OUTLIER DETECTION IN CYLINDRICAL DATA

NURUL HIDAYAH BINTI SADIKON

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

OUTLIER DETECTION IN CYLINDRICAL DATA

NURUL HIDAYAH BINTI SADIKON

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (MATHEMATICS)

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

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: NURUL HIDAYAH BINTI SADIKON

Registration/Matric No.: SGP140002

Name of Degree: MASTER OF SCIENCE (MATHEMATICS)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

OUTLIER DETECTION IN CYLINDRICAL DATA

Field of Study: CIRCULAR STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

OUTLIER DETECTION IN CYLINDRICAL DATA

ABSTRACT

A cylindrical data set consists of a circular and a linear variables. Few distributions have been proposed for such data pioneered by Johnson and Wehrly (1978). In this study, we look at two problems of detecting outliers in cylindrical data. Firstly, we define outlier in cylindrical data and propose a new test of discordancy to detect outlier in cylindrical data generated from Johnson-Wehrly distribution. Secondly, we focus on detecting outliers in Johnson-Wehrly circular-linear regression model. In both cases, the outlier detection procedures are developed using the k-nearest neighbor distance. The cut-off points are obtained and the performance of the new statistic is examined via simulation. A practical example is presented using the wind data set from the Malaysian Meteorological Department. The findings of the study should lead to better inferences, model fitting and forecasting of cylindrical data sets.

Keywords: cylindrical data, cylindrical regression model, detection procedure, k-nearest neighbor's method, outlier

PENGESANAN CERAPAN TERSISIH DALAM DATA SILINDER

ABSTRAK

Data silinder adalah data yang mengandungi satu pembolehubah linear dan satu pembolehubah bulatan. Beberapa taburan untuk data silinder telah dicadangkan yang dipelopori oleh Johnson dan Wehrly (1978). Kajian ini mempertimbangkan dua masalah untuk mengesan kewujudan cerapan tersisih di dalam data silinder. Pertamanya, kami mendefinasikan cerapan tersisih di dalam data silinder dan seterusnya mencadangkan satu ujian sejajar yang baru untuk mengesan kewujudan cerapan tersisih dalam data silinder yang di jana daripada taburan Johnson-Wehrly. Keduanya, kami fokus untuk mengesan kewujudan cerapan tersisih di dalam model regresi bulatan-linear Johnson-Wehrly. Bagi kedua-dua kes tersebut, prosedur pengesanan cerapan tersisih dibina menggunakan teori jarak jiran k-terdekat. Titik potongan diperolehi dan prestasi bagi kedua-dua ujian tersebut diperiksa secara simulasi. Contoh praktikal dipersembahkan menggunakan set data angin daripada Jabatan Meteorologi Malaysia. Hasil daripada kajian ini seharusnya membawa kesan yang lebih baik kepada inferens, pemadanan model dan peramalan set data silinder.

Kata kunci: data silinder, model regresi silinder, prosedur pengesanan, kaedah jiran k-terdekat, cerapan tersisih

ACKNOWLEDGEMENTS

Bissmillahirrahmanirrahim,

Alhamdulillah. Thanks to Allah SWT, whom with His wills gives me the opportunity to complete this Master Degree study. First and foremost, I would like to thank my supervisors, Dr. Adriana Irawati Nur Binti Ibrahim and Prof. Dr. Ibrahim Bin Mohamed for their guidance, help, advices and encouragement throughout my study. They are my inspiration and the persons I consult the most during the period of this study. I must admit it was quite a bit of challenge to grasp with the knowledge that is quite new in my field of study. Without their persistent help, this dissertation would not have been possible. Not only they give me advises on this study, they also give me advises on how to be more presentable and confident.

Secondly, I would like to convey my special thanks to Dr. Adzhar as he also have helped me a lot in completing this study. Last but not least, to my family and my fellow friends who have always given me plenty of supports and for being understanding during the period of finishing this study. Honestly, this is an achievement I would not succeed by myself. Thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

k-NN : k-nearest neighbor.

BFGS : Broyden-Fletcher-Goldfarb-Shanno method.

BHHH : Berndt-Hall-Hall-Hausman method.

CG : Conjugate gradients method.

CN : Circular normal distribution.

ESD : Extreme Studentized Deviate.

hPa : Hectopascals.

JW : Johnson-Wehrly.

MLE : Maximum likelihood estimation.

MSE : Mean squared error.

NM : Nelder-Mead method.

NR : Newton-Raphson method.

RMSE : Root mean squared error.

SANN : Simulated annealing method.

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Linear data set is commonly found in the real world albeit of different forms in various areas. At the same time, there are also variables measured in degrees or radians. Data with such variables are referred as directional data. Measure of directions of birds' migration and wind are examples of directional variables. The simplest form of directional data is the univariate circular data which can be depicted on a circle. Higher order of directional data and a mixture of directional/linear variables requires the development of different statistical tools from that for linear data and become the focus of this study.

A circular observation refers to a point on a circumference of a circle or a vector in a plane. Each circular observation can be specified by the angle between the initial point to the corresponding observation point on a circle. Circular data usually deals with direction and bounded in the range of $[0, 2\pi)$. Hence, we need to consider special statistical methods in analyzing such data, which cannot be carried out using the corresponding linear statistics. Circular data arise in many fields including those corresponding to two circular measuring instruments, the clock and the compass. The data measured using these measurements include the migration or homing of birds, the wind and wave directions, the direction of earthquake displacement and the arrival times (in 24-hour clock) of patient at an emergency unit in a hospital.

Meanwhile, cylindrical data is a form of bivariate data, where one component is measured on the circular scale while the other is linear. The variables can be represented in two different form. First, they can be represented in three dimensional polar coordinate (r, θ_i, x_i). Second, they can be represented in Cartesian coordinate (x_i, y_i, z_i) where $x_i = r\cos\theta_i$, $y_i = r\sin\theta_i$ and $z = x_i$. The joint distribution between the random vector of

linear and circular components is called a cylindrical distribution or a distribution on the cylinder.

The analysis of circular data as well as cylindrical data has attracted the interest of statisticians and researchers from different scientific fields due to the emergence of such data. Cylindrical data can arise in various applications, such as

1. Meteorology: wind direction and speed or ozone or temperature. For example, analysis and parameterizations of wind profiles in the low atmosphere (Perez et al., 2005) , analysis of wind speed and temperature (Perez et al., 2007) and wind speed and direction for wind energy analysis (Carta et al., 2008).
2. Biology: plant or animal migration patterns. For example, directions and distances moved by small blue periwinkles (Fisher & Lee, 1992).
3. Industrial applications: quantifying the imbalance of rotating parts as a direction of imbalance and magnitude in automotive industry with wheel, brake and engine component. For example, the relationship between the imbalance of rotating parts and the magnitude (Anderson-Cook, 2000).

There are different types of distributions available for the linear case such as uniform, normal, gamma and exponential. Normal distribution is often used in the linear case due to its valuable properties. Similarly, there are various literature in circular and cylindrical distributions. In the circular case, various distributions such as von Mises, wrapped Cauchy, wrapped normal, and Fisher are available in the literature. The von Mises distribution is the most common distribution used for circular data. This distribution can be considered as important as the normal distribution for the linear case. Meanwhile, in the cylindrical case, there have been limited attempts to handle such data. However, there are a few researchers who have been working on cylindrical models for the past few decades. Johnson and Wehrly

(1978) developed four circular-linear distributions based on the principle of maximum entropy and one circular-linear distribution with specified marginal distribution. In this study, we will use one of the distributions from Johnson and Wehrly (1978). On the same year, Mardia and Sutton (1978) also proposed a model for cylindrical variables. Then, an extension to the Mardia and Sutton (1978) model has been proposed by Anderson-Cook (1997). Anderson-Cook (2000) once again extended Mardia and Sutton first order model to form a second order model. Kato and Shimizu (2008) proposed a model which could be useful to fit cylindrical data with asymmetry and/or bimodality of marginal circular component based on the principle of maximum entropy. Their model is an extension from the Mardia and Sutton model.

In statistical modeling, regression analysis is one of the most important methods to estimate the relationship among the variables. For the linear case, linear regression can be found in various literature. In circular regression, the regression can be divided into three categories; namely (i) circular-circular regression: both the dependent and independent variables are circular (ii) circular-linear regression: the linear variable depends on the circular variable and (iii) linear-circular regression: the circular variable depends on the linear variable (Jammalamadaka & SenGupta, 2001). The regression for cylindrical data can be considered as the circular-linear regression or the linear-circular regression since both types use the circular and linear variables.

The existence of unexpected or outlying observations has been a concern in statistical analysis. These are often seen as contaminating the data and may affect the information that can be obtained from the data set. Hence, it is natural to find a way to identify the outliers and treat them accordingly. Outlier is an observation having extremely large or small values (Barnett & Lewis, 1994). Meanwhile outlier in circular data is defined as an observation having large value of circular distance from the value of its two neighboring

observations on a unit circle (Mohamed et al., 2016). Just like the linear case, the existence of outliers in circular data will affect the parameter estimation as well as the forecasted values. Hence, it is very important to handle outlier problem by developing appropriate methods of identifying them.

There are several tests that can be used to detect outlier in circular data. Collett (1980) presented four different discordancy tests in circular data namely L, C, D and M statistics. Then Abuzaid et al. (2009) proposed a new test known as A statistic. On the other hand, there are also a few methods available in handling outliers in circular regression, particularly in circular-circular regression. Ibrahim et al. (2013) proposed the COVRATIO statistic and Abuzaid et al. (2013) and Rambli et al. (2016) proposed the Mean Circular Error (MCE) statistic using a row-deletion method with different approach. Abuzaid et al. (2013) used the circular distance between two circular observations while Rambli et al. (2016) transformed the residuals into linear scales using a trigonometric function. However, there is no literature available yet on the outlier detection in cylindrical data and cylindrical regression model, and hence these become the objectives of this study.

1.2 Problem Statement

The existence of outlier in data sets is one of the most common problems that may occur in any statistical analysis. Outlier in circular data can be defined as an observation having large value of circular distance from the value of its two neighboring observations on a unit circle (Mohamed et al., 2016). Similar problems have also been explored with circular regression models. While different outlier detection procedures have been developed for circular samples and regression models, no such work has been done on cylindrical data and the corresponding regression model. Hence, we intend to develop new theories on outliers in cylindrical data, and later to develop numerical methods of identifying outliers in this type of data and the corresponding regression model.

1.3 Objective

Based on the statement of the problem above, we have outlined the following objectives for this study:

1. To develop a discordancy test of identifying outliers in cylindrical data.
2. To propose a new procedure of outlier detection in the JW circular-linear regression model for cylindrical data.
3. To apply the proposed methods on real data sets.

1.4 Research Outline

This research attempts to handle the problem of outliers in cylindrical data and the corresponding regression model by proposing two new statistical methods. The research is outlined as follows:

Chapter two gives the literature review on the linear, circular and cylindrical data. The review on the linear, circular and cylindrical distributions are also presented as well as the outlier detection methods for these data. Then, we review the circular regression models and the outlier detection methods in the regression models. Some of the theories on the method used in the construction of the discordancy tests are also presented.

Chapter three discusses the properties of JW distribution and its parameter estimation. Some theories of the optimization method for estimation purposes are presented. Then, through simulation study, we investigate the accuracy of the estimation.

Chapter four presents the development of a new discordancy test of identifying outlier in the cylindrical data. We present the distance between two points on the surface of a cylinder. Through simulation, we obtain the cut-off points and study the performance of the test for a single outlier. The statistic is then applied on the wind direction and wind speed data.

Chapter five presents a new test statistic to detect outlier in the regression model for cylindrical data. We discuss the theory of the JW circular-linear regression model and the effect of outliers on the model. We obtain the cut-off points and investigate the performance of the test for a single outlier through simulation. The statistic is then applied on the wind data set.

Chapter six presents the summary of the research work. We also provide suggestions for extending the research work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, we review the theories related to linear, circular and cylindrical data. First, we look at the topics in circular data; the descriptive statistics and the distance measurement. Then, we review the circular and cylindrical distributions as well as the circular regression models. Next, we look at the outlier detection methods in circular data and circular regression. Finally, we review the theory of the k-nearest neighbor method which will be employed in the outlier detection in cylindrical data and its regression problems.

2.2 Outliers

In statistical analysis, one of the most common problems that arises is the existence of unexpected observations in data set which are called outliers. The existence of outlier can affect the results and violate model assumptions. However, outlier might provide useful information about the data when we look into the response of the study. Outlier can occur due to many reasons including error during recording, reading and calculating. If the collection of the data is correct, the outlier represents a rare event. When the existence of outlier are questionable to the data set, the pertaining values should be dealt with accordingly. Therefore, appropriate treatment of outlier has become an important discussion by many authors. Outlier refers to observation that stands out remarkably from other observations (Barnett & Lewis, 1994). However, the detection of outlier in circular data necessitate different method from the linear case. In circular data, the outlier is defined as an observation having large value of circular distance from the value of its two neighboring observations on a unit circle (Mohamed et al., 2016).

Many outlier detection methods have been developed for different types of data including linear and directional data. Various tests according to different distributions are discussed in Barnett and Lewis (1994). Iglewicz and Hoaglin (1993) reviewed five different tests for normal distribution, including generalized Extreme Studentized Deviate (ESD), Shapiro-Wilk and Dixon test.

In linear regression, there are many outlier detection methods that have been proposed. For the case of single outlier, Srikantan (1961) and Barnett and Lewis (1994) used residuals from the least square fit. Cook (1977) presented a new distance measure based on two maximum likelihood estimates, where one with full data set and one with reduced data set by removing a specific observation. Srivastava and Rosen (1998) proposed a likelihood ratio test for detecting single outlier in multivariate regression models. For the case of multiple outlier, Hadi and Simonoff (1993) proposed procedures and tests to detect outliers in univariate linear regression model. Barrett and Ling (1992) presented general classes of multivariate influence measure for a univariate regression based on Cook's influence measure.

In circular data, graphical and numerical methods are the most common tools used to investigate the outliers. The graphical methods include rose diagram, circular histogram, P-P plot and circular plot. To date, there are several tests that can be used to detect outlier in circular data. Collett (1980) pointed that the outlier detection in circular data highly depends on the concentration parameter. It is easier to find an outlier when the data is concentrated towards a particular direction than those with smaller concentration. The author also suggested that an observation with maximum value of circular distance $d(\theta_i, \theta_j)$ is expected to be a candidate of an outlier. The author also presented four different discordancy tests in circular data namely L, C, D and M statistics. Abuzaid et al. (2009) presented a new test of discordancy for univariate circular data, known as A statistic. The

author developed this statistic based on the summation of circular distance of the point interest to all other points.

The outlier detection in circular regression mainly focus on the circular-circular regression models. A number of circular regression models have been proposed by various authors. The earliest model was proposed by Laylock (1975) where the regression takes the form of multiple regression model with complex entries. Jammalamada and Sarma (1993) also proposed a regression model of two circular random variables by using the definition of characteristic function of a complex number. Then Rivest (1997) proposed a model to predict the y-direction based on the rotation of the decentred x-angle. On the other hand, Downs and Mardia (2002) applied a mapping method on their regression model to relate the variables and Hussin (2004) proposed a simple circular regression model with one independent circular variable only. Then, Kato et al. (2008) proposed a regression model where the regression curve is expressed as a form of Mobius circle transformation. Recently, SenGupta and Kim (2016) proposed a new circular-circular regression model for studying two circular genomes using Mobius transformation.

Abuzaid et al. (2011) and Ibrahim et al. (2013) extended the COVRATIO statistic method that is used in linear regression for circular-circular regression models to detect outlier. For circular-circular regression models, Abuzaid et al. (2013) and Rambli et al. (2016) proposed new outlier detection methods in this type of regression models called Mean Circular Error (MCE) statistic by using a row-deletion method. Rambli et al. (2016) transformed the residuals into linear scales using a trigonometric function while Abuzaid et al. (2013) used the circular distance between two circular observations.

Meanwhile, circular regression where the predictors are scalars was first introduced by Gould (1969). His model follows closely the concept of linear regression and the iterative method is used to find the parameter estimate via the corresponding log-likelihood function.

However, his model has infinite number of curves on the surface of an infinite cylinders. Mardia (1972) and Laylock (1975) also proposed regression model of circular variate on linear variates. Johnson and Wehrly (1978) improved Gould method by proposing three different class of models where two of them are the regression of a circular variate on linear variate. Then, Fisher and Lee (1992) proposed a regression model where the mean direction and dispersion of a von Mises variate are related to the explanatory variables by a general link function. The mode basically transforms the linear variables into circular using the chosen general link function. George and Ghosh (2006) developed a regression of a circular variable on a linear predictor using a Bayesian approach where the regression coefficients are assumed to be nonparametric.

Finally, Johnson and Wehrly (1978) proposed a regression of a linear variate on other linear and circular variates in which the model follow closely the linear regression. The least square method is used to find the parameter estimates. Then SenGupta and Ugwuowo (2006) proposed three different models of circular-linear regression for multivariate data based on both circular and linear predictors. These models can be used to deal with both symmetric and asymmetric model forms. Qin et al. (2011) proposed a nonparametric regression model for circular-linear multivariate regressors using a kernel-weighted local linear method.

In this study, we propose a new test of outlier detection in regression model for cylindrical data.

2.3 Linear Data

2.3.1 Cartesian Coordinate

The Cartesian plane is a plane in rectangular coordinate system with points in the form of a pair of numbers. A Cartesian coordinate is a point on the plane specified by a pair of ordered real numbers where their distances to the origin is measured in a unit length. Each

reference line is called as an axis of the system where two perpendicular real axes in the plane describe a Cartesian coordinate system in two-dimensional space while a pairwise perpendicular axes describe a Cartesian coordinate system in three-dimensional space. In the two-dimensional space, the horizontal axis is called x-axis and the vertical axis is called y-axis. Meanwhile, in three-dimensional space the axis is usually labeled by x, y and z.

2.3.2 Euclidean Distance

Euclidean distance is a distance between two points in the Euclidean n-space. In one dimension, the Euclidean distance is simply an absolute value of the differences between two points say u and v such that

$$d = \sqrt{(u-v)^2} = |u-v|.$$

In two dimensional Euclidean space, the Euclidean distance between two points $\boldsymbol{u} = (u_1, u_2)$ and $\boldsymbol{v} = (v_1, v_2)$ is given by

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2}.$$

On the other hand, the Euclidean distance in three-dimensional Euclidean space between points $\boldsymbol{u} = (u_1, u_2, u_3)$ and $\boldsymbol{v} = (v_1, v_2, v_3)$ is given by

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2}.$$

Hence in general, given Cartesian coordinates $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ in Euclidean n-space, the distance between these two points is given by

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2)} = \sqrt{\sum (u_i - v_i)^2}.$$

2.3.3 Euclidean Length or Magnitude

Magnitude is a measure of the vector length. A length of vector $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$ can be measured by

$$|\boldsymbol{u}| = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}$$

2.4 Circular Data

Observations measured as angles in radian or degree are referred to as directional data. A two-dimensional direction can be represented as angles measured between a starting points, i.e, 0° moving either clockwise/anticlockwise in [0°,360°).

2.4.1 Descriptive Statistic for Circular Data

Descriptive statistic is used to describe the basic features of the data by providing a simple summary about the sample. Some commonly used descriptive statistics in circular data are described below (Jammalamadaka & SenGupta, 2001):

(i) Mean direction, μ

In circular statistics, the mean is generally referred to mean direction. Let $\theta_1, \theta_2, \ldots, \theta_n$ be a sample of circular data. The resultant length, R and the mean direction, u is given by, respectively,

$$R = \sqrt{C^2 + S^2},$$

where

$$C = \sum_{i=1}^{n} \cos\theta_i, \quad S = \sum_{i=1}^{n} \sin\theta_i$$

and

$$\mu = \begin{cases} \tan^{-1}\frac{S}{C}, & \text{if } C > 0, S > 0, \\ \frac{\Pi}{2}, & \text{if } C = 0, S > 0, \\ \tan_{-1}\frac{S}{C} + \pi, & \text{if } C < 0, \\ \tan_{-1}\frac{S}{C} + 2\pi, & \text{if } C = 0, S = 0, \\ undefined, & \text{if } C = 0, S = 0. \end{cases}$$

(ii) Mean resultant length, $\bar{R}$

The mean resultant length is defined as

$$\bar{R} = \frac{R}{n}.$$

Mean resultant length is used to measure how concentrated the data is towards the mean direction, μ. The value lies in the range [0,1]. When $\bar{R}$ is close to 1, all the directions in the data set are almost similar which means the data are concentrated or have small dispersion.

(iii) Median direction

Mardia (1975) defined the median direction as any point ϕ, where

(i) Half of the sample points lie in the arc $[\phi, \phi + \pi)$

(ii) Majority of the sample are nearer to ϕ than to $\phi + \pi$.

Fisher (1993) defined the median direction as the point ϕ which minimize the

summation of circular distance of all observations given by

$$d(\phi) = \pi - \sum_{i=1}^{n} |\pi - |\theta_i - \phi||.$$

(iv) Sample circular variance

The sample circular variance is given by

$$V = 1 - \bar{R},\ 0 \leq V \leq 1.$$

The smaller the value of circular variance, the more concentrated the sample is said to be. However, this measure is rarely used compared to other measures of circular concentration.

(v) Sample circular standard deviation

The sample circular standard deviation is given by

$$\begin{aligned} v &= \sqrt{-2 \log 1 - V} \\ &= \sqrt{-2 \log \bar{R}}, \quad 0 < v < \infty. \end{aligned}$$

(vi) Circular range

The circular range is defined as the length of the smallest arc which consist all the sample observations. Let $\theta_1, \theta_2, \cdots, \theta_n$ be the sample observation and $\theta_{(1)}, \theta_{(2)}, \cdots, \theta_{(n)}$ is the linear order statistic. Then, the arc lengths between the adjacent points are

$$T_i = \theta_{(i+1)} - \theta_{(i)},\ i = 1, 2, \cdots, n$$

$$T_n = 2\pi - \theta_{(n)} + \theta_{(1)}.$$

The circular range w is

$$w = 2\pi - max(T_i, \cdots, T_n).$$

(vii) Concentration parameter

The standard measure for dispersion for circular data is the concentration parameter, denoted by κ. Best and Fisher (1981) gave the maximum likelihood estimates of the concentration parameter κ as

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R^3} + \frac{5}{6}\bar{R^5}, & \text{if } \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \frac{0.43}{1-\bar{R}}, & \text{if } 0.53 \leq \bar{R} < 0.85 \\ (\bar{R^3} - 4\bar{R^2} + 3\bar{R})^{-1}, & \text{if } \bar{R} \geq 0.85 \end{cases}$$

where $\bar{R}$ is the mean resultant length.

