PARAMETER ESTIMATION AND OUTLIER DETECTION FOR SOME TYPES OF CIRCULAR MODEL

SITI ZANARIAH BINTI SATARI

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

PARAMETER ESTIMATION AND OUTLIER DETECTION FOR SOME TYPES OF CIRCULAR MODEL

SITI ZANARIAH BINTI SATARI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

ABSTRACT

This study focuses on the parameter estimation and outlier detection for some types of the circular model. We first look at the concentration parameter of von Mises distribution. The von Mises distribution is the most commonly used probability distribution of a circular random variable, and the concentration of a circular data set is measured using the mean resultant length. We propose a new and efficient approximation of the concentration parameter estimates using two approaches, namely, the roots of a polynomial function and minimizing the negative value of the log-likelihood function in this study.

Secondly, we consider the construction of confidence interval for the unknown parameter of a type of circular regression model, namely the model by Down and Mardia (2002). The parameters being considered in this study is the error concentration parameter. The confidence interval of the error concentration parameter is not straight forward due to the complexity of getting a closed form and the wrap-around nature of the data. In this study, we propose an alternative method of constructing a confidence interval based from the distribution of the estimated value of error concentration parameter obtained from the Fisher information matrix.

Thirdly, a new functional relationship model for circular variables, which is an extended version of a circular regression model as proposed by Down and Mardia (2002), is developed in this study. Both the dependent and independent variables in the model are subjected to errors. We derive the maximum likelihood estimation of parameters as well as the variance-covariance of parameters. Later, we assess the

performance of confidence interval for error concentration parameter for the new functional relationship model via simulation study.

Lastly, we consider the problem of detecting multiple outliers in circular regression models based on the clustering algorithm. We develop the clustering-based procedure for the predicted and residual values obtained from the Down and Mardia model fit of a circular-circular data set. Here, we introduce a measure of similarity based on the circular distance and obtain a cluster tree using the single linkage clustering algorithm. Then, a stopping rule for the cluster tree based on the mean direction and circular standard deviation of the tree height is proposed. We classify the cluster group that exceeds the stopping rule as potential outliers.

Model verification of all method and model proposed in this study are examined using the simulation study. As illustration, applications are displayed using wind and wave circular data sets.

ABSTRAK

Kajian ini memberi tumpuan kepada penganggaran parameter dan ujian pengesanan titik terpencil dalam beberapa jenis model bulatan. Pertama, kita akan melihat parameter menumpu bagi taburan von Mises. Taburan von Mises adalah taburan kebarangkalian yang paling kerap digunakan bagi pemboleh ubah rawak bulat dan penumpuan bagi set data bulatan dikira menggunakan min paduan panjang. Dalam kajian ini, nilai anggaran yang baharu dan efisien bagi parameter menumpu data membulat tersebut telah dicadangkan dengan menggunakan dua pendekatan iaitu, nilai punca fungsi polinomial dan meminimumkan nilai negatif bagi fungsi log-kemungkinan.

Kedua, kami mempertimbangkan pembinaan selang keyakinan untuk parameter yang tidak diketahui bagi sejenis model regresi membulat, iaitu model Down dan Mardia (2002). Parameter yang dipertimbangkan dalam model ini adalah parameter menumpu bagi ralat. Selang keyakinan parameter menumpu bagi ralat tidak boleh ditulis secara langsung kerana kerumitan mendapatkan bentuk yang tertutup dan membalut sekitar sifat data. Dalam kajian ini, kami mencadangkan satu kaedah alternatif dalam membina selang keyakinan berasaskan taburan bagi nilai anggaran parameter menumpu ralat yang diperolehi daripada Matriks Fisher Bermaklumat.

Ketiga, satu model hubungan berfungsi yang baharu bagi pembolehubah bulatan, yang mana merupakan lanjutan daripada model regresi bulatan yang dicadangkan oleh Down dan Mardia (2002) dibangunkan dalam kajian ini. Kedua-dua pembolehubah bersandar dan tak bersandar dalam model ini adalah tertakluk kepada ralat. Kami memperoleh nilai anggaran kemungkinan maksimum bagi parameter

beserta parameter varians-kovarians. Seterusnya, kami juga menilai prestasi selang keyakinan untuk parameter menumpu nilai ralat bagi model hubungan berfungsi yang baharu ini melalui kajian simulasi.

