Promotors
Prof. Dr. Ir. O.Thas & Prof. Dr. J.P. Ottoy
Department of Applied Mathematics, Biometrics and Process Control
Ghent University

Dean
Prof. Dr. ir. H. Van Langenhove

Rector
Prof. Dr. P. Van Cauwenberge

Heidi Wouters

Goodness-of-fit tests and graphical diagnostical tools for circular and linear data

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (Ph.D.) in Applied Biological Sciences

Dutch translation of the title:
Toetsen voor aanpassing en grafische diagnostische technieken voor circulaire en lineaire data

Please refer to this work as follows:
Heidi Wouters, 2007. Goodness-of-fit tests and graphical diagnostical tools for circular and linear data. Ph.D. thesis, Ghent University, Gent, Belgium.

ISBN-number: 987-90-5989-195-1

The author and the promotor give the authorisation to consult and to copy parts of this work for personal use only. Every other use is subject to the copyright laws. Permission to reproduce any material contained in this work should be obtained from the author.

Dankwoord

Het werk dat gepresenteerd is in deze thesis, is het resultaat van vele jaren onderzoek. Het is vaak een zaak geweest van zoeken naar kleine spelden in grote hooibergen, van door de bomen het bos niet meer zien, en dus ook van bijhorende teleurstellingen. Maar gelukkig werden de laatste jaren ook mooie successen geboekt waardoor het einde toch nabij leek. Ik ben nu ontzettend blij dat ik uiteindelijk bereikt heb waar ik al zo lang naar uitkeek. Ik wil dan ook iedereen bedanken die hier op een of andere manier heeft toe bijgedragen. Zonder jullie steun was dit boekje er nooit gekomen. Een aantal mensen wil ik echter in het bijzonder vermelden.

In eerste instantie wens ik mijn promotoren Prof. Olivier Thas en Prof. Jean-Pierre Ottoy te bedanken voor de kans die ze me gegeven hebben om onderzoek te verrichten. Ik dank hen ook voor het vertrouwen en de mogelijkheden die ze me hebben gegeven om naar het buitenland te gaan in het kader van mijn onderzoek. Olivier bedank ik in het bijzonder om me met veel enthousiasme aan te moedigen en zonder veel woorden telkens opnieuw te motiveren om door te gaan.

Thanks to all the members of the jury for careful reading and evaluating this work. In particular, I would like to thank Prof. Dr. John Rayner for making useful comments and suggestions.

Veel hulp en steun kreeg ik van mijn bureaugenoten Ellen, Lieven, Peter, Valerie, Lieve en Kristof. De boeiende gesprekken en de vele leuke momenten maakten het werk veel aangenamer. Verder ook dank aan de mensen van het secretariaat en alle andere collega's van biomath voor hun gezelschap en hulpvaardigheid.

Familie en vrienden dank ik voor de motiverende woordjes, steun, goeie raad, medeleven en begrip. Mijn vier broers en zus dank ik in het bijzonder voor het luisterend oor en de vele leuke en gezellige gezinsmomenten. Een speciaal woord van dank reserveer ik voor mijn ouders, die er overal en altijd voor mij zijn en me onvoorwaardelijk hebben gesteund en gemotiveerd.

Tenslotte wil ik mijn vriend Gert bedanken omdat hij me elke dag opnieuw zijn steun, begrip en liefde geeft!

Gent, September 2007
Heidi Wouters

Contents

Dankwoord **v**

1 Outline **1**

2 Introduction through examples **5**
2.1 Linear data 6
2.1.1 Lottery data 6
2.1.2 Lew data 9
2.1.3 Chemical Concentration data 10
2.1.4 Fastfood data 15
2.1.5 Old Faithful geyser data 17
2.2 Circular data 17
2.2.1 Birth time data 19
2.2.2 Homing pigeons 22
2.2.3 Turtles data 27
2.2.4 Ants data 27
2.2.5 Direzione data 30
2.2.6 Arrival data 31

