Supported By
tetfund
TERTIARY EDUCATION TRUST FUND

MODELING OUTLIERS IN GAUSSIAN AND NON-GAUSSIAN DISTRIBUTIONS: THE WAVELET APPROACH

Efuwape B. T. [1]*, Aideyan D. O [2], Abdullah K-K. A.[3] and Efuwape T. O [4]
[1,3]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria.
[2]Department of Mathematical Sciences, Kogi State University, Ayingba, Nigeria.
[4]Academic Planning Unit, Bells University of Technology, Ota, Nigeria.
***Corresponding author:** efuwape.biodun@oouagoiwoye.edu.ng efuwapebt@yahoo.com

Received: December 13, 2021 Accepted: March 20, 2022

ABSTRACT Aberrant Observations (AOs) are observations that deviate significantly from the majority. They may be generated by a different mechanism corresponding to normal data and may be due to sensor noise, process disturbances, instrument degradation or human related errors. Otherwise, decisions on suspected aberrant observations might be inappropriate. In this paper, we present aberrant observations modeling approach based on wavelet analysis in Gaussian Normal Distribution (ND) and Non-Gaussian Distributions - Contaminated Normal Distribution (CND) and Laplace Distribution (LD). In order to characterize these distributions, a simulation of 508,1020 and 2040 data sets from normal distribution and contaminated with four, four and eight aberrant observations while two real data University College Hospital Ibadan Diabetic Data (UCHDD) and Zadakat Data (ZD) from a local mosque in Ibadan of 128 observations each were analyzed, since wavelet analysis is dyadic. The Mallat algorithm was used to reduce the sizes of the data into smaller resolutions while preserving the desired statistics. In the first three (simulated) series, it was observed that the CND has highest Akaike Information Criterion (AIC) estimates followed by ND and LD hence LD is the most efficient in modeling data in the presence of aberrant observations. From series A (UCHDD) and B (ZD) which are real datasets, the observations were the same as that of simulated datasets except that it was observed that the more the observations, the lower the LD are in modeling aberrant observations.

Keywords: Modeling; Resolution; Wavelet Analysis; Akaike Information Criteria; Data

INTRODUCTION

Wavelet development can be linked to several separate trains of thoughts. It started with Haar's work in the twentieth century. Notable contributions to wavelet theory can be attributed to Zweig's discovery of the continuous wavelet transform in 1975 (originally called the Cochlear transform and discovered while studying the reaction of the ear to sound) (Chiann & Morettin, 1999): Pierre Goupillard, Grossmann and mortlet's formulation of what is now known as the continuous wavelet transform, (Antoniadis, 1997) early work on discrete wavelet. (Daubechies, 1988): Orthogonal wavelets with compact support (1988): Mallat's multi-resolution framework (Mallat 1989a and b), (Daubechies, 1988), time-frequency interpretation of the continuous wavelet transform, (Gencay, Selcuk, & Whitcher, 2001)**:** Hormonic wavelet transform and many others. Wavelet transform though very new, would appear to be very appropriate for analyzing non – stationery signals (Chui, 1999) and a link between wavelet and the difference operator was made in (Jewerth & Sweldens, 1994). Wavelets are mathematical tools used for analyzing time series which we take to be any sequence of observations associated with an ordered independent variable t which can assume either a discrete set of values or a continuum of value. Examples of both include time, depth, or distance along a line. Wavelets are a synthesis of older ideas with new elegant mathematical results and efficient computational algorithms. In some cases it complements the existing analysis techniques like correlation, and spectral analysis and capable of solving problems for which little progress had been made prior to the introduction of wavelets.

Its application is now appearing each year and with the total number of over sixteen thousand articles being published to date in diverse field of study (Addison, 2004; Crowley, 2005). The theoretical underpinning of wavelet were completed in the late eighty's, whereas the 1990's witnessed a rapid increase in the number of different practical applications. These applied fields include (Nason & Von-Sachs, 1999), among others signal and image processing, data compression, astronomy, acoustics (scientific study of sound) partial differential equations optics, and nuclear physics. At the moment, they are entering mainstream Econometrics (Gencay, Selcuk, & Whitcher, 2001) with some applications in different fields of finance and economics (Ramsey, 2001). Wavelet based methods offer a viable alternative to the ubiquitous Fourier analysis (Walter & Shen, 2000).

There are two main waves of wavelets. The first known as continuous wavelet transform (CWT) designed to work with time series defined over the entire real line; the second is the Discrete Wavelet Transform (DWT) which deals with series defined essentially over a range of integers (usually t = 0,1,…….N-1 where it denotes the number of values in the time series)

There are varieties of wavelet and wavelet scaling functions but the most common ones are

- Haar wavelet
- Daubechies wavelet
- Mayer wavelet
- Shanon wavelet
- Spline wavelet
- Coiflets
- Biorthogonal Wavelets

For the purpose of this research work, we shall apply Haar Wavelet Transform.

