CHAPTER 3

OUTLIER IN UNIVARIATE CIRCULAR DATA

3.1 Introduction

In this study, we review four statistical methods to detect possible outliers in the circular data. One of them is proposed by Abuzaid (2009), the *A* statistic, which has been shown to perform better than other methods for data from the *VM* distribution. In this chapter, we apply the methods to detect possible outliers by assuming that the data follow wrapped normal (*WN*) distributions. We obtain the cut-off points of the statistics and measure its power of performance through simulation studies.

3.2 Outlier Detection Methods in Univariate Circular Data

As in the linear case, the existence of outliers in circular data is expected to affect the estimation of parameters and weaken the accuracy of forecast. Thus, it is very important that methods of identifying outliers in circular data are developed for proper handling of the data. Graphical and numerical methods are the most common tools used in investigating the existence of outliers in circular data. We may use circular plots as given in Figures 3.1-3.2 to suggest the existence of outliers in a sample. Other graphical plots including circular histogram, rose diagram and P-P plot may also be used. In addition, Jammalamadaka and SenGupta (2001) defined circular distance between any two points as the smaller of two arc lengths between the points along the circumference which can also be used in detecting outliers. The circular distance between the mean direction $\bar{\theta}$ and each observation θ_i is defined as

$$d_i = \min\left(\theta_i - \bar{\theta}, 2\pi - \left(\theta_i - \bar{\theta}\right)\right) = \pi - \left|\pi - \left|\theta_i - \bar{\theta}\right|\right|$$

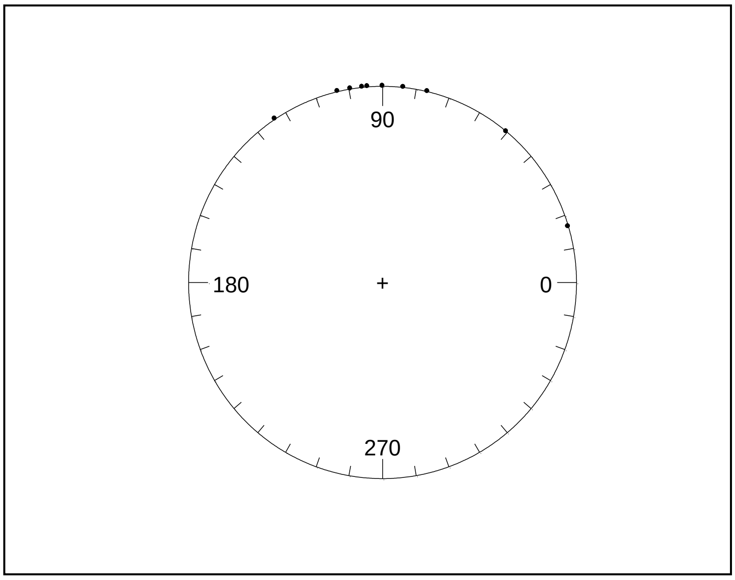


Figure 3.1: Circular plot of univariate data

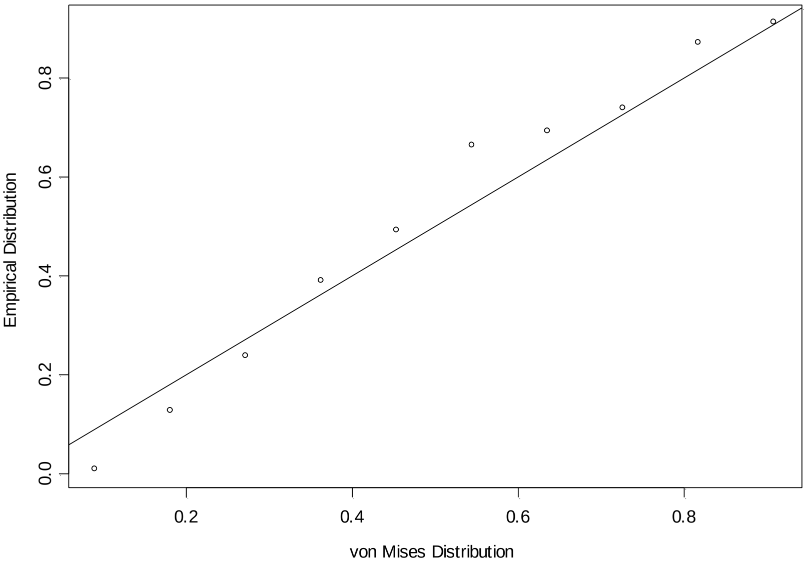


Figure 3.2: P-P plot of univariate data

Collett (1980) suggested that an observation with the maximum of d_i will be a candidate for outlier. He also pointed out that the identification of outliers in circular data highly depends on the concentration parameter. It is easier to identify an outlier in high concentrated circular samples than those with smaller concentration. He presented four different numerical tests of discordancy in circular data, namely *C*, *D*, *L* and improved *M* statistics.