3 Some goodness-of-fit tests **33**
3.1 GOF tests for linear and circular data 34
3.1.1 Statement of the one-sample GOF problem 34
3.1.2 Circular distributions 35
3.1.3 Additional issues for circular GOF testing 40
3.1.4 Additional issues for general GOF testing 42
3.2 Pearson's χ^2 GOF tests 45

3.2.1 An illustration of the original construction of Pearson's test through Mendel's data 45
3.2.2 Pearson's test for circular data 47
3.3 Smooth GOF tests 50
3.3.1 General construction 50
3.3.2 Simple null hypothesis 52
3.3.3 Composite null hypothesis 54
3.3.4 Data-driven smooth tests 56
3.3.5 Smooth test for circular uniformity 59
3.4 EDF GOF tests 61
3.4.1 General Construction 62
3.4.2 Supremum EDF tests 65
3.4.3 Integral tests 67
3.5 Link between Pearson, EDF and smooth tests 71
3.5.1 Pearson χ^2 versus smooth tests 71
3.5.2 EDF versus smooth tests 72
3.5.3 EDF versus Pearson χ^2 tests 74
3.6 Other GOF tests 74
3.7 Non-parametric density estimation 77
3.7.1 Kernel density estimator 78
3.7.2 Orthonormal series density estimator 79

4 Smooth goodness-of-fit tests on the circle 85
4.1 General Construction 86
4.1.1 Complex values on the circle 87
4.1.2 Construction of the smooth model 87
4.1.3 Construction of the smooth test 89
4.1.4 Origin-invariance 91
4.2 Simple null hypothesis 92
4.3 Composite null hypothesis of circular normality 93
4.3.1 MLE and MME 94
4.3.2 The orthonormal polynomials 95
4.3.3 The efficient score test 96
4.3.4 Relation to the BarCox test 100
4.4 Data-driven smooth tests 100
4.5 Nonparametric density estimation 101
4.6 Examples 103
4.6.1 Turtles Data 104
4.6.2 Ants Data 105
4.6.3 Direzione Data 105
4.6.4 Arrival Data 106

4.7 Simulation Study 108
4.8 Discussion 113

5 Localised Pearson χ^2 Test 115
5.1 Introduction 116
5.2 The linear SSPc test 118
    5.2.1 Construction of the test statistics 118
    5.2.2 Computational formulae 120
    5.2.3 Asymptotic theory 121
    5.2.4 Empirical levels 121
    5.2.5 Limiting distribution of SSP3 statistic under contiguous alternatives 121
5.3 Limiting behaviour of the SSP*c* test under contiguous alternatives 130
5.4 Data-driven Test 138
    5.4.1 Construction of the test statistic 138
    5.4.2 Asymptotic theory 143
5.5 Examples 144
    5.5.1 Lottery data 144
    5.5.2 Lew data 145
    5.5.3 Birth time data 146
5.6 Simulation study 147
5.7 SSPc test for composite null hypothesis 152
    5.7.1 Examples 155
5.8 The circular SSP*c* test 158
    5.8.1 Construction of the test statistics 158
    5.8.2 Asymptotic theory 159
    5.8.3 Computational formulae 160
    5.8.4 Data-driven Test 161
    5.8.5 Examples 162
    5.8.6 Simulation Study 164
    5.8.7 Circular SSP*c* test for composite null hypothesis 166
5.9 Discussion 166

6 The Interval-based PP-plot 173
6.1 Construction of the IBPP-plot 174
6.2 The formal IBPP-plot 177
6.3 Limiting behaviour of IBPP-plots under contiguous alternatives . 179
6.4 Examples 181
    6.4.1 Birth time data 187
    6.4.2 Homing pigeons data 188
    6.4.3 Turtles data 191

6.4.4 Ants data 191
6.4.5 Direzione data 193
6.4.6 Arrival data 193
6.4.7 Contaminated Lottery data 193
6.4.8 Old Faithful geyser data 196
6.5 Simulation Study 198
6.6 The IBPP-plot as a tool for comparing density estimates 207
6.6.1 Construction 208
6.6.2 Fastfood data 209
6.6.3 Arrival Data 210
6.7 Discussion 213

