Wavelet Method for Detecting and Modeling Anomalous Observations in Gaussian and Non - Gaussian Distributions

Shittu Olarewanju Ismail[1] University of Ibada, Department of Statistics, Ibadan, Oyo State, Nigeria
Aideyan Donald Osaro[2], Kogi state University, Department of Mathematical Sciences, Anyigba, Kogi State, Nigeria.

ABTRACT: Wavelet analysis has been applied recently for analyzing data completely due to its potential. In this paper, we present aberrant observation detection and modeling approach based on wavelet analysis in Gaussian and Non-Gaussian distributions. In order to characterize these distributions, a simulation of 1020 data set from normal distribution and contaminated with four normal data and later with four aberrant observations since wavelet analysis is dyadic. It was discovered that Normal (Gaussian) distribution with aberrant observations is the most efficient in detecting aberrant observations while Laplace (Non-Gaussian) distribution is the optimal distribution in modeling aberrant observations using the three distributions.

Index Terms: Wavelets, Outliers, resolution, Residuals, Distributions, Gaussian, Discrete, Analysis

---------------------------------------♦---

1 INRODUCTION

Aberrant observations (outliers) are defined as data points that are distinctly separate from the rest of the data. It is an observation that lies an abnormal distance from other values in a set of data. In statistics, an aberrant observation is an observation that is numerically distinct from the rest of the data. They can occur by chance in any distribution but are often indicative either of measurement error or that the population is heavy tailed. It can also indicate faulty data, erroneous procedures, etc. Section 2 looks at the overview of wavelet analysis which uses both resolution and location in analyzing data completely. Section 3 describes how these outliers will be detected using these distributions which are the main goal of this paper. Section 4 discusses the analysis for these residuals while Section 5 interprets the results, conclusion and informed us of areas of further work.

2 OVERVIEW OF WAVELET ANALYSIS

Wavelet analysis is a statistical tool that can be used to extract information from any kind of data and are generally needed to analyze data fully at different resolution (scale) and location.

Discrete Wavelet Transform re - expresses a time series in terms of coefficients that are associated with a particular time and a particular dyadic scale 2^J. These coefficients are fully equivalent to the original series from its Discrete Wavelet Transform coefficients.

The Discrete Wavelet Transform allows us to partition (decompose) the information in a time series into pieces that are associated with different scales and time. This decomposition is very close to the statistical technique known as the Analysis of variance (ANOVA), so DWT leads to a scaled – based ANOVA that is quite analogous to the frequency – based ANOVA provided by the power spectrum R. Todd Ogden (Dec., 1996).

It effectively decorrelates a wide variety of time series that occurs quite commonly in physical applications. This property is the key to the use of the DWT in the statistical methodology L. Li and G.Lee (2003).

Illustration

We begin with a set of discrete sequence of data $y = y_1, y_2 \dots \dots y_n$ Where each of y_i is a real number and i is an integer ranging from 1to n. we assume that the length of our sequence n is a power of two, $n = 2^J$ for some $J \geq 0$. This should not be seen as a restriction as this can be modified for other n Abraham Maslow (Dec., 2008). We call the sequence Where $n = 2$ dyadic one. The key information we extract is the "detail" in the sequence at different scale and different locations. By detail we mean the degree of the difference or variation between successive observations of the vector that is, $d_1 = y_2 - y_1$ at the given scale and location.

$d_k = detail\ at\ location\ k$

$$d_k = (y_{2k} - y_{2k-1}) \quad \dots\dots\dots\dots (2.1)$$

$For\ k = 1,2, \dots \dots \dots, {}^{n}/_{2}$

e.g

$d_1 = y_2 - y_1, d_2 = y_4 - y_3, d_3 = y_6 - y_5,\ etc$

In equation (2.1) if the detail in $y_{2k} - y_{2k-1}$ are similar, then the coefficient d_k will be very small; if they are exactly the same, $d_k$is zero and if very large the coefficient will be very large. d_k encodes the difference between successive pairs of observations in the original y vector. d_k Is known as the finest scale detail Abraham Maslow (Dec., 1998)

$[d_k]_{k=1}^{n/2}$ Is not the conventional first difference vector since difference such as $y_3 - y_2$ are missing from $\{d_k\}$ location. d_k only gives information about $2k$ and its neighbor at the finest possible scale of detail G.P. Nason (2008)