Recently, Abuzaid (2009) proposed a new test of discordancy in circular data called the *A* statistic which is developed based on the summation of circular distances of the point of interest to all others points. In the following section, we explore the performance of all tests on samples from the *WN* distribution.

3.3 Numerical Tests

In addition to the graphical methods, there are four numerical tests available to identify outliers in univariate circular sample; *C*, *D*, *M*, and *A* statistics. Through simulation study, we obtain the cut-off points for each statistic when the data come from the *WN* distribution.

i. *C* Statistic

The mean resultant length of circular data set is given by $\bar{R} = \frac{R}{n}$. By omitting the *i*th observation, the mean resultant length is given by $\bar{R}_{(-i)} = \frac{R_{(-i)}}{n-1}$. Therefore,

$$C = \max_i \left\{ \frac{\bar{R}_{(-i)} - \bar{R}}{\bar{R}} \right\}, \tag{3.1}$$

can be considered as a test statistic. Values of C statistic will then be compared with the cut-off points for the corresponding sample size n and estimated concentration parameter κ. If C is larger than the cut-off point, we reject the null hypothesis so that the *i*th observation is identified as an outlier.

ii. *D* Statistic

The *D* statistic uses the relative arc length based on the ordered observation of a circular sample $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(n)}$. Let T_i be the arc length between consecutive observations given by $T_i = \theta_{(i+1)} - \theta_{(i)}$, $i = 1,2,\ldots,n$ and $T_n = 2\pi - \theta_{(n)} + \theta_{(1)}$. Define $D_i = \dfrac{T_i}{T_{i-1}}$, $i = 1,2,\ldots,n$ and $T_0 \equiv T_n$. Let $D_k = \dfrac{T_k}{T_{k-1}}$ corresponds to the greatest arc containing a single observation θ_k. Note that D_k is two tailed. Collet (1980) suggested working in terms of

$$D = \min(D_k, D_k^{-1}) \tag{3.2}$$

where $0 < D < 1$. The observation θ_k can be considered as an outlier if the value of *D* is larger than the cut-off point.

iii. *M* Statistic

Mardia (1975) suggested a statistic of discordancy which is given by $M' = \min_i \left\{ \dfrac{n - 1 - R_{(-i)}}{n - R} \right\}$. Later, Collett (1980) reformulated the M' statistic in terms of

$$M = 1 - M' = \max_i \left\{ \frac{R_{(-i)} - R + 1}{n - R} \right\} = \frac{R_q - R + 1}{n - R}, \tag{3.3}$$

where $R_q = \max_i \{R_{(-i)}\}$. He stated the asymptotic distribution of the *M* statistic for large values of κ as follows: As the value of κ increases, the von Mises distribution will be approximated by a standard normal distribution. On the other hand, the *M* statistic can be approximated by $\frac{n(b^*)^2}{n-1}$, where $b^* = \max_i \left\{ \frac{|x_i - \bar{x}|}{\sum_j (x_j - \bar{x})^2} \right\}$ is the test statistic used to identify discordancy in normal data. Percentage points for b^* are given in Pearson and Hartley (1966).

iv. *A* Statistic

Suppose $\theta_1, \theta_2, ..., \theta_n$ are (i.i.d) circular observations located on the circumference of a unit circle. Rao (1969) defined the circular distance between θ_i and θ_j as

$$d_{ij} = 1 - \cos(\theta_i - \theta_j)$$

where d_{ij} is a monotone increasing function of $(\theta_i - \theta_j)$ and $d_{ij} \in [0,2]$. The summation of all circular distances of the point of interest θ_j to all other points is given by

$$D_j = \sum_{i=1}^{n} \left(1 - \cos(\theta_i - \theta_j)\right), \ j = 1,2,\ldots,n$$

If the observation θ_j is an outlier, then the value of D_j will increase. Thus, the average circular distance given by $\frac{D_j}{n-1}$ can be used to identify possible outliers in the circular sample. Abuzaid (2009) proposed the *A* statistic as

$$A = \max_j \left\{ \frac{D_j}{2(n-1)} \right\}, \ j = 1,2,\ldots,n \qquad (3.4)$$

where $A \in [0,1]$ is a linear measure. The average circular distance is divided by 2 in order to standardize the values of statistic *A*. The proposed statistic is based on the relative decrease in the summation of circular distances by omitting the point of interest θ_j.