7 Conclusions and further research **215**

A Orthonormal polynomials **221**
A.1 Hermite polynomials 221
A.2 Legendre polynomials 222

B Proofs **223**
B.1 Proof of Theorem 5.1 223
B.2 Proof of Theorem 5.2 224
B.3 Proof of Lemma 5.1 225
B.4 Proof of Lemma 5.2 228
B.5 Proof of Theorem 5.7 235
B.6 Proof of Theorem 5.8 236

C Datasets **237**
C.1 Lottery data 237
C.2 Lew data 238
C.3 Chemical concentration data 238
C.4 Fastfood data 238
C.5 Old Faithful geyser data 238
C.6 Birth time data 239
C.7 Homing pigeons data 239
C.8 Turtles data 239
C.9 Ants data 239
C.10 Direzione data 240
C.11 Arrival data 240

Bibliography **241**

Summary **249**

Samenvatting **253**

Notations and abbreviations

AD	Anderson-Darling
AIC	Akaike's information criterion
BIC	Baysian in formation criterion
CDF	cumulative distribution function
CLT	central limit theorem
CiPP	Circular PP
CN or VM	circular normal or von Mises
CU	circular uniform
CvM	Cramér-von Mises
EDF	empirical distribution function
GOF	goodness-of-fit
IBPP	interval-based PP
i.i.d.	independentely and identically distributed
KS	Kolmogorov-Smirnov
LL	$\ln - \ln$
LOF	lack-of-fit
MLE	maximum likelihood estimate
MISE	mean integrated squared error
PDF	probability density function
PIT	probability integral transformation
PP	probability-probability
SSP	sample space partition
UCV	unbiased cross validation
$\wedge$	minimum operator
$\vee$	maximum operator

CHAPTER 1

Outline

This thesis is situated in the area of goodness-of-fit (GOF) testing for the one-sample problem. For this type of statistical problems we are interested in whether a sample of observations possibly comes from a given distribution or not. The null hypothesis of such a statistical problem is that the data follow that particular distribution, which is referred to as the hypothesised or null distribution. In this area a large number of statistical methods and tests have been developed since the introduction of the well known Pearson χ^2 statistic in 1900. Pearson's test, which is in principle only applicable to discrete data, often has been used for categorised continuous data problems as well. Nowadays, however, many other more competitive GOF methods have been developped particularly for continuous data problems, and the application of Pearson's test is only recommended in the discrete case. Nevertheless, Pearson's statistic is still often used as a basis to construct test statistics for continuous data. The reason is probably that the idea behind Pearson's statistic is very simple. The class of sample space partition (SSP) tests, which was originally developed by Thas (2001), forms an example hereof. In particular, this family of statistics is constructed by taking the average of the Pearson's statistics over all possible partitions of the sample space.

Test statistics of this class can be written in terms of the empirical distribution function (EDF). In such, they are related to a second, quite large class of GOF tests, the so-called EDF tests. That class includes e.g. the Anderson-

Darling (AD) test and the Kolmogorov-Smirnov (KS) test.

A third class consists of the smooth tests, which were introduced by Neyman (1937). The test statistics of this type are particularly useful since they can often be decomposed into meaningful components which give information on how the true distribution deviates from the hypothesised. This deviation is usually expressed in terms of moments, e.g. a difference in skewness which is related to the third moment. An issue with smooth tests is how to choose the number of components included in the test statistic, which is in turn related to the number of parameters in a smooth model. The choice of the order is particularly important as choosing an inappropriate order might result in a power loss. In order to overcome the problem of choosing the order, Ledwina 1994 proposed a data-driven version of the smooth test by making an optimal choice based on the data.

The above described statistics, Pearson's, EDF and smooth tests are considered as the three main classes of GOF statistics for linear data.

