Detection of Outliers in the Unreplicated Linear Circular Functional Relationship Model via Functional Form

By:

Abdul Ghapor Hussin, Ali Abu Zaid and I. Mohamed

(Paper presented at the ***International Conference on Nonparametric Methods for Measurement Error Models and Related Topics*** held on 3-5 Mei 2009 in Ottawa, Canada)

DETECTION OF OUTLIERS IN THE UNREPLICATED LINEAR CIRCULAR FUNCTIONAL RELATIONSHIP MODEL VIA FUNCTIONAL FORM

*ABDUL GHAPOR HUSSIN, ALI ABU ZAID & I MOHAMED

University of Malaya

Summary

In this paper we consider the problem of outliers for the functional relationship model of circular variables by transforming the circular data to continuous or real line data set via complex form. The ***COVRATIO*** statistic is extended from the linear regression models to the proposed model to detect any possible outliers. The cut-off points are obtained and the power of performance is examined by simulation studies. The model is illustrated with an application to the analysis of wind direction data recorded by two different techniques and the detection procedure of outliers is implied.

Key word: Circular variable; COVRATIO statistic; Outlier; Straight line fitting; Wind data.

1. Introduction

The functional relationship model is part of the general class of error-in-variables model (EIVM), (e.g. Anderson (1984)) which may be describes as follows. Suppose that we were given two sets of n observations $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$ and that these observations correspond to measurements of certain unobserved underlying variables (X_j, Y_j), $j = 1,2,...n$ which is deterministic or fixed, made with random errors $(\delta_j, \varepsilon_j)$, $j = 1,2,...n$. Further $x_j = X_j + \delta_j$, $y_j = Y_j + \varepsilon_j$ for $j = 1,2,...n$. When the data

*Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.
E-mail: ghapor@um.edu.my.

is linear (i.e. takes values on the real line, as in the case of wind speed), an adequate statistical method for fitting a linear functional relationship model were described, for instance Fuller (1987) and Kendall & Stuart (1973). Some of the main differences between ordinary regression and EVIM can be summarized as follows: firstly, for any pair of observation (x,y) ordinary regression assumes the x value is mathematical, whereas in EIVM both x and y are observed with error. Secondly, in EIVM there is no distinction between "explanatory" and "response" variables, unlike ordinary linear regression. Lastly, ordinary linear regression is more appropriate if the aim is to predict one variable from the other rather than to look at the underlying relationship between the two variables. In some practical problems the variables are no longer linear, for instance to compare the measurements of wind direction using two different instrument. In this case the variables are known as circular variables, where the variables are taking values between $(0, \pi)$ radians or $(0°,306°)$. Due to the bounded closed space of the circular variables, different techniques from those appropriate for linear variables must be used. An example of regression model of circular variables is given by Hussin *et. al.* (2004). Further the comprehensive statistics of circular variables is also available in Fisher (1993).

The circular variables can be treated as complex variables. Hussin (1998) proposed the unreplicated complex linear functional relationship model (UCLFR) to fit a straight line when both variables are circular and subject to errors. However, there is no published work has been found in the literature on the problem of outliers in the circular functional relationship models even though any statistical data is commonly subjected to be contaminated by some outliers.

In this paper, we investigate the occurrence of outliers in the UCLFR model by transforming the circular data to continuous or real line data set via complex form. Further we extended the row deletion approach of the ***COVRATIO*** statistic which was firstly proposed by Belsley et al (1980) to identify outlier in the linear regression models to UCLFR model.

This paper is organized as follows: the following section reviews the UCLFR models and discusses the maximum likelihood estimates and the asymptotic properties of model parameters. Section 3 introduces the ***COVRATIO*** statistic for the proposed model. Section 4 presents the cut-off points and the power of performance through extensive simulation studies. The application of the such model is given in Section 5 followed by some remarks on the proposed technique.

2. The linear functional relationship model for circular variables

For any two circular variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ with observation of $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$ where $0 \le x_j, y_j < 2\pi$ for j=1,..,n may be denoted by a series of complex numbers $(\cos x_1 + i\sin x_1, \cos y_1 + i\sin y_1),...,(\cos x_n + i\sin x_n, \cos y_n + i\sin y_n)$.