Akhir sekali, kami mempertimbangkan masalah mengesan titik terpencil berganda dalam model regresi bulatan berdasarkan algoritma berkelompok. Kami membangunkan prosedur berasaskan kelompok ini untuk nilai ramalan dan reja yang diperoleh daripada penyuaian model Down dan Mardia bagi set data bulatan. Di sini, kami memperkenalkan ukuran persamaan berasaskan jarak bulatan, dan seterusnya membina pokok kelompok dengan menggunakan algoritma hubungan kelompok tunggal. Kemudian, kami mencadangkan satu nilai potongan untuk pokok kelompok berdasarkan min terarah dan sisihan piawai bulatan bagi ketinggian pokok tersebut. Kami mengklasifikasikan kumpulan data yang melebihi titik potongan ini sebagai titik terpencil.

Pengesahan model bagi semua kaedah dan model yang dicadangkan dalam kajian ini diuji menggunakan kajian simulasi. Sebagai contoh, aplikasi dipaparkan menggunakan data arah angin dan gelombang.

ACKNOWLEDGEMENTS

"In the name of Allah the All-Merciful, the All-Compassionate."

All praise is to Allah for giving me the strength and chance to complete this thesis. A very great gratitude and appreciation is expressed to all those who contributes towards the completion of the thesis entitle "*Parameter Estimation and Outlier Detection for Some Types of Circular Model"*, either directly or indirectly. Particularly, I would like to express my grateful and sincere appreciation to my supervisors, Associate Professor Dr. Yong Zulina Zubairi and Professor Dr. Abdul Ghapor Hussin for their support, guidance, commentaries and advices. Also huge thanks, for their trust and confidence on my capability to complete this thesis.

My gratitude and special thanks also go to my family for their support, love and encouragement throughout the course of my study in University of Malaya (UM). Thank you for always being patient and considerate. Last but not least, a huge thank you to my fellows' postgraduate friends for their support, assistance, and cooperation. My sincere appreciation also extends to all my friends and others who have provided assistance at various occasions. Their view and tips are useful indeed. With blessing from everyone, I am very grateful to eventually complete this study.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

LIST OF SYMBOLS

κ	Concentration parameter
$\hat{\kappa}$	Estimate value of concentration parameter
$\mathbf{x}$	A unit vector in the plane
θ	An angle on the circle (A circular data)
z	A unit complex number on a circle
R	Resultant length
$\bar{R}$	Sample Mean resultant length
$\bar{\theta}$	Sample Mean direction
μ	Mean direction
$\hat{\mu}$	Estimate value of Mean direction
ρ	Mean resultant length (precision parameter)
$A(\kappa)$	Mean resultant length for von Mises distribution
n	Sample sizes
$\tilde{\theta}$	Sample Median direction
V	Sample circular variance
υ	Sample circular standard deviation
$\hat{\delta}$	Sample circular dispersion
$d_0(\alpha)$	Circular mean deviation
$\bar{D}_0$	Circular mean difference
$I_p(\kappa)$	The modified Bessel function of the first kind and order p
B	Bootstrap sample sizes

y	Dependent (response) variable
x	Independent (predictor) variable
e	Angular error
$\hat{e}$	Fitted error
ε	Circular random error for dependent variable
δ	Circular random error for independent variable
λ	Ratio of error concentration parameters in a functional relationship model
h	Height of a cluster tree
$\overline{h}$	Average height of a cluster tree
s_h	Sample standard deviation of the heights of a cluster tree
d	Circular distance
$\boldsymbol{d}_{ij}$	Distance between observation i and j
$\overline{\hat{\kappa}}$	means of concentration parameter
v	Dependent (response) variable for Down and Mardia model
u	Independent (predictor) variable for Down and Mardia model
$\hat{v}$	Predicted (fitted) values
s	Number of simulation
Y	Dependent random variable for a functional relationship model
X	Independent random variable for a functional relationship model
$\tilde{\kappa}$	Corrected estimator of concentration parameter
$\mathbf{I}$	Fisher information matrix