3.4 The cut-off point for the *C, M, D* and *A* Statistics for *WN* Distribution

In this section, we consider the use of the statistics to detect outlier in data generated from the wrapped normal distribution. In order to investigate the power of performance for the statistics, firstly we have to compute the cut-off points for each statistic. To compute the cut-off points, we design a simulation study to find the percentage points of the null distribution of no outliers in the circular data set. We consider eleven values of measure of concentration parameter in the range of 0.1 to 0.975 and different sample sizes from 5 to 150. The process is carried out 2000 times for each combination of n and ρ. We generate sample from $WN(\mu = 0, \rho)$. All the statistics in each generated random sample are calculated based on Eq. (3.1)-Eq. (3.4) respectively. We wish to estimate the percentage points of the statistics at the 10%, 5% and 1% upper percentiles when no outlier presents in the sample. Tables 3.1-3.4 show the cut-off points of the four statistics. The following results are observed for all four statistics:

1. As the measure of concentration parameter increases, the cut-off points decreases for the three levels of percentiles. This is expected as the circular data are more concentrated with larger ρ resulting in a smaller difference between two largest values of the statistics.
2. As the sample size increases, the value of the cut-off point decreases. Again, this should be true as the sample size increases, the distance between the circular observations in circular plot become smaller.

Table 3.1: Table of cut-off points for the *C* statistic

n	Level of percentile	ρ					
		0.1	0.2	0.4	0.6	0.8	0.9
5	10%	1.711	1.577	1.171	0.686	0.241	0.109
	5%	2.379	2.195	1.581	0.866	0.313	0.133
	1%	4.867	4.306	2.776	1.622	0.506	0.216
10	10%	1.114	0.990	0.526	0.303	0.136	0.063
	5%	1.542	1.354	0.666	0.339	0.165	0.075
	1%	3.668	2.664	1.343	0.465	0.211	0.104
20	10%	0.688	0.520	0.239	0.143	0.073	0.037
	5%	0.967	0.732	0.274	0.154	0.081	0.043
	1%	2.119	1.592	0.439	0.176	0.098	0.057
30	10%	0.551	0.373	0.156	0.097	0.052	0.026
	5%	0.739	0.496	0.177	0.102	0.058	0.030
	1%	1.398	1.132	0.233	0.111	0.070	0.037
40	10%	0.424	0.239	0.111	0.071	0.040	0.020
	5%	0.635	0.312	0.125	0.074	0.044	0.023
	1%	1.764	0.628	0.156	0.082	0.052	0.030
50	10%	0.365	0.200	0.087	0.057	0.034	0.017
	5%	0.528	0.257	0.095	0.059	0.037	0.019
	1%	1.398	0.535	0.111	0.064	0.043	0.023
60	10%	0.291	0.164	0.070	0.047	0.028	0.014
	5%	0.418	0.210	0.076	0.048	0.031	0.016
	1%	0.845	0.397	0.091	0.052	0.035	0.020
70	10%	0.286	0.134	0.060	0.040	0.024	0.013
	5%	0.409	0.162	0.065	0.041	0.026	0.014
	1%	0.954	0.309	0.073	0.043	0.030	0.018
80	10%	0.263	0.113	0.051	0.035	0.022	0.011
	5%	0.366	0.134	0.054	0.036	0.023	0.013
	1%	0.855	0.237	0.061	0.038	0.026	0.016
90	10%	0.222	0.100	0.045	0.031	0.019	0.010
	5%	0.321	0.122	0.048	0.032	0.021	0.012
	1%	0.694	0.201	0.054	0.034	0.023	0.014
100	10%	0.202	0.089	0.040	0.028	0.018	0.010
	5%	0.283	0.107	0.042	0.029	0.019	0.011
	1%	0.544	0.179	0.046	0.030	0.021	0.013
150	10%	0.130	0.056	0.026	0.019	0.012	0.007
	5%	0.177	0.065	0.027	0.019	0.013	0.007
	1%	0.378	0.093	0.030	0.019	0.014	0.009