Hussin (1998) proposed a UCLFR model when both circular variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are subject to errors and only the unreplicated data is considered. For any fixed X_j we assume x_j and y_j have been measured with errors δ_j and ε_j respectively. The complex linear functional relationship model is given by

$$(\cos x_j + i\sin x_j) = (\cos X_j + i\sin X_j) + \delta_j, \text{ and } (\cos y_j + i\sin y_j) = (\cos Y_j + i\sin Y_j) + \varepsilon_j, \quad (1)$$

where

$$(\cos Y + i\sin Y) = \alpha + \beta(\cos X_j + i\sin X_j), \quad (2)$$

where δ_j are independently distributed from the bivariate complex Gaussian distribution, with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\sum_x = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$ respectively. Further ε_j are also independently distributed from the bivariate complex Gaussian distribution, with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$

and covariance matrix $\sum_y = \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ respectively. Hussin (1998) used the same assumption to the continuous linear variables to overcome the problem of unbounded likelihood function. Thus, the ratio of two error variances is assumed to be known, i.e. $\frac{\sigma_1^2}{\sigma_2^2} = \lambda$. The log likelihood function for model (2) is given by

$$\log L\left(\alpha, \beta, \sigma_1^2, X_1, \ldots, X_n; \lambda, x_1, \ldots, x_n, y_1, \ldots, y_n\right) =$$

$$-2n\log(\pi) - n\log\left(\lambda\sigma_1^2\right) - \frac{1}{\sigma_1^2}\sum_j \left|\delta_j\right|^2 - \frac{1}{\lambda\sigma_1^2}\sum_j \left|\varepsilon_j\right|^2$$

$$= -2n\log(\pi) - n\log\left(\lambda\sigma_1^2\right) - \frac{1}{\sigma_1^2}\sum\left\{2 - 2\cos x_j \cos X_j - 2\sin x_j \sin X_j\right\}$$

$$-\frac{1}{\lambda\sigma_1^2}\sum\left\{1 + \alpha^2 + \beta^2 + 2\alpha\left(\beta\cos X_j - \cos y_j\right)\right\}$$

$$+\frac{2\beta}{\lambda\sigma_1^2}\sum\left(\cos y_j \cos X_j + \sin y_j \sin X_j\right).$$

The maximum likelihood estimates of the parameters are given by

$$\hat{\alpha} = \frac{1}{n}\sum\left(\cos y_j - \hat{\beta}\cos\hat{X}_j\right),$$

$$\hat{\beta} = \frac{1}{n}\sum\left(\cos y_j \cos\hat{X}_j + \sin y_j \sin\hat{X}_j - \hat{\alpha}\cos\hat{X}_j\right),$$

$$\hat{X}_j = \tan^{-1}\left\{\frac{\lambda\sin x_j + \hat{\beta}\sin y_j}{\lambda\cos x_j + \hat{\beta}\cos y_j - \hat{\alpha}\hat{\beta}}\right\}, \text{ for } j = 1, \ldots, n$$

and

$$\sigma_1^2 = \frac{1}{n}\Sigma\left(2 - 2\cos x_j \cos \hat{X}_j - 2\sin x_j \sin \hat{X}_j\right) + \frac{1}{\lambda n}\Sigma\left(1 + \hat{\alpha}^2 + \hat{\beta}^2 + 2\hat{\alpha}\left(\hat{\beta}\cos \hat{X}_j - \cos y_j\right)\right)$$
$$- \frac{2\hat{\beta}}{\lambda n}\Sigma\left(\cos y_j \cos \hat{X}_j + \sin y_j \sin \hat{X}_j\right).$$

Due to the absence of the close-form for $\hat{\alpha}, \hat{\beta}, \hat{X}_j$ and $\hat{\sigma}_1^2$ the estimates may be obtained iteratively. The asymptotic properties of $\hat{\alpha}$ and $\hat{\beta}$ are obtained from Fisher's information matrix and given by

$$Var(\hat{\alpha}) = \frac{a_1 - b_3}{(a_1 - b_1)(a_1 - b_3) - (a_2 - b_2)^2},$$

$$Var(\hat{\beta}) = \frac{a_1 - b_1}{(a_1 - b_1)(a_1 - b_3) - (a_2 - b_2)^2},$$

and

$$Cov(\hat{\alpha}, \hat{\beta}) = \frac{b_2 - a_2}{(a_1 - b_1)(a_1 - b_3) - (a_2 - b_2)^2},$$