In this thesis we are particularly interested in GOF techniques for circular data. Circular data occur in many fields and are essentially observations measured on a circle. Two typical examples of circular data are wind directions and arrival times, for which the measuring instrument is the compass and the clock, respectively. A question that often arises with such data is whether the directions or time measurements are uniformly distributed over the circle. This is a GOF problem for which the hypothesised distribution is the circular uniform (CU) distribution. In the context of GOF problems on the circle, it is necessary to use appropriate techniques which take the special structure of the circle into account. In particular, the proposed methods should be invariant to the arbitrary chosen origin and rotation direction.

Some of the above GOF methods have been adapted for circular data. For example, the Kuiper test (1960) is the circular analogue to the KS test. The smooth test for linear uniformity is adapted for circular uniformity by Bogdan et al. (2002). Although the same three classes of GOF tests exist also for circular data, the spectrum of GOF tests for circular data is not as large as that of the GOF tests for linear data. Most GOF techniques on the line have been developed for simple null hypotheses as well as for composite null hypotheses. A simple null hypothesis refers to a completely specified null distribution, while composite null hypothesis refers to a null distribution for which some parameters are unknown, and are therefore to be estimated in practice. However, the extension of GOF tests to composite null hypotheses is not always straightforward. In particular, while Bogdan et al. (2002) proposed the smooth test for the simple null hypothesis of circular uniformity, a smooth test for composite hypothesis is harder and has not been developed yet to our knowledge.

In this thesis, we introduce a general framework for the construction of smooth tests on the circle. We use the complex-valued representation of circular data and generalise the theory of smooth tests for linear data by Rayner and Best (1989) to complex-valued circular data. This allows us to construct a smooth test for any circular composite distribution. We explicitly develop the theory in the case of testing for circular normality, which is the circular analogue to the normal distribution on the line. Similarly as in the linear case, we present a data-driven version of the circular smooth test. As a by-product of the data-driven smooth test, a non-parametric density estimator can be constructed, which immediately may give a visual impression of how the true distribution deviates from the hypothesised.

We also present new results on the class of localised Pearson χ^2 tests, which is closely related to the class of SSP tests, introduced by Thas (2001). The class of statistics is shown to have some interesting asymptotic properties. In particular, they are powerful against "local" alternatives, which are alternatives that deviate from the null distribution only in small intervals of the sample space. Our initial discussion will deal with the linear case, but we extend this class of tests to include circular localised Pearson χ^2 tests.

Apart from GOF tests, we know that explorative analysis of the distributional properties of the data is something that should not be omitted. Often such analyses reveal interesting features in the data. Most of the common explorative graphs are subjective when it comes to interpretation. Moreover, they depend on the choice of certain parameters. For example the construction of a histogram depends on the choice of the bin width. In this thesis we focus on visualisation techniques that are related to statistical tests. The probability-probability plot (PP-plot) is such an example since it is related to the KS test. In fact, the PP-plot is a visual representation of the information that is used by the KS test in making its decision between the null and the alternative hypothesis. This implies that conclusions can be drawn from this plot which are formal and objective. The PP-plot, however, is not useful for circular data since it is not origin-invariant. We therefore develop in this thesis a new graphical tool that is related to the circular analogue of the KS test, i.e. the Kuiper test. Two versions of the graph are presented and are referred to as the Circular PP-plot (CiPP) and the Interval-based PP-plot (IBPP). The graphical diagnostic tool is particularly useful for localising the region where the true distribution deviates from the hypothesised, which is also called the *location* of the lack-of-fit (LOF).

We also illustrate how an adapted version of the IBPP-plot can be used to assess how well the empirical distribution is fitted by the non-parametric density estimator associated with the data-driven smooth test. This density estimator is essentially an orthonormal series density estimator and could be compared based on the IBPP-plot to, for example, a kernel density estimator.