Table 3.2: Table of cut-off points for the *M* statistic

n	Level of percentile	ρ					
		0.1	0.2	0.4	0.6	0.8	0.9
5	10%	0.762	0.785	0.808	0.843	0.865	0.866
	5%	0.838	0.851	0.878	0.896	0.914	0.913
	1%	0.934	0.932	0.954	0.967	0.973	0.975
10	10%	0.349	0.374	0.443	0.506	0.552	0.581
	5%	0.389	0.415	0.498	0.560	0.603	0.645
	1%	0.487	0.519	0.602	0.692	0.709	0.744
20	10%	0.154	0.169	0.217	0.269	0.323	0.347
	5%	0.167	0.182	0.234	0.297	0.352	0.385
	1%	0.194	0.212	0.277	0.345	0.420	0.460
30	10%	0.096	0.107	0.141	0.180	0.228	0.253
	5%	0.102	0.114	0.153	0.195	0.246	0.282
	1%	0.116	0.130	0.178	0.230	0.318	0.346
40	10%	0.069	0.078	0.102	0.136	0.180	0.201
	5%	0.073	0.082	0.108	0.146	0.201	0.224
	1%	0.080	0.092	0.122	0.171	0.240	0.269
50	10%	0.054	0.060	0.080	0.110	0.150	0.163
	5%	0.057	0.063	0.085	0.116	0.164	0.179
	1%	0.061	0.069	0.094	0.128	0.192	0.220
60	10%	0.044	0.049	0.066	0.092	0.126	0.141
	5%	0.046	0.051	0.070	0.097	0.138	0.155
	1%	0.048	0.056	0.076	0.107	0.163	0.182
70	10%	0.037	0.042	0.056	0.079	0.108	0.125
	5%	0.038	0.043	0.059	0.083	0.118	0.138
	1%	0.041	0.047	0.063	0.093	0.134	0.169
80	10%	0.032	0.036	0.048	0.069	0.097	0.113
	5%	0.033	0.038	0.050	0.073	0.103	0.124
	1%	0.036	0.041	0.054	0.079	0.122	0.146
90	10%	0.028	0.032	0.043	0.061	0.087	0.101
	5%	0.029	0.033	0.045	0.064	0.093	0.112
	1%	0.031	0.036	0.047	0.070	0.105	0.134
100	10%	0.025	0.029	0.038	0.055	0.080	0.093
	5%	0.026	0.029	0.040	0.057	0.085	0.103
	1%	0.028	0.031	0.043	0.062	0.096	0.124
150	10%	0.016	0.019	0.025	0.037	0.055	0.065
	5%	0.017	0.019	0.026	0.038	0.058	0.070
	1%	0.018	0.020	0.028	0.041	0.066	0.081

Table 3.3: Table of cut-off points for the *D* statistic

n	Level of percentile	ρ					
		0.1	0.2	0.4	0.6	0.8	0.9
5	10%	0.872	0.857	0.811	0.587	0.277	0.171
	5%	0.937	0.931	0.898	0.737	0.337	0.198
	1%	0.990	0.984	0.984	0.909	0.487	0.275
10	10%	0.857	0.872	0.814	0.611	0.268	0.150
	5%	0.924	0.933	0.911	0.762	0.323	0.185
	1%	0.980	0.986	0.980	0.930	0.504	0.246
20	10%	0.859	0.872	0.843	0.653	0.253	0.138
	5%	0.922	0.928	0.916	0.815	0.312	0.170
	1%	0.979	0.984	0.986	0.954	0.444	0.243
30	10%	0.872	0.875	0.814	0.695	0.259	0.132
	5%	0.937	0.926	0.902	0.847	0.335	0.165
	1%	0.988	0.985	0.981	0.966	0.494	0.226
40	10%	0.876	0.859	0.863	0.718	0.253	0.128
	5%	0.935	0.923	0.936	0.837	0.322	0.157
	1%	0.985	0.983	0.986	0.967	0.462	0.235
50	10%	0.873	0.864	0.849	0.750	0.279	0.129
	5%	0.938	0.939	0.923	0.857	0.358	0.159
	1%	0.981	0.988	0.986	0.965	0.582	0.214
60	10%	0.842	0.862	0.852	0.749	0.266	0.127
	5%	0.924	0.921	0.922	0.862	0.355	0.158
	1%	0.984	0.983	0.982	0.979	0.508	0.209
70	10%	0.873	0.869	0.861	0.758	0.252	0.125
	5%	0.931	0.939	0.934	0.876	0.334	0.152
	1%	0.989	0.985	0.985	0.965	0.555	0.218
80	10%	0.861	0.867	0.873	0.791	0.262	0.133
	5%	0.926	0.929	0.939	0.873	0.328	0.161
	1%	0.986	0.985	0.987	0.966	0.583	0.228
90	10%	0.857	0.857	0.845	0.791	0.259	0.130
	5%	0.933	0.930	0.923	0.893	0.324	0.156
	1%	0.989	0.979	0.991	0.968	0.517	0.225
100	10%	0.864	0.861	0.870	0.779	0.264	0.127
	5%	0.930	0.930	0.930	0.882	0.341	0.160
	1%	0.984	0.984	0.980	0.974	0.541	0.219
150	10%	0.861	0.869	0.863	0.819	0.269	0.118
	5%	0.922	0.928	0.932	0.905	0.353	0.142
	1%	0.981	0.986	0.985	0.982	0.550	0.200