2.4.2 Circular Distance

Jammalamadaka and SenGupta (2001) defined circular distance between any two points as the smaller of the two arc lengths between the points along the circumference. The circular distance between θ_i and θ_j is given by

$$\begin{aligned} d(\theta_i, \theta_j) &= \min(\theta_i - \theta_j, 2\pi - (\theta_i - \theta_j)) \\ &= \pi - |\pi - |\theta_i - \theta_j||. \end{aligned}$$

where $d(\theta_i, \theta_j) \in [0, 2\pi)$. For example, the circular distance $d(\theta_i, \theta_j)$ between any two points of circular data can be illustrated using the following set of circular data: $10°, 15°, 45°$ and $355°$. Using the equation above, the distance between observation $10°$ to the other

points are 5°, 35° and 15° respectively. Another circular distance between θ_i and θ_j (Rao, 1969) is given by

$$d(\theta_i, \theta_j) = 1 - \cos(\theta_i - \theta_j),$$

where $d(\theta_i, \theta_j)$ is a monotone increasing function of $(\theta_i - \theta_j)$ and $d(\theta_i, \theta_j) \in [0, 2]$.

2.4.3 Circular Distributions

(i) The von Mises Distribution

The von Mises distribution was introduced by von Mises (1918) to study the deviations of measured atomic weight from integral values. It is the most common distribution used in circular statistics for unimodal samples of circular data. Von Mises distribution denoted by $VM(\mu, \kappa)$, has probability distribution function given by

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos\theta - \mu\},$$

where $0 \leq \theta < 2\pi, \kappa > 0, 0 \leq \mu < 2\pi$ and $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero given by

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{\kappa \cos\theta\} d\theta. \tag{2.4.1}$$

This distribution also called circular normal (CN) distribution to emphasize its importance and similarity to normal distribution in linear data. The parameters for $VM(\mu, \kappa)$ are the mean direction, μ, and concentration parameter, κ. The distribution is unimodal and symmetric about the direction μ. It appears as a normal distribution that is truncated at ±180°. When $\kappa = 0$, the von Mises distribution reduces to the uniform distribution. As κ gets large, the von Mises distribution approaches the normal distribution, $N\left(\mu, \frac{1}{\kappa^{\frac{1}{2}}}\right)$. Higher value of κ indicates that the sample has

higher concentration towards the sample mean direction μ.

(ii) The Wrapped Cauchy Distribution

A wrapped Cauchy distribution is obtained by wrapping the Cauchy distribution on the real line around the circle. The density of Cauchy distribution is given by

$$f(x) = \left(\frac{1}{\pi}\right)\frac{\sigma}{\sigma^2 + (x-\mu)^2}, \ -\infty < x < \infty$$

and the probability density function for wrapped Cauchy distribution is

$$\begin{aligned} g(\theta) &= \frac{1}{2\pi}\left(1 + 2\sum_{k=1}^{\infty}\rho^k \cos k(\theta - \mu)\right) \\ &= \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2 - 2\rho\cos(\theta-\mu)}, \end{aligned}$$

where $0 \leq \theta < 2\pi$ and $\rho = e^{-\sigma}$. The equality of the equations above is verified by equating the real parts of the geometric series identity as shown below

$$\sum_{k=1}^{\infty} a^k = \frac{a}{1-a}$$

with $a = \rho e^{-i(\theta-\mu)}$. The wrapped Cauchy distribution is unimodal and symmetric.

2.5 Cylindrical Data

Cylindrical data is a form of bivariate data, where one component is measured on the angular scale while the other is linear. The angular component is a directional variable which follows the circular statistics and is bounded in $[0,2\pi)$. The joint distribution between the random vector from linear and angular components is called a cylindrical distribution or a distribution on the cylinder. This type of data arises in many fields such as biology, meteorology, geology and industry. For example, consider a set of data comprising wind

speed (m/s) and wind direction (radian). The height of the cylinder is labeled to be the wind speed which is the linear variable while the direction is considered to be a circular distribution where the data is ranging from 0° to 360° as illustrated in Figure 2.1.

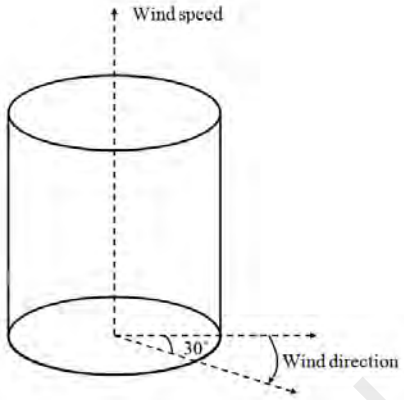


Figure 2.1: Illustration of wind data on cylinder

2.5.1 Cylindrical Coordinate

Cylindrical coordinate is a coordinate for three-dimensional space. In cylindrical coordinate system, point P as illustrated in Figure 2.2 is specified by (r, θ, z) where $r = \sqrt{x^2 + y^2},\ 0 \leq r < \infty$, is the radius between the center and point P, $\theta = \tan^{-1} \frac{y}{x},\ 0 \leq \theta < 2\pi$ is the angle measured from the x-axis and $z,\ -\infty < z < \infty$ is the height scale.

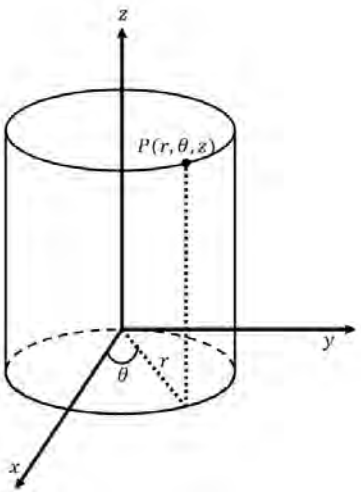


Figure 2.2: Cylindrical coordinate

2.5.2 Cylindrical Distributions

There are several models for cylindrical data that have been introduced. Johnson and Wehrly (1978) developed the angular-linear distributions based on the maximum entropy principle and by the specification of the marginal distributions. Mardia and Sutton (1978) proposed a cylindrical distribution using Fisher's idea of constructing a von Mises distribution as a conditional distribution of bivariate normal on a unit circle. The later distributions are mostly the extension from Johnson-Wehrly and Mardia-Sutton distributions.

Anderson-Cook (1997, 2000, 2001) proposed distributions with general type of circular-linear association or known as C-association and a second order models. Both models are the extension from the Mardia and Sutton model. Then in 2001,the author proposed an alternate model for limited range data. Later, Fernandez-Duran (2007) proposed a cylindrical model based on the nonnegative trigonometric sums.

Kato and Shimizu (2008) proposed two new distributions; (i) a generalized form of the distribution by Johnson and Wehrly, and (ii) a flexible model which is procured as a maximum entropy distribution or conditional distribution of a trivariate normal distribution. Their second distribution can also be regarded as an extension to the distribution proposed by Mardia and Sutton. Recently, Abe and Ley (2016) proposed a tractable, parsimonious and highly flexible model for cylindrical data. The distribution is constructed by combining the sine-skewed von Mises distribution from the circular part with the Weibull distribution from the linear part. The focus of this study will be based on the cylindrical model introduced by Johnson and Wehrly (1978).

2.6 k-Nearest Neighbor Method

The k-nearest neighbor (k-NN) is widely used in statistical estimation and classification. It is a simple algorithm that classifies new cases based on a distance measure. A new case

is identified by measuring the distance between each observations and the majority among the most common from its k-NN goes to the same class. The optimal value of k can be chosen by inspecting the data set.

Fukunage and Narendra (1975) implemented a method of branch and bound in the k-NN algorithm to increase the calculation speed by eliminating the requirement of calculating many distances. Hand (1981) provided a good review on the k-NN methods. A fuzzy algorithm of $k-$NN has been introduced by Keller et al. (1985) to overcome the problem in deciding the class of the samples regardless of their "typicalness". Tran et al. (2006) developed a density-based clustering algorithm using k-NN kernel for a higher dimensional data. Weinberger and Saul (2009) proposed a distance metric for k-NN classification known as Mahalanobis metric.

2.7 Summary

In this chapter, we discuss the existence of outlier in different types of data and their regression models as well as the methods to detect the outlier. We also highlight the differences between linear, circular and cylindrical data. Due to the different nature of their properties, we need different methods to analyze these data sets. Some descriptive statistics and several distributions are reviewed. Next, we have reviewed the k-nearest neighbor method. In our study, we consider the cylindrical distribution by Johnson and Wehrly (1978) and the von Mises distribution for circular distribution in studying the development of new statistical tests in cylindrical data.

CHAPTER 3: CYLINDRICAL DATA - JW DISTRIBUTION

3.1 Introduction

This chapter reviews the theory on cylindrical distribution model based on Johnson-Wehrly (JW) (1978) distribution. Then, we look at an alternative procedure to find the maximum likelihood estimation using optimization. Next, we review the criteria to evaluate the performance of the maximum likelihood estimation in order to estimate the parameters of the JW distribution. Then, we obtain the parameter estimate of the distribution via simulation study. Lastly, for illustration, we apply the distribution on wind data set obtained from Malaysian Meteorological Department.

3.2 The JW Distribution for Cylindrical Data

The JW cylindrical density function (Johnson & Wehrly, 1978) of (Θ,X), denoted by JW(μ, κ, λ) is given by

$$f(\theta, x) = \frac{\sqrt{\lambda^2 - \kappa^2}}{2\pi} exp(-\lambda x + \kappa x cos(\theta - \mu)) \tag{3.2.1}$$

where $0 \leq \theta < 2\pi$, $x > 0$, $0 < \kappa < \lambda$, $\lambda > 0$ and $0 \leq \mu < 2\pi$. Equation (3.2.1) is developed based on the maximum entropy distribution subject to $E(X)$, $E(X\cos\Theta)$ and $E(X\sin\Theta)$, taking specified values which are consistent with expectation with respect to the distribution (3.2.1). Then, we define a_1, a_2 and a_3 by

$$E(X) = a_1, \quad E(Xcos\Theta) = a_2, \quad E(Xsin\Theta) = a_3$$

where the expectations are taken with respect to the distribution (3.2.1).

3.2.1 Marginal Distribution

For the JW probability density function given in equation (3.2.1), the marginal density of x can be calculated by taking the integral of the joint distribution $f(x, \theta)$ with respect to θ,

$$\begin{aligned} f_1(x) &= \int_0^{2\pi} \frac{\sqrt{\lambda^2 - \kappa^2}}{2\pi} \, exp\{-\lambda x + \kappa x \, cos(\theta - \mu)\} \, d\theta \\ &= \frac{\sqrt{\lambda^2 - \kappa^2}}{2\pi} \, e^{-\lambda x} \int_0^{2\pi} exp\{\kappa x \, cos(\theta - \mu)\} \, d\theta \\ &= \sqrt{\lambda^2 - \kappa^2} \, e^{-\lambda x} \left[\frac{1}{2\pi} \int_0^{2\pi} exp\{\kappa x \, cos(\theta - \mu)\} \right] d\theta \\ &= \sqrt{\lambda^2 - \kappa^2} \, I_0(\kappa x) \, e^{-\lambda x} \end{aligned} \tag{3.2.2}$$

where $I_0(\kappa x)$ is the modified Bessel function of first kind and order zero as given in equation (2.4.1).

Meanwhile, the marginal density of θ (Johnson & Wehrly, 1978) is given by

$$f_2(\theta) = \frac{1}{2\pi} \, \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \, cos(\theta - \mu)}, \; 0 \leq \theta < 2\pi. \tag{3.2.3}$$

In summary, the marginal distribution (3.2.2) does not follow a familiar distribution while the marginal distribution (3.2.3) is the wrapped Cauchy distribution, where μ is the mean direction and ρ is the mean resultant length.

3.2.2 Conditional Distributions

Now, we derive the conditional probability distribution of JW distribution. The conditional distribution for x given θ is

$$g_1(x|\theta) = \frac{\sqrt{\lambda^2 - \kappa^2}\ (2\pi)^{-1}\ exp\{-\lambda x + \kappa x\ cos(\theta - \mu)\}}{\sqrt{\lambda^2 - \kappa^2}\ (2\pi)^{-1}\ [\lambda - \kappa\ cos(\theta - \mu)]^{-1}}$$

$$= \lambda - \kappa\ cos(\theta - \mu)\ exp\{-\lambda x + \kappa x\ cos(\theta - \mu)\}. \tag{3.2.4}$$

Comparing equation (3.2.4) with the exponential distribution function, we can see that equation (3.2.4) is actually the pdf of the exponential distribution with mean $[\lambda - \kappa\ cos(\theta - \mu)]^{-1}$. Meanwhile, the conditional distribution for θ given x is given by

$$g_2(\theta|x) = \frac{\sqrt{\lambda^2 - \kappa^2}\ (2\pi)^{-1}\ exp\{-\lambda x + \kappa x\ cos(\theta - \mu)\}}{\sqrt{\lambda^2 - \kappa^2}\ I_0(\kappa x)\ e^{-\lambda x}}$$

$$= \frac{1}{2\pi\ I_0(\kappa x)}\ exp\{\kappa x\ cos(\theta - \mu)\}. \tag{3.2.5}$$

Notice that, equation (3.2.5) follows von Misses distribution with mean direction μ and concentration parameter κx or $VM(\mu, \kappa x)$.

3.3 Parameter Estimation of JW Distribution

Maximum likelihood estimation (MLE) method is used to find the parameter estimates of JW distribution. The MLE of μ, κ, and λ are obtained by maximizing the log-likelihood function of the JW distribution, $JW(\mu, \kappa, \lambda)$. The likelihood equation for $JW(\mu, \kappa, \lambda)$ is given by

$$L = \prod_{i=1}^{n} f(x_i, \theta_i)$$

$$= \frac{(\lambda^2 - \kappa^2)^{\frac{n}{2}}}{(2\pi)^n}\ exp\left[-\lambda \sum_{i=1}^{n} x_i + \kappa \sum_{i=1}^{n} x_i\ cos(\theta_i - \mu)\right],$$

and the log-likelihood is

$$\log(L) = \log\left\{\frac{(\lambda^2-\kappa^2)^{\frac{n}{2}}}{2\pi^n}\ exp\left[-\lambda\sum_{i=1}^{n}x_i + \kappa\sum_{i=1}^{n}x_i\ cos(\theta_i-\mu)\right]\right\}$$
$$= \frac{n}{2}\log(\lambda^2-\kappa^2) - n\log 2\pi - \lambda\sum_{i=1}^{n}x_i + \kappa\sum_{i=1}^{n}x_i\cos(\theta_i-\mu).$$

However it is quite difficult to separate the unknown parameters for the JW distribution, so we are unable to find the closed-form or exact maximum likelihood estimator for each parameter. Hence, we are using optimization method on the log-likelihood function to estimate the parameters.

This can be done by using *optim* package in *R* statistical software. The default method is an implementation of Nelder and Mead (1965) that uses only function values and is robust but relatively slow. It will work reasonably well for non-differentiable functions.

3.3.1 Parameter Estimation using Optimization

Optimization is the selection of a best element from some set of available alternatives with regards to some criteria. Optimization problems consist of finding the maximum or the minimum value that a function can take by choosing initial values from within an allowed set. Optimization take place in the absolute extrema where the maximum or the minimum value that a function would take on an interval. An optimization problem can be represented as follow:

Given $f_0(x) \in R$,

i. minimize $f_0(x)$ subject to $f_i(x) \leq b_i$ or

ii. maximize $f_0(x)$ subject to $f_i(x) \geq b_i$ for $i = 1, \ldots, n$

where b_i is a constant.

There are several methods or algorithms of optimization that can be used to find the MLE. Nelder-Mead method is often used to find the parameter estimations of a function.

This method can be found in *optim*, *optimx*, *maxLik* or *mle* packages which are available in the *R* statistical software. The maximum likelihood estimation is obtained by minimizing the negative log-likelihood function of a distribution, namely here, JW(μ, κ, λ). Some available algorithms in these packages are (Henningsen & Toomet, 2011):

(i) Newton-Raphson method (NR).

This method is based on quadratic approximation and uses both gradient and Hessian of the function.

(ii) Berndt-Hall-Hall-Hausman method (BHHH).

A version of NR where the Hessian is approximated by information equality (only works for maximizing log-likelihood).

(iii) Broyden-Fletcher-Goldfarb-Shanno method (BFGS).

Also known as a variable metric algorithm which is another quasi-Newton method with a different approximation of Hessian. This method uses function values and gradients to build up a picture of the surface to be optimized.

(iv) Conjugate gradients method (CG).

A method which only uses gradients and function values and does not approximate the Hessian. May be useful (but slow) for large problems. Conjugate gradient methods will generally be more fragile than the BFGS method, but as they do not store a matrix they may be successful in much larger optimization problems.

(v) Simulated annealing method (SANN).

This method only uses function values. It is a stochastic optimization method which may be used if the objective function has secondary maximums, and in case of non-differentiable functions.

(vi) Nelder-Mead method (NM).

A simplex-based method using only the function values. It work reasonably well for

non-differentiable functions.

(vii) Brent.

It is useful for one-dimensional problems only.

For this research, we find the maximum likelihood estimation (MLE) using Nelder-mead method as available in *optim* package in *R*.

3.3.2 Nelder-Mead Method

A simplex method for finding a local minimum of a function of several variables has been developed by Nelder and Mead (1965). Simplex is a generalization of the notion of a triangle or tetrahedron to arbitrary n dimensions. Their method minimizes a function of n variables, which depends on the comparison of function values at the $(n + 1)$ vertices of a general simplex, followed by the replacement of the vertex with the highest value by another point.

For $n = 2$, a simplex is a triangle. The method is a pattern search that compares function values at the three vertices of a triangle. The worst vertex, where $f(x, y)$ is largest is rejected and replaced with a new vertex. A new triangle is formed and the search is continued. The process generates a sequence of triangles (could have different shapes), for which the function values at the vertices get smaller and smaller. The size of the triangles is reduced and the coordinates of the minimum point are found.

3.3.3 Criteria to Evaluate the MLE Method

(i) Mean squared error (MSE).

Mean squared error is a measure of how close data points to the fitted line. Another definition is it measures the average of the squares of all of the errors, which is the differences between the true values and the estimated values. Given Y_i is the

observed values and $\hat{Y}_i$ is the estimates of the observed values, the MSE can be estimated by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(ii) Bias.

Bias is an error that makes the estimated values wrong by certain amount. Bias is defined as the difference between the estimated value and the true value of the parameter, $(Y_i - \hat{Y}_i)$.

3.3.4 Simulation Study

For simulation, we generate random samples of size $n = 20, 50, 100, 200$ from JW distribution as in equation (3.2.1) and estimate the parameters using the *optim* package in *R*. For each sample, we repeat the simulation for different iteration = 100, 500, 1000, 5000. Table 3.1 and Table 3.2 give the parameter estimates of μ, κ, and λ given the true values for each parameter.

Table 3.1 gives the parameter estimates from a sample generated from JW distribution with the true value of $\mu = 0, \kappa = 1$, and $\lambda = 2$. Different sample size are generated to compare the accuracy of the estimation. We can see that the estimated parameter values of μ, κ, and λ are closer to their true values as the sample size increases. Notice that, the values of the estimated parameters are also closer to their true values as the number of iteration for each simulation increase. The MSE and bias for each sample size are calculated. The MSE and bias are decreasing (closer to 0) as the sample size and iteration increase. Similar results are observed for the case $\mu = 1.85, \kappa = 0.22$, and $\lambda = 0.40$ as given in Table 3.2.

Table 3.1: Parameter estimates of JW$(0, 1, 2)$.

iteration	n	estimated parameter			bias			mse		
		$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$
100	20	0.36	1.20	2.35	0.36	0.20	0.35	0.94	0.31	0.59
500		0.33	1.24	2.34	0.33	0.24	0.34	0.70	0.44	0.64
1000		0.34	1.27	2.35	0.34	0.27	0.35	0.77	0.46	0.62
5000		0.34	1.31	2.36	0.34	0.31	0.36	0.76	0.55	0.61
100	50	0.19	1.08	2.15	0.19	0.08	0.15	0.43	0.11	0.14
500		0.20	1.12	2.13	0.20	0.12	0.13	0.22	0.13	0.15
1000		0.20	1.11	2.13	0.20	0.11	0.13	0.21	0.13	0.15
5000		0.20	1.10	2.13	0.20	0.10	0.13	0.16	0.13	0.15
100	100	0.14	1.04	2.05	0.14	0.04	0.05	0.03	0.06	0.07
500		0.14	1.05	2.06	0.14	0.05	0.06	0.05	0.06	0.06
1000		0.14	1.06	2.06	0.14	0.06	0.06	0.04	0.06	0.06
5000		0.14	1.05	2.06	0.14	0.05	0.06	0.04	0.06	0.06
100	200	0.10	1.00	2.05	0.10	0.00	0.05	0.01	0.02	0.03
500		0.09	1.02	2.03	0.09	0.02	0.03	0.01	0.02	0.03
1000		0.10	1.02	2.03	0.10	0.02	0.03	0.01	0.02	0.03
5000		0.10	1.02	2.03	0.10	0.02	0.03	0.02	0.03	0.03
100	500	0.06	1.01	2.00	0.06	0.01	0.00	0.01	0.01	0.01
500		0.06	1.01	2.01	0.06	0.01	0.01	0.01	0.01	0.01
1000		0.06	1.01	2.01	0.06	0.01	0.01	0.01	0.01	0.01
5000		0.06	1.01	2.01	0.06	0.01	0.01	0.01	0.01	0.01
100	1000	0.04	1.00	2.01	0.04	0.00	0.01	0.00	0.01	0.01
500		0.04	1.00	2.01	0.04	0.00	0.01	0.00	0.01	0.01
1000		0.04	1.01	2.01	0.04	0.01	0.01	0.00	0.01	0.00
5000		0.04	1.00	2.01	0.04	0.00	0.01	0.00	0.01	0.00

Table 3.2: Parameter estimates of JW$(1.85, 0.22, 0.40)$**.**

iteration	n	estimated parameter			bias			mse		
		$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\lambda}$
100	20	1.61	0.27	0.47	0.24	0.05	0.07	0.17	0.02	0.02
500		1.58	0.29	0.47	0.27	0.07	0.07	0.18	0.02	0.02
1000		1.59	0.28	0.47	0.26	0.06	0.07	0.16	0.02	0.03
5000		1.59	0.28	0.47	0.26	0.06	0.07	0.16	0.02	0.03
100	50	1.73	0.26	0.43	0.12	0.04	0.03	0.04	0.01	0.01
500		1.71	0.24	0.42	0.14	0.02	0.02	0.05	0.01	0.01
1000		1.71	0.24	0.43	0.14	0.02	0.03	0.05	0.01	0.01
5000		1.71	0.24	0.43	0.14	0.02	0.03	0.06	0.01	0.01
100	100	1.76	0.23	0.41	0.09	0.01	0.01	0.02	0.00	0.00
500		1.77	0.23	0.41	0.08	0.01	0.01	0.02	0.00	0.00
1000		1.77	0.23	0.41	0.08	0.01	0.01	0.03	0.00	0.00
5000		1.77	0.23	0.41	0.08	0.01	0.01	0.03	0.00	0.00
100	200	1.80	0.22	0.40	0.05	0.00	0.00	0.01	0.00	0.00
500		1.81	0.22	0.40	0.04	0.00	0.00	0.02	0.00	0.00
1000		1.81	0.22	0.40	0.04	0.00	0.00	0.02	0.00	0.00
5000		1.81	0.22	0.40	0.04	0.00	0.00	0.02	0.00	0.00
100	500	1.83	0.23	0.40	0.02	0.01	0.00	0.02	0.00	0.00
500		1.83	0.22	0.40	0.02	0.00	0.00	0.03	0.00	0.00
1000		1.83	0.22	0.40	0.02	0.00	0.00	0.03	0.00	0.00
5000		1.83	0.22	0.40	0.02	0.00	0.00	0.02	0.00	0.00
100	1000	1.85	0.22	0.40	0.00	0.00	0.00	0.00	0.00	0.00
500		1.85	0.22	0.40	0.00	0.00	0.00	0.01	0.00	0.00
1000		1.84	0.22	0.40	0.01	0.00	0.00	0.01	0.00	0.00
5000		1.84	0.22	0.40	0.01	0.00	0.00	0.02	0.00	0.00

3.4 Application on Real Data Set

For illustration, we consider the wind direction and wind speed data taken from Malaysian Meteorological Department. In total, we have 31 observations taken from Bayan Lepas, Penang, Malaysia in January 2005 with pressure of 850 hPa. The data are given in Table 3.3.