where

$$a_1 = \frac{2n}{\lambda\hat{\sigma}_1^2}, \qquad a_2 = \frac{2\Sigma\cos \hat{X}_j}{\lambda\hat{\sigma}_1^2},$$

$$b_1 = \frac{\Sigma W_{j1}^2 R_{jj}}{\Sigma R_{jj}}, \qquad b_2 = \frac{\Sigma W_{j2} R_{jj} W_{j1}}{\Sigma R_{jj}}, \qquad b_3 = \frac{\Sigma W_{j2}^2 R_{jj}}{\Sigma R_{jj}}$$

and also,

$$W_{j1} = \frac{2\hat{\beta}\sin \hat{X}_j}{\lambda\hat{\sigma}_1^2}, \qquad W_{j2} = \frac{2}{\lambda\hat{\sigma}_1^2}\left(\cos y_j \sin \hat{X}_j - \sin y_j \cos \hat{X}_j - \hat{\alpha}\sin \hat{X}_j\right),$$

$$R_{jj} = \frac{2}{\hat{\sigma}_1^2}\left(\cos x_j \cos \hat{X}_j + \sin x_j \sin \hat{X}_j\right) + \frac{2\hat{\beta}}{\lambda \hat{\sigma}_1^2}\left(\sin y_j \sin \hat{X}_j + \cos y_j \sin \hat{X}_j - \hat{\alpha} \cos \hat{X}_j\right)$$

The asymptotic properties for the $\hat{\sigma}_1^2$ is given by

$$Var\left(\hat{\sigma}_1^2\right) = S^{-1},$$

where

$$S = \frac{2}{\lambda \hat{\sigma}_1^6} \Sigma \left\{1 + \hat{\alpha}^2 + \hat{\beta}^2 + 2\hat{\alpha}\left(\hat{\beta}\cos \hat{X}_j - \cos y_j\right) - 2\hat{\beta}\left(\cos y_j \cos \hat{X}_j + \sin y_j \sin \hat{X}_j\right)\right\}$$
$$+ \frac{4}{\hat{\sigma}_1^6} \Sigma \left\{1 - \cos x_j \cos \hat{X}_j - \sin x_j \sin \hat{X}_j\right\} - \frac{n}{\hat{\sigma}_1^4}.$$

For large values of n, these estimates are distributed normally and they can be used to estimate the standard error of $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}_1^2$.

3. COVRATIO statistic for the UCLFR model

For linear regression models, Belsley *et al.* (1980) suggested identifying outliers based on the determination ratio of covariance matrixes, which is given by $COVRATIO_{(-i)} = \frac{|COV_{(-i)}|}{|COV|}$, where COV is the covariance matrix for full data set and $COV_{(-i)}$ is the covariance matrix for the reduced data set by excluding the i^{th} row. Any data point with $\left|COVRATIO_{(-i)} - 1\right|$ value close to or larger than $\left(\frac{3p}{n}\right)$ indicates that the i^{th} observation is a candidate to be an outlier, where p is the number of estimated

coefficients and n is the sample size. If the ratio is close to the unity then there is no significant difference between the covariance matrices, i.e. the i^{th} observation is consistent with the other observations. The determinant of coefficients covariance matrix for model (2) can be expressed by

$$|COV| = \frac{1}{(a_1 - b_1)(a_1 - b_3) - (a_2 - b_2)^2}, \qquad (3)$$

where a_1, a_2, b_1, b_2 and b_3 are given in Section 2. In the following section we will obtain the cut-off points for the $|COVRATIO_{(-i)} - 1|$ statistic and investigate its performance.

4. Simulation studies

4.1. The percentiles of COVRATIO statistic

The percentile points are obtained by using Monte Carlo simulation method. Six different sample sizes of n=30,50,70,100,120 and 150 are used. For each sample size n, two sets of bivariate complex Gaussian distribution with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variances $\sum_x = \sum_y$ respectively are generated where σ_1^2=0.2, 0.5 and 1. The percentiles are obtained as follows:

Step 1. Generate $\boldsymbol{X}$ variable of size n from von Mises distribution, $VM\left(\frac{\pi}{4}, 3\right)$. Without loss of generality the parameters of the unreplicted complex linear functional relationship model are fixed at α=0 and β=1.