Table 3.4: Table of cut-off points for the *A* statistic

n	Level of percentile	ρ					
		0.1	0.2	0.4	0.6	0.8	0.9
5	10%	0.924	0.928	0.920	0.892	0.726	0.549
	5%	0.944	0.948	0.946	0.927	0.783	0.592
	1%	0.971	0.978	0.979	0.962	0.879	0.705
10	10%	0.859	0.872	0.893	0.888	0.751	0.569
	5%	0.880	0.890	0.914	0.911	0.802	0.612
	1%	0.913	0.927	0.949	0.943	0.869	0.691
20	10%	0.811	0.832	0.872	0.886	0.774	0.602
	5%	0.831	0.849	0.888	0.903	0.808	0.640
	1%	0.861	0.875	0.912	0.929	0.875	0.725
30	10%	0.787	0.813	0.865	0.889	0.796	0.616
	5%	0.803	0.827	0.877	0.901	0.826	0.657
	1%	0.829	0.852	0.897	0.925	0.890	0.722
40	10%	0.771	0.802	0.856	0.890	0.809	0.628
	5%	0.784	0.816	0.867	0.902	0.837	0.665
	1%	0.808	0.839	0.885	0.918	0.895	0.742
50	10%	0.760	0.790	0.851	0.889	0.826	0.638
	5%	0.774	0.802	0.860	0.898	0.860	0.671
	1%	0.793	0.824	0.878	0.914	0.909	0.729
60	10%	0.750	0.784	0.848	0.892	0.829	0.644
	5%	0.761	0.794	0.857	0.901	0.857	0.678
	1%	0.779	0.813	0.872	0.912	0.909	0.751
70	10%	0.743	0.778	0.845	0.891	0.828	0.654
	5%	0.754	0.787	0.855	0.899	0.858	0.689
	1%	0.774	0.809	0.867	0.914	0.906	0.756
80	10%	0.740	0.775	0.843	0.892	0.837	0.661
	5%	0.750	0.786	0.850	0.899	0.863	0.695
	1%	0.770	0.804	0.862	0.912	0.918	0.760
90	10%	0.735	0.771	0.841	0.892	0.838	0.667
	5%	0.744	0.781	0.848	0.899	0.865	0.700
	1%	0.767	0.802	0.861	0.910	0.906	0.759
100	10%	0.734	0.770	0.840	0.890	0.846	0.674
	5%	0.744	0.779	0.846	0.897	0.871	0.706
	1%	0.761	0.794	0.862	0.908	0.913	0.765
150	10%	0.720	0.760	0.833	0.892	0.857	0.682
	5%	0.729	0.768	0.839	0.897	0.880	0.708
	1%	0.746	0.779	0.852	0.908	0.926	0.767

3.5 Performance of the Statistics for *WN* Distribution

Collett (1980) applied selected measures to test the performances of several statistics in outlier detection for circular samples. In this section, we use similar measures to compare the performance of the *A*, *C*, *D*, and *M* statistics. David (1970, p. 185), and Barnett and Lewis (1978, pp. 64–68), stated that a good test should have: (i) a high power function; (ii) a high probability of identifying a contaminating value as an outlier when it is in fact an extreme value, where an extreme value is defined as a point with the maximum circular deviation; and (iii) a low probability of wrongly identifying a good observation as discordant.

Let P1 = 1 − β be the power function where β is the Type-II error; P3 denotes the probability that the contaminant point is an extreme point and is identified as discordant; while P5 denotes the probability that the contaminant point is identified as discordant given that it is an extreme point. For a good test, we expect to have (i) high P1, (ii) high P5, and (iii) low P1 − P3.

To study the performance of all numerical tests, we use 2000 simulations based on different sizes of *n* and ρ. The samples are generated in such a way that $(n-1)$of the observations come from the $WN(\alpha,\rho)$ and one observation from $WN(\alpha+\lambda\pi,\rho)$, where λ is the degree of contamination and $0 \le \lambda \le 1$. The *C, D, M* and *A* statistics in each random sample are then calculated.

Figures 3.3-3.5 illustrate the power of performance of the statistics for different cases. The following results are observed:

1. In Figures 3.3(a) and (b), for *n* = 50 and ρ = 0.90, the *A* statistic performs better than the others for all contamination levels λ since the P1 and P5 curves are

