minima is the correct one to localize the q_{dist} value of the distribution? As shown in the Inset, the points where ΔK_q intersects the zero line for the first time always correspond to the first minimum points in the main graphs. This first minimum serves as an indicator of the transition point of ΔK_q from positive to negative. From this point onwards, the analytical value of q-kurtosis overcomes its numerical value. After the second minimum, the behavior changes once again. The hump present in the main graphs is actually a negative value region.We know the value of q_{dist} for each dataset as we generate synthetic data using the q-Box-Muller method. This verifies that the appropriate q value corresponds to the first minimum point. We calculate ΔK_q (both numerically and analytically) as a function of q_{ref} to find the value of q_{dist} within the range of our precision of increment (0.01).

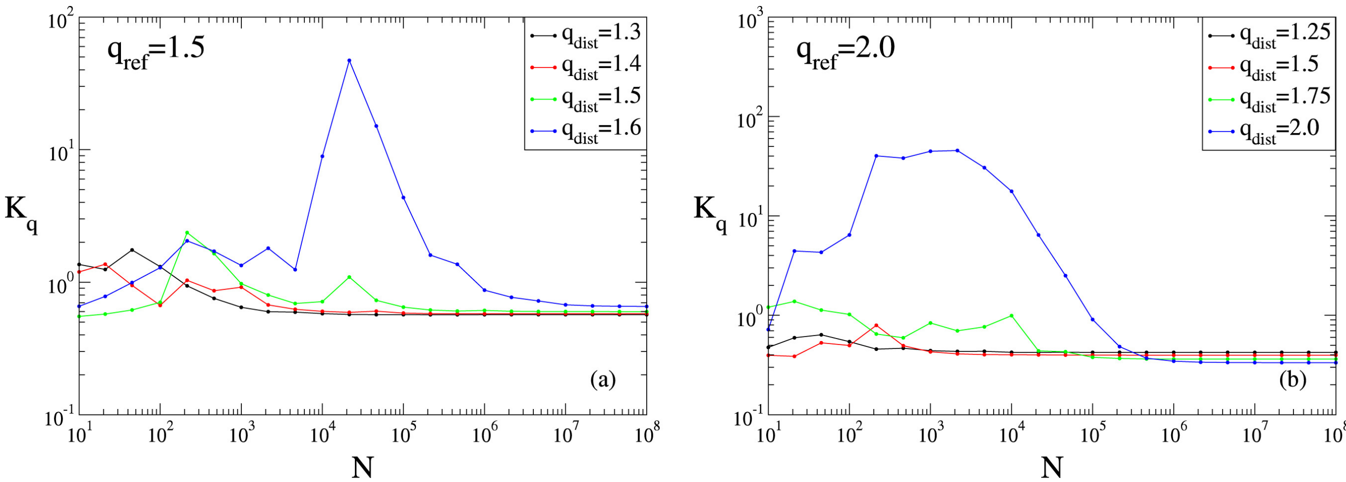


Fig.9 Plots of K_q v/s N for different q_{dist} values, (a) q_{ref}=1.25, and (b) q_{ref}=1.25

3.8 Conclusion

In our study, we aimed to clarify when it is appropriate to use skewness and kurtosis to analyze deviations from normal distributions. We also see that insufficient sampling of the original distribution can mislead the conclusions, as we have previously shown. Additionally, we observed that as the sample size (N) increases, the kurtosis value also increases until it reaches its saturation when $1 < q < 7/5$. This saturated value of N increases as the non-Gaussian distribution diverges further from the Gaussian distribution.

Our results suggest that if all moments up to the fourth order exist, the kurtosis converges to a fixed value as the sample size increases. However, if at least one moment up to the fourth is infinite, the kurtosis does not converge to a fixed value for any N. Therefore, comparing distributions using kurtosis may yield incorrect results for distributions with finite fourth moments if the dataset sizes are not sufficiently large. It is crucial to ensure that the dataset length allows the kurtosis to stabilize to a constant value; otherwise, one might erroneously interpret a distribution with longer tails as being closer to Gaussian than a distribution with shorter tails.