Table 3.3: The Wind Data.

Wind Speed	Wind Direction (°)	Wind Speed	Wind Direction (°)
14.9	85	2.1	125
5.1	85	1.5	185
4.6	140	1	190
6.2	100	0.5	70
3.6	135	4.6	135
1.5	310	2.6	125
2.1	340	3.6	90
4.6	120	2.1	200
4.6	130	3.6	5
5.1	120	2.6	30
4.6	150	3.1	165
2.6	80	3.1	260
1	205	4.6	325
0.5	60	3.6	325
5.1	110	2.6	345
3.1	125		

In Figure 3.1, the data were plotted in 3-dimensional plot to illustrate a cylindrical shape. The horizontal axis is the linear component (wind speed) while x and y axes are the transformation from the circular component (wind direction). Figure 3.3 shows the scatter plot of the data with the fitted JW contour plot. Here, we can see that the data have roughly follow the JW distribution.

The parameter estimates using *optim* package in *R* are given in Table 3.4. The mean direction, $\hat{\mu} = 106.1°$, the concentration parameter, $\hat{\kappa} = 0.22$ and $\hat{\lambda} = 0.40$. It can be seen

from Figure 3.1 and Figure 3.2 that most of the data are concentrated towards the mean direction μ. Hence, the wind actually blew to the East. However, there is one observation located further away from the rest. This observation is a candidate of outlier. Hence, further investigation is needed to understand this observation.

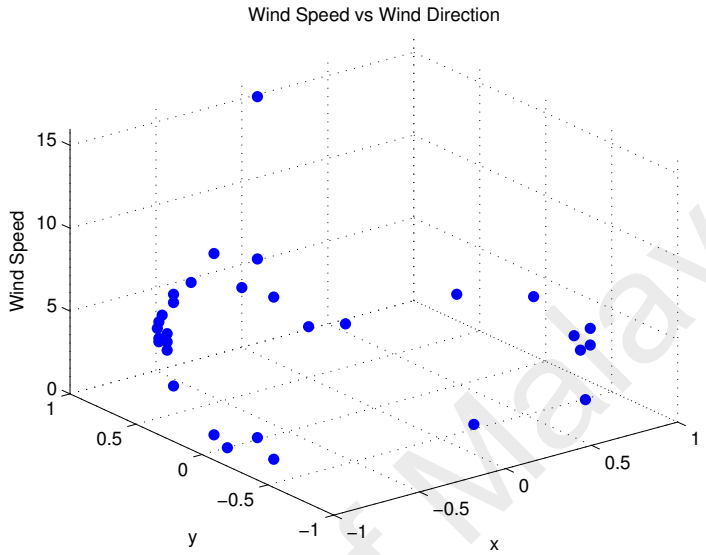


Figure 3.1: 3-D graph of wind speed vs wind direction

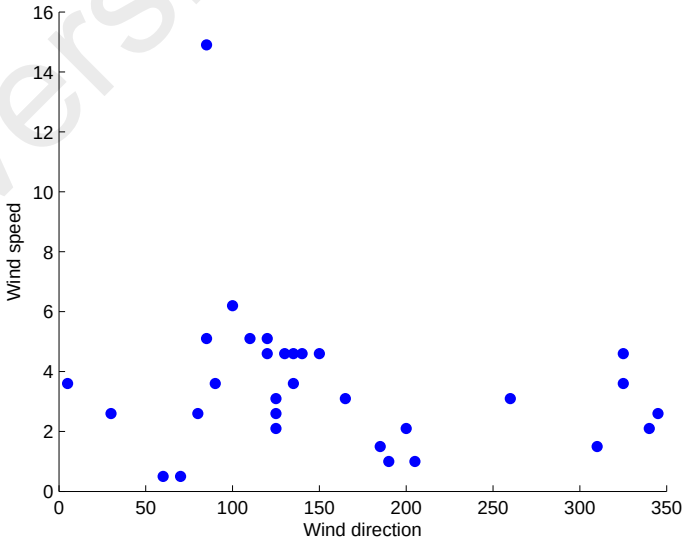


Figure 3.2: The scatter plot of the data

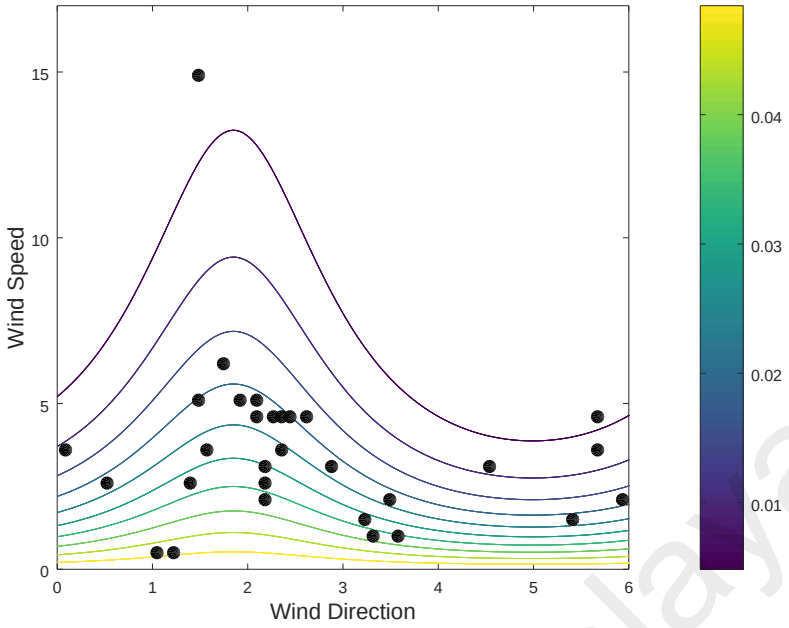


Figure 3.3: Scatter plot of the wind speed and wind direction data, together with the fitted JW contour plot.

Table 3.4: Parameter Estimation of JW distribution

	Parameter Estimates
Mean Direction ($\hat{\mu}$)	106.1°
Concentration ($\hat{\kappa}$)	0.22
$\hat{\lambda}$	0.40

3.5 Summary

In this chapter, we have discussed about cylindrical data and the theory of JW distribution. Next, we reviewed on the parameter estimation of JW distribution and the alternative method to find the parameter estimates. Then, the criteria for choosing the best method for MLE is discussed and the simulation study is obtained using the *optim* package in *R*. As an illustration, a real data set is used to obtain the parameter estimates of JW distribution.

CHAPTER 4: A NEW TEST OF DISCORDANCY IN CYLINDRICAL DATA

4.1 Introduction

In this chapter, we propose a measure of distance for detecting outliers in cylindrical data. The distance between two points on a cylindrical surface is measured based on law of cosines. Then, we developed a new test statistic to detect an outlier in cylindrical data based on k-nearest neighbor (k-NN) distance. We obtain the critical values and study the performance of the test using simulation. We also apply the test on the wind data set which obtained from Malaysian Meteorological Department.

4.2 Outlier in Cylindrical data

Outlier refers to observation that stands out remarkably in certain ways from other observations. However, the detection of outlier in directional data necessitate different method from the linear case. In circular data, the outlier is defined as an observation having large value of circular distance from the value of its two neighboring observations on a unit circle (Mohamed et al., 2016). In cylindrical set up, outlier can be categorized by the type of variables; namely circular, linear or both. We divide the outlier detection into three categories: (1) outlier in circular part; (2) outlier in linear part; (3) outlier in both circular and linear parts. Hence, we define outlier in cylindrical data as an observation that satisfies at least one of the outlier definitions in circular or linear or both. In other words, when an observation is located far away from the rest of the observations in the linear or circular direction, the observation will become a candidate of outlier in cylindrical data.

4.3 Distance on a Cylinder

A combination of linear and circular data form a three-dimensional data called cylindrical data. Distance between two cylindrical points can be calculated as a distance between two vectors. Suppose we have cylindrical observations $(\theta_i, x_i), i = 1, 2, \ldots, n$. For each

observations (θ_i, x_i), we transform the data into Cartesian coordinate $\boldsymbol{v}_i = (v_{i1}, v_{i2}, v_{i3})$ where

$$v_{i1} = r\cos\theta_i,$$

$$v_{i2} = r\sin\theta_i,$$

$$v_{i3} = x_i.$$

Next, we standardize $\boldsymbol{v}_i = (v_{i1}, v_{i2}, v_{i3})$ to form a new standardized set of data $\boldsymbol{w}_i = (w_{i1}, w_{i2}, w_{i3})$ using

$$w_{ij} = \frac{v_{ij} - \bar{v}_j}{s_j},\ j = 1, 2, 3, \tag{4.3.1}$$

where $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_{ij} - \bar{v}_j)^2}$ and $\bar{v}_j = \sum_{i=1}^{n}\frac{v_{ij}}{n}$. This standardization is needed to eliminate the influence of radius, r, for the outlier detection purposes.

Hence, given $\boldsymbol{w}_1 = (w_{11}, w_{12}, w_{13})$ and $\boldsymbol{w}_2 = (w_{21}, w_{22}, w_{23})$, the standardized distance between two vectors on cylinder is given by

$$\begin{aligned} d(\boldsymbol{w}_1, \boldsymbol{w}_2) &= (\boldsymbol{w}_1 - \boldsymbol{w}_2)\cdot(\boldsymbol{w}_1 - \boldsymbol{w}_2) \\ &= \|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2 - 2\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|\cos\theta \\ &= \|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2 - 2(\boldsymbol{w}_1\cdot\boldsymbol{w}_2) \end{aligned}$$

where θ is the angle between two vector $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ and $\boldsymbol{w}_1\cdot\boldsymbol{w}_2 = \|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|\cos\theta$.

In general, for any two vectors $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, the cylindrical distance is given by

$$d(\boldsymbol{w}_i, \boldsymbol{w}_j) = \|\boldsymbol{w}_i\|^2 + \|\boldsymbol{w}_j\|^2 - 2(\boldsymbol{w}_i\cdot\boldsymbol{w}_j),\ i, j = 1, 2, \ldots, n.$$

Note that this distance is a linear distance between vectors $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, not the distance on

the surface of a cylinder. For example, given three cylindrical points (θ_1, x_1)=(100°,6.2), (θ_2, x_2)=(135°,3.6), and (θ_3, x_3)=(340°,2.1). The distance between (θ_1, x_1) and (θ_2, x_2) is 2.13 which is logically smaller than the distance between (θ_1, x_1) and (θ_3, x_3) which is 9.26.

4.4 The k-Nearest Neighbor's Distance

We denote $d(w_i, w_1), d(w_i, w_2), \ldots, d(w_i, w_n)$ as the distances between the i-th observation with the other observations, given $i = 1, 2, \ldots, n$, while $d_{(1)}(w_i, w_1), d_{(2)}(w_i, w_2), \ldots, d_{(k)}(w_i, w_n)$ are the corresponding ordered distances from each pair of observations. The nearest distance for the i-th observation, say x_i, is defined as the smallest distance or the distance at the first position in the ordered distance (Rambli, 2015) given by

$$C_{1i} = d_{(1)}(\boldsymbol{w}_i, \boldsymbol{w}_j) \quad \text{for} \quad i, j = 1, 2, \ldots, n, \quad i \neq j \tag{4.4.1}$$

Note that $\{C_{1i}, i = 1, 2, \ldots, n\}$ gives a sequence of distances between successive observations on the p-dimensional surface. We will use equation (4.4.1) to detect outliers in cylindrical data. Hence, we define C_{1i} as the k-nearest neighbor distance for the i-th observation, $k = 1, 2, 3, \ldots$ and $i = 1, 2, \ldots, n$, given by

$$C_{ki} = d_{(k)}(\boldsymbol{w}_i, \boldsymbol{w}_j) \quad \text{for} \quad j = 1, 2, \ldots, n, \quad i \neq j \tag{4.4.2}$$

We develop a new test of discordancy to detect outlier in cylindrical data using equation (4.4.2) which will be shown in the next section.

4.5 A New Test of Outlier detection in Cylindrical Data

Suppose we have a sample vector of 3-dimensional variables, $\boldsymbol{v}_i(x_i, y_i, z_i), i = 1, 2, \ldots, n$, with sample generated from the JW distribution. We define a test statistic C_n^k such that

$$C_n^k = \max_i\{C_{ki}\}. \tag{4.5.1}$$

where n is the sample size, $k > 0$ is the $k-NN$ and C_{ki} given in equation (4.4.2). Following is the steps required to detect outlier in cylindrical data:

Step 1: Transform the cylindrical data, (θ_i, x_i) into Cartesian data $\boldsymbol{v_i} = (\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \boldsymbol{v}_{i3}), i = 1, 2, \ldots, n$.

Step 2: Then standardize $\boldsymbol{v}_i$ to $\boldsymbol{w}_i = (\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, \boldsymbol{w}_{i3})$ as described in equation (4.3.1) .

Step 3: Calculate $C_{ki}, k > 0, i = 1, 2, \ldots, n$ as given in equation (4.4.2).

Step 4: Then we define the test statistic C_n^k given in equation (4.5.1).

Step 5: If the value of C_n^k exceeds the cut-off point, say a_c, then the i^{th} observation corresponding to $\max_i\{C_{ki}\}$ is identified as an outlier.

4.6 The cut-off point of C_n^k Statistic

We design a simulation study to find the percentage points under the null hypothesis of no outlier present in the cylindrical data. This simulation study is developed using *R* statistical package. We focus on the C_n^k statistic when $k = 1$ for the case of single outlier. The generation of cut-off points is based on various values of sample size n with parameters μ, κ, and λ as given in Table 4.1. For various combination n, κ, λ with $\mu = \pi/2$, we generate a sample from JW distribution and find the distance between each observation. We then sort the distance from the smallest to largest to find the respective ordered distance. Next, we set $k = 1$ to find the 1^{st} nearest distance to obtain the C_n^1 statistic as given in

equation (4.5.1). This process is repeated for 2000 times and the estimated percentage points of the C_n^1 statistic at 10%, 5% and 1% upper percentiles are obtained.

The cut-off points of C_n^1 statistic are tabulated in Table 4.1. For the combination of parameter κ, λ and the percentile level, the values of the cut-off point are an increasing function of κ. Different combination of κ and λ will give different values of the cut-off point. In addition, the cut-off points of C_n^1 statistic are similar for any value of μ, as the distance calculated between two points are independent of the value of μ. The exact cut-off point can be obtained for any combination of estimated κ and λ, and sample size n. In fact, we expect to have a more accurate cut-off point by using higher number of simulation as pointed by Verma et al. (2017).

4.7 Performance of C_n^1 Statistic

The focus is on C_n^1 statistic which is useful to find a single outlier. We want to find an observation (θ_i, x_i) that is located furthest away from the rest of the sample, which can be attributed as an outlier in the cylindrical data. Simulation method is used to investigate the power of performance of the C_n^k statistic when applied to the JW distribution. We perform the simulation based on three categories: (1) outlier in circular component; (2) outlier in linear component; (3) outlier in linear-circular component, to study the performance of C_n^1 statistic.

Barnett and Lewis (1994) and David (1981) stated that a reasonable measure of the performance of the test are based on the probabilities of $P1$, $P3$ and $P5$ where $P1 = 1 - \beta$ is the power function where β is the type-II error, $P3$ is the probability that the contaminant point is an outlier and identified as discordant, and lastly $P5$ is the probability that the contaminant point is an outlier given that it is identified as discordant. A good test can be characterized by having: (i) high $P1$; (ii) high $P5$; and (iii) low $P1 - P3$.

Table 4.1: Cut-off points for C_n^1 statistic.

	λ	1.35	2.52	4.54	6.86	8.45	9.38	10.80	20.34
	κ	1.21	2.32	4.31	6.72	8.33	9.26	10.68	20.15
n	Significance Level								
5	10%	8.48	8.71	9.06	9.44	9.57	9.61	9.65	9.71
	5%	9.37	9.52	9.77	10.00	10.09	10.12	10.13	10.17
	1%	10.42	10.58	10.79	11.02	11.06	11.06	11.11	11.18
10	10%	9.46	9.71	10.60	12.38	13.16	13.34	13.64	14.05
	5%	10.52	10.82	11.82	14.16	14.63	14.93	15.23	15.51
	1%	13.00	13.89	14.71	16.21	16.79	16.96	17.13	17.40
15	10%	9.97	10.43	11.56	14.77	15.31	15.61	16.12	16.73
	5%	11.56	11.90	13.27	15.85	17.09	17.49	17.96	18.62
	1%	14.58	14.84	16.15	19.64	20.71	21.05	21.37	21.59
20	10%	9.37	10.37	12.01	16.08	17.72	18.21	18.68	19.48
	5%	11.57	12.27	14.27	18.22	19.84	20.03	20.45	21.16
	1%	14.87	16.59	17.43	21.12	23.76	24.44	25.09	25.98
25	10%	9.21	9.90	11.87	16.65	19.58	20.10	20.90	21.80
	5%	11.35	12.42	14.37	19.77	22.36	22.52	23.08	23.99
	1%	15.15	16.65	18.57	24.56	25.92	26.40	27.53	28.56
30	10%	8.27	8.79	10.73	16.79	19.89	20.70	22.14	23.20
	5%	11.04	11.29	14.04	19.87	23.66	24.40	25.08	26.31
	1%	15.74	16.26	19.67	25.07	28.42	29.11	29.73	30.45
50	10%	7.91	7.93	9.09	14.61	18.86	19.64	21.36	23.86
	5%	11.47	11.67	12.95	18.86	24.04	25.18	26.66	29.12
	1%	19.11	19.20	19.97	25.97	33.98	35.87	38.83	38.92
70	10%	7.73	7.85	8.34	11.42	15.60	17.15	19.15	21.79
	5%	11.54	11.72	11.92	16.35	20.61	22.27	25.20	27.89
	1%	22.47	22.48	23.37	26.22	29.87	32.86	36.32	41.24
100	10%	8.31	8.40	8.41	9.62	11.63	13.15	13.97	16.28
	5%	12.59	12.78	12.81	14.03	16.03	17.06	18.84	21.28
	1%	25.89	25.94	26.31	26.32	30.03	31.34	33.35	35.57

The performance is constructed based on the null hypothesis that the sample comes from the JW distribution. We generate samples from JW distribution based on different sample size $n = 30, 50$ and 100 with concentration parameter values $\kappa = 0.5, 1, 5, 10, 20$ and also $\lambda = 1, 2, 6, 11, 21$. The C_n^1 statistic for each random sample is calculated, if the C_n^1 value is greater than the cut-off point at upper 5% significance level, then we have correctly detected the outlier. Note that the cut-off point for each category is the same. The simulation generates 2000 replications of n samples from JW distribution and the proportion of correct detection is obtained. We will study the performance of the proposed statistic when the outlier appear in circular or linear or both components in the following sections.

4.7.1 Outlier in Circular Component

The samples are generated in such a way that all n observations come from JW distribution with fixed value of parameter $\mu = \pi/2$. The outlier is generated by altering

$$(\theta_n', x_n') = (\theta_n + \Delta_\theta, x_n)$$

where $\Delta_\theta = 0, 0.3, 0.5, 0.7, 1$ is the contamination level. The C_n^1 statistic for each random sample is calculated and if the C_n^1 value greater than the cut off point, then we have correctly detected the outlier. The process is repeated 2000 times and the results are obtained.

Table 4.2 shows the proportion of the correct detection of outlier for various values of sample size n and κ. It can be seen that, the proportion of correct detection are almost zero for all cases considered. This indicates that the performance is very weak. This is due to the nature of the JW distribution that is spread along the circumference of a circle even at large value of κ. Hence, when we introduce an outlier by adding the value of the n-th observation by Δ_θ on the circular component, this observation might not become an outlier.