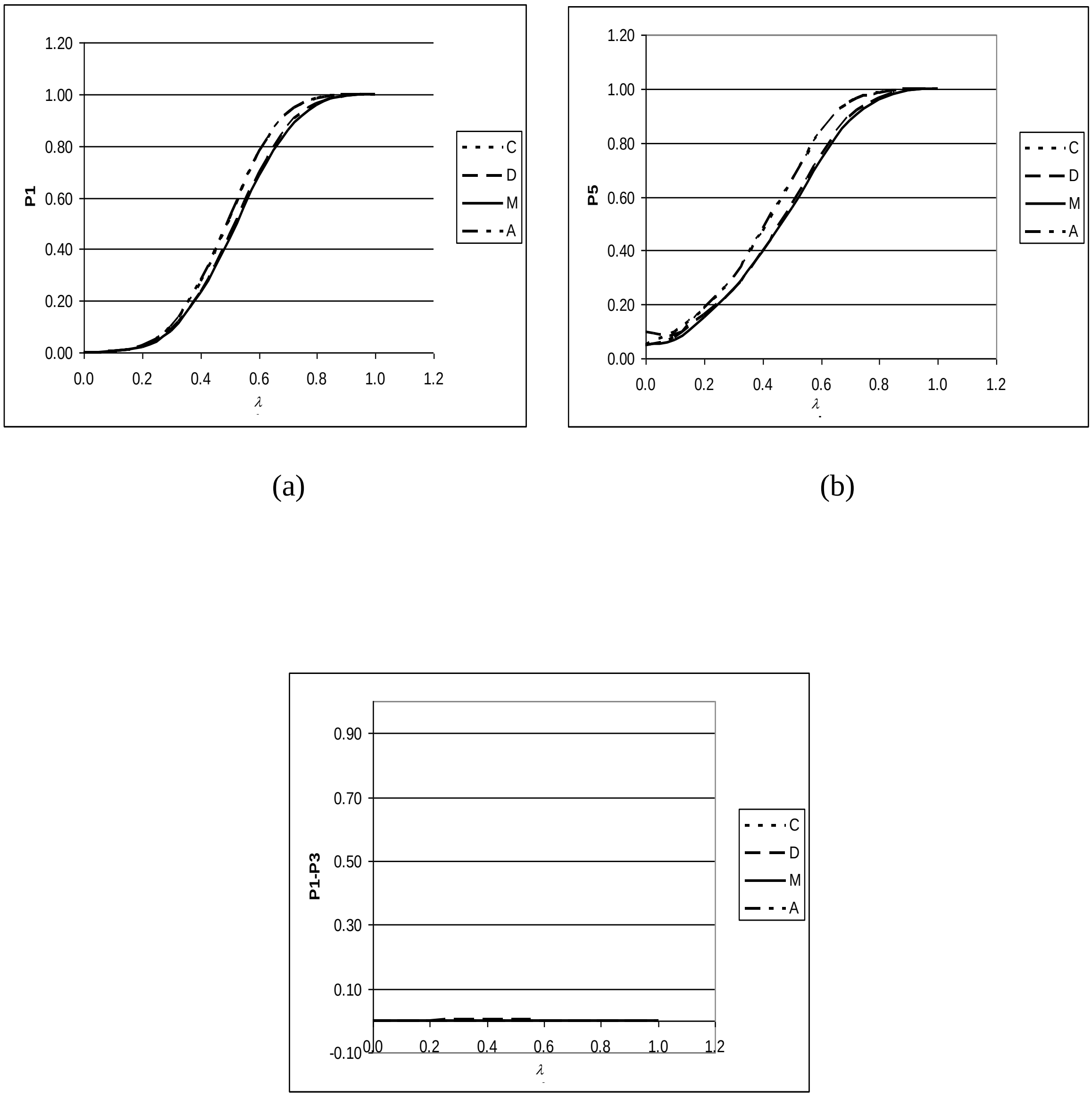
always greater than the others. As in Figure 3.3 (c), the four statistics have low P1-P3 since the curves are almost 0.

2. From Figure 3.3(a) and Figure 3.4(a), P1 reaches its maximum value when $\lambda = 0.8$ and $\lambda = 0.6$ respectively. Similar results are observed for P5 as shown in Figure 3.3(b) and Figure 3.4(b). This suggest that as ρ gets larger, the four statistics show better performance of detecting outliers at lower contamination level.
3. From Figure 3.3(a) and Figure 3.5(a), when *n* becomes smaller, the *A* statistic performs better than *C* and *D* statistics, but performs much better than the *M* statistic. Similar results are observed for P5 as shown in Figure 3.3(b) and Figure 3.5(b).

The four numerical methods have been investigated by Abuzaid (2009) when data follow the von Mises distribution. The cut-off points for *C* and *D* statistics can be obtained from Collet (1980) while *M* and *A* statistics are available in Mardia (1975) and Abuzaid (2009) respectively. Several results have been observed by Abuzaid (2009) on the superiority of *A* statistic over the other statistics, summarized as follows:

1. In case of small sample size and small measure of concentration parameter, the values of P1 are better for *M* statistic compared to others for all contamination λ levels. However, as *n* gets larger, the *A* statistic performs better. A similar trend is observed for P5 and P1-P3.
2. For large sample size and large measure of concentration parameter, the *A* statistic performs slightly better in terms of P1, P5 and P1-P3 compared to *C* and *D* statistics but they are much better than *M* statistic.

In conclusion, the result for the *VM* and *WN* are similar except for the case when *n* is small. In this particular case, the *A* statistic performs better than others in terms of P1 and P5 for the *WN* distribution but the *M* statistics performs better for the *VM* distribution.