Step 2. Calculate the observed values of the response variable $\boldsymbol{Y}$ based on model (2).

Step 3. Fit the generated circular data by using model (2).

Step 4. Calculate $|COV|$ by using equation (3).

Step 5. Exclude the i^{th} row from the sample, where $i = 1,...,n$.

Repeat steps 3 to 5 to obtain $\left|COV_{(-i)}\right|$ for all i.

Step 6. Compute the value of $\left|COVRATIO_{(-i)}-1\right|$ for all i and specify the maximum value.

The process is repeated 2000 times for each combination of sample size n and variance σ_I^2. Then the 10%, 5% and 1% upper percentiles of the maximum values of $\left|COVRATIO_{(-i)}-1\right|$ are calculated.

The results are tabulated in Table 1. The 10%, 5% and 1% upper percentile values are given in the first, second and third rows respectively of each cell. The results show that the cut-off points are decreasing function of the sample size n. Further it is also notice that the cut-off point is increasing as we increases the error variances. Figure 1 display the cut-off points for n=100 at different values of the error variance σ_I^2. It is obvious that the cut-off points are increasing function of the error variance σ_I^2.

4.2. The power of performance of COVRATIO statistic

Monte Carlo simulation method is used to examine the performance of $\left|COVRATIO_{(-i)}-1\right|$ statistic for detecting outliers in model (2). Four different sample sizes n=30,50,70 and 100 with one value of error variance σ_I^2=0.5 is used. We follow similar procedure described in Subsection 4.1 to generate the data. In addition, we let the observation at the position d, say $y[d]$, be contaminated and examined separately for the real and imaginary part of response variable $\boldsymbol{Y}$ to investigate the impact of contamination on the power of performance. The real part is contaminated at position d such that:

$$\cos\left(y^*[d]\right)+i\sin\left(y^*[d]\right)=\cos(y[d])+i\sin(y[d])+\lambda\pi.$$

On the other hand, the imaginary part is contaminated as follows.

$$\cos\left(y^*[d]\right)+i\sin\left(y^*[d]\right)=\cos(y[d])+i\sin(y[d])+i\lambda\pi,$$

where $y^*[d]$ is the value of $y[d]$ after contamination and λ is the degree of contamination in the range $0 \le \lambda \le 1$. Although the form of contaminated model is linear but we contaminate by adding $\lambda\pi$ in order to study the power of performance within close interval of contamination $[0, \pi]$.

The generated data of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are then fitted by using model (2) and $|COV|$ is calculated using equation (3). Consequently, by excluding the i^{th} row from sample, for $i = 1,...,n$ and refitting the reduced data we specify the maximum value of the $|COVRATIO_{(-i)} - 1|$ statistic.

The process is repeated for 2000 times. The power of performance of the $|COVRATIO_{(-i)} - 1|$ statistic is examined by computing the percentage of correct detection of the contaminated observation at position $[d]$. Figure 2(a) shows the power of performance of the $|COVRATIO_{(-i)} - 1|$ statistic for different sample size n with error variance σ_1^2=0.5. It is obvious that the performance is decreasing function of the sample size n. There is no significance difference in the power of performance by contaminating the real or imaginary parts with same level of contamination for specific sample size n and error variance σ_1^2 as shown in Figure 3. Generally, the performance increases as the level of contamination λ increases.

5. Practical example

A total of 129 observations of wind directions (in radians) are recorded over the period of 22.7 days along the Holderness coastline (the Humberside coast of the North Sea, United Kingdom) by using two different instruments; HF radar system and anchored wave buoy. Figure 4 shows the scatter plot of wind direction data. Note that the scale is artificially broken at 0 (or equivalently 2π) radians. Two points seem to be outliers at the top left of the plot. However, they are actually consistent with the rest of the observations as they are close to other observations at the top right or left bottom due to wrap-around measurement

from 2π back to 0. Figure 4 shows that it is reasonable to assume a linear relationship between the direction measured by an anchored wave buoy and by HF radar.