Advantages of Wavelet Analysis

i. Wavelet analysis appears best suited to exploratory data analysis of complex, non-stationary data which summarizes their main characteristics in easy-to-understand form of visual graphs without using a statistical model or having formulated a hypothesis proposed by John Turkey to encourage statisticians visually to examine their data set e.g. Box-plot (Dahlhaus, 1997).

ii. Statistically, wavelets can be viewed as non-parametric orthogonal series estimators with new elegant statistical results and efficient computational algorithms, (Burrus, Gopinath, & Guo, 1997) that can effectively handle the discontinuities caused by different regime shift (characteristic conditions) that typically plague the economic and financial data.

iii. They (non-stationary) are especially suitable to the comprehensive multi-decision analysis of disaggregate (scaled) series; the process of data aggregation and concept of equispaced series do not play any fundamental role in the context of wavelet analysis.

MATERIALS AND METHODS

Five sets of data were used in the course of this research. Series W, X and Y are simulated datasets from normal distribution using R-Software while series A and B are real datasets.

Series W: Simulated series (n=512) observations with four aberrant observations injected randomly.

Series X: Simulated series (n=1024) observations with four aberrant observations injected randomly.

Series Y: Simulated series (n=2048) observations with eight aberrant observations injected randomly.

Series A: UCH Diabetic Data (n=128).

Series B: Zadakat Data (n=128).

Wavelet Parametric Approach

We considered the Gaussian (ND) and Non – Gaussian Distributions. The Non –Gaussian distributions are LD and CND (when the Normal distribution is contaminated with aberrant observations).

For Distribution with shape parameter, yet when it is not exponentially distributed, no simple closed form solution can be found. However, it is known that as long as the distribution of is not singular, p (I) must have a larger kurtosis value than Gaussian distribution i.e. it is fat tail. In fact, the larger the variance of the distribution, the greater the kurtosis value. To allow for shape values it is better to use a generalized Gaussian distribution, to model wavelet sub-band (at different resolution) coefficients.

The generalized Gaussian distribution has the probability density function by J. Armando Dominguez Molona *et al*.,(2003).

$$A(\rho,\sigma)=\left[\frac{\sigma^2\Gamma\,1/2}{\Gamma\,3/2}\right]^{1/2}$$

$$\mu\in\Re,\rho,\sigma\rangle 0$$

$$f(x;\mu,\sigma,\rho)=\frac{1}{2\Gamma(1+1/\rho)A(\rho,\sigma)}\ell^{-\left\{\left|\frac{X-\mu}{A(\rho,\sigma)}\right|^{\ell}\right.}$$

(1)

The parameter μ is the mean, the function $A(\rho,\sigma)$ is a scaling factor which allows variance (X) = σ^2 and ρ is the shape parameter which controls the shape of the distribution and skewness.

As $\rho \to \infty$, the Gaussian Distribution approaches the uniform distribution.

When $\rho = 2$, the Generalized Gaussian Distribution becomes Gaussian Distribution.

and

If $\rho = 1$, the Generalised Gaussian Distribution becomes a Laplace Distribution.

Laplace Distribution

Let the Probability density function (pdf) of Laplace distribution be given as:

$$f(y, x, \lambda) = \frac{1}{\sqrt{\frac{2}{\lambda}}} \ell^{-(2\lambda)^{\frac{1}{2}}(y-\mu)} \quad (2)$$

$$-\infty < y, \mu < \infty, \lambda > 0$$

The mean of the response variable is given as:

$$E(y) = \frac{1}{\sqrt{2\lambda}} \quad (3)$$

The variance of the response variable as:

$$V(y) = \frac{1}{2\lambda} \quad (4)$$

Eq. (3) and Eq. (4) are required in wavelet analysis as location and scale parameters respectively.

DATA ANALYSIS

Table 1: AIC estimates of 512 observations with 4 AOs injected by resolution levels

Resolution Levels (No of Observations)	Uncontaminated Normal Dist.	Contaminated Distributions	
		Contaminated Normal Distribution	Laplace Distribution
9 (512)	1402.113	4677.203	1822.043
8 (256)	748.4688	3868.213	998.632
7 (128)	371.899	1736.954	536.7024
6 (64)	187.5164	775.0375	288.5806
5 (32)	97.5664	457.8793	156.8911

Table 2: AIC estimates of 1024 observations with 4 AOs injected by resolution levels

Resolution Levels (No of Observations)	Uncontaminated Normal Dist.	Contaminated Distributions	
		Cont. Normal Distribution	Laplace Distribution
10 (1024)	2874.0970	4818.2140	3317.6830
9 (512)	1486.9720	2413.0950	1759.5760
8 (256)	731.7840	1202.4400	924.6532
7 (128)	361.2020	607.1812	484.7221
6 (64)	175.4640	304.7123	248.7585

5 (32)	83.9990	150.5107	127.7655

Table 3: AIC estimates of 2048 observations with 8 AOs injected by resolution levels