Table 4.2: The proportion of correct detection of C_n^1 statistic in circular component.

n	κ / Δ_θ	0.5	1	5	10	20
30	0	0.00	0.00	0.00	0.00	0.00
	0.3	0.00	0.00	0.00	0.00	0.00
	0.5	0.00	0.00	0.00	0.01	0.01
	0.7	0.00	0.00	0.01	0.02	0.02
	1	0.00	0.00	0.01	0.02	0.03
50	0	0.00	0.00	0.00	0.00	0.00
	0.3	0.00	0.00	0.00	0.00	0.00
	0.5	0.00	0.00	0.00	0.00	0.00
	0.7	0.00	0.00	0.00	0.00	0.01
	1	0.00	0.00	0.00	0.01	0.02
100	0	0.00	0.00	0.00	0.00	0.00
	0.3	0.00	0.00	0.00	0.00	0.00
	0.5	0.00	0.00	0.00	0.00	0.01
	0.7	0.00	0.00	0.00	0.01	0.01
	1	0.00	0.00	0.00	0.01	0.02

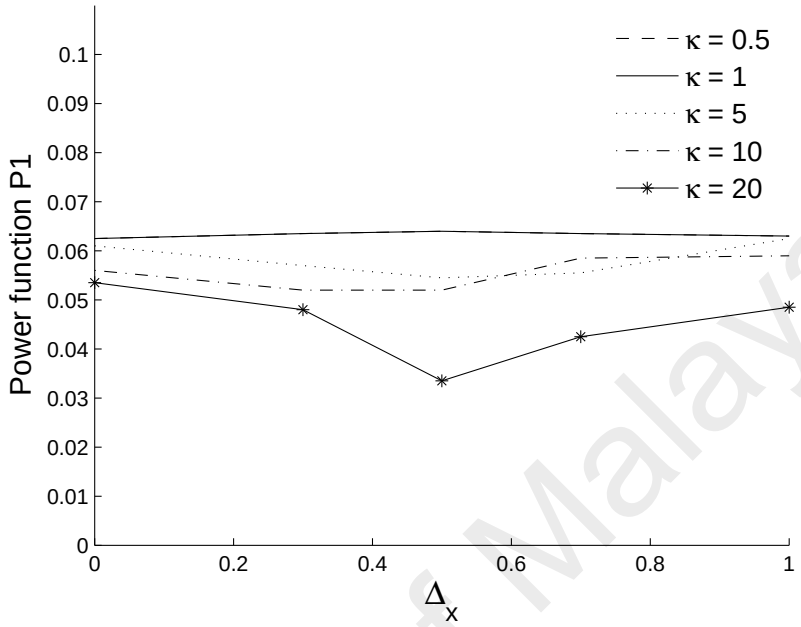
Similarly, the results based on the performance $P1$ and $P5$ for various values of κ and n as given in Figures 4.1 and 4.2 are all unsatisfying. In conclusion, for JW distribution, outlier in circular component is almost impossible to occur due to the nature of the distribution that is spread along the circumference of a circle even at large value of κ.

4.7.2 Outlier in Linear Component

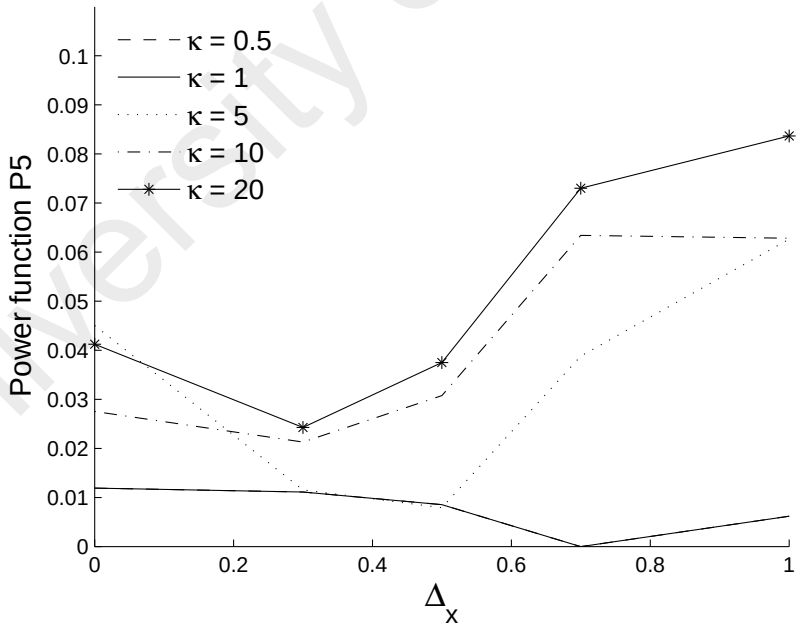
The samples are generated in which all n observations are coming from the JW distribution with fixed value of parameter $\mu = \pi/2$. The outlier is introduced at the n^{th} simulated value such that

$$(\theta_n', x_n') = (\theta_n, x_n + \Delta_x)$$

where $\Delta_x = 0, 3, 5, 7, 10, 15, 20$ is the contamination level. This process is repeated for 2000 times and the results are obtained.

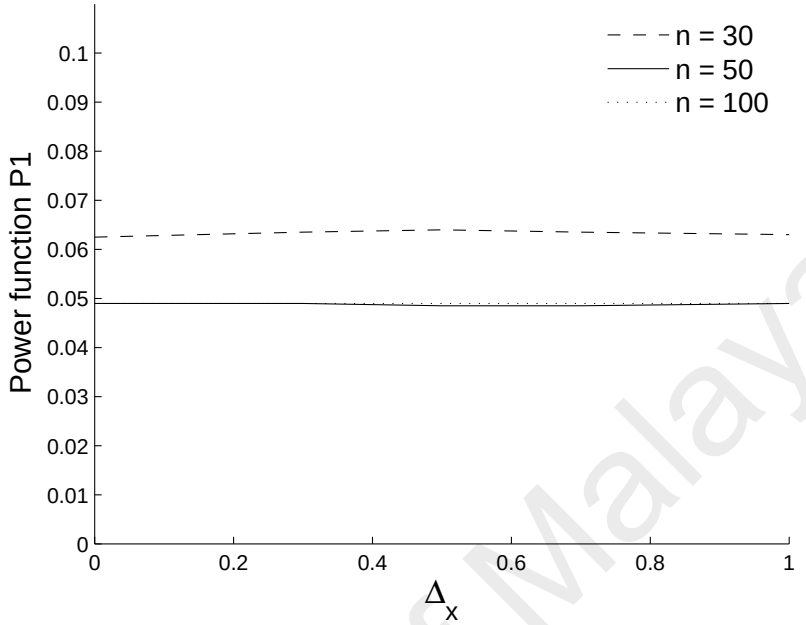


(a) Values of P1

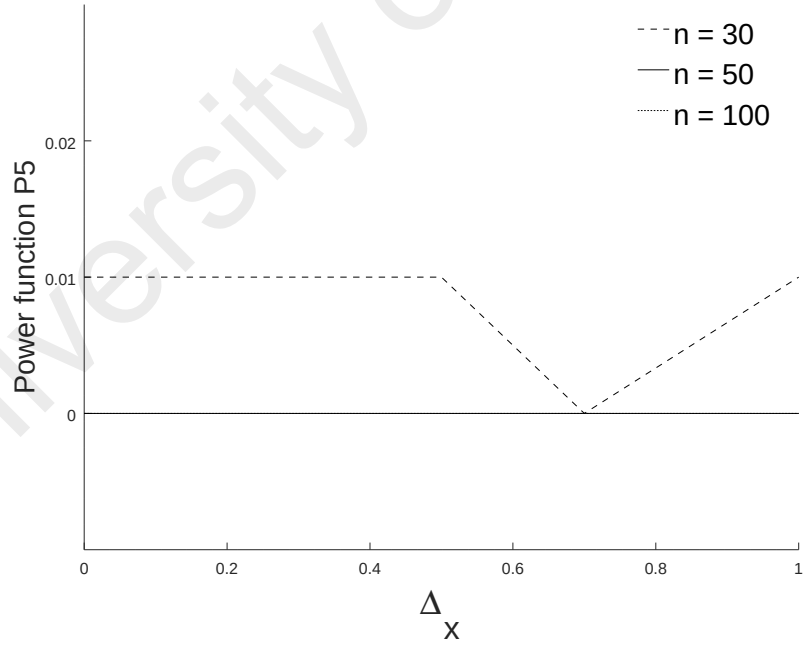


(b) Values of P1

Figure 4.1: The performance of the C_{30}^{1} statistic for different values of κ when $n = 30$.



(a) Values of P1



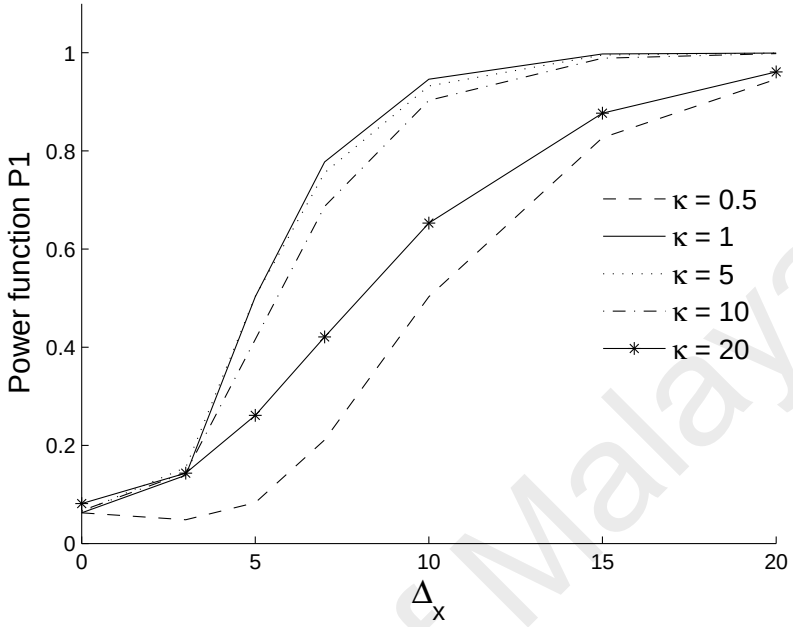
(b) Values of P5

Figure 4.2: The performance of the C_{30}^{1} statistic for different values of sample size n when $\kappa = 1$.

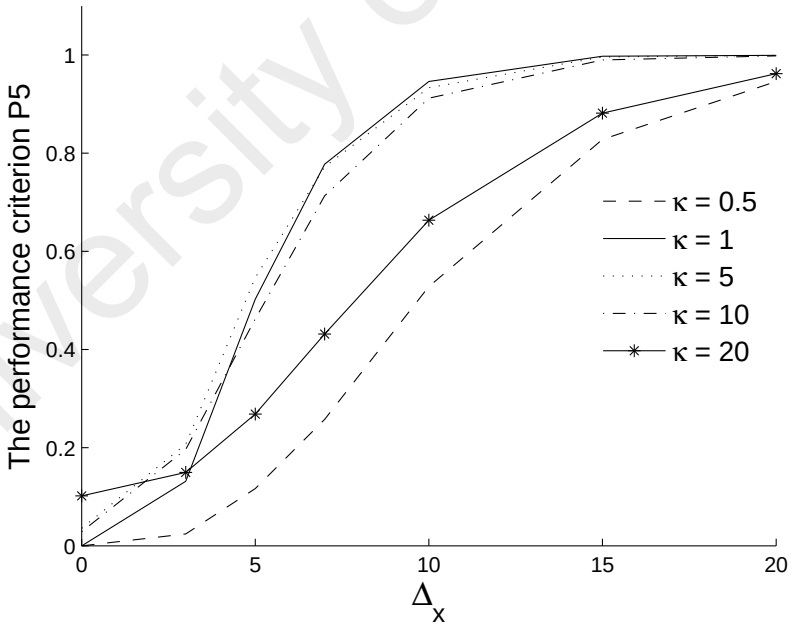
Table 4.3 shows the proportion of correct detection of outlier. It can be seen that the proportion is increasing as sample size increases. Figure 4.3(a)-(b) show the graphs of the probability $P1$ and $P5$ against the contamination level Δ_x using the upper 5% significance level of C_n^1 statistic for various values of κ when $n = 30$. These curves imply a rapid increase in the probability $P1$ and $P5$ as the contamination level Δ_x increases. From these figures, it can be seen that for $\kappa \geq 1$, the values of $P1$ and $P5$ show similar behaviour when the value rapidly increasing after $\Delta_x = 4$. This is because the data are very concentrated, therefore it is easier to detect any outlier. However, when κ is very small ($\kappa \leq 1$), the increase in values is slow. This is because, when κ is small, the data are more dispersed, therefore it is harder to detect any outliers.

Table 4.3: The proportion of correct detection of C_n^1 statistic in linear component.

n	κ / Δ_x	0.5	1	5	10	20
30	0	0.00	0.00	0.00	0.00	0.01
	3	0.01	0.18	0.15	0.12	0.07
	5	0.07	0.53	0.50	0.39	0.19
	7	0.21	0.78	0.75	0.67	0.36
	10	0.50	0.95	0.93	0.89	0.60
	15	0.83	1.00	1.00	0.99	0.84
	20	0.95	1.00	1.00	1.00	0.93
50	0	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.05	0.07	0.07	0.06
	5	0.02	0.36	0.36	0.35	0.28
	7	0.10	0.74	0.72	0.70	0.62
	10	0.36	0.95	0.94	0.94	0.90
	15	0.79	1.00	1.00	1.00	0.99
	20	0.95	1.00	1.00	1.00	1.00
100	0	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.02	0.04	0.05	0.06
	5	0.01	0.24	0.28	0.29	0.27
	7	0.04	0.68	0.70	0.69	0.68
	10	0.24	0.95	0.95	0.95	0.95
	15	0.76	1.00	1.00	1.00	1.00
	20	0.95	1.00	1.00	1.00	1.00



(a) Values of P1



(b) Values of P5

Figure 4.3: The performance of the C_{30}^{1} statistic for different values of κ when $n = 30$.

Figures 4.4(a)-(b) show the graphs of the probability $P1$ and $P5$ against the contamination level Δ_x using the upper 5% significance level of C_n^1 statistic for various values of n when $\kappa = 1$. These curves show that smaller sample size has slightly better performance for smaller value of Δ_x. But in general, $P1$ and $P5$ show similar behaviour for $n = 30, 50$ and 100. However for larger values of κ, the performance of the C_n^1 statistic is actually slightly better for larger sample size. As the value of the concentration parameter increases, the performance actually is getting better at large sample size.

The values of the probability $P1$ and $P5$ for various sample size, n when $\kappa = 20$ are given in Figures 4.5(a)-(b). When κ is large, the performance of the C_n^1 statistic is an increasing function of sample size, n. In addition, the differences between $P1$ and $P3$ (Appendix C) are also approximately close to 0.

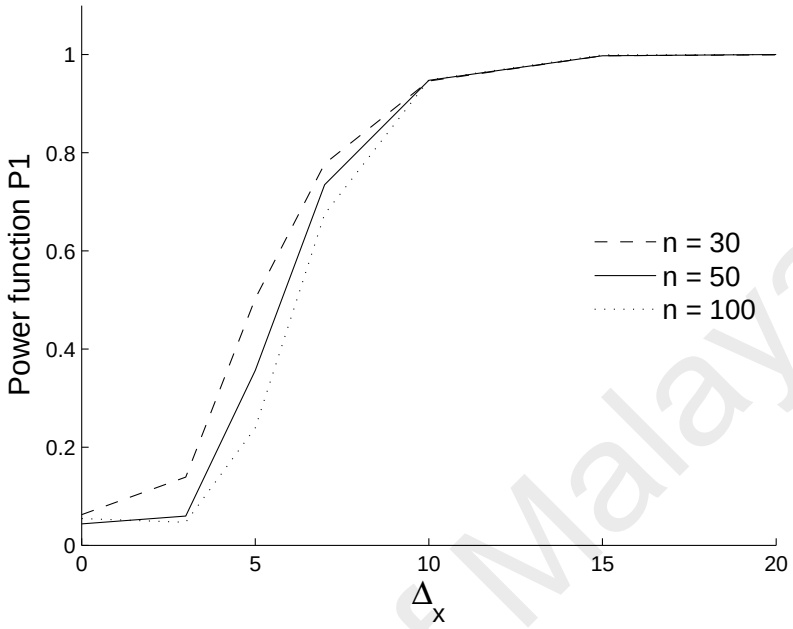
4.7.3 Outlier in Linear-Circular Component

Similar to the previous section, the outlier is generated by altering the n^{th} simulated value in both components such that

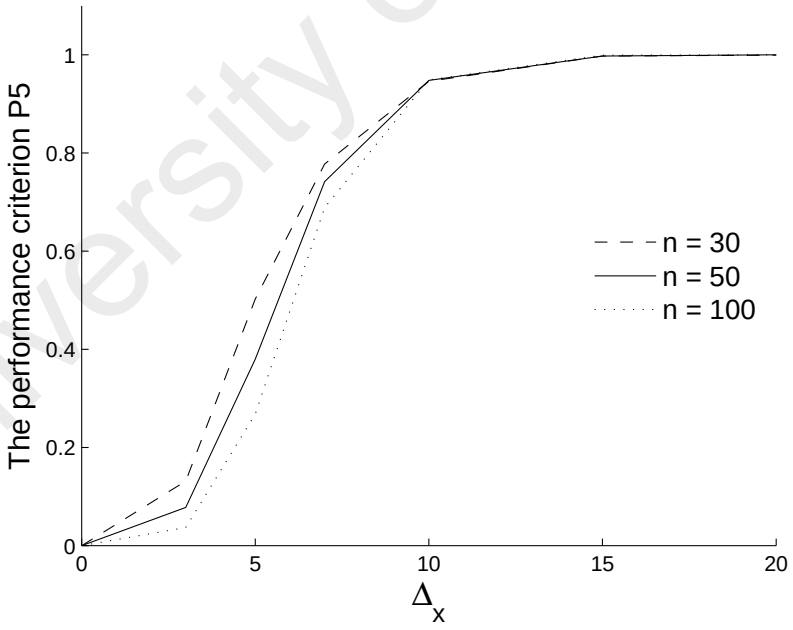
$$(\theta_n', x_n') = (\theta_n + \Delta_\theta \pi, x_n + \Delta_x)$$

where $\Delta_\theta = 0, 0.3, 0.5, 0.7, 1$ and $\Delta_x = 0, 3, 5, 7, 10, 15, 20$ are the contamination levels for θ and x respectively.

The power function $P1$ for the case when $n = 30$ are plotted in Figures 4.6(a)-(e), each having different values of contamination level Δ_θ. These curves show a rapid increase in the power function $P1$ of the C_{30}^1 statistic as the value of the contamination level Δ_x increases for $\kappa \geq 5$. The $P1$ values of the C_{30}^1 statistic show similar behaviour for all Δ_θ. For smaller value of κ, the increase in power is slower as the data are more dispersed and it is harder to detect the outlier. Hence, the power function $P1$ is an increasing function of κ.

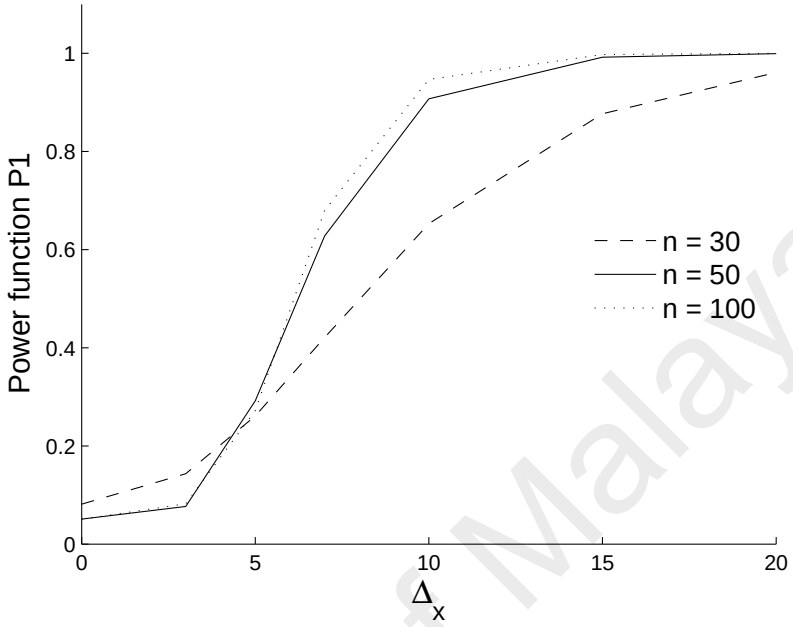


(a) Values of P1

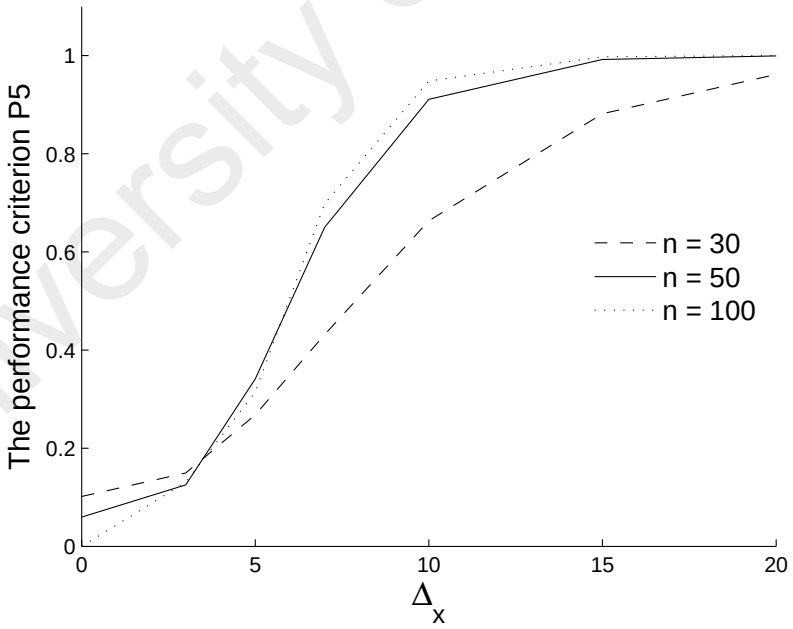


(b) Values of P5

Figure 4.4: The performance of the C_n^1 statistic when $\kappa = 1$.

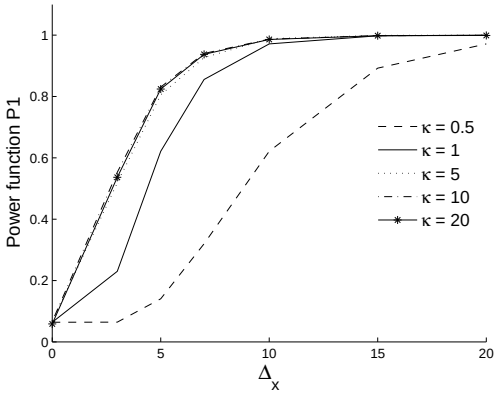


(a) Values of P1

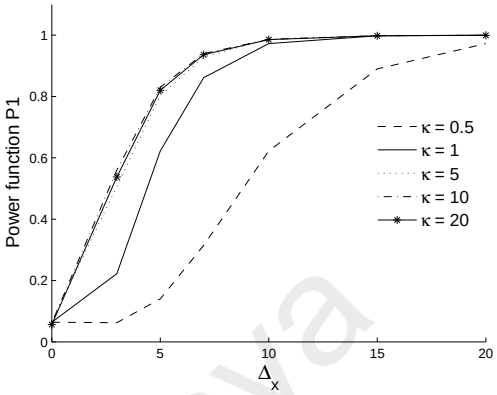


(b) Values of P5

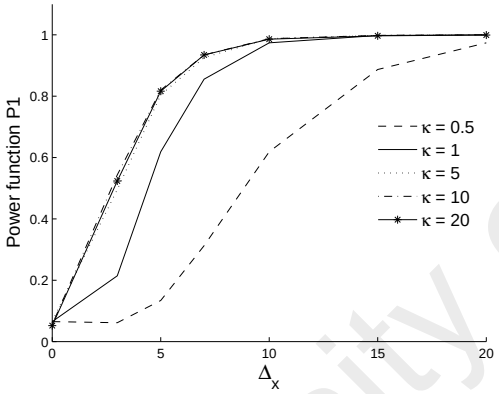
Figure 4.5: The performance of the C_n^1 statistic when $\kappa = 20$.



