

**OUTLIER DETECTION IN CIRCULAR DATA AND  
CIRCULAR-CIRCULAR REGRESSION MODEL**

**ADZHAR RAMBLI**

**FACULTY OF SCIENCE  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2011**

**OUTLIER DETECTION IN CIRCULAR DATA AND  
CIRCULAR-CIRCULAR REGRESSION MODEL**

**ADZHAR RAMBLI**

**DISSERTATION SUBMITTED IN FULFILLMENT OF  
THE REQUIREMENT FOR THE DEGREE  
OF MASTER OF SCIENCE**

**INSTITUTE OF MATHEMATICAL SCIENCES  
FACULTY OF SCIENCE  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2011**

## ABSTRAK

Secara keseluruhan, kajian ini mengenai lanjutan daripada kaedah-kaedah berangka untuk mengesan nilai tersisih bagi taburan sampel bulatan dan model regresi bulatan. Perkara yang pertama adalah melihat kepada taburan yang berlainan untuk sampel bulatan dan model regresi bulatan. Parameter-parameter yang dapat menggambarkan setiap taburan bulatan akan dibincang dan ilustrasi menggunakan data yang sebenar. Bagi model regresi bulatan, kita akan mempertimbangkan model regresi bulatan DM yang dikemukakan oleh Down dan Mardia (2002). Kemudian kami menggunakan kaedah Anggaran Kebolehjadian Maksimum (MLE) untuk menganggarkan parameter-parameter yang didapati dalam model tersebut.

Perkara yang kedua adalah mempertimbangkan kehadiran nilai tersisih dalam data bulatan yang dijana daripada taburan keluarga “ $\alpha$ -stable wrapped”, iaitu taburan “wrapped normal”. Empat kaedah-kaedah berangka statistik  $C$ ,  $M$ ,  $D$  and  $A$  dipertimbangkan untuk mengesan kehadiran nilai tersisih. Titik potongan bagi kesemua statistik diperolehi berdasarkan kajian simulasi. Kajian simulasi menunjukkan bahawa statistik  $A$  mempunyai prestasi lebih baik berbanding statistik-statistik yang lain. Kajian simulasi, didapati bahawa titik potongan amat bergantung kepada ukuran penumpuan sampel bulatan. Kaedah-kaedah ini telah diaplikasikan kepada data arah angin di Kuantan dan telah berjaya mengesan sampel yang terletak jauh daripada sampel-sampel yang lain sebagai nilai tersisih.

Ketiganya adalah masalah untuk mengesan kewujudan nilai tersisih dengan menggunakan model regresi bulatan DM berdasarkan kepada dua jenis statistik;  $COVRATIO$  statistik dan  $DMCEs$  statistik. Titik potongan dan prestasi bagi kedua-dua statistik ini dikaji secara simulasi. Aplikasi bagi kedua-dua statistik ini dilakukan dengan

menggunakan data arah angin laut dan irama biologi sirkadian. Sekali lagi, kedua-dua statistik ini berjaya untuk mengenal pasti sepasang sampel sebagai nilai tersisih.

## ABSTRACT

This study focuses on studying several numerical methods used to detect outliers in circular data and circular-circular regression models. Firstly, we will look at different distributions for circular variable and different regression models involving bivariate circular data. The parameters that describe each circular distribution will be explained in detail and illustrated using real data sets. As for the circular regression, we will consider the DM circular regression model proposed by Down and Mardia (2002). We employ the maximum likelihood estimation method to estimate the parameters of the model.

Secondly, we consider the occurrence of outliers in circular data generated from a family of  $\alpha$ -stable wrapped distribution, that is, the wrapped normal distribution. Four numerical methods; the  $C$ ,  $M$ ,  $D$  and  $A$  statistics are considered to detect the outliers. The cut-off points of the statistics are obtained via simulation. In our case, we show through simulation that the  $A$  statistic performs better in detecting outliers than the other statistics. The methods are applied on the Kuantan wind direction data set and are able to detect observations further away from the rest as outliers.

Thirdly, we consider the problem of detecting a single influential observation in the DM circular regression based on two different statistics; the  $COVRATIO$  statistic and  $DMCEs$  statistic. The cut-off points and the performance of both procedures are studied via simulation. The application of the procedures is illustrated by the ocean wind direction data and circadian biological rhythm data respectively. Again, the procedures are able to identify outlying observation as possible influential observation.

## ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisors, Assoc. Prof. Dr. Ibrahim bin Mohamed and Assoc. Prof. Dr. Abdul Ghapor bin Hussin for their guidance, invaluable help and encouragement throughout the period of my research.

Secondly, huge thanks go to my mother, brothers and sisters for their support and encouragement at all stages of this study. With their blessing and prayer, I am very grateful to eventually complete this study.

Next, I would like to gratefully acknowledge and the support from Assoc. Prof. Dr. Mohd Omar, Head of Department of the Institute of Mathematical Sciences (ISM), University of Malaya (UM) and all staff members of ISM, UM especially Puan Budiyah Yeop and En. Abd Malek Osmani. I would also like to thank the University of Malaya for the financial support.

Special thanks go to my friends; Ali, Safwati, Syuhada, Mardziah, Norli, Siti, and Amir for their assistance and support during the course of study.

# CONTENTS

	Page
<b>ABSTRAK</b>	i
<b>ABSTRACT</b>	iii
<b>ACKNOWLEDGEMENT</b>	iv
<b>CONTENTS</b>	v
<b>LIST OF TABLES</b>	viii
<b>LIST OF FIGURES</b>	ix
<b>CHAPTER 1</b>	
1.1 Background of the Study	1
1.2 Problem Statement	5
1.3 Objectives	5
1.4 Thesis outline	6
<b>CHAPTER 2</b>	
2.1 Introduction	8
2.2 Descriptive Statistics of Univariate Circular Data	8
2.3 Circular Graphs	11
2.4 The Distributions of Circular Data	13
2.4.1 The circular uniform distribution	14
2.4.2 The von Mises ( <i>VM</i> ) distribution	14
2.4.3 The general Wrapped Stable ( <i>WS</i> ) distribution	16
2.4.4 The wrapped Cauchy ( <i>WC</i> ) distribution	17
2.4.5 The wrapped normal ( <i>WN</i> ) distribution	19
2.4.6 Discussion	20
2.5 Goodness of Fit Test	21

2.6	Practical Example	23
2.7	Summary	24

### **CHAPTER 3**

3.1	Introduction	25
3.2	Outlier Detection Methods in Univariate Circular Data	25
3.3	Numerical Tests	27
3.4	The cut-off point for the C, M, D and A Statistics for <i>WN</i> Distribution	30
3.5	Performance of the Statistics for <i>WN</i> Distribution	35
3.6	Practical Example	40
3.7	Summary	41

### **CHAPTER 4**

4.1	Introduction	42
4.2	Down and Mardia Circular Regression (DM) Model	44
4.3	The Maximum Likelihood Estimation of Parameters in DM Model	45
4.4	Covariance Matrix of Circular Regression Model	46
4.5	Practical Example	47
	4.5.1 Ocean wind direction data	47
	4.5.2 Circadian data	51
4.6	Summary	54



## **CHAPTER 5**

5.1	Introduction	55
5.2	Robustness of Maximum Likelihood Estimation (MLE) Method	55
5.3	Graphical Techniques	58
5.4	COVRATIO Statistic	60
5.5	Sampling Behaviour of the COVRATIO Statistic	61
5.6	Power of Performance of COVRATIO Statistic	63
5.7	Practical Example	65
5.8	Summary	67

## **CHAPTER 6**

6.1	Introduction	68
6.2	Circular Residuals	68
6.3	Mean Circular Error	69
6.4	Sampling Behavior of the <i>DMCEs</i> Statistic	70
6.5	Power of Performance of <i>DMCEs</i> Statistic	73
6.6	Practical Example	74
6.7	Summary	76