In **Chapter 2** we introduce the GOF problems for linear and circular data in an informal way through some real data examples. We also carry out a preliminary exploration of the data through some classical graphical techniques and descriptive summary statistics. The examples described in this chapter will be used in later chapters to illustrate the developed methods.

An overview of the three main classes of GOF tests for circular as well as for linear data is presented in **Chapter 3**. Since there exists a wide spectrum of GOF tests, it is not the intention of this thesis to be complete in this matter. The GOF techniques that are included in the overview are relevant to our proposed methods.

In **Chapter 4** we present the general framework for the construction of smooth tests on the circle. In this framework a smooth test for any circular distribution can be developed. It is shown that for the special case of circular uniformity, the smooth test of Bogdan et al. (2002) is obtained. Considerable attention will be paid in this chapter to the smooth test for composite circular normality.

The characteristics and extensions to the Localised Pearson χ^2 tests are in **Chapter 5** and the new graphical diagnostic tool to detect the location of the LOF is presented in **Chapter 6**.

The applicability of the proposed methods is widely demonstrated by the various examples throughout this thesis.

In **Chapter 7** we summarise the conclusions of the thesis, and we provide some topics for further research.

Finally, we want to stress that the use of the term “local” throughout this thesis, may not be confused with the use of this term in the context of “local alternatives”. We will therefore always use the equivalent term “contiguous alternatives” instead. With “local” we always mean a small subset of the sample.

CHAPTER 2

Introduction through examples

In this chapter we introduce in an informal way the *one-sample goodness-of-fit (GOF) problem* for *linear* and *circular* data. In particular, we present the real datasets that will be used in the next chapters as examples for demonstrating the proposed methods. Since we will propose new GOF techniques for linear as well as for circular data, both types of data will be considered, respectively in Section 2.1 and Section 2.2. Since we only discuss methods to analyse univariate data, we sometimes use the terms data and the realisations of a random variable (rv) interchangeably. If confusion is possible, we will be more explicit. To provide some insight in the data, we add an exploratory analysis for each example, using summary statistics and classical graphical tools. The descriptive summary statistics include appropriate estimates for location, scale, skewness and kurtosis. As graphical tools we consider visualisation techniques for the raw data, such as a histogram or a kernel density estimate. Additionally, we consider plots that provide more information concerning the fit of a particular distribution. The PP-plot is an example of such a tool and has the advantage that it is related to a formal statistical test (the KS test). However, these plots are usually not appropriate for circular data since they are not origin-invariant. In Chapter 6 we propose a new graphical tool to localise the *lack-of-fit* (LOF) that is appropriate for both linear and circular data.

2.1 Linear data

Linear data, which is the most common type of data, can be measured on the real line or on some interval of the real line. The examples in this section introduce the *linear one-sample GOF problem.* All corresponding datasets can be found in Appendix C.

2.1.1 Lottery data

The following dataset is a reference dataset provided by the National Institute of Standards & Technology (NIST) and available from the StRD webpage (`http://www.itl.nist.gov/div898/strd/index.html`). It consists of 218 3-digit numbers (from 000 to 999) resulting from the state of Maryland's Pick-3 Lottery. The data were collected for the 32-week period from September 3, 1989 until April 14, 1990. One 3-digit random number was drawn per day, 7 days per week for most weeks, and 6 or 5 days per week for other weeks. An interesting data-analytic question involving this dataset is whether the lottery numbers are uniformly distributed. The answer to that question can be given by GOF tests. The testing problem described here is called the *simple one-sample GOF problem.* Here, *simple* refers to the hypothesised distribution which is completely specified. Indeed, the probability mass function of the uniform distribution over numbers from 0 to 999 is completely determined and does not require specification of any parameters. A general formulation of the simple one-sample null hypothesis is given by

H_0: the underlying distribution of the data is F_0, where F_0 is completely specified.

Classical GOF techniques to test this null hypothesis are described in Chapter 3. Note that the Lottery data in fact concerns observations from a *discrete* random variable. However, the number of possibilities within the interval [0,999] is assumed to be large enough to approach continuity, such that continuous GOF methods are justified.