At Coarser Scale; for coarser detail

$$C_k = y_{2k} + y_{2k-1} \dots\dots\dots\dots\dots\dots\dots (2.2)$$

$[C_k]_{k=1}^{n/2}$ is the sum of scaled average (scaled because it is not divided by 2). The information in $[C_k]$ is a roughing of that original y vector. The operation that turns $[y_i]$ to $[C_k]$ is similar to the moving average smoothening operation except that the differencing does not overlap consecutive pairs. A. Dainotti, A. Pescape and G. Viorgio (2006)

Each C_k contains information originating from y_{2k} and y_{2k-1} (adjacent observations)

The original sequence y consist of 2^J observations $\{d_k\}$ consist of ${}^{n}/_{2} = 2^{J-1}$ observations

If $j = J - 1$ then d_k can be written as $d_{j,k}$ and the first level averages or smooth C_k are renamed to becomes C_{j-1}, k written as $C_{j,k}$

To obtain the next coarsest detail, we repeat the operation of equation (3.1) to the finest level averages C_{j-1}, k as follows.

2.1 SCALE/LEVEL TERMINOLOGY

The scale in the quantity 2^J where $j = J - 1$ and the level for the intergral quantity while k is the locations Larger j (positive) corresponds to finer scale and smaller j refers to the coarser scale in the contents of this work from equation 2.2 $C_{j-1,k}$

$$C_{j-1,2L} = C_{j-1,2L} + C_{j-1,2L-1} \dots\dots\dots\dots\dots (2.3)$$

For $l = 1,2, \dots \dots \dots \dots \dots, {}^{n}/_{4}$

From the original vector y for $l = 1$

$$C_{j-1,} l = (y_{4L+2} + y_{4L-3}) - (y_{4L} + y_{4L-1})$$
$$= (y_2 + y_1) - (y_4 + y_3)$$
$$= y_1 + y_2 + y_3 + y_4$$

This is a kind of moving average except that it is not divided by ${}^{1}/_{4}$

$d_{j,k}$ "detailed" coefficients are wavelet coefficients and $C_{j,k}$ coefficients are known as father wavelet or scaling function coefficients.

This general pyramid algorithm is called Haar wavelet transform.

The inverse of the original sequence can be reconstructed exactly by using wavelet coefficients $d_{j,k}$ and last C_{00} W. Lu and I. Traore (2005)

2.2 Sparsity

The behavior of sparsity is a characteristic of wavelet: piece wise smooth functions have sparse representation G.P. Nason (2008).

To conserve information we change equation (2.1) and (2.2) by introducing α as follows

$$d_k = \alpha(y_{2k} - y_{2k-1}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.4)$$

$$C_k = \alpha\ (y_{2k} + y_{2k-1}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.5)$$

The inputs are (y_{2k}, y_{2k-1}) transformed into the output $(d_k,\ C_k)$ and the (squared) norm of the output.

ISSN 2229-5518

$$d_k^2 + c_k^2 = \alpha^2 (y_{2k}^2 + 2y_{2k}y_{2k-1} + y_{2k-1})$$

$$+\alpha^2 (y_{2k} + 2y_{2k}y_{2k-1} + y_{2k-1})$$

$$= 2\alpha^2 (y_{2k}^2 + y_{2k-1}^2) \ldots\ldots (2.6)$$

Where $y_{2k}^2 + y_{2k-1}^2$the squared norm of the input coefficients hence is to wish the norm of output equals norms of input

Let$2\alpha^2 = 1$ therefore

$\alpha = 2^{-1/2}$

Then the discrete wavelet coefficients is

$$d_k = (y_{2k} - y_{2k-1}) / \sqrt{2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.7)$$

Equation 2.7 can be rewritten as

$$d_k = g_0 y_{2k} + g_1 y_{2k-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.8)$$

Where $g_0 = 2^{-1/2}$ and $g_1 = -2^{-1/2}$

In general

$$d_k = \sum_{L=\infty}^{\infty} g_L y_{2k-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.9)$$

$$\text{Where} \quad g_L = \begin{cases} 2^{-1/2}\ for & L = 0 \\ 2^{-1/2}\ for & L = 1 \\ 0 & otherwise \end{cases} \ldots\ldots (2.10)$$