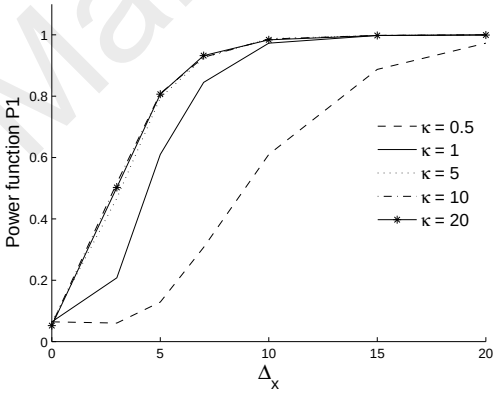
(a) $\Delta_\theta = 0$



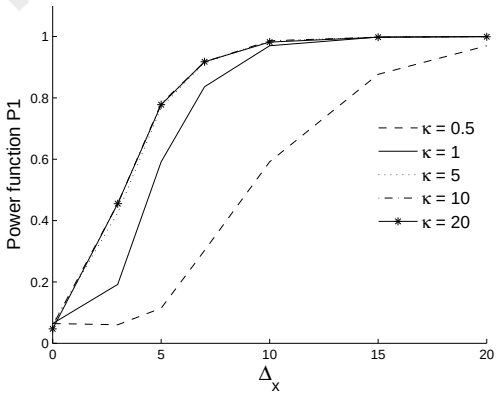
(b) $\Delta_\theta = 0.3$



(c) $\Delta_\theta = 0.5$



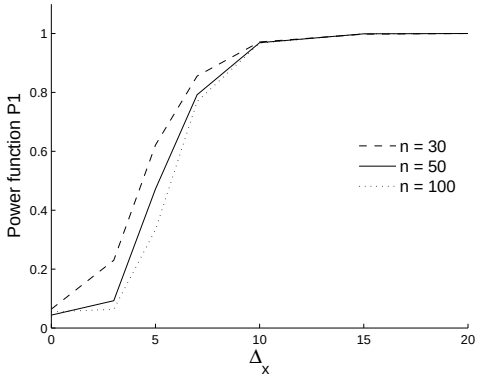
(d) $\Delta_\theta = 0.7$



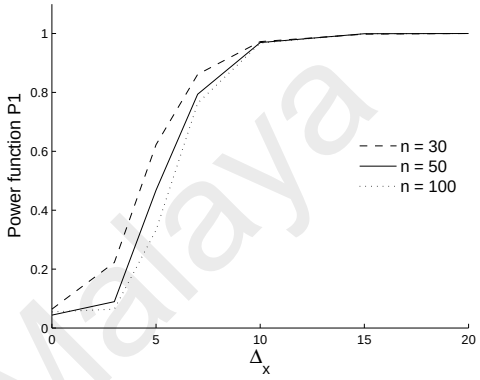
(e) $\Delta_\theta = 1$

Figure 4.6: Power function, P1 of the C_{30}^1 statistic for different values of κ when $n = 30$.

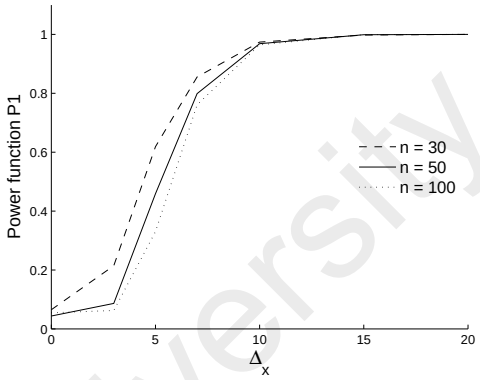
The power performances of C_n^1 statistic for various value of n when $\kappa = 1$ are given in Figures 4.7(a)-(e). From these curves, it can be seen that small sample size has slightly better performance as the contamination level Δ_x increases. However, for very concentrated data ($\kappa = 20$), the power of the statistic is similar for all values of n as shown in Figures 4.8(a)-(e).



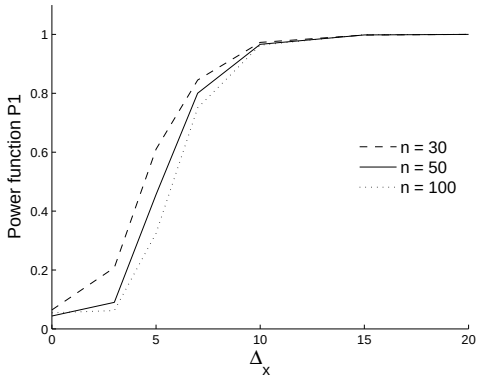
(a) $\Delta_\theta = 0$



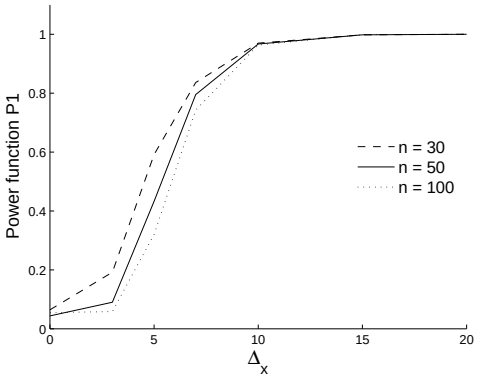
(b) $\Delta_\theta = 0.3$



(c) $\Delta_\theta = 0.5$

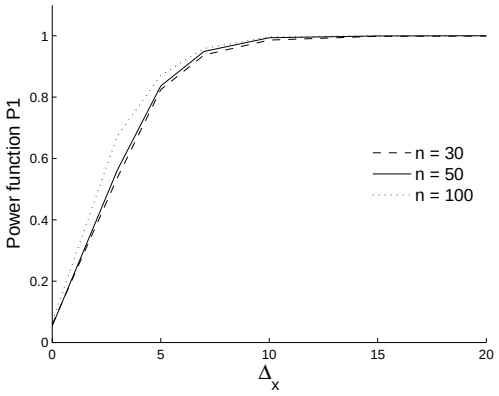


(d) $\Delta_\theta = 0.7$

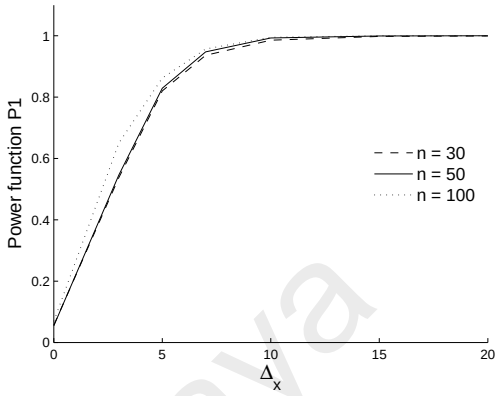


(e) $\Delta_\theta = 1$

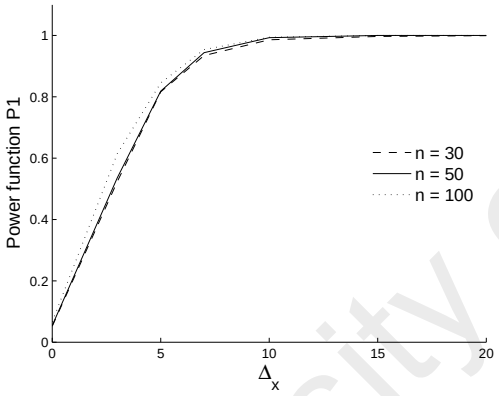
Figure 4.7: Power function, P1 of the C_{30}^1 statistic when $\kappa = 1$.



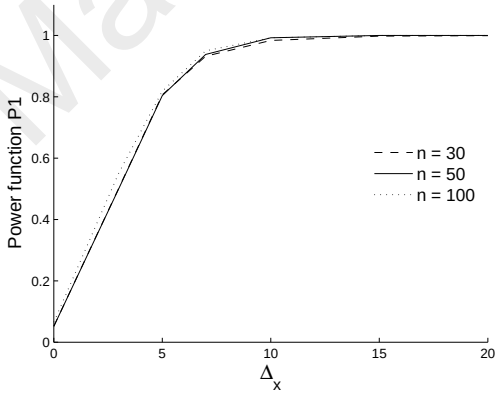
(a) $\Delta_\theta = 0$



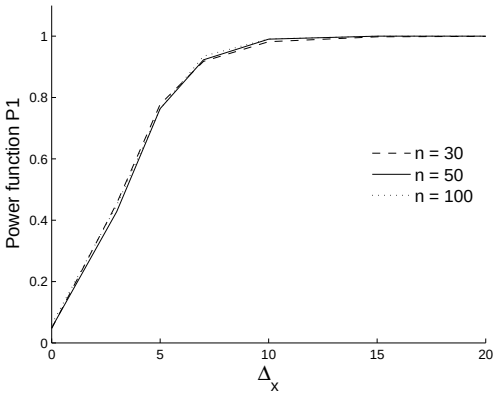
(b) $\Delta_\theta = 0.3$



(c) $\Delta_\theta = 0.5$



(d) $\Delta_\theta = 0.7$



(e) $\Delta_\theta = 1$

Figure 4.8: Power function, P1 of the C_{30}^1 statistic when $\kappa = 20$.

The performance is the same for the other performance criteria $P3$ and $P5$. In addition, the differences between $P1$ and $P3$ (Appendix E) also close to 0. Based on the results obtained, we conclude that C_n^1 statistic performs well in detecting outlier in a cylindrical data.

4.8 Practical Example

For a practical example, we use wind direction (in degrees) and wind speed (in m/s) data obtained from the Malaysian Meteorological Department, which were measured at Bayan Lepas, Penang, in January 2005 with pressure of 850 hPa around 12:00 am. The data is given in Table 3.3.

We obtain the parameter values of the JW distribution where $\hat{\mu} = 106°$, $\hat{\kappa} = 0.22$ and $\hat{\lambda} = 0.40$ using the iterative MLE method. Low value of concentration parameter κ implies that the data are dispersed with most of the wind blows to the east of Bayan Lepas. The data were plotted in Figure 4.9. The figure shows a 3D plot of the data to illustrate a cylindrical shape where the data are scattered around the circumference of the cylinder. From the figure, its obviously shown that there is only one observation located further away from the rest of the observations. This observation is notably an outlier from the linear component. Hence, further investigation is needed to test the discordancy using C_n^k statistic.

Based on the estimated parameter values, we obtain the critical values for the C_{31}^1 statistic given in Table 4.4. The value of the test statistic which corresponds to observation 1 is $C_{31}^1 = 11.49$. We reject the null hypothesis at 5% significance level which is 10.19, given in Table 4.4. Hence, the C_n^k statistic has identified the 1st observation as an outlier. The deletion of the 1st observation from the original data set changes the values of $\hat{\mu}$ to 113°, $\hat{\kappa}$ to 0.20 and $\hat{\lambda}$ to 0.41.

Table 4.4: The critical values for C_{31}^1 statistic.

Upper Level	C_{31}^1
10%	8.04
5%	10.19
1%	15.16

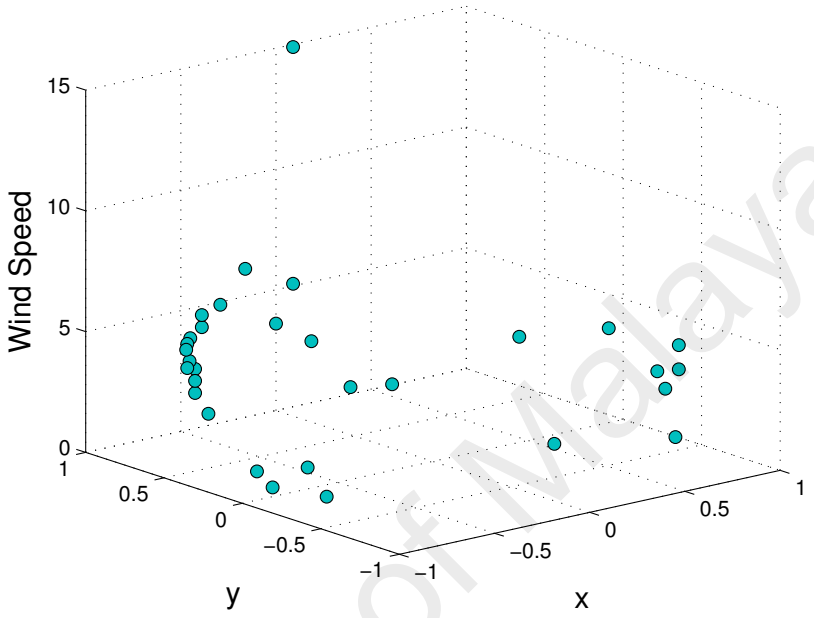


Figure 4.9: Scatter plot of wind speed vs wind direction.

4.9 Summary

In this chapter, we propose a new statistical test for detecting outlier in cylindrical data based on the theory of k-nearest neighbor. For the case of 1-nearest neighbor, we have shown that the C_n^1 statistic performs well in detecting single outlier especially for the case of outlier in linear component and outlier in both circular and linear components. Moreover, we have shown that the C_n^1 statistic can be used on wind data set by detecting an outlier in linear component as expected.

CHAPTER 5: REGRESSION FOR CYLINDRICAL DATA

5.1 Introduction

Cylindrical regression is a type of regression when both the circular and linear variables present in the model. There are three different classes of regression in circular data, namely circular-circular regression, circular-linear regression and linear-circular regression. Hence, we can treat both circular-linear regression and linear-circular regression as a type of cylindrical regression. In this chapter, we propose a new test statistic to detect outliers in the regression for cylindrical data. We then try to obtain the critical values and study the performance of the test statistic via simulation. As an illustration, the wind data set is used for the purpose of outlier detection in cylindrical data.

5.2 Johnson & Wehrly (JW) Circular-Linear Regression Model

Johnson and Wehrly (1978) proposed three different models of circular regression. One of the model is known as a regression of a linear variate on other linear and circular variates or known as circular-linear regression. This model will be referred as the JW circular-linear regression model. The model is constructed from the conditional distribution of $\boldsymbol{x}_1 = (x_1, \ldots, x_r)'$ given $\boldsymbol{x}_2$ and $\boldsymbol{\theta}$, $f(\boldsymbol{x}_1|\boldsymbol{x}_2, \boldsymbol{\theta})$ which is the r-dimensional normal distribution with mean $\lambda_1 + \Sigma_{12}\Sigma_{22}^{-1}[\boldsymbol{x}_2 - (\lambda_2 + \boldsymbol{a}_2(\boldsymbol{\theta}))]$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ from the joint density $f(\boldsymbol{\theta}, \boldsymbol{x})$ such that

$$f(\boldsymbol{\theta}, \boldsymbol{x}) = c \cdot \exp\{-\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x} + \lambda'\Sigma^{-1}\boldsymbol{x} + \boldsymbol{a}(\boldsymbol{\theta})'\Sigma^{-1}\boldsymbol{x}\}$$

where c is a constant of integration, $\boldsymbol{a}(\boldsymbol{\theta})' = (a_1(\boldsymbol{\theta}), \ldots, a_q(\boldsymbol{\theta}))$ given by

$$a_i(\boldsymbol{\theta}) = \sum_{j=1}^{p}\sum_{k=1}^{n}\Big[u_{ijk}\cos(k\theta_j) + v_{ijk}\sin(k\theta_j)\Big], \quad i = 1, \ldots, q,$$

where $\boldsymbol{x} \in \mathbb{R}^q, \boldsymbol{\theta} \in [0, 2\pi)^p, \lambda \in \mathbb{R}^q, \Sigma^{-1}$ is positive definite while u_{ijk} and v_{ijk} are constant. Let us partition $\boldsymbol{x} = (\boldsymbol{x}_1'|\boldsymbol{x}_2')'$ and hence λ, Σ and $\boldsymbol{a}(\boldsymbol{\theta})$ accordingly. The conditional distribution of $f(\boldsymbol{x}|\boldsymbol{\theta})$ is q-dimensional multivariate normal with mean $\lambda + \boldsymbol{a}(\boldsymbol{\theta})$ and covariance matrix Σ. To predict $\boldsymbol{x}_1$ for a given $\boldsymbol{x}_2$ and $\boldsymbol{\theta}$, consider the conditional expectation of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ and $\boldsymbol{\theta}$ such that

$$
\begin{aligned}
E(\boldsymbol{x}_1|\boldsymbol{x}_2, \boldsymbol{\theta}) &= \lambda_1 + \Sigma_{12}\Sigma_{22}^{-1}[\boldsymbol{x}_2 - (\lambda_2 + \boldsymbol{a}_2(\boldsymbol{\theta}))] \\
&= \lambda_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2 - \Sigma_{12}\Sigma_{22}^{-1}\Big(\lambda_2 + \boldsymbol{a}_2(\boldsymbol{\theta})\Big) \\
&= \lambda_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2 - \Sigma_{12}\Sigma_{22}^{-1}\Big[\lambda_2 + \sum_{i=r+1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\Big(u_{ijk}\cos(k\theta_j) \\
&\quad + v_{ijk}\sin(k\theta_j)\Big)\Big] \\
&= \Big(\lambda_1 - \Sigma_{12}\Sigma_{22}^{-1}\lambda_2\Big) + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2 - \Big[\Sigma_{12}\Sigma_{22}^{-1}\sum_{i=r+1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} \\
&\quad \Big(u_{ijk}\cos(k\theta_j) + v_{ijk}\sin(k\theta_j)\Big)\Big] \\
&= \Big(\lambda_1 - \Sigma_{12}\Sigma_{22}^{-1}\lambda_2\Big) + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2 - \Sigma_{12}\Sigma_{22}^{-1}\sum_{i=r+1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} \\
&\quad (u_{ijk}\cos(k\theta_j) - \Sigma_{12}\Sigma_{22}^{-1}\sum_{i=r+1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} v_{ijk}\sin(k\theta_j) \\
&= \beta_0 + \beta_2\boldsymbol{x}_2 + \sum_{i=r+1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\Big[\gamma_{ijk}\cos(k\theta_j) + \delta_{ijk}\sin(k\theta_j)\Big]
\end{aligned}
$$

where

$$
\boldsymbol{\beta}_0 = \boldsymbol{\lambda}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\lambda}_2, \quad \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1},
$$

$$
\boldsymbol{\gamma}_{ijk} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{u}_{ijk}, \quad \boldsymbol{\delta}_{ijk} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{v}_{ijk}.
$$

For each component $x_i, i = 1, \ldots, r$ of $\boldsymbol{x}_1$, the mean is given by

$$\beta_0 + \sum_{i=r+1}^{q} \beta_i x_i + \sum_{i=r+1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} \Big[\gamma_{ijk} \cos(k\theta_j) + \delta_{ijk} \sin(k\theta_j)\Big].$$

Hence, the regression for the cylindrical data of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ and $\boldsymbol{\theta}$ can be shown in the form of

$$\boldsymbol{x}_1 = \beta_0 + \beta_2 \boldsymbol{x}_2 + \sum_{k=1}^{n} \Big[\gamma \cos(k\boldsymbol{\theta}) + \delta \sin(k\boldsymbol{\theta})\Big] + \boldsymbol{\epsilon}, \tag{5.2.1}$$

where β_0, β_2, γ and δ are the coefficients which represent the relationship between the variables, k is the angular frequency. This model is basically reduced to a standard method of predicting a linear variable from a mixture of linear and circular variables.

In the next section, we use the simple form of the model given in equation (5.2.1) with one linear variable and one circular variable with the frequency $k = 1$. The model takes the form of

$$x_{1i} = \beta_0 + \beta_2 x_{2i} + \gamma \cos(\theta_i) + \delta \sin(\theta_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{5.2.2}$$

$\epsilon_i \sim N(0, \sigma^2)$. The estimation of the parameters β_0, β_2, γ and δ can be obtained using the least square estimation method.

5.3 Estimation of JW Circular-Linear Regression Model

The matrix notation of the regression is given by

$$\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & \cos(k\theta_1) & \sin(k\theta_1) \\ 1 & x_{22} & \cos(k\theta_2) & \sin(k\theta_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cos(k\theta_n) & \sin(k\theta_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

For simplicity. we define

$$\boldsymbol{x_1} = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix}, \quad \boldsymbol{x_2} = \begin{bmatrix} 1 & x_{21} & \cos(k\theta_1) & \sin(k\theta_1) \\ 1 & x_{22} & \cos(k\theta_2) & \sin(k\theta_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cos(k\theta_n) & \sin(k\theta_n) \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \gamma \\ \delta \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

In order to estimate $\boldsymbol{b}$, we need to minimize the sum of squares of the residual as given by

$$\begin{aligned} \boldsymbol{E} = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\boldsymbol{x_1} - \boldsymbol{x_2 b})'(\boldsymbol{x_1} - \boldsymbol{x_2 b}) \\ &= \boldsymbol{x_1}'\boldsymbol{x_1} - \boldsymbol{x_1}'\boldsymbol{x_2 b} - \boldsymbol{x_2}'\boldsymbol{b}'\boldsymbol{x_1} + \boldsymbol{x_2 b}'\boldsymbol{x_2 b} \\ &= \boldsymbol{x_1}'\boldsymbol{x_1} - 2\boldsymbol{b}'\boldsymbol{x_2}'\boldsymbol{x_1} + \boldsymbol{b}'\boldsymbol{x_2}'\boldsymbol{x_2 b} \end{aligned}$$

then differentiate $\boldsymbol{E}$ with respect to $\boldsymbol{b}$ and set the derivative equal to 0.

$$\begin{aligned} \frac{\partial(\boldsymbol{E})}{\partial(\boldsymbol{b})} = -2\boldsymbol{b}'\boldsymbol{x_2}'\boldsymbol{x_1} + 2\boldsymbol{x_2}'\boldsymbol{x_2 b} &= 0 \\ \boldsymbol{x_2}'\boldsymbol{x_2 b} &= \boldsymbol{x_2}'\boldsymbol{x_1} \\ \hat{\boldsymbol{b}} &= (\boldsymbol{x_2}'\boldsymbol{x_2})^{-1}\boldsymbol{x_2}'\boldsymbol{x_1} \end{aligned}$$

Therefore, the maximum likelihood estimation of $\boldsymbol{b}$ is $\hat{\boldsymbol{b}} = (\boldsymbol{x_2}'\boldsymbol{x_2})^{-1}\boldsymbol{x_2}'\boldsymbol{x_1}$.