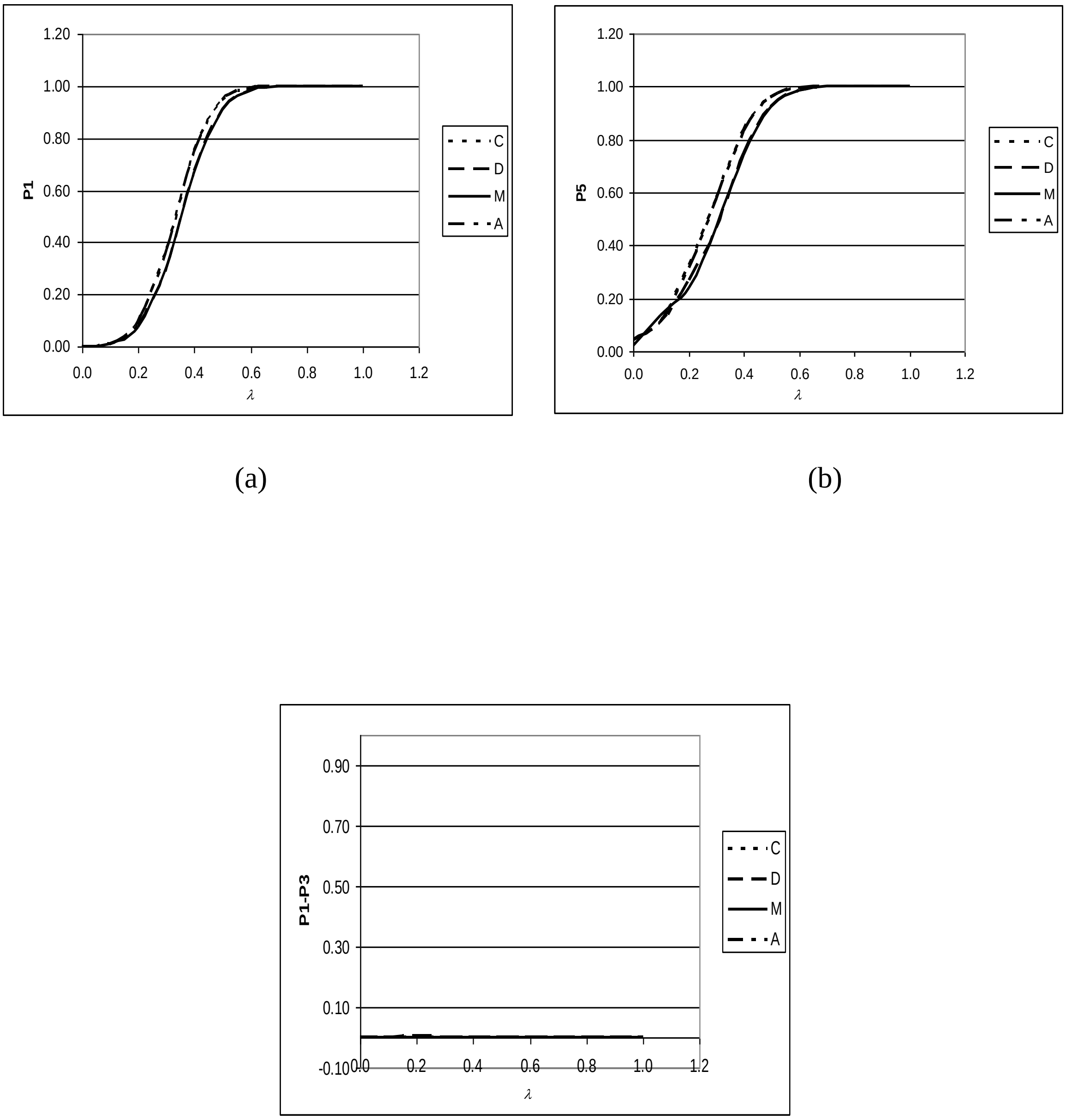


(a)

(b)

(c)