LIST OF ABBREVIATIONS

AC	Agglomerative coefficient
AEB	Absolute Estimated Bias
cdf	Cumulative distribution function
CI	Confidence interval
cov	Covariance
DM	Down and Mardia
DMCE	Difference Mean Circular Error
DMCEs	Difference Mean Circular Error Statistic
EAE	Estimated Absolute Error
EIV	Error in variables
ERE	Estimated Relative Error
ESE	Estimated Standard Error
ERMSE	Estimated Root Mean square Error
JS	Jammalamadaka and SenGupta
LS	Least square
ME1	Method 1
ME2	Method 2
MLE	Maximum likelihood estimation
MS	Minimum sum method
ms	Minimum sum function
pdf	Probability density function
Prob	Probability
pmask	Masking error

pswamp	Swamping error
polyroot	Polyroot method
pout	“Success” probability of detecting outliers
se	Standard error
var	Variance
VM	Von Mises distribution

CHAPTER 1 : RESEARCH FRAMEWORK

1.1 Background of the Study

Circular statistics is a branch of statistics that involve circular data in the form of direction or cyclic time. Circular data are measured in degree $\left(0°, 360°\right]$ or radian $\left(0, 2\pi\right]$. Examples of circular data include the days of the week and compass direction. Since the data are cyclic, Monday is said to be closer to Sunday than to Wednesday, while 350^{o} closer to 1^{o} than to 300^{o}. We can found the applications of circular statistics in various areas such as in biology, geology, geography and medical. For example, biologist used circular statistics to study the orientation of an animal while meteorologist used this method to study the direction of the wind. Mardia and Jupp (2000) stated that circular data can also be identified in waves of sound, the human perception under various conditions, the orientation of ridges of fingerprints, the orientation of sand grains from a beach, the death due to a disease at various times in a year, and astronomical observations.

Circular data are usually measured using a compass, clock, protractor or other circular measuring instruments. We represent a single circular observation $\theta°$ as a point on a circle of unit radius or unit vector. Generally, the researchers will choose an initial direction and an orientation of the circle based on the needs and nature of the studies. This situation is described and illustrated in Figure 1.1. The idea is to choose an orthogonal coordinate system on the plane. Then, each circular observation can be specified by the angle from the initial direction to a point on the circle corresponding to the observation (Mardia and Jupp, 2000).

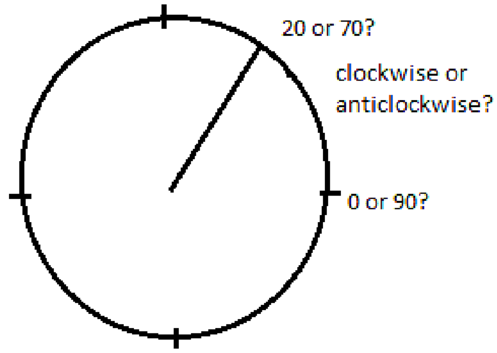

Figure 1.1: Initial direction and orientation of the circle

New theories and statistical methods of circular data are developed over the years. However, we believe that many statistical aspects can further be improved and refined. In particular, we may focus on the parameter estimation of the von Mises distribution. The von Mises distribution is the most common continuous probability distribution on the circle and known as the circular normal distribution because of its close relationship to the normal distribution in a real line. It plays a major role in many statistical inferences on the circle such as the sampling distribution, confidence interval, and prediction. One of the parameters being considered in the von Mises distribution is the concentration parameter. According to Mardia and Jupp (2002, p39-41) and Jammaladaka and SenGupta (2001, p35-42), the concentration parameter in von Mises distribution describes how concentrated the data is towards the center. Several solutions can be found in the literature since the parameter estimation of the concentration parameter cannot be obtained analytically.

Apart from univariate circular data, researchers sometimes deal with bivariate circular data. Many researchers show strong interests in circular regression models and circular functional relationship models. According to Down and Mardia (2002), circular regression methods have been used in many diverse applications since 1957.