5.4 Outlier Detection in a Regression Model for Cylindrical Data using k-NN statistic

The new outlier detection in a regression model for cylindrical data using JW circular-linear regression model is constructed based on the k-NN approach. The method is then

applied to the distance measure between two residual. The residual is given by

$$e_i = x_{1i} - \hat{x}_{1i}, \quad i = 1, \ldots, n. \tag{5.4.1}$$

Given e_i and e_j, the distance between two residual is given by

$$d(e_i, e_j) = |e_i - e_j|, \quad i = 1, 2, \ldots, n.$$

Using the same k-NN approach given in section 4.4, the k-NN distance for this case is given as

$$L_{ki} = d_{(k)}(e_i, e_j), \quad k = 1, 2, 3, \ldots, \quad i, j = 1, 2, \ldots, n, \quad i \neq j. \tag{5.4.2}$$

Therefore, the test statistic is given by

$$L_n^k = \max_i\{L_{ki}\}, \tag{5.4.3}$$

where n is the sample size and k is the k^{th}-nearest neighbor. The complete steps to detect the outlier in regression for cylindrical data are given below:

Step 1: Fit the circular-linear regression x_1 given in equation (5.2.2).

Step 2: Calculate the residual, e for each observations using equation (5.4.1).

Step 3: Choose any $k = 1, 2, 3, \ldots$ for the k-nearest neighbor distance, then calculate the distance between each residuals, L_{ki} as given in equation (5.4.2).

Step 4: Then define the test statistic L_n^k as given in equation (5.4.3).

Step 5: If the value of L_n^k exceeds the cut-off point, say a_L, then the i^{th} observation corresponding to $\max_i\{L_{ki}\}$ is identified as an outlier.

We note that L_n^k statistic can also be used to detect a patch of outliers. For example, when $k = 1$, it can be used to detect an outlier while when $k = 2$, it can be used to detect a patch of 2 outliers. For multiple outliers, we usually need to repeat the L_n^k statistic iteratively for $k = 1, 2, 3, \ldots$ until no outliers are detected.

5.5 Cut-off Points of the Test Statistics

We design a simulation study for L_n^k statistic to obtain the cut-off points using the *R* statistical package based on the null hypothesis that there are no outliers present in the cylindrical data set. The focus of this study is for single outlier and also two outliers. The generation of the cut-off points are based on the sample size n and residual standard deviation σ.

In our study, the cut-off points are generated from various values of sample size, n and σ as shown in Table 5.1-5.4. We generate samples x_2 from Normal distribution, $N(5, 2)$ and θ from von Mises distribution, $VM(\pi, 2)$. Then, we generate e of size n from $N(0, \sigma)$. For each samples, we find the estimated value of the parameters $\beta_0 = 0.306, \beta_2 = 1, \gamma = 1$ and $\delta = 1$ and obtain the variable x_1 using equation (5.2.2). Next, we compute the fitted value $\hat{x}_1$ and calculate the residual, e. We then calculate the distance between each residuals and sort the values from smallest to the largest to find the respective ordered distance. Next, we set $k = 1$ for single outlier or $k = 2$ for two outliers to find the k^{th} nearest distance to obtain the L_n^k statistic as given in equation (5.4.3). The process is repeated 2000 times and the estimated cut-off points at 10%, 5% and 1% upper percentile are collected.

The cut off-points of L_n^k statistic for $k = 1, 2$ are tabulated in Tables 5.1 - 5.4 respectively. It can be seen that for each sample size n, the cut off points are increases as the value of σ increases. High values of σ indicates that the residuals are spread out over the wider range from the mean and resulting in higher values of cut-off points. On the other hand, the cut-off points are a decreasing function of sample size n. The exact cut-off point can

Table 5.1: Cut-off points for L_n^1 statistic when $0.05 \leq \sigma \leq 0.4$.

n	Significance Level	σ					
		0.05	0.08	0.1	0.2	0.3	0.4
10	10%	0.06	0.10	0.12	0.24	0.36	0.48
	5%	0.07	0.11	0.14	0.28	0.43	0.57
	1%	0.09	0.15	0.18	0.37	0.55	0.73
20	10%	0.06	0.09	0.12	0.23	0.35	0.47
	5%	0.07	0.11	0.14	0.28	0.41	0.55
	1%	0.09	0.14	0.18	0.36	0.54	0.72
30	10%	0.06	0.09	0.11	0.22	0.33	0.44
	5%	0.07	0.11	0.13	0.26	0.40	0.53
	1%	0.09	0.14	0.18	0.35	0.53	0.71
50	10%	0.05	0.08	0.10	0.21	0.31	0.42
	5%	0.06	0.10	0.13	0.26	0.39	0.52
	1%	0.09	0.14	0.18	0.36	0.53	0.71
80	10%	0.05	0.08	0.10	0.20	0.30	0.40
	5%	0.06	0.10	0.12	0.24	0.37	0.49
	1%	0.08	0.13	0.16	0.32	0.48	0.65
100	10%	0.05	0.08	0.10	0.20	0.31	0.41
	5%	0.06	0.10	0.12	0.24	0.36	0.48
	1%	0.08	0.12	0.16	0.31	0.47	0.62

Table 5.2: Cut-off points for L_n^1 statistic when $0.5 \leq \sigma \leq 1$.

n	Significance Level	σ					
		0.5	0.6	0.7	0.8	0.9	1
10	10%	0.60	0.72	0.84	0.96	1.08	1.20
	5%	0.71	0.85	0.99	1.14	1.28	1.42
	1%	0.91	1.10	1.28	1.46	1.65	1.83
20	10%	0.59	0.70	0.82	0.94	1.06	1.17
	5%	0.69	0.83	0.97	1.11	1.24	1.38
	1%	0.90	1.08	1.26	1.44	1.62	1.80
30	10%	0.55	0.66	0.77	0.88	0.99	1.11
	5%	0.66	0.79	0.93	1.06	1.19	1.32
	1%	0.89	1.06	1.24	1.42	1.59	1.77
50	10%	0.52	0.63	0.73	0.84	0.94	1.05
	5%	0.64	0.77	0.90	1.03	1.16	1.29
	1%	0.89	1.07	1.25	1.43	1.60	1.78
80	10%	0.51	0.61	0.71	0.81	0.91	1.01
	5%	0.61	0.73	0.85	0.98	1.10	1.22
	1%	0.81	0.97	1.13	1.29	1.45	1.62
100	10%	0.51	0.61	0.71	0.82	0.92	1.02
	5%	0.60	0.71	0.83	0.95	1.07	1.19
	1%	0.78	0.94	1.09	1.25	1.40	1.56

Table 5.3: Cut-off points for L_n^2 statistic when $0.05 \leq \sigma \leq 0.4$.

n	Significance Level	σ					
		0.05	0.08	0.1	0.2	0.3	0.4
10	10%	0.06	0.12	0.12	0.31	0.36	0.48
	5%	0.07	0.14	0.14	0.36	0.43	0.57
	1%	0.09	0.18	0.18	0.45	0.55	0.73
20	10%	0.06	0.12	0.12	0.30	0.35	0.47
	5%	0.07	0.14	0.14	0.34	0.41	0.55
	1%	0.09	0.17	0.18	0.41	0.54	0.72
30	10%	0.06	0.11	0.11	0.29	0.33	0.44
	5%	0.07	0.13	0.13	0.33	0.40	0.53
	1%	0.09	0.17	0.18	0.42	0.53	0.71
50	10%	0.05	0.11	0.10	0.27	0.31	0.42
	5%	0.06	0.13	0.13	0.31	0.39	0.52
	1%	0.09	0.16	0.18	0.40	0.53	0.71
80	10%	0.05	0.10	0.10	0.26	0.30	0.40
	5%	0.06	0.12	0.12	0.30	0.37	0.49
	1%	0.08	0.15	0.16	0.37	0.48	0.65
100	10%	0.05	0.10	0.10	0.25	0.31	0.41
	5%	0.06	0.12	0.12	0.29	0.36	0.48
	1%	0.08	0.15	0.16	0.37	0.47	0.62

Table 5.4: Cut-off points for L_n^2 statistic when $0.5 \leq \sigma \leq 1$.

n	Significance Level	σ					
		0.5	0.6	0.7	0.8	0.9	1
10	10%	0.78	0.72	0.84	1.25	1.08	1.56
	5%	0.90	0.85	0.99	1.44	1.28	1.80
	1%	1.13	1.10	1.28	1.80	1.65	2.25
20	10%	0.75	0.70	0.82	1.19	1.06	1.49
	5%	0.84	0.83	0.97	1.35	1.24	1.69
	1%	1.03	1.08	1.26	1.65	1.62	2.07
30	10%	0.71	0.66	0.77	1.14	0.99	1.43
	5%	0.81	0.79	0.93	1.30	1.19	1.63
	1%	1.04	1.06	1.24	1.66	1.59	2.08
50	10%	0.67	0.63	0.73	1.08	0.94	1.35
	5%	0.78	0.77	0.90	1.25	1.16	1.57
	1%	1.01	1.07	1.25	1.61	1.60	2.01
80	10%	0.64	0.61	0.71	1.03	0.91	1.28
	5%	0.75	0.73	0.85	1.19	1.10	1.49
	1%	0.93	0.97	1.13	1.48	1.45	1.85
100	10%	0.63	0.61	0.71	1.01	0.92	1.26
	5%	0.72	0.71	0.83	1.16	1.07	1.45
	1%	0.93	0.94	1.09	1.49	1.40	1.86

be obtained for any combination of estimated σ and sample size n. In fact, we expect to have a more accurate cut-off point by using higher number of simulation as pointed by Verma et al. (2017).

5.6 The Performance of L_n^k statistic

5.6.1 The Performance of L_n^1 Statistic

To investigate the performance of L_n^1 statistic, we use similar procedures as given in section 4.7. The focus on this statistic is when $k = 1$ for single outlier. From Barnett and Lewis (1994) and David (1981), $P1 = 1 - \beta$ is the power function where β is the type-II error; $P3$ is the probability that the contaminant point is an outlier and it is identified as discordant; and $P5$ is the probability that that the contaminant point is an outlier given that it is identified as discordant. A good test should have (i) high $P1$; (ii) high $P5$ and (iii) low $P1 - P3$.

The performance of L_n^1 statistic is conducted using simulation method. The samples are generated from various samples $n = 20, 50, 80, 100$ from Normal distribution, $x_2 \sim N(5, 2)$ and von Mises distribution, $\theta \sim VM(\pi, 2)$ with different values of $\sigma = 0.2, 0.3, 0.5, 0.8, 1, 2$. Using the generated data of x_2 and θ, the values of the response variable x_1 is obtained using equation (5.2.2). Then, The outlier is generated by altering

$$x'_{1,n} = x_{1,n} + \Delta,$$

where $\Delta > 0$ is the contamination level. Next, the generated cylindrical data of x_1, x_2, and θ are fitted to JW circular-linear regression to find the estimates of $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}$ and $\hat{\delta}$. Then, we apply the L_n^1 statistic for the detection of outlier in each samples. If the value of the L_n^1 statistic is greater than the specified cut-off points, then we have correctly detected the outlier. The process is repeated for 2000 times and the values of $P1, P3$ and $P5$ are

obtained.

Table 5.5 shows the proportion of correct detection of outlier when applied to L_n^1 statistic. The results shows that the proportion is an increasing function of n. The proportion is decreasing as the value of σ increasing. It is expected since small σ indicates that the samples are very close to the mean, hence make it easier to detect an outlier. For large value of σ, the proportion are increasing as the value of the contamination level increases.

The results for the samples when $n = 20$ and $n = 100$ are plotted in Figure 5.1 and Figure 5.2 respectively. From both figures, the performance of $P1$ and $P5$ shows similar behaviour. It can be seen that the performance of L_n^1 statistic depend on the value of σ. The performance is better as the value of σ decreases. Hence, the performances are a decreasing function of σ. However, $n = 100$ has a better performance compared to when $n = 20$. When n is large, the distance between the residuals is expected to be shorter resulting in lower values of L_n^1 statistic as illustrated by the smaller cut-off points as shown in Table 5.1. Hence, when outlier occurs in large sample size, we detect the corresponding observation easier as its respective distance will be relatively longer compared to the case in smaller sample size.

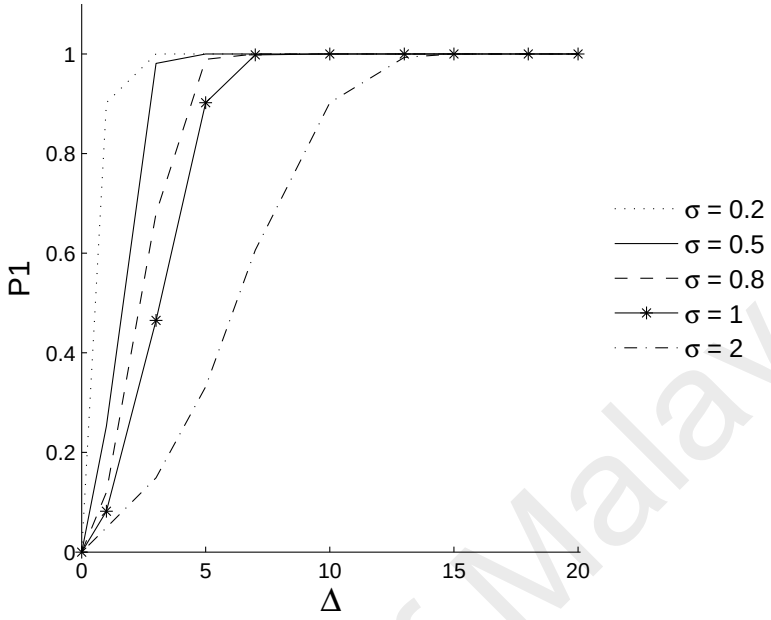
From Tables 5.3 - 5.5, we can see that the performance of $P1$ and $P5$ for different values of n shows similar behaviour. However, we note that the large sample size approaching 1 slightly faster. In addition, the differences between $P1$ and $P3$ (Appendix F) are also approximately close to 0.

5.6.2 The Performance of L_n^2 Statistic

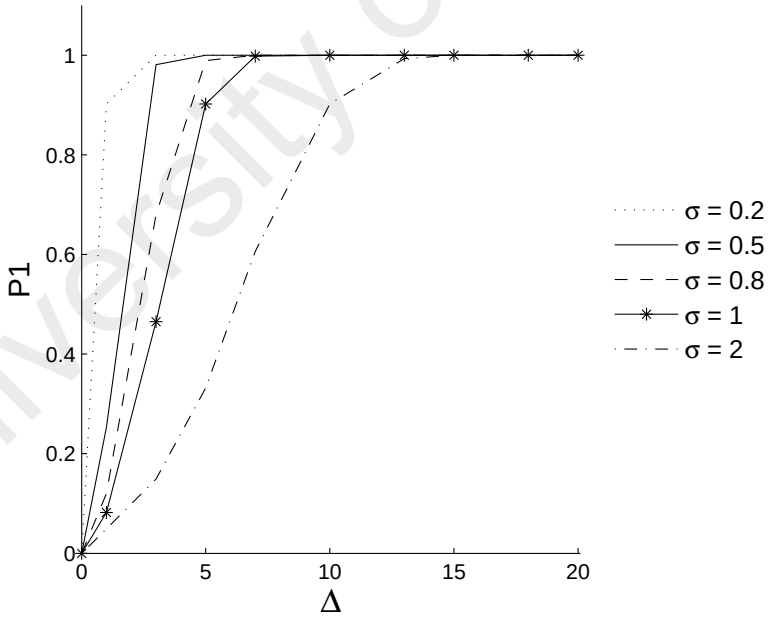
To investigate the performance of L_n^2 statistic for multiple outlier, the samples are generated from various samples $n = 20, 50, 80, 100$ from normal distribution, $x_2 \sim N(5, 2)$ and von Mises distribution, $\theta \sim VM(\pi, 5)$ with different values of $\sigma = 0.2, 0.5, 0.8, 1, 2$.

Table 5.5: The proportion of correct detection of outlier for L_n^1 statistic.

n	Δ	σ					
		0.2	0.3	0.5	0.8	1	2
20	0	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.78	0.38	0.09	0.02	0.01	0.00
	3	0.99	0.97	0.89	0.49	0.28	0.04
	5	0.99	0.99	0.97	0.91	0.78	0.18
	7	0.99	0.99	0.99	0.97	0.94	0.43
	10	1.00	0.99	0.99	0.99	0.97	0.78
	13	1.00	1.00	0.99	0.99	0.99	0.92
	15	1.00	1.00	0.99	0.99	0.99	0.96
	18	1.00	1.00	0.99	0.99	0.99	0.97
	20	1.00	1.00	0.99	0.99	0.99	0.97
50	0	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.87	0.40	0.07	0.02	0.01	0.01
	3	1.00	1.00	0.97	0.54	0.30	0.03
	5	1.00	1.00	1.00	0.98	0.87	0.17
	7	1.00	1.00	1.00	1.00	0.99	0.46
	10	1.00	1.00	1.00	1.00	1.00	0.87
	13	1.00	1.00	1.00	1.00	1.00	0.99
	15	1.00	1.00	1.00	1.00	1.00	1.00
	18	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00
80	0	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.86	0.36	0.06	0.01	0.01	0.00
	3	1.00	1.00	0.97	0.51	0.25	0.02
	5	1.00	1.00	1.00	0.98	0.86	0.14
	7	1.00	1.00	1.00	1.00	1.00	0.42
	10	1.00	1.00	1.00	1.00	1.00	0.86
	13	1.00	1.00	1.00	1.00	1.00	0.99
	15	1.00	1.00	1.00	1.00	1.00	1.00
	18	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00
100	0	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.87	0.36	0.06	0.01	0.00	0.00
	3	1.00	1.00	0.98	0.52	0.26	0.02
	5	1.00	1.00	1.00	0.99	0.87	0.13
	7	1.00	1.00	1.00	1.00	1.00	0.43
	10	1.00	1.00	1.00	1.00	1.00	0.87
	13	1.00	1.00	1.00	1.00	1.00	0.99
	15	1.00	1.00	1.00	1.00	1.00	1.00
	18	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00

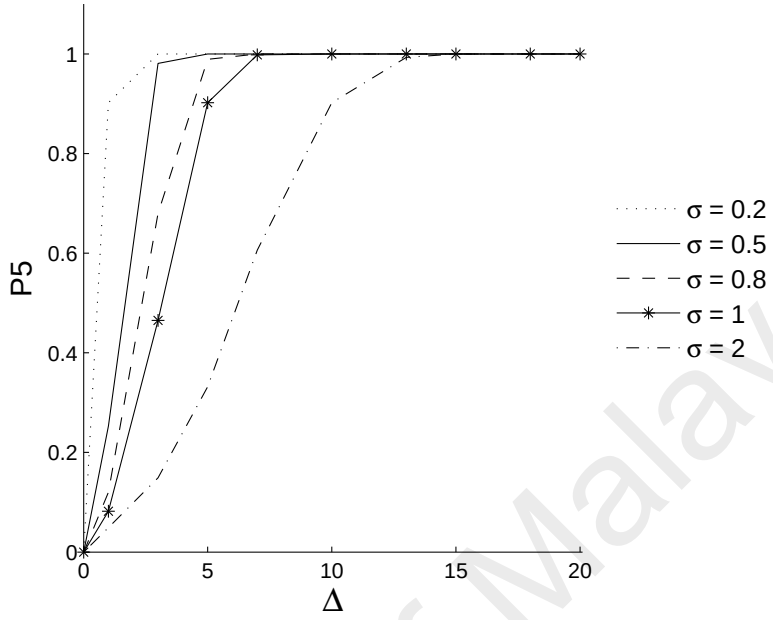


(a)

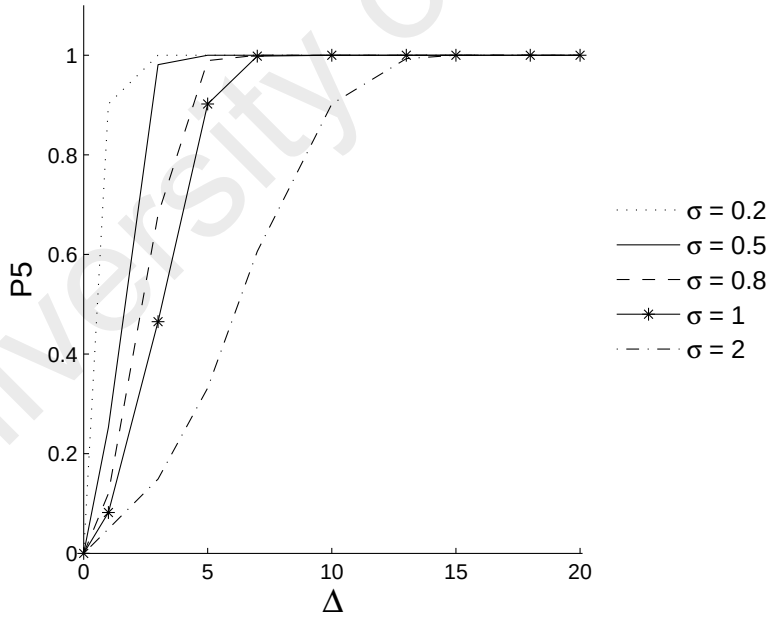


(b)

Figure 5.1: Sampling behaviour of the L_n^1 statistic for different values of σ when $n = 20$.