Equation 2.9 is similar to a filtering operation with coefficient of $\{g_L\}_{l=\infty}^{\infty}$[10,11]

That is the input sequence can be thought to possess a norm as defined by

$$\|y\|^2 = \sum_{L=1}^{n} y_i^2$$

Another interesting component of the filter object is the H component which is equal to the vector operation $\left(2^{-1/2}, 2^{-1/2}\right)$ which is involved in the filtering operation analogous to that equation 3.12 that produce C_k as

$$C_k = \sum_{L=\infty}^{\infty} h_L y_{2k-l} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.11)$$

$$h_L = \begin{cases} 2^{-1/2}\ for & L = 0 \\ 2^{-1/2}\ for & L = 1 \\ 0 & otherwise \end{cases} \ldots\ldots\ldots\ldots\ldots\ldots (2.12)$$

2.3 MATRX REPRESENTATION

Like the orthonormal discrete Fourier transform, the discrete wavelets transform (DWT) of X_t is an orthonormal transform [5]. Let $[W_n; n = 0 \ldots\ldots N-1]$ be the DWT coefficients then, we can write $W = w_n$ where W is a column vector of length $N = 2^2$ whose n^{th} DWT and satisfying $w^T w = I_N$ orthornormality implies that $X = w^T w$ and $\|W\|^2 = \|X\|^2$. Hence W_n^2 represents the contribution to the energy attributable to the DWT coefficient with index n.

Whereas ODFT coefficients are associated with frequencies the n^{th} wavelet coefficient W_n is associated with a particular scale and with a particular set of times H. Nayyar and Ali. A. Ghorbani (2006).

Explicitly, the rows of this mature for n=0, 8, 12, 14, and 15 are

$$w_0^T = \left[-1/\sqrt{2}, 1/\sqrt{2}, \underbrace{0 \ldots\ldots\ldots\ldots 0}_{14\ zero}\right]$$

$$w_8^T = \left[-1/2, -1/2, 1/2, 1/2, \underbrace{0 \ldots\ldots\ldots 0}_{12\ zero}\right]$$

$$w_{12}^T = \left[-1/\sqrt{8}, \ldots\ldots\ldots\ldots, -1/\sqrt{8}, 1/\sqrt{8} \ldots\ldots\ldots 1/\sqrt{8}, \underbrace{0 \ldots\ldots 0}_{8\ zero}\right]$$

$$w_{14}^T = \left[-1/\sqrt{4}, \ldots\ldots, -1/\sqrt{4}, 1/\sqrt{4} \ldots\ldots\ldots 1/\sqrt{4}\right]$$

$$w_{15}^T = \left[1/\sqrt{4} \ldots\ldots\ldots 1/\sqrt{4}\right]$$

The remaining eleven rows are shifted version of the above;

ISSN 2229-5518

$$w_1 = T^2 w_0, \quad w_2 = T^4 w_0 \ldots\ldots w_7 = T^{14} w_0$$

$$w_9 = T^4 w_8, \qquad w_{10} = T^8 w_8 \qquad w_{11} = T^{12} w_8$$

$$w_{13} = T^8 w_{12}$$

Let us now, define exactly what the notation of scale means for a positive integer k let

$$\bar{X}_t(k) = \frac{1}{k}\sum_{l=0}^{k-1} X_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.13)$$

Donald B. Percival, Andrew T. Walden (2000)

3 OUTLIER DETECTION

In this section, we assume that the higher the value of the residuals, the more anomalous the data Wei Lu, Mahbod Tavallaee and Ali A. (2008). As a result, in order to identify these outliers the residuals of these distributions at different resolutions will be obtained and compared to identify their rate of detection J. McHugh (2000) and P. Barford, J. Kline, D. Plonka and A. Ron (2002).

4 ANALYSIS OF RESIDUALS

The purpose for analyzing the residuals of these distributions is to support our assumption in section (3). The data analyzed were simulated from Normal distribution involving 1020 data set. Since Wavelet analysis is dyadic, we introduced four data within the maximum and minimum values in the data set and analyzed it as Normal distribution without aberrant observations (NO) at different resolution (j). These four values were removed and four aberrant observations were now introduced and further analysis using Normal (NW), Laplace and Cauchy distributions were used to analyze the contaminated data set at different resolutions. The mean and standard deviation (residual) were obtained at different resolutions using the Maximum Likelihood estimate which of course, is more efficient than the conventional method. Since Wavelet analysis is dyadic, the data were analyze at different band size (1024, 512, 256, 128, 64, 32) and at different resolutions (j= 10 9,8,7,6 and 5) respectively.