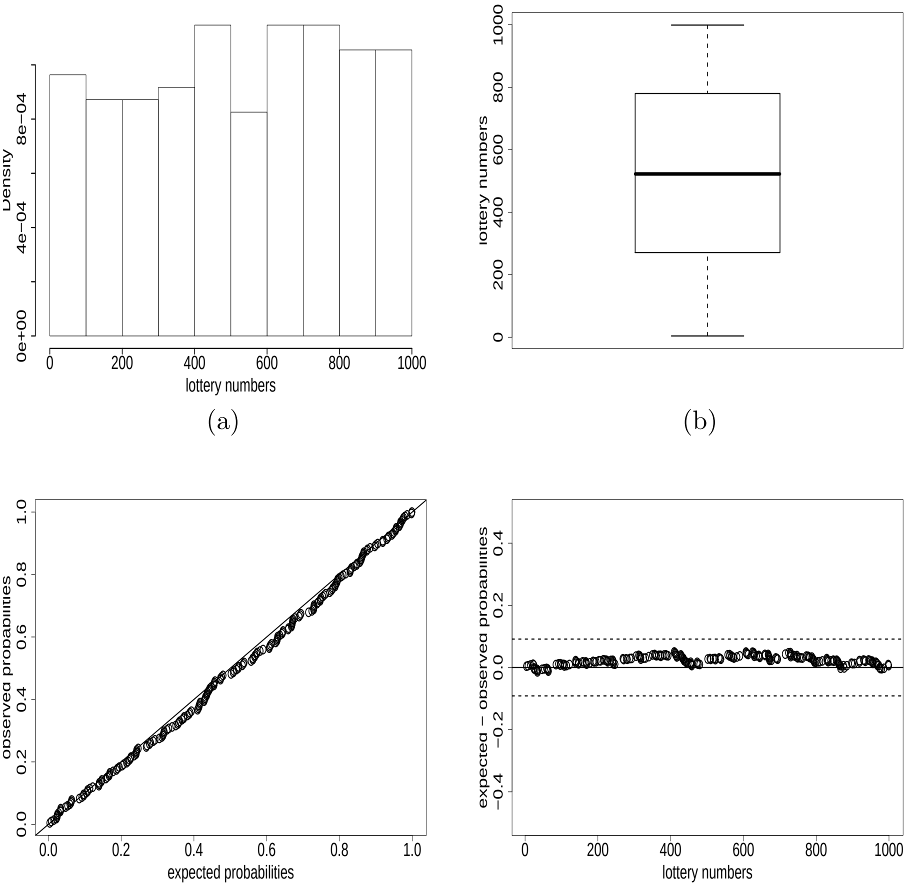
An appropriate GOF test for continuous uniformity is for example the Kolmogorov - Smirnov (KS) test, which is based on the supremum of the deviations between the observed and expected probabilities for uniformity (see Section 3.4.2). For this dataset, the KS test statistic has value $D_n = 0.048\sqrt{n}$ ($n = 218$) which results in a p-value of 0.689, indicating a non-significant result. For the application of other GOF tests to the Lottery data, we refer to Chapters 5 and 6.

Besides using *formal* statistical GOF tests, we stress that it is also important to use *explorative* tools in order to recognise features and get more insight in the data. We use the term *explorative* to refer to the plots and the descriptive summary statistics, since their interpretation is subjective. On the other hand, the term *formal* refers to GOF techniques from which an objective conclusion can be drawn. Statistical hypothesis tests form an example of these.

Figure 2.1 shows explorative plots for the Lottery data. The histogram in panel (a) indicates that the data is probably uniformly distributed up to some small deviations. The box plot, in panel (b), shows that the three quartiles divide the data in four roughly equal parts and no outliers are present. The PP-plot in panel (c) plots the observed probabilities versus the expected probabilities in case the data follows the uniform distribution. The points are scattered at random around the solid line. Hence, there is no evidence to reject the null hypothesis.