CHAPTER 1 : RESEARCH FRAMEWORK

1.1 Background of the Study

Circular statistics is a branch of statistics that involve circular data in the form of direction or cyclic time. Circular data are measured in degree $\left(0°, 360°\right]$ or radian $\left(0, 2\pi\right]$. Examples of circular data include the days of the week and compass direction. Since the data are cyclic, Monday is said to be closer to Sunday than to Wednesday, while 350^{o} closer to 1^{o} than to 300^{o}. We can found the applications of circular statistics in various areas such as in biology, geology, geography and medical. For example, biologist used circular statistics to study the orientation of an animal while meteorologist used this method to study the direction of the wind. Mardia and Jupp (2000) stated that circular data can also be identified in waves of sound, the human perception under various conditions, the orientation of ridges of fingerprints, the orientation of sand grains from a beach, the death due to a disease at various times in a year, and astronomical observations.

Circular data are usually measured using a compass, clock, protractor or other circular measuring instruments. We represent a single circular observation $\theta°$ as a point on a circle of unit radius or unit vector. Generally, the researchers will choose an initial direction and an orientation of the circle based on the needs and nature of the studies. This situation is described and illustrated in Figure 1.1. The idea is to choose an orthogonal coordinate system on the plane. Then, each circular observation can be specified by the angle from the initial direction to a point on the circle corresponding to the observation (Mardia and Jupp, 2000).

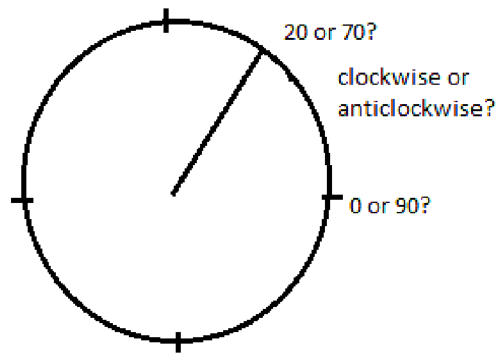

Figure 1.1: Initial direction and orientation of the circle

New theories and statistical methods of circular data are developed over the years. However, we believe that many statistical aspects can further be improved and refined. In particular, we may focus on the parameter estimation of the von Mises distribution. The von Mises distribution is the most common continuous probability distribution on the circle and known as the circular normal distribution because of its close relationship to the normal distribution in a real line. It plays a major role in many statistical inferences on the circle such as the sampling distribution, confidence interval, and prediction. One of the parameters being considered in the von Mises distribution is the concentration parameter. According to Mardia and Jupp (2002, p39-41) and Jammaladaka and SenGupta (2001, p35-42), the concentration parameter in von Mises distribution describes how concentrated the data is towards the center. Several solutions can be found in the literature since the parameter estimation of the concentration parameter cannot be obtained analytically.

Apart from univariate circular data, researchers sometimes deal with bivariate circular data. Many researchers show strong interests in circular regression models and circular functional relationship models. According to Down and Mardia (2002), circular regression methods have been used in many diverse applications since 1957.

Some examples include Gould (1969) who proposed an angular response regression model and applied the model to a set of data arising from a study of movements of intertidal gastropods. Mardia (1975) developed a nonparametric rank correlation coefficient for circular data.

In particular, the circular functional relationship model is known as errors-in-variables models or measurement errors models which account for measurement errors in the independent variables. The problem of estimating the parameters are usually considered for both types of the statistical model. One of the parameters being considered in the circular model is the error concentration parameter.

In most model building, the presence of influential observation, outliers and missing values cannot be ignored. Outliers may occur in circular data and become unnoticed. It may be a result of keypunch errors, misplaced decimal points, recording or transmission error, or exceptional phenomena such as earthquakes and natural disasters. If outliers in circular regression model remain undetected, it can lead to erroneous parameter estimates and inferences from the model. In this study, our focus is on multiple outlier detection procedures in circular regression models.

1.2 Problem Statement

In this study, our main focus is on circular regression model and circular functional relationship model for von Mises distribution. For the von Mises distribution, procedures for obtaining the parameter estimates are not straight forward as it is mathematically intractable. In particular, the maximum likelihood estimation of the concentration parameter κ involves inverting the mean resultant length. In the literature, some estimates of the concentration parameter are better for small κ values

CHAPTER 3 : A NEW EFFICIENT APPROXIMATION OF CONCENTRATION PARAMETER

3.1 Introduction

The aim of this chapter is to propose an efficient approximation for mean resultant length $A(\kappa)$ and concentration parameter $\hat{\kappa}$. For that reason, a brief discussion on the general construction of mean resultant length $A(\kappa)$ is given in Section 3.2, and the existence approximation solution of the concentration parameter is listed for large and small κ in Section 3.3. A new formula of $A(\kappa)$ is constructed using the reconstruction of summation series of $I_0(\kappa)$ using two approaches namely, piecewise approximation and maximum likelihood estimator. The new approximation method is given in Section 3.4.