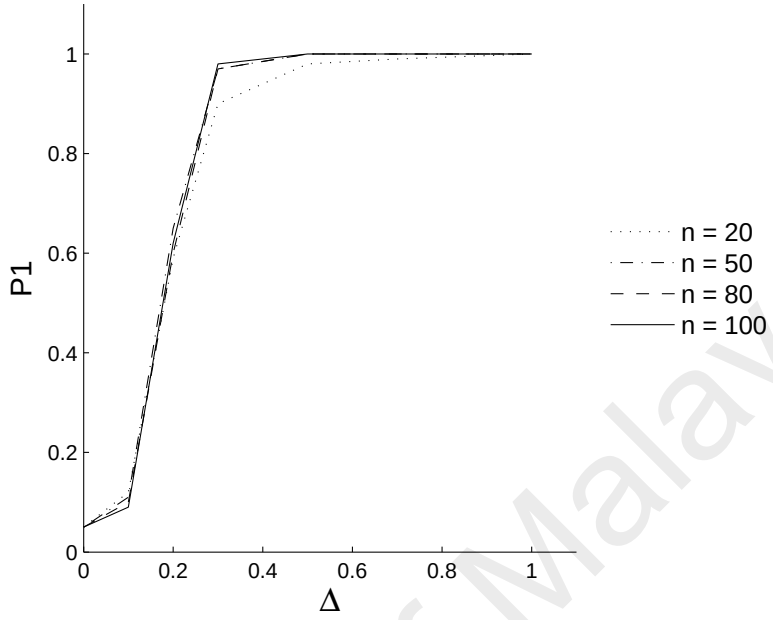


(a)

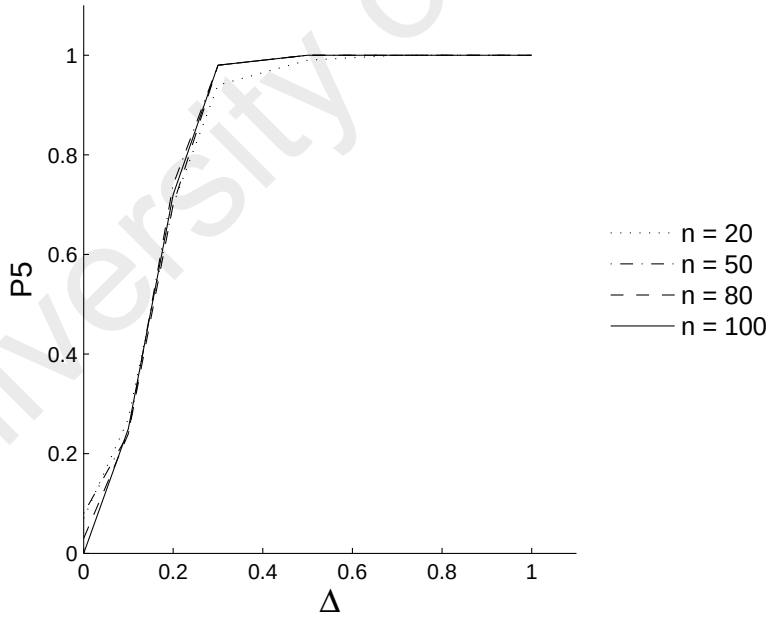


(b)

Figure 5.2: Sampling behaviour of the L_n^1 statistic for different values of σ when $n = 100$.

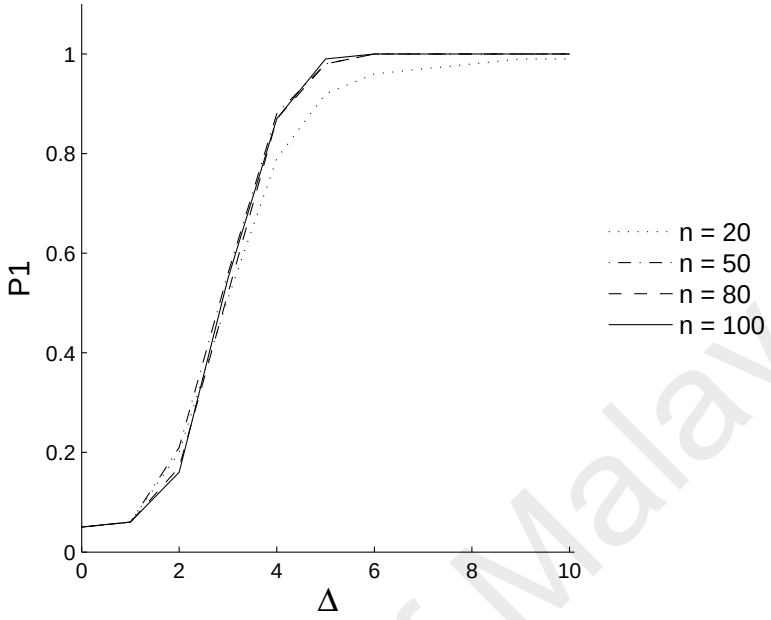


(a)

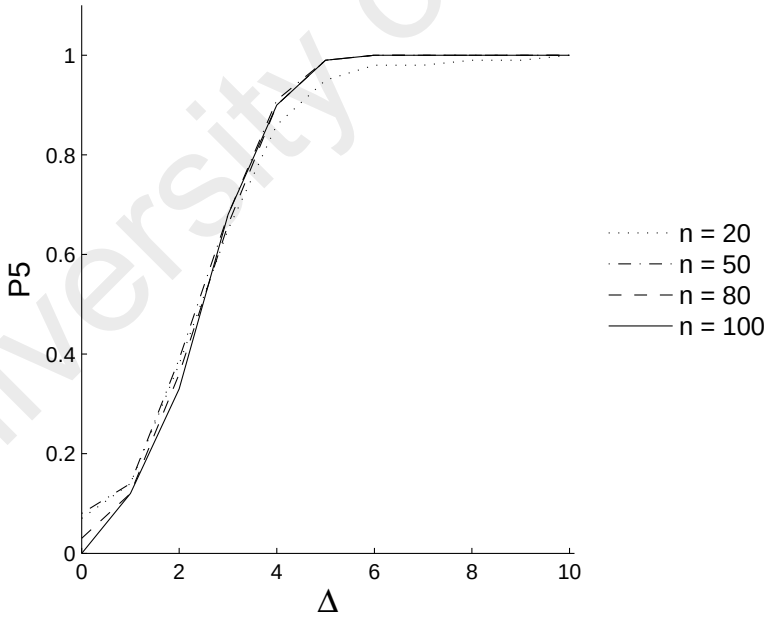


(b)

Figure 5.3: Sampling behaviour of the L_n^1 statistic for different values of n when $\sigma = 0.05$.

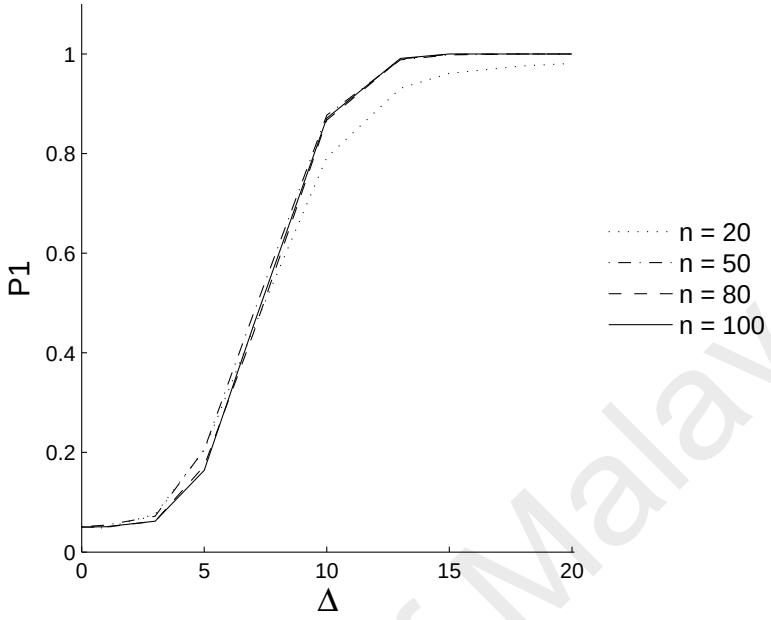


(a)

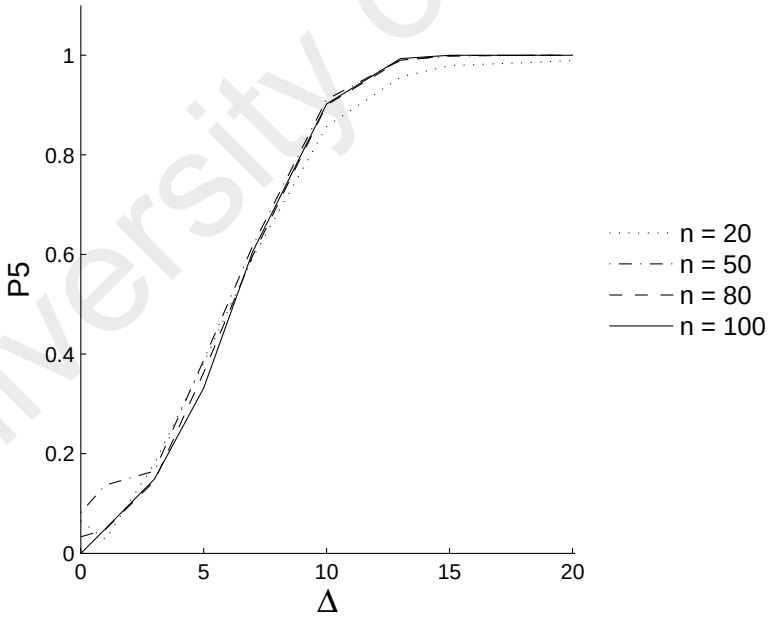


(b)

Figure 5.4: Sampling behaviour of the L_n^1 statistic for different values of n when $\sigma = 0.8$.



(a)



(b)

Figure 5.5: Sampling behaviour of the L_n^1 statistic for different values of n when $\sigma = 2$.

Then, the outliers are generated by altering

$$x'_{1,n} = x_{1,n} + \Delta$$

$$x'_{1,n-1} = x_{1,n-1} + \Delta$$

where $\Delta \geq 0$ is the contamination level. Then, similar procedure as the performance of L_n^1 is used.

The performance of L_n^2 statistic when $n = 50$ and $n = 100$ are given in Figure 5.6 and Figure 5.7 respectively. It can be seen that the performance is increasing as the value of σ is decreasing. Hence, the performance is a decreasing function of σ. Meanwhile, Figure 5.8 and Figure 5.9 shows that the performances of the test statistic at various sample size n are similar. However, the performance is smaller when the sample size is small. On the other hand, the differences between $P1$ and $P3$ (Appendix H) are also approximately close to 0. Generally, the performance of L_n^2 statistic shows a similar behaviour to L_n^1 statistic.

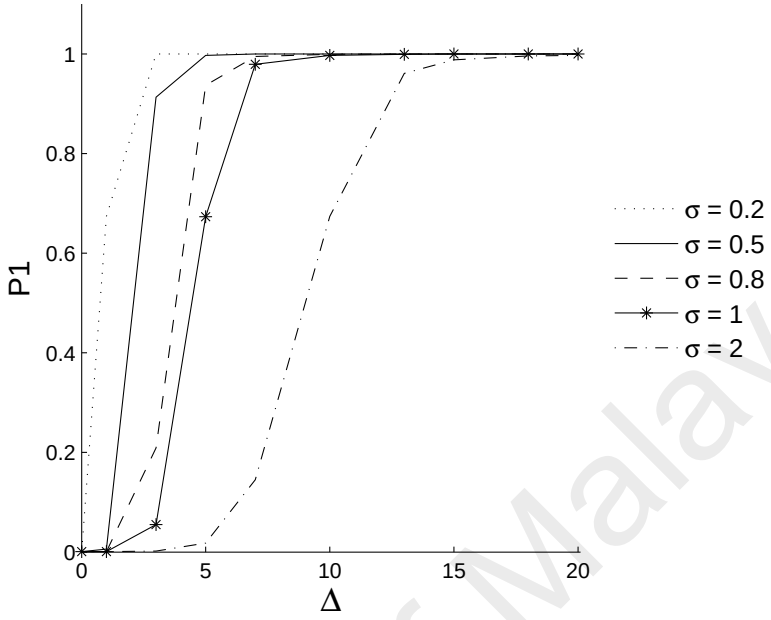
5.7 Practical Example

We now apply the JW circular-linear regression model on a real data set. The same data used in section 4.8 is applied here. However, we add another linear variable which is Temperature ($^\circ C$) to be fitted by JW circular-linear regression. The data is given in Table 5.6.

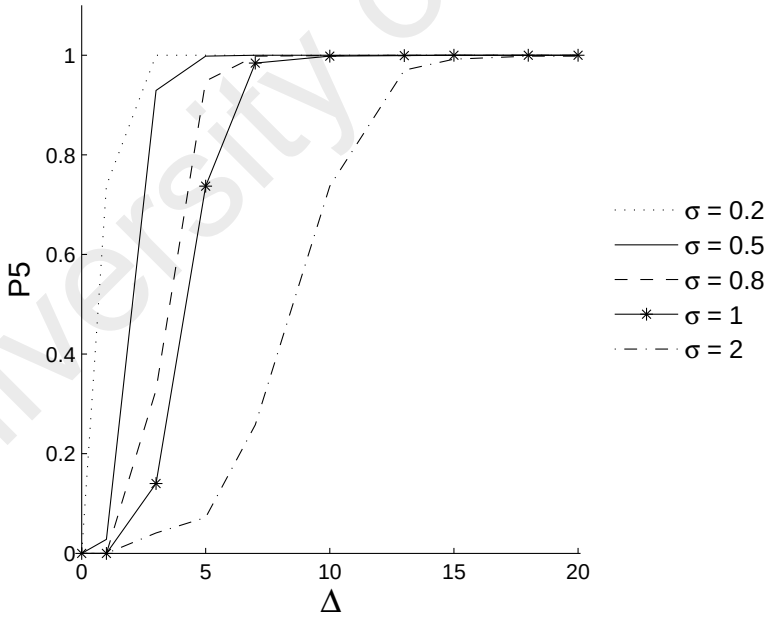
The parameter estimates of the JW circular-linear regression model are $\hat{\beta}_0 = 5.26, \hat{\beta}_2 = -0.12, \hat{\gamma} = 0.31$, and $\hat{\delta} = 1.49$ with the fitted $\hat{x_1}$ is given by

$$\hat{x_1} = 5.26 - 0.12x_2 + 0.31\cos\theta + 1.49\sin\theta$$

The regression plot of the variables shows a possibility of occurrence of outlier in the data

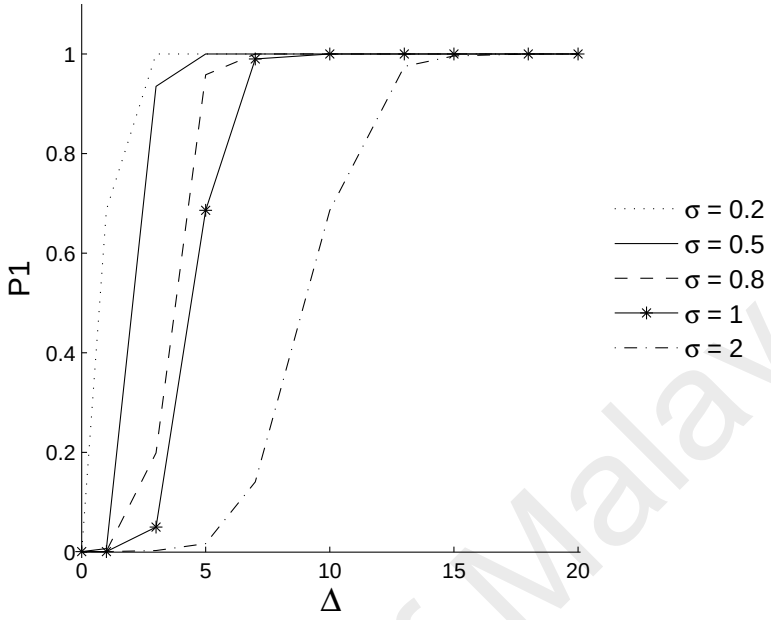


(a)

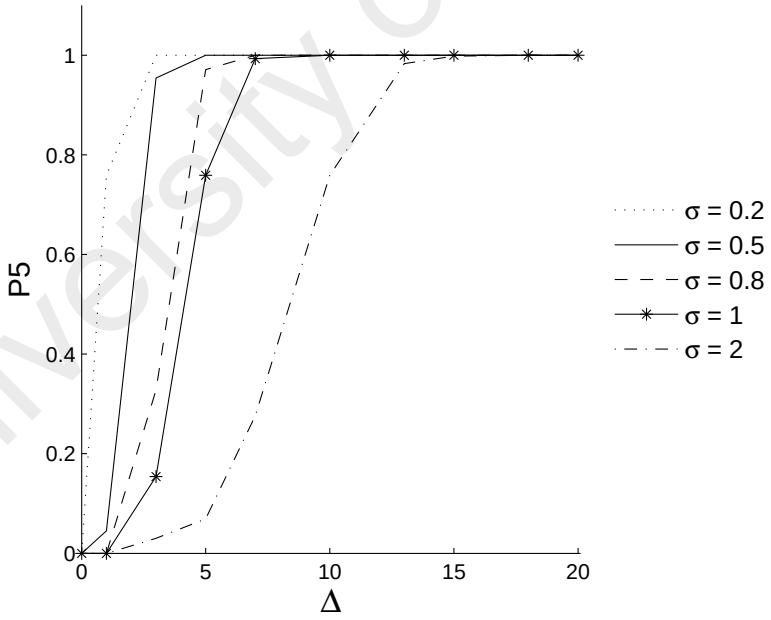


(b)

Figure 5.6: Sampling behaviour of the L_n^2 statistic for different values of σ when $n = 50$.

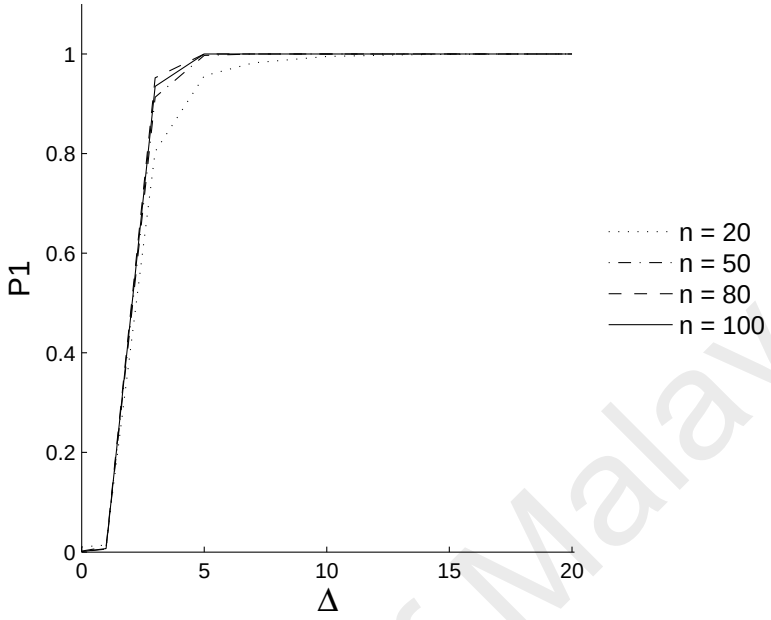


(a)

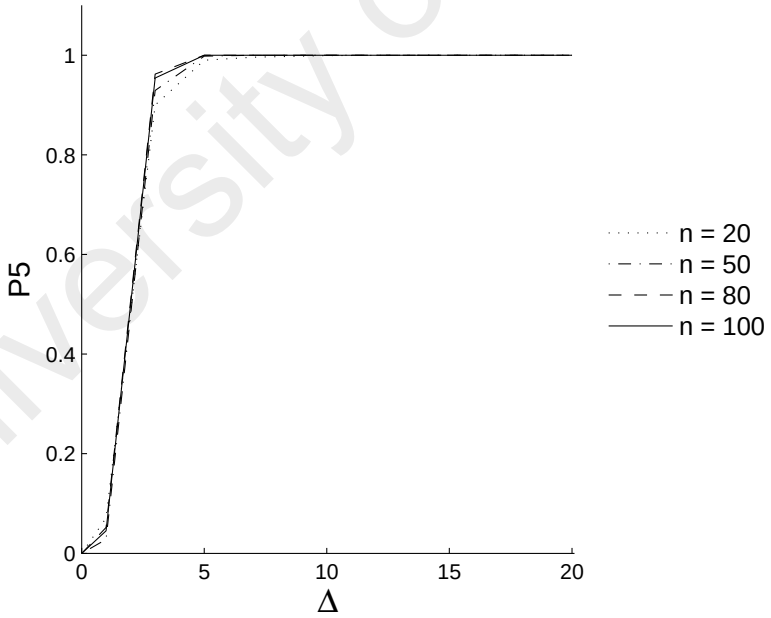


(b)

Figure 5.7: Sampling behaviour of the L_n^2 statistic for different values of σ when $n = 100$.

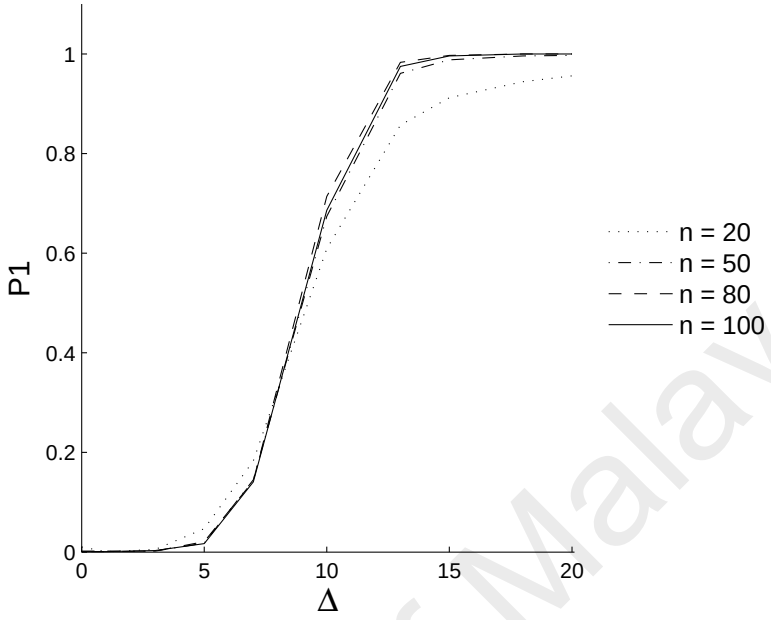


(a)

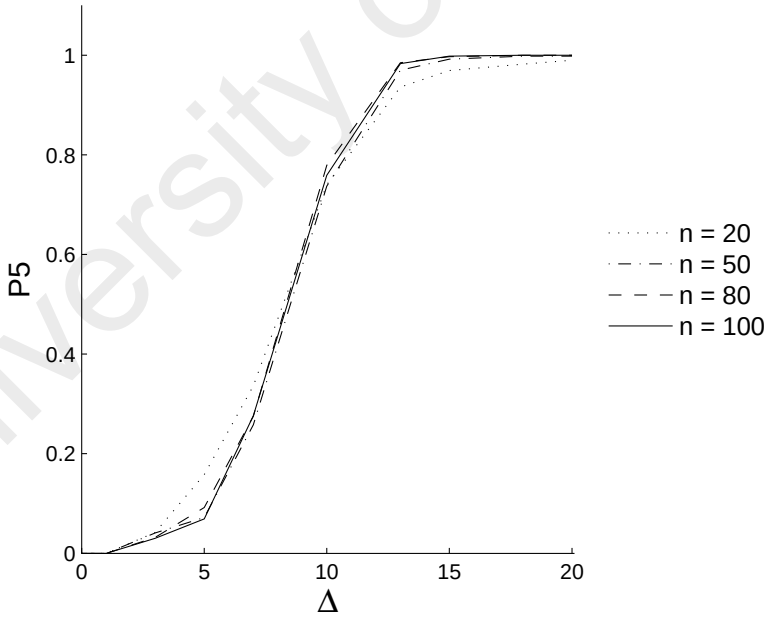


(b)

Figure 5.8: Sampling behaviour of the L_n^2 statistic for different values of n when $\sigma = 0.5$.