The absolute distance between the points and the line in the PP-plot is in fact the basis for the KS test. In particular, the supremum of these absolute deviations results in the KS statistic. Alternatively, the distances between the observed and the expected probabilities can be plotted versus the observations in a *detrended* PP-plot. If the KS test is significant at the α-level, then the detrended PP-plot may show where the LOF is located by indicating the values that exceed the α-level critical value of the KS test. Panel (d) of Figure 2.1 shows the detrended PP-plot for the lottery data. The two horizontal dotted lines indicate $-u_\alpha/\sqrt{n}$ and $u_\alpha/\sqrt{n}$, where $u_\alpha = 0.091\sqrt{n}$ is the 5% critical value of the KS test. The region outside the two dotted lines is the rejection region of the KS test. If the KS test is significant, at least one point lies in that rejection region. The coordinates on the horizontal axis of the points in the rejection region then indicate the location of the LOF. For the Lottery data, no significant result was found and hence no location of LOF is detected. We may conclude that the selection of the numbers occurs completely at random (uniformly distributed). In Section 5.5, we will apply other GOF techniques to these data. Additionally, in the same section, we will create a corrupted sample from the lottery data. In particular, the numbers between 800 and 875 are changed. For this new corrupted data, we demonstrate that the localised Pearson χ^2 tests are useful to detect such a deviation from uniformity.

The relation between the (detrended) PP-plot and the KS test is particularly interesting. The reason is that usually the drawback of explorative tools is that they are subjective. However, if the plot is related to a formal test, the conclusions of that test may be derived from the plot. Therefore, such a tool combines the advantages of a visualization technique with the objectivity of a test. A graphical tool which corresponds in a similar way to the Kuiper test for circular distributions will be proposed in Chapter 6.



(a) (b)

(c) (d)

Figure 2.1: Histogram (a), box plot (b), PP-plot (c) and detrended PP-plot (d) for the Lottery data. The region between (outside) the dotted lines in the detrended PP-plots, shows the acception (rejection) region of the KS test at the 5% significance level.

Table 2.1: Descriptive summary statistics for the Lottery data together with the population characteristic for uniformity.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
observed	4.0	272.8	522.5	779.2	999.0
expected	0.0	250.0	500.0	750.0	999.0
	Mean	Variance	Skewness	Kurtosis	
observed	519.0	85088.73	-0.093	-1.193	
expected	500.0	83333.33	0.000	-1.200	

Finally, we give some descriptive summary statistics of the Lottery data in Table 2.1. The table also includes the corresponding population characteristics for the uniform distribution on [0,999]. The sample skewness and sample kurtosis in Table 2.1 are calculated from the data. Let m_2, m_3 and m_4 be the second, third and fourth central sample moments, i.e.

$$m_j = \frac{1}{n}\sum_{j=1}^{n}(X_i - \overline{X})^j \quad j = 2, 3, 4. \tag{2.1}$$

Then, sample skewness and the sample kurtosis are computed by

$$g_1 = \frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{m_2^{3/2}} \text{ and } g_2 = \frac{n^2((n+1)m_4 - 3(n-1)m_2^2)}{(n-1)(n-2)(n-3)}\frac{(n-1)^2}{n^2m_2^2}, \tag{2.2}$$

respectively. These are consistent estimators for the population skewness and the population kurtosis. For the lottery data, it is seen that the sample characteristics are fairly close to the population characteristics of the uniform distribution.