Furthermore, new approximation solutions of $\hat{\kappa}$ is also proposed using two approaches in Section 3.5. First, we consider the power series expansion of the mean resultant length and the estimate of the concentration parameter may be obtained by the roots of a polynomial function. Detailed description is given in Section 3.6. Secondly, we consider the power series expansion of the reciprocal of a Bessel function in the log-likelihood function of the concentration parameter and the estimate of concentration parameter may be obtained by minimizing the negative value of the log-likelihood function. Detailed description is given in Section 3.7.

The concentration parameter from both approaches may be estimated, for example, using the polyroot function and minimum sum function in the `SPlus` package. The efficiency of new proposed method is then tested using a simulation study with random data, and again with applications data. Detailed descriptions are given in Section 3.8 to Section 3.10. Result, discussion and conclusion of new proposed method are provided at the end of this chapter.

3.2 The General Construction of Mean Resultant Length

A brief discussion on the general construction of mean resultant length of von Mises distribution can be found in Mardia and Jupp (2000), and Jammalamadaka and SenGupta (2001). Recall that, the von Mises distribution is symmetrical about mean direction μ and has pdf

$$f(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)}, \quad 0\le\theta<2\pi, \tag{2.12}$$

where $0\le\mu<2\pi$ and $\kappa\ge0$ are parameters. Thus, the distribution is invariant under the transformation

$$\theta\mapsto\mu-\theta, \tag{3.1}$$

and its density has the following property,

$$f(\theta-\mu)=f(\mu-\theta). \tag{3.2}$$

Generally, the value of mean direction $\bar{\theta}$ and mean resultant length $\bar{R}$ are obtained from the first trigonometric moment about zero direction. The p^{th} trigonometric moment about zero direction is given by

$$\phi_p=\alpha_p+i\beta_p=E\left(e^{ip\theta}\right)=\int_0^{2\pi}e^{ip\theta}dF(\theta), \quad p=0,\pm1,\pm2,\ldots \tag{3.3}$$

where

$$\phi_0 = 1, \qquad |\phi_p| \le 1,$$

complex conjugate of ϕ_p, $\overline{\phi}_p = \phi_{-p}$,

$$\alpha_p = E(\cos p\theta) = \int_0^{2\pi} \cos p\theta dF(\theta),$$

$$\beta_p = E(\sin p\theta) = \int_0^{2\pi} \sin p\theta dF(\theta),$$

$F(\theta) = \int_0^{2\pi} \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)} d\theta$ is cumulative density function,

and $\alpha_{-p} = \alpha_p, \quad \beta_{-p} = \beta_p, \quad |\alpha_p| \le 1, \quad |\beta_p| \le 1$.

Hence, for $p \ge 0$,

$$\phi_p = \rho_p e^{i\mu_p}, \tag{3.4}$$

where $\rho_p = \sqrt{\alpha_p^2 + \beta_p^2}$, $\mu_p = \tan^{-1}\left(\frac{\beta_p}{\alpha_p}\right)$ and $\rho_p \ge 0$. When $p = 1$,

$$\phi_1 = \rho_1 e^{i\mu_1} = \rho e^{i\mu}. \tag{3.5}$$

Also, the p^{th} trigonometric moment about mean direction is defined by

$$\overline{\phi}_p = E\left(e^{ip(\theta-\mu)}\right) = \int_0^{2\pi} e^{ip(\theta-\mu)} dF(\theta-\mu) = \overline{\alpha}_p + i\overline{\beta}_p, \tag{3.6}$$

where

$$\overline{\alpha}_p = E(\cos p(\theta-\mu)) = \int_0^{2\pi} \cos p(\theta-\mu) dF(\theta-\mu), \text{ and}$$

$$\overline{\beta}_p = E(\sin p(\theta-\mu)) = \int_0^{2\pi} \sin p(\theta-\mu) dF(\theta-\mu).$$