## **CHAPTER 7**

7.1	Summary of the Study	77
7.2	Significance of the Study	78
7.3	Further Research	79

<b>APPENDIX</b>	80
-----------------	----

<b>REFERENCES</b>	93
-------------------	----

## LIST OF TABLES

	Page
Table 2.1: Kuantan wind direction data	23
Table 2.2: Descriptive statistics	23
Table 3.1: Table of cut-off point of C statistic	31
Table 3.2: Table of cut-off point of M statistic	32
Table 3.3: Table of cut-off point of D statistic	33
Table 3.4: Table of cut-off point of A statistic	34
Table 3.5: Result based on <i>C</i> , <i>M</i> , <i>D</i> and <i>A</i> statistics	40
Table 4.1: Descriptive statistics for the ocean wind direction data	50
Table 4.2: Descriptive statistics for the circadian data	52
Table 5.1: Estimate of parameters and Biasness (True value $\alpha = 0.5$ , $\beta = 0.5$ , $\omega = 0.2$ )	56
Table 5.2: Estimate of parameters and Biasness (True value $\alpha = 2.5$ , $\beta = 2.5$ , $\omega = 0.5$ )	57
Table 5.3: Estimate of parameters and Biasness (True value $\alpha = 1.5$ , $\beta = 1.5$ , $\omega = 0.4$ )	57
Table 5.4: Cut-off point of COVRATIO statistic	62
Table 5.5: Result based on COVRATIO statistic	66
Table 5.6: Effect of influential observation on parameter estimates	67
Table 6.1: Cut-off point of the DMCEs statistic	72
Table 6.2: Effect of influential observation on parameter estimates	75

## LIST OF FIGURES

	Page
Figure 1.1: Linear plot	2
Figure 1.2: Circular plot	2
Figure 2.1: Rose histogram	12
Figure 2.2: Circular histogram	12
Figure 2.3: Arrow histogram	12
Figure 2.4: Raw data plot	12
Figure 2.5: The von Mises distribution; $n=20, \kappa=3$	15
Figure 2.6: The von Mises distribution; $n=20, \kappa=5$	15
Figure 2.7: The von Mises distribution; $n=20, \kappa=10$	16
Figure 2.8: The wrapped Cauchy distribution; $n=20, \rho = 0.3$	18
Figure 2.9: The wrapped Cauchy distribution; $n =20, \rho = 0.7$	18
Figure 2.10: The wrapped Cauchy distribution; $n =20, \rho = 0.975$	18
Figure 2.11: The wrapped normal distribution; $n =20, \rho = 0.3$	20
Figure 2.12: The wrapped normal distribution; $n =20, \rho = 0.7$	20
Figure 2.13: The wrapped normal distribution; $n =20, \rho = 0.975$	20
Figure 2.14: Circular histogram of Kuantan wind data	24
Figure 2.15: Circular plot of Kuantan wind data	24
Figure 3.1: Circular plot of univariate data	26
Figure 3.2: P-P plot of univariate data	26
Figure 3.3: Performance of the Statistics ( $WN$ ), (a), (b) and (c) for $n=50$ and $\rho=0.90$	37
Figure 3.4: Performance of the Statistics ( $WN$ ),	38

(a), (b) and (c) for $n=50$ and $\rho=0.975$	
Figure 3.5: Performance of the Statistics ( $WN$ ),	39
(a), (b) and (c) for $n=20$ and $\rho=0.90$	
Figure 3.6: Circular plot of Kuantan wind data	40
Figure 4.1: Circular Histogram for HF	49
Figure 4.2: Circular Histogram for AB	49
Figure 4.3: Q-Q plot for HF and AB	49
Figure 4.4: Plot for finding the precision parameter	50
Figure 4.5: P-P plot of the residuals	51
Figure 4.6: Circular Histogram for S1	52
Figure 4.7: Circular Histogram for S2	52
Figure 4.8: Q-Q plot for S1 and S2	52
Figure 4.9: Plot for finding the precision parameter	53
Figure 4.10: P-P plot of the residuals	54
Figure 5.1: Circular histogram	59
Figure 5.2: Circular residuals versus index	59
Figure 5.3: Q-Q plot for circular residuals	59
Figure 5.4: Circular boxplot	59
Figure 5.5: A sample of Spoke plot for variables $u$ and $v$	59
Figure 5.6: Power of performance of COVRATIO at $n=70$	64
Figure 5.7: Power of performance of COVRATIO for $\kappa=10$	64
Figure 5.8: Spoke plot of wind data	65
Figure 5.9: Plot $p$ versus index for 1 <sup>st</sup> iteration	66
Figure 5.10: Plot $p$ versus index for 2 <sup>nd</sup> iteration	67
Figure 6.1: Power of performance of DMCEs at $n=70$	74

Figure 6.2: Power of performance of DMCEs at $\kappa=10$	74
Figure 6.3: Spoke plot of Circadian data	75