Resolutions Level(No of Observations)	Uncontaminated Normal Dist.	Contaminated Distributions	
		Cont. Normal Distribution	Laplace Distribution
11 (2048)	5758.3330	10557.1000	6824.0680
10 (1024)	2980.4130	5280.9330	3640.4620
9 (512)	1466.7610	2630.6690	1926.3770
8 (256)	729.4609	1327.1120	1021.2780
7 (128)	357.9632	1634.8270	532.7774
6 (64)	173.4225	628.5099	279.1899
5 (32)	92.9444	439.8291	154.1428

Table 4: AOC estimates of UCHDD with 128 observations by resolution levels

Resolution Levels (No of Observations)	Contaminated Normal Distributions	Laplace Distribution
7 (128)	2157.1960	2143.276
6 (64)	1140.9800	420.543
5 (32)	518.1440	207.5651
4 (16)	240.4674	100.0472

Table 5: AIC estimates of ZD with 128 observations by resolution levels

Resolutions Level(No of Observations)	Contaminated Normal Dist.	Laplace Distribution
7 (128)	2319.0200	2435.2860
6 (64)	1160.9090	862.9327
5 (32)	535.7745	424.9759
4 (16)	392.6204	188.4098

RESULTS AND DISCUSSION

Table 1 shows the AIC estimates by resolution levels. It can be observed that LD has the lowest values which confirm that it is the best method of modeling aberrant observation among the three distributions.

Table 2 are the AIC estimates obtained from the distributions. . It is also observed that LD has the lowest values which confirm that it is the best method of modeling aberrant observation among the three distributions.

Table 3 shows the AIC estimates obtained from the distributions, it is observed that the LD has the lowest values which confirm that it is the best method of modeling aberrant observation among the three distributions.

Table 4 just as in the standard deviation estimates, the AIC estimates of the, LD has the lower values in all resolutions which confirm that it is the better method of modeling aberrant observation among the two distributions.

Table 5 just as in Table 4, it can be observed that LD has a lower in all resolution values which confirm that it is the best method of modeling aberrant observation among the two distributions.

CONCLUSION

In series W,X and Y data sets 512,1024 and 2048 were simulated using Normal distribution, their mean, standard deviation, likelihood and the Akaike Information Criterion estimates were obtained at different resolution levels . The value so obtained from the mean, and standard deviation were approximately 0 and 1 respectively at each resolution level. Thereafter, 4, 4 and 8 aberrant observations were injected randomly into the series and their corresponding mean, standard deviation, likelihood and the Akaike Information Criterion estimates were obtained using CND, ND and LP. In the first three (simulated) series, which has aberrant observations injected, it was observed that the CND has higher values followed by ND and LD has the least Akaike Information Criterion estimates hence is more efficient in modeling data in the presence of aberrant observations.

From series A (UCH Diabetic datasets) and B (Zadakat datasets) which are real life data, the observations were the same as that of simulated data except that it was observed that the more the observations, the lower the Laplace Distribution is in modeling aberrant observations in real life datasets.

ACKNOWLEDGEMENT

The authors would like to acknowledge the reviewers and editors of FUW Trends in Science & Technology Journal.

CONFLICT OF INTEREST

No conflict of interest has been declared by the authors.

REFERENCES

Addison, A. (2004, April 2-6). Constructing the Landscape of Southwest Paper. United Arab Emirates.

Antoniadis, A. (1997). Wavelength in Statistics: A Review. *Journal of the Italian Statistical Society, 2*, 97-108.

Burrus, C. S., Gopinath, R. A., & Guo, H. (1997). *Introduction to Wavelets and Wavelet Transforms.* Upper Saddle River, NJ: Prentice Hall.

Chiann, C., & Morettin, P. A. (1999). A Wavelet Analysis for Time Series. *Journal of Nonparametric Statistics, 10*, 1-46.

Chui, C. (1999). *An Introduction to Wavelets.* London: Academic Press.

Crowley, P. (2005). *An intuitive guide to wavelets for economists.* Germany: Econometrics 0503017, University Library of Munich.

Dahlhaus, R. (1997). Fitting Time Series Models to NonStationary Processes. *The Annals of Statistics, 25*(1), 1-37.

Daubechies, I. (1988). Orthnormal Bases of Compactly Supported Wavelets. *Communication on Pure and Applied Mathematics, 41*(7), 906-996.

Gencay, R., Selcuk, F., & Whitcher, B. (2001). *An Introduction to Wavelets and Other Filtering Methods in Finance and Economic.* San Diego: Academic Press.

Jewerth, B., & Sweldens, W. (1994). An Overview of Wavelet Based Multiresolution Analysis. *SIAM Review, 36*(3), 377-412.

Mallat, S. (1989a). Multiresolution Approximation and Wavelet Orthonormal Bases of L2(R). *Transactions of the American Mathematical Society, 315*(1), 69-87.

Mallat, S. G. (1989b). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 674-693.

Nason, G. P., & Von-Sachs, R. (1999). Wavelets in Time Series. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 357*(1760), 2511-2526.

Ramsey, J. (2001). *The Elements of Statistics with Applications to Economics and Social Sciences.* Duxbury_Thomson Lrarning.

Walter, G., & Shen, X. (2000). *Wavelets and other Orthogonal Systems.* CRC Press.