(a)



(b)

Figure 5.9: Sampling behaviour of the L_n^2 statistic for different values of n when $\sigma = 2$.

Table 5.6: The Wind Data.

Wind Speed (m/s)	Temperature ($^{\circ}C$)	Wind Direction ($^{\circ}$)
14.9	17.6	85
5.1	18.0	85
4.6	18.2	140
6.2	18.0	100
3.6	18.2	135
1.5	17.6	310
2.1	18.4	340
4.6	18.2	120
4.6	17.6	130
5.1	17.4	120
4.6	19.0	150
2.6	17.6	80
1.0	18.4	205
0.5	18.6	60
5.1	18.4	110
3.1	18.0	125
2.1	19.0	125
1.5	17.8	185
1.0	17.4	190
0.5	17.2	70
4.6	16.6	135
2.6	17.2	125
3.6	18.2	90
2.1	16.6	200
3.6	18.0	5
2.6	17.2	30
3.1	17.2	165
3.1	18.6	260
4.6	18.0	325
3.6	17.4	325
2.6	17.8	345

set. From Figures 5.10 and 5.11, it can be seen that there is an observation that is located far away from the rest of the data. This observation affect the Q-Q plot as the point deviate far away from the other points. This shows that a possible outlier is present in the data set. Hence, further investigation is needed to test whether the 1st observation is an outlier by applying the outlier detection method for circular-linear regression using L_n^1 statistic.

The root mean squared error (RMSE) for this data set is 2.55 and value of test statistic $L_{31}^1 = 8.14$. Given the values of $n = 31$ and the estimated $\sigma = 2.55$, we obtain the cut-off point using simulation. The value of the cut-off point for L_{31}^1 statistic at 5% significance

level is 3.30. Clearly, the value of the L_{31}^{1} statistic for the 1st observation is greater than the cut-off point. Hence, the observation is identified as an outlier. We apply again the procedure on the reduced data set by removing the 1st observation and no outlier is detected.

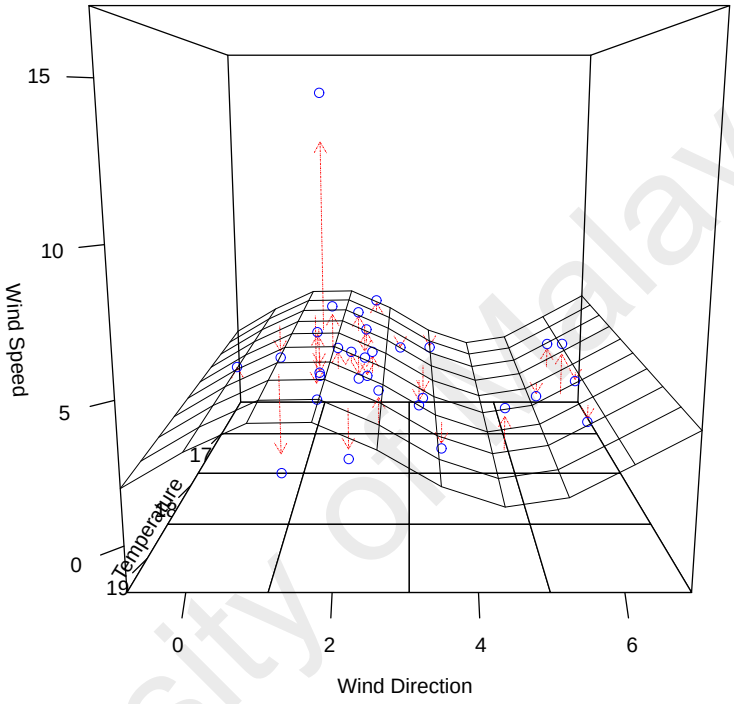


Figure 5.10: The regression plot of wind speed, temperature and wind direction

The removal of the 1st observation from the data set notably changes the value of $\hat{\beta}_0, \hat{\beta}_2, \hat{\gamma}, \hat{\delta}$ and σ. The results are shown in Table 5.7. Before the outlier is removed, it significantly affect the regression plane as shown in Figure 5.10. Thus, the removal of the 1st observation resulting in a better model fitting to the data set since the estimation is more accurate as shown in Figure 5.12. From Table 5.7, it can be seen that the standard error of the parameter estimates is reducing when the 1st observation is removed. Besides, the Q-Q plot of the residuals without the 1st observation from the wind data is shown in Figure 5.13. The points are now much closer to the straight line indicating a better fit for the data.

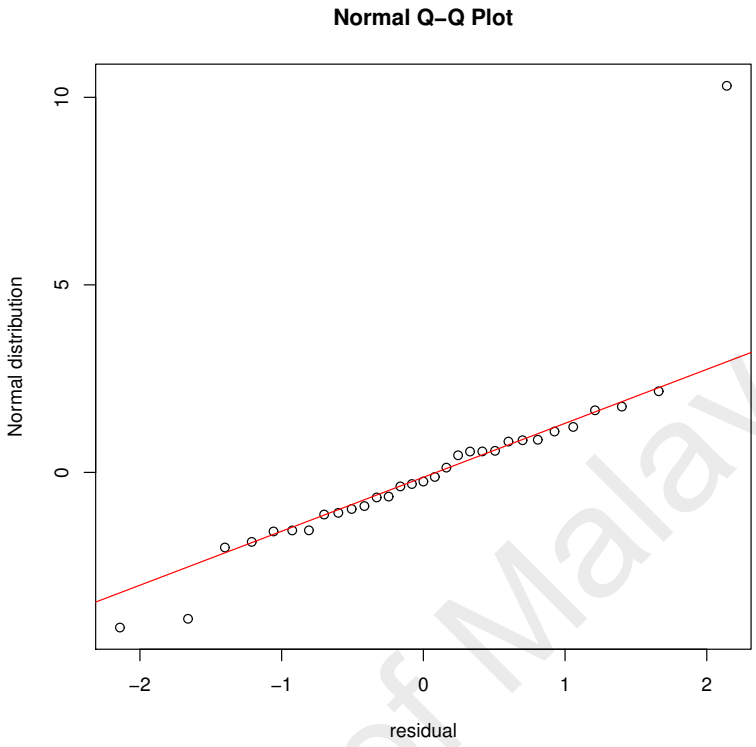


Figure 5.11: Q-Q normal plot of the residuals

Table 5.7: The summary of the effect of outlier removal from the wind data set.

Parameters	Full data		Data after excluding the 1st observation	
	Estimate	Standard Error	Estimate	Standard Error
$\hat{\beta}_0$	5.26	13.76	-0.14	8.19
$\hat{\beta}_2$	-0.12	0.77	0.17	0.46
$\hat{\gamma}$	0.31	0.70	-0.03	0.42
$\hat{\delta}$	1.49	0.76	0.81	0.46
σ	2.55	-	1.51	-

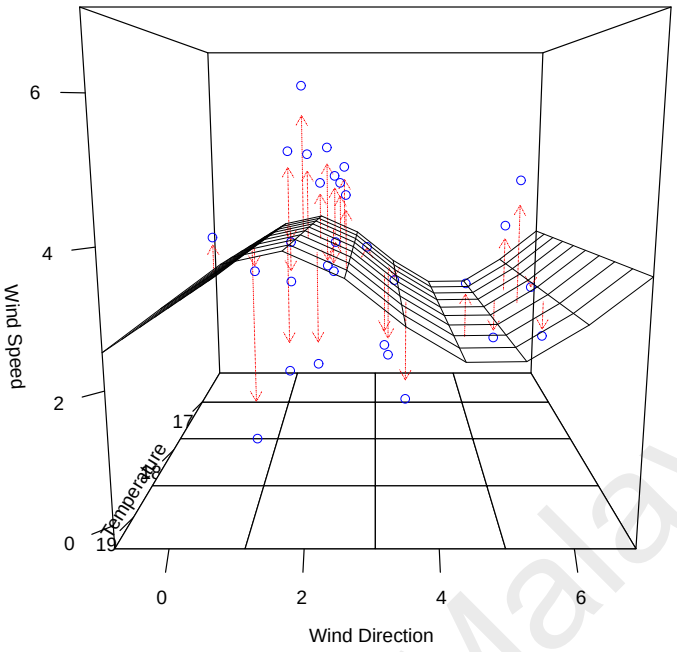


Figure 5.12: The regression plot of wind speed, temperature and wind direction after removing the 1st observation

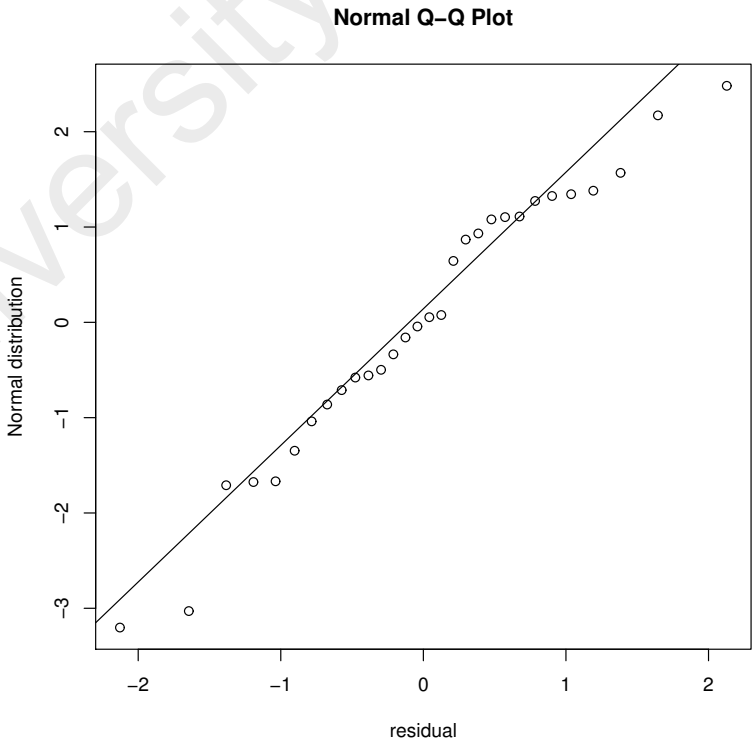


Figure 5.13: Q-Q plot of the residuals without the 1st observation

5.8 Summary

In this chapter, we have discussed the theory of JW circular-linear regression model and proposed a new discordancy test using this regression model based on the theory of k-nearest neighbor. The cut-off points of the test is obtained and the performance is examined via simulation study. The outlier detection procedure is developed to identify outliers in JW circular-linear regression model. However, the proposed procedure should work for other circular-linear regression models with the corresponding cut-off points obtained from simulation.

CHAPTER 6: CONCLUSION

6.1 Summary of the Study

The study looks at the outlier detection in cylindrical data and their regression. In this study, we focus on the JW cylindrical model and also JW circular-linear regression model.

We focus on the outlier detection in cylindrical data where the sample are generated from the JW cylindrical model. At present, there is no outlier detection method have been developed for the cylindrical data. Thus, we propose a new test called C_n^k statistic based on the k-nearest neighbor method. The outlier in cylindrical data can be divided into three categories; (i) outlier in the circular part; (ii) outlier in the linear part; (iii) outlier in the linear-circular part. The new test were conducted through simulation for the case of single outlier. Through simulation study, the C_n^1 statistic shows a good performance for a single outlier in the case of outlier in the linear part and in the linear-circular part. However, the performance of the new test statistic in the case of circular variable does not meet the desirable result like the other two cases. This is because outlier in circular variable almost impossible to occur due to the nature of the JW cylindrical model where the observations are spread along the circumference of a circle. The method have been applied on the wind data set and are able to detect observation that is located further away from the rest of the data. The proposed method should work for other cylindrical distribution with its corresponding generated cut-off points.

We then look at the outlier problem in the regression for the cylindrical data. We propose a new discordancy test to detect outliers in the regression model using JW circular-linear regression model called L_n^k statistic. This new test statistic also use the k-nearest neighbor method but on the residuals. The focus of the study is to detect single and two outliers where the performance shows that the test statistic can be used to detect one and two

outliers. From the study, we conclude that the new test shows a good performance.

In conclusion, we look into an outlier problems using JW cylindrical distribution and JW circular-linear regression model proposed by Johnson and Wehrly (1978). The work is very significant in providing information on the outliers in the cylindrical data.

6.2 Contributions

The study has contributed to cylindrical data analysis in the following ways:

1. Using the k-nearest neighbor theory, the C_n^k statistic is used to detect outlier in the cylindrical data. The cut-off points are generated via simulation. Through simulation, the test statistic performs well in detecting outlier in the linear component and linear-circular component.
2. Using the k-nearest neighbor theory on the residual, the L_n^k statistic is used to detect outlier in the regression model for cylindrical data. A full table of cut-off points for the statistic is generated through simulation. This statistic also proven to perform well in identifying outlier in the regression model.
3. We apply the outlier detection methods in cylindrical data and circular-linear regression using the wind data sets. The test statistics are able to detect outliers in the corresponding data sets.

6.3 Further Research

There are various possibilities for further research in this area. Some suggestion are given as follows:

1. To extend the C_n^k statistic to detect multiple or a patch of outliers in the cylindrical data.
2. To extend the work for MLE for circuar-linear regression model with outlier occurs in the circular variable.

3. To develop a better statistic to detect outlier in circular component of cylindrical data.

We are aware that there are still many problems ready to be explored in the directional statistics especially in the cylindrical case.

REFERENCES

Abe, T., & Ley, C. (2016). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, 1-14.

Abuzaid, A., Hussin, A., & Mohamed, I. (2013). Detection of outliers in simple circular regression models using mean circular error statistic. *Journal of Statistical Computation and Simulation*, *83*(2), 269-277.

Abuzaid, A., Mohamed, I., & Hussin, A. (2009). A new test of discordancy in circular data. *Communication is Statistics - Simulation and Computation*, *38*(4), 682-691.

Abuzaid, A., Mohamed, I., Hussin, A. G., & Rambli, A. (2011). Covratio statistic for simple circular regression model. *Chiang Mai Journal of Science*, *38*(3), 321-330.

Anderson-Cook, C. (1997). An extension to modeling cylindrical variables. *Statistics and Probability Letters*, *35*(3), 215-223.

Anderson-Cook, C. (2000). An industrial example using one-way analysis of circular-linear data. *Computational Statistics and Data Analysis*, *33*(1), 45-57.

Anderson-Cook, C. (2001). An alternate model for cylindrical data. *Nonlinear Analysis*, *47*(3), 2011-2022.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (2nd ed.). John Wiley and Sons, New York.

Barrett, B. E., & Ling, R. F. (1992). General classes of influence measures for multivariate regression. *Journal of the American Statistical Association*, *87*(417), 184-191.

Best, D., & Fisher, N. (1981). The bias of the maximum likelihood estimators of the von mises-fisher concentration parameters. *Communication in Statistics - Simulations and Computations*, *10*(5), 394-502.

Carta, J., Ramirez, P., & Bueno, C. (2008). A joint probability density function of wind speed and direction for wind energy analysis. *Energy Conversion and Management*, *49*, 1309-1320.

Collett, D. (1980). Outliers in circular data. *Journal of the Royal Statistical Society*, *29*, 50-57.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Econometrics*, *19*, 15-18.

David, H. A. (1981). *Order statistic*. Wiley, New York.

Downs, T. D., & Mardia, K. V. (2002). Circular regression. *Biometrika*, *89*(3), 683-697.

Fernandez-Duran, J. (2007). Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics*, *63*(2), 579-585.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press, London.

Fisher, N. I., & Lee, A. J. (1992). Regression models for an angular response. *Biometrics*, *48*, 665-677.

Fukunage, K., & Narendra, P. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Computer Society*, *24*(7), 750-753.

George, B., & Ghosh, K. (2006). A semiparametric bayesian model for circular-linear regression. *Communications in Statistics - Simulation and Computation*, *35*, 911-923.

Gould, A. L. (1969). A regression technique for angular variates. *Biometrics*, *25*(4), 683-700.

Hadi, A. S., & Simonoff, J. S. (1993). Detection of influential observations in linear regression procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*, 1264-1272.

Hand, D. (1981). *Discrimination and classification*. John Wiley and Sons.

Henningsen, A., & Toomet, O. (2011). maxlik: A package for maximum likelihood

estimation in R. *Computational Statistics*, *26*(3), 443-458. Retrieved from http://www.maxlik.org/ doi: 10.1007/s00180-010-0217-1

Hussin, A. G. (2004). Linear regression model for circular variables with application to directional data. *Journal of Applied Science and Technology*, *9*(1), 1-6.

Ibrahim, S., Rambli, A., Hussin, A., & Mohamed, I. (2013). Outlier detection in a circular regression model using covratio statistic. *Communications in Statistics - SImulation and Computation*, *42*(10), 2272-2280.

Iglewicz, B., & Hoaglin, D. (1993). *How to detect and handle outliers*. ASQC Quality Press.

Jammalamada, S. R., & Sarma, Y. R. (1993). Circular regression. *Statistical Theory and Data Analysis*, 109-128.

Jammalamadaka, S., & SenGupta, A. (2001). *Topics in circular statistics*. World Scientific Press, Singapore.

Johnson, R., & Wehrly, T. (1978). Some angular-linear distribution and related regression models. *Journal of the American Statistical Association*, *73*(363), 602-606.

Kato, S., & Shimizu, K. (2008). Dependent models for observations which include angular ones. *Journal of Statistical Planning and Inference*, *138*(11), 3538 - 3549.

Kato, S., Shimizu, K., & Shich, G. S. (2008). A circular-circular regression model. *Statistica Sinica*, *18*(2), 633-645.

Keller, J., Gray, M., & Givens, J. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*, 580 - 585.

Laylock, P. J. (1975). Optimal design: Regression models for directions. *Biometrika*, *62*(2), 305-311.

Mardia, K. V. (1972). *Statistics of directional data* (Z. W. Birbaum & E. Lukacs, Eds.). Academic Press.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society. Series B*, *37*(3), 349-393.

Mardia, K. V., & Sutton, T. (1978). Model for cylindrical variables with applications. *Journal of the Royal Statistical Society. Series B*, *40*(2), 229-233.

Mohamed, I., Rambli, A., Khaliddin, N., & Ibrahim, A. (2016). A new discordancy test in circular data using spacings theory. *Communications in Statistics - Simulation and Computation*, *45*(8), 2904-2916.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308-313.

Perez, I., Garcia, M., Sanchez, M., & de Torre, B. (2005). Analysis and parameterisation and wind profiles in the low atmosphere. *Solar Energy*, *78*(6), 809-821.

Perez, I., Garcia, M., Sanchez, M. L., & de Torre, B. (2007). Analysis of directional meteorological data by means of cylindrical models. *Renewable Energy*, *32*(3), 459-473.

Qin, X., Jiang, S., & Xiao, D. (2011). A nonparametric circular-linear multivariate regression model with a rule-of-thumb bandwidth selector. *Computers and Mathematics with Applicationss*, *62*(8), 3048-3055.

Rambli, A. (2015). *A half-circular distribution and outlier detection procedures in directional data* (Unpublished doctoral dissertation). University of Malaya.

Rambli, A., Ali, A., Mohamed, I., & Hussin, A. G. (2016). Procedure for detecting outliers in circular regression model. *PLoS ONE*, *11*(4).

Rao, J. (1969). *Some contributions to the analysis of circular data* (Unpublished doctoral dissertation). Indian Statistical Institute, Calcutta, India.

Rivest, L. P. (1997). A decentres predictor fot circular-circular regression. *Biometrika*, *84*(3), 717-726.

SenGupta, A., & Kim, S. S. (2016). Statistical inference for homologous gene pairs between two circular genomees: a new circular-circular regression model. *Statistical*

Methods and Applications, *25*(3), 421-432.

SenGupta, A., & Ugwuowo, F. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics*, *13*(3), 299-309.

Srikantan, K. S. (1961). Testing for the single outlier in a regression models. *Sankhya Series A*, *23*, 251-260.

Srivastava, M. S., & Rosen, D. (1998). Outliers in multivariate regression models. *Journal of Multivariate Analysis*, *65*(2), 195-208.

Tran, T., Wehrens, R., & Buydens, L. (2006). Knn-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics and Data Analysis*, *51*(2), 513-525.

Verma, S., Díaz-González, L., Pérez-Garza, J., & Rosales-Rivera, M. (2017). Erratum to: Quality control in geochemistry from a comparison of four central tendency and five dispersion estimators and example of a geochemical reference material. *Arabian Journal of Geosciences*, *10*(2), 24.

von Mises, R. (1918). Uber die "ganzzahligkeit" der atomicwichte und verwandte. *Fragen. Phys. Z.*, *19*, 490-500.

Weinberger, K., & Saul, L. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207-244.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Sadikon, N., Ibrahim, A. I. N., Mohamed, I. B., & Shimizu, K. (in press). A new test of discordancy in cylindrical data. *Communications in Statistics – Simulation and Computation*.

A NEW TEST OF DISCORDANCY IN CYLINDRICAL DATA

Nurul Hidayah Sadikon[1], Adriana I. N. Ibrahim[1], Ibrahim Mohamed[1], Kunio Shimizu[2]
[1]Institute of Mathematical Science, University of Malaya, Kuala Lumpur, Malaysia
[2]School of Statistical Thinking, The Institute of Statistical Mathematics, Tokyo, Japan
adrianaibrahim@um.edu.my

Key Words: Cylindrical data; k-nearest neighbor's distance; outlier.

ABSTRACT

Cylindrical data are bivariate data from the combination of circular and linear variables. However, up to now no work has been done on the detection of outlier in cylindrical data. We introduce a definition of outlier for cylindrical data and present a new test of discordancy to detect outlier in this type of data, based on the k-nearest neighbor's distance. Cut-off points of the new test statistic based on the Johnson-Wehrly distribution are calculated and its performance is examined using simulation. A practical example is presented using wind speed and wind direction data obtained from the Malaysian Meteorological Department.

1 INTRODUCTION

Cylindrical data are bivariate data where one component is measured on the circular scale while the other is linear. The circular component is a directional variable where it is bounded in $[0,2\pi)$. This type of data arises in many fields such as biology, meteorology, agriculture, geology and industry. The most common example in meteorology is the wind direction with wind speed, temperature, level of humidity, amount of rainfall, or level of pollutants (Pérez et al.,2007; Jammalamadaka & Sengupta, 2001). Another example is the time of birth within a cycle of 24 hours where the circular variable is the time in 24 hours while the linear variable is the number of births (Jammalamadaka & Sengupta, 2001).

In statistical analysis, the most common problem that arises is the existence of unexpected observations in data set which are usually called outliers. The existence of outliers can affect