Figure 3.3: Performance of the statistics for n = 50 and ρ = 0.90

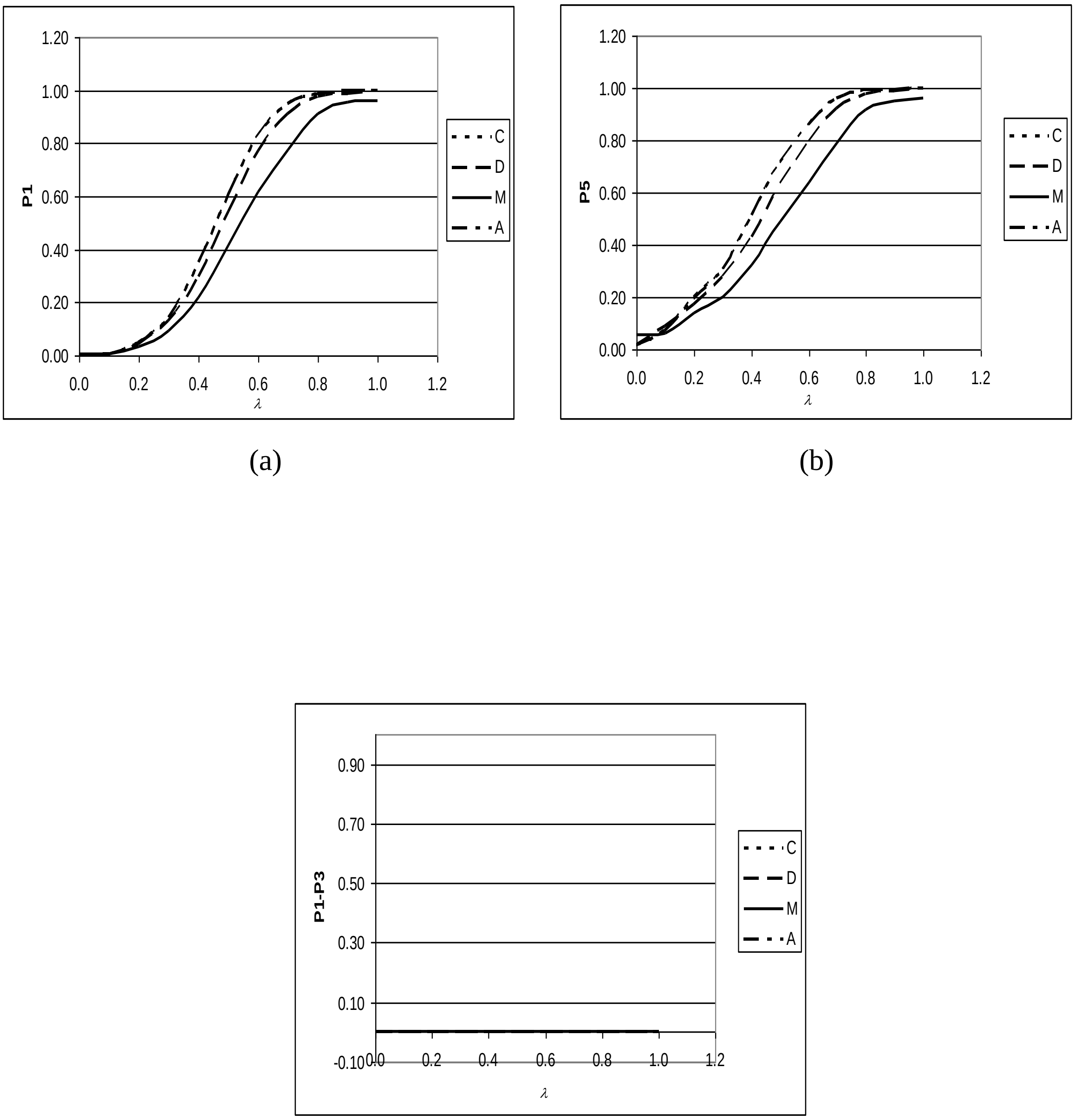


(a)

(b)

(c)

Figure 3.4: Performance of the statistics for n = 50 and ρ = 0.975



(a)

(b)

(c)

Figure 3.5: Performance of the statistics for *n* = 20 and ρ = 0.90

3.6 Practical Example

We refer to the Kuantan wind direction data described in Chapter 2. The circular plot is reproduced in Figure 3.6. From the circular plot, we noticed there is one observation located a bit separated from the rest of the observation. In this section, we apply all four discordancy tests on the data.

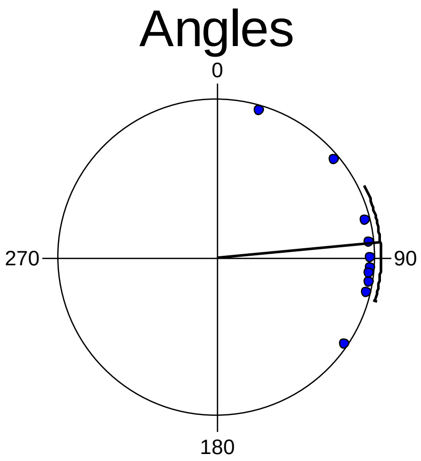


Figure 3.6: Circular plot of Kuantan wind data

Here, we have n = 10 and ρ = 0.88. Table 3.5 gives the value of the test statistics, the cut-off point for n = 10 and ρ = 0.9, as well as the decision for each statistic. It can be seen that only the M statistic successfully detects the outlying observation described earlier as an outlier. Also note that the values of A and C statistics are very close to their respective cut-off points. Thus, it warrants further investigation on the observation.

Table 3.5: Result based on C, M, D and A statistics

Test	Test value	Cut-off point	Decision
C	0.07	0.08	Not an outlier
M	0.59	0.19	outlier
D	0.13	0.64	Not an outlier
A	0.60	0.61	Not an outlier

3.7 Summary

To summarize, we reviewed four numerical tests for identifying the existence of outliers in circular data. The cut-off points for the *C*, *M*, *D* and *A* statistics for the wrapped normal distribution are obtained via simulation studies. We have compared the performance of the statistics for the *VM* and the *WN* distributions. As an illustration, we apply the statistics to identify the presence of outlier on Kuantan wind direction data.