The COVRATIO statistic is applied to detect any possible outlier in such data. The determinant of the coefficients covariance matrix for full data set $|COV|$ is computed as well as the corresponding cut-off point. The $|COVRATIO_{(-i)} - 1|$ statistic values for the data are plotted in Figure 5. From Table 1, by comparing for n=129 and $\hat{\sigma}_1^2 = 0.1486$ for the given data set, it can be seen that the $|COVRATIO_{(-i)} - 1|$ statistic values for observations number 38 and 111 exceed the cut-off point. Thus, the proposed *COVRATIO* statistic for simple circular regression model has successfully identified both observations as outliers. This is very much agree with the results obtained by Abuzaid *et al.* (2008) when they discussed the identification of outliers in circular regression model by using a new definition of circular residuals via different graphical and numerical methods for the same data set.

6. Conclusions

In this paper, we have considered the unreplicted complex linear functional relationship model and proposed a technique to identify possible outliers in the circular variables. The ***COVRATIO*** statistic based on row deletion approach has been extended to the such model. The cut-of points are obtained and the power of performance is examined through extensive simulation study. It is found that the cut-off points are highly depends on the sample size and the error variance. As an application the wind direction data is fitted using the proposed model. Further, the ***COVRATIO*** statistics was able to identify two outliers in the wind direction data that have been measured by two different techniques and the same points were detected in Abuzaid *et al.* (2008) by using other alternative approach. Since there is no such literature available, the proposed technique is very significant in the problem of detecting outliers for functional model when the variables are circular

References

ABUZAID, A. H., HUSSIN, A. G. & MOHAMED, I. B. (2008). Identifying single outlier in linear circular regression model based on circular distance, *Journal of Applied Probability and Statistics*, **3** (1), 107–117.

ANDERSON, T. W. (1984). Estimating linear statistical relationships. *The Annals of Statistics,* **12,** 1-45.

BELSLEY, D. A., KUH, E. & WELSCH, R. E. (1980). *Regression Diagnostics: Identifying influential*

data and sources of collinearity, John Wiley & Sons (New York; Chichester).

FISHER, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

FULLER, W. A. (1987). *Measurement Error Models*. John Wiley, New York.

HUSSIN, A. G. (1998). The unreplicated complex linear functional relationship model and its application, *Bull. Malaysian Math. Soc.* (Second Series) **21**, 79-86.

HUSSIN, A. G., FIELLER, N. R. J. & STILLMAN, E. C. (2004). Linear regression for circular variables with application to directional data. *Journal of Applied Science and Technology*, **8** (1 & 2), 1-6.

KENDALL, M. G. & STUART, A. (1973). *The Advanced Theory of Statistics*. Hafner, New York.

Table 1: The cut-off points for COVRATIO statistic

n	σ^2		
	0.2	**0.5**	**1**
30	0.2534 0.3022 0.5180	0.2921 0.3393 0.5266	0.3861 0.4328 0.5598
50	0.1674 0.2444 0.3332	0.1691 0.2679 0.5155	0.2433 0.3858 0.5448
70	0.1169 0.1886 0.2576	0.1426 0.2573 0.4732	0.2192 0.3421 0.5124
100	0.1078 0.1263 0.1610	0.1185 0.2433 0.4094	0.1932 0.3213 0.4935
120	0.0637 0.1024 0.1167	0.0960 0.1948 0.3567	0.1562 0.2953 0.4536
150	0.0436 0.0923 0.1001	0.0568 0.1398 0.3105	0.1359 0.2125 0.3684

* The 10%, 5% and 1% upper percentile values are given in the first, second and third rows respectively