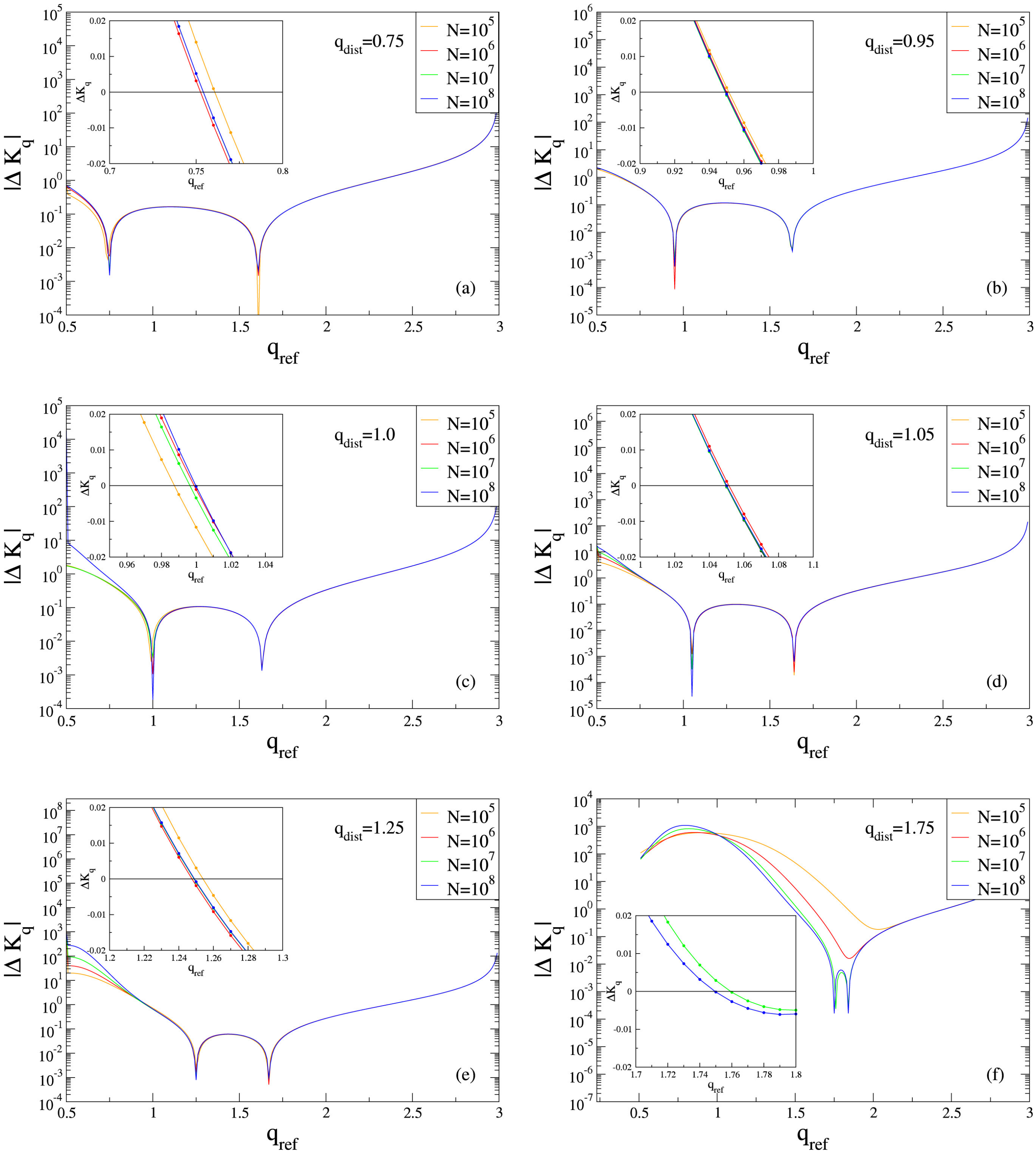


Fig.11 Plots for $|\Delta K_q|$ v/s q. The insets in the graph show a plot between ΔK_q and q without absolute value of ΔK_q.

Moreover, for distributions with an infinite fourth moment, such as q-Gaussians with $7/5 < q < 3$, comparing them using standard kurtosis is useless. This difficulty arises from the misuse of kurtosis, which was originally intended to discern whether a distribution is Gaussian. Hence, even for finite kurtosis values, misinterpretations may occur. Similarly, each q-Gaussian is characterized by a single q-kurtosis value, making comparisons meaningful only when using this specific value. Using an arbitrary q_{ref} value different from the distribution's q value renders the calculated kurtosis meaningless.
In light of these insights, we must exercise caution in using and interpreting kurtosis and skewness, not only in the standard case but also in the realm of q-statistics. The standard definitions of moments may be inadequate for certain non-Gaussian distributions. Therefore, it may be necessary to modify moment definitions, as demonstrated in our analysis of q-Gaussian distributions. It's essential to note that appropriate moment definitions must be determined for other classes of non-Gaussian distributions.

3.9 Future works:

- We can find a new way to compare q-Gaussian distribution when $7/5 < q < 1$, as the method provided in the paper is not satisfactory.
- Generalization of ICC using Tsallis entropy.

3.10 Refrences

- Skewness and kurtosis analysis for non-Gaussian distributions, By:- Ahmet Celikoglu, Ugur Tirnakli
- Universal relation between skewness and kurtosis in complex dynamics, By:- Matthieu Cristelli, Andrea Zaccaria, and Luciano Pietronero
- Influence of the interaction range on the thermostatistics of a classical many-body system, By:- L.J.L. Cirto, V.R.V. Assis, C. Tsallis