Table 1: Mean and Standard Deviations of the Distributions

Resolution level j	Band Size	NO		NW		LAPLACE		CAUCHY	
		Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
10	1024	0.04545	0.9869	0.0727	1.4858	0.0711	0.8613	0.0671	0.6255
9	512	-0.0581	0.9470	0.0160	1.5503	0.1752	0.8681	0.1701	0.6326
8	256	0.0573	0.9738	0.0547	1.5059	0.0012	0.8975	0.0114	0.6599
7	128	-0.2162	0.9549	-0.0163	1.5041	0.2338	0.9059	0.2465	0.6198
6	64	-0.0989	0.6676	-0.1550	1.5739	-0.1842	0.8498	-0.1572	0.5021
5	32	-0.1220	0.6514	-0.2535	1.4314	-0.0332	0.8505	-0.0036	0.4851

Key

NO: Normal distribution without aberrant observations

NW: Normal distribution with aberrant observations

5 EXPERIMENTAL EVALUATIONS

From the above, the mean and standard deviations for the coefficients of Normal distribution without outliers(NO) at different resolutions(j) or band size with approximately mean = 0 and standard deviation = 1 confirms the absence of outliers. Also for the other three distributions (Normal with outliers(NW), Laplace and Cauchy), it was observed that the Laplace distribution has a standard deviation closer to that of Normal without outliers, followed by the Cauchy distribution and finally the Normal distribution with aberrant observations(NW).

Since Normal (Gaussian) distribution with aberrant observations has the highest standard deviation at all resolutions from Normal, we conclude that among the three distributions, it is the most efficient in detecting aberrant observations. While on the other hand, Laplace (Non-Gaussian) distribution whose standard deviations at different resolution is closest to the Normal distribution without aberrant observations is regarded as the optimal distributions for modeling aberrant observations among these distributions.

References

[1] Abraham Maslow, Histogram Smoothing Via The Wavelet Transform. Journal of Computational and Graphic Statistics, Vol.7, No.4 (Dec., 1998)

[2] A. Dainotti, A. Pescape and G. Viorgio "Wavelet-based Detection of Dos Attack"
Proceedings of IEEE Global Telecommunication Conference, San Francisco, 2006

[3] Donald B. Percival, Andrew T. Walden, "Wavelets Methods for Time Series Analysis." Cambridge University Press (2000)

[4] G.P. Nason, Wavelet Methods In Statistics With R Springer (2008)

[5] H. Nayyar and Ali. A. Ghorbani "Approximate Autoregressive Modeling for Network Attack Detection," Proceedings of the 4th Annual Conference on Privacy, Security and Trust, pp. 175-184, Markham, Canada, 2006

[6] J. McHugh, " Testing Intrusion Detecting System: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory ," CAM Transition on Information and System Security, 3 (4):262-294, 2000

[7] L. Li, and G.Lee, " DDoS attack Detection and Wavelets" Proceedings of 12th International Conference on Communication and Networks, pp. 421-427, Texas, 2003

[8] P. Barford, J. Kline, D. Plonka and A. Ron, "A Signal Analysis of Network Traffic Anomalies" A Proceedings of Internet Workshop 2002, Marseille, France, 2002.

[9] R. Todd Ogden "Essential Wavelets for Statistical Application and Data Analysis" Birkhauser Boston (Dec., 1996)

[10] W. Lu and I. Traore "A Novel Unsupervised Anomaly Detection Framework for Detecting Network Attacks in Real Time", Lecture Note in Computer Science, Vol., 3810, pp. 96-109, Springer, 2005, Y.G. Desmedt et al (Eds.)

[11] Wei Lu, Mahbod Tavallaee andAli A. "Ghorbani Detecting Network Anomalous Using Wavelet Basis Functions" CNSR, pg 149-156, IEEE Computer Science (2008)

First Arthur
Shittu Olarewanju Ismail*
P.D.S.,B.Sc, M.Sc, M.Phil,Ph.D

Second Arthur
Aideyan Donald Osaro**
P.D.S.,B.Sc, M.Sc.