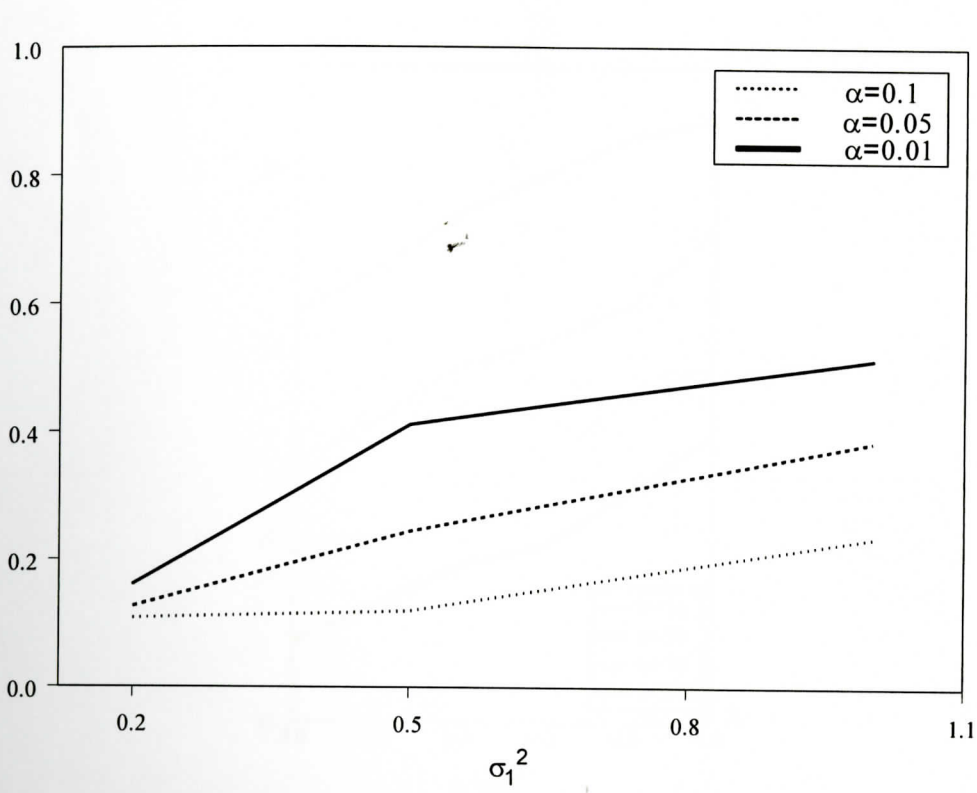

Figure 1: The Cut-off points at three level of significant α, for ***COVRATIO*** statistic, when sample size n=100 for different error variance σ_1^2 values.

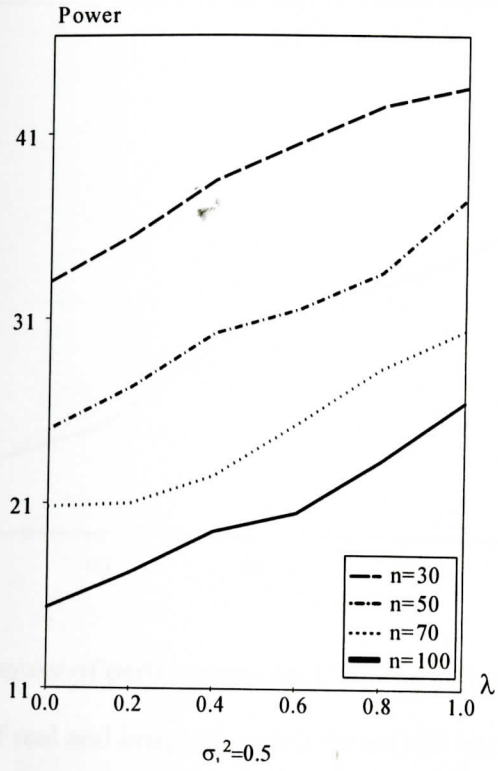

Figure 2: The power of performance for COVRATIO statistic, for the error variance σ_1^2=0.5 with different sample size.

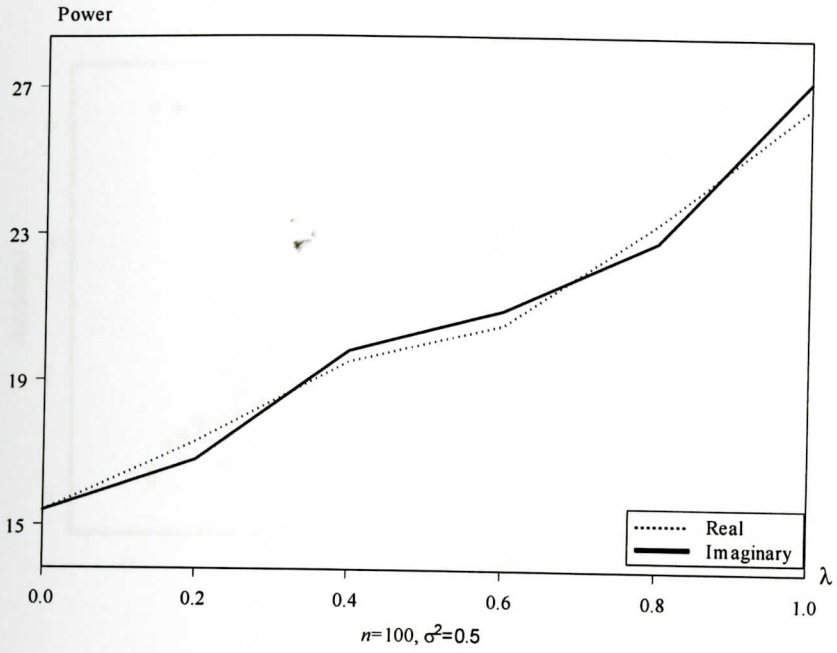

Figure 3: The power of performance for COVRATIO statistic after the contamination of real and imaginary parts for n=100 and σ_1^2=0.5.

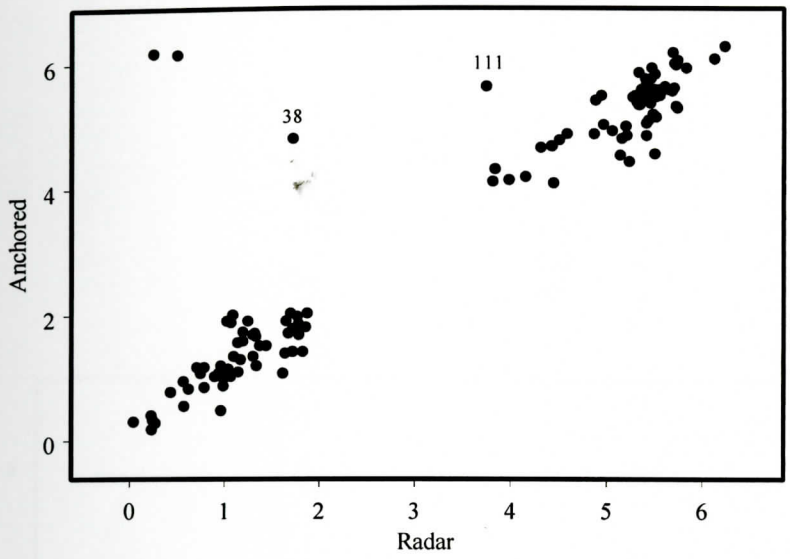

Figure 4: A scatter plot of wind direction data measured by HF radar system and anchored wave buoy

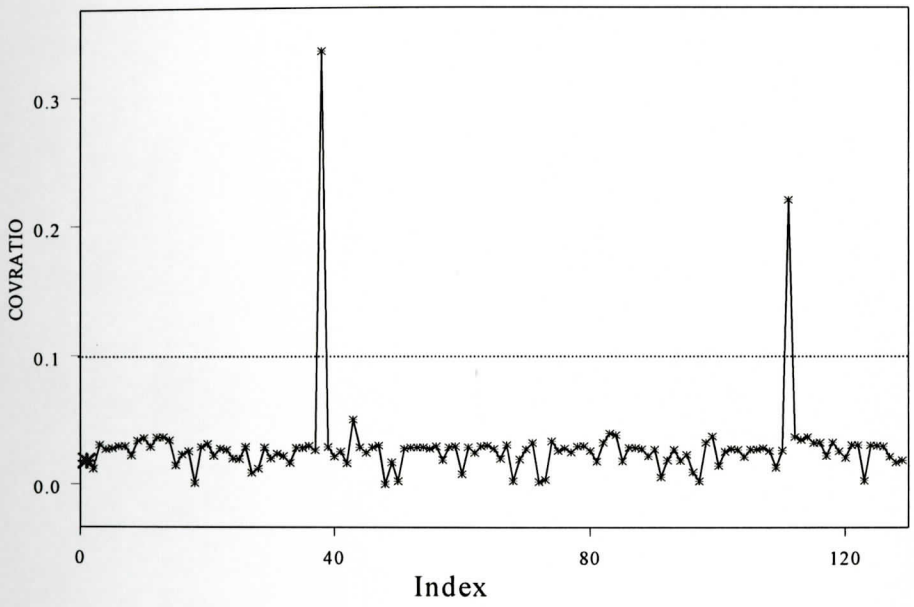

Figure 5: The values of the $\left|COVRATIO_{(-i)} - 1\right|$ statistic for wind direction data