2.1.2 Lew data

The Lew dataset is also taken from the NIST StRD webpage and is the result of a study by H. S. Lew of the Structures Division of the Center for Building Technology at the NIST. The purpose of the study was to characterise the physical behavior of steel-concrete beams under periodic load. The variable that was measured is deflection from a rest point of the steel-concrete beam. The 200 observations are equispaced in time. The researchers want to know whether this beam deflection point is uniformly distributed over the interval [-580,301], which is assumed to be the sample space. This is again a simple one-sample GOF problem. The KS GOF test for uniformity gives a value $D_n = 0.113\sqrt{n}$

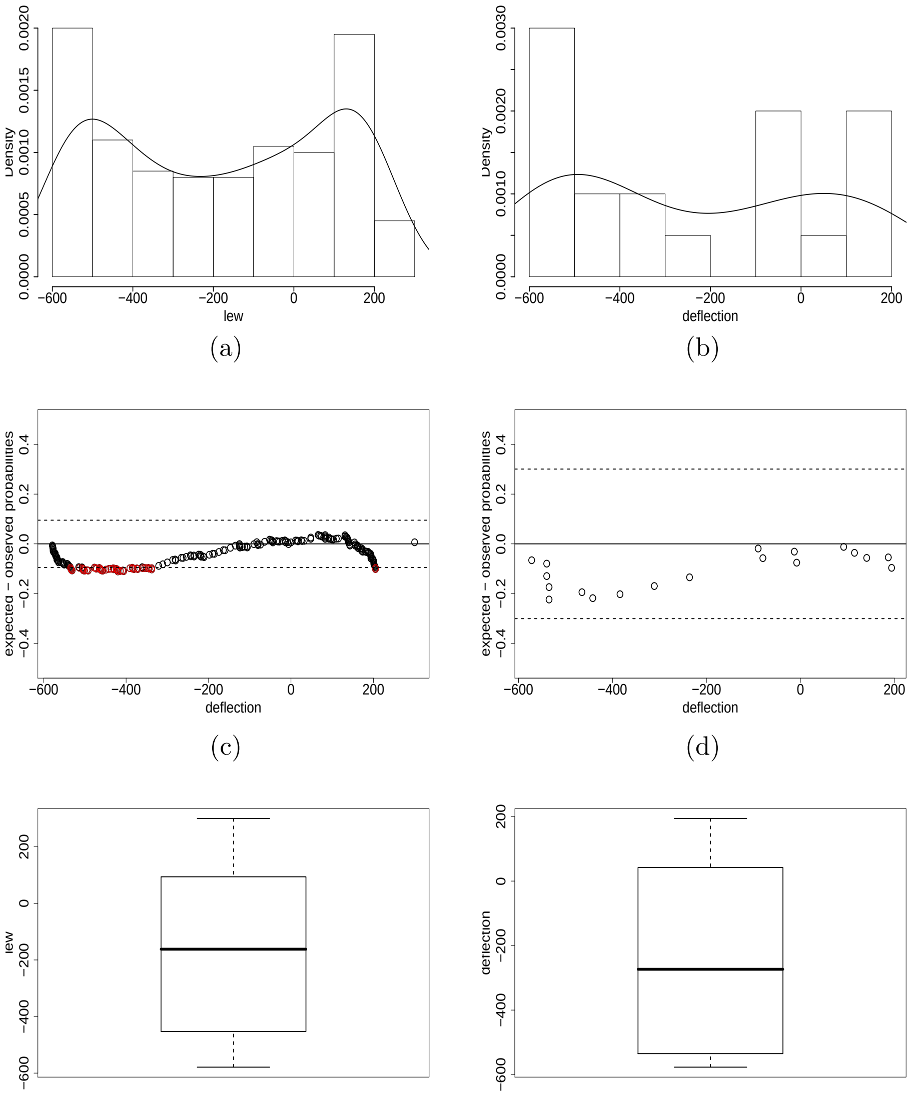
($n = 200$) which is a significant result (p=0.011) at the 5% level. Hence, the data shows evidence against uniformity.

The histogram of the data in panel (a) of Figure 2.2 shows two modes near the limits of the interval. This graph also includes the kernel density estimate, which is obtained by a Gaussian kernel with bandwidth determined by means of the unbiased cross-validation method (Silverman (1986) and Venables and Ripley (1997)). The bimodal pattern is apparent in this kernel density estimate as well. The red points in the detrended PP-plot in panel (c) of Figure 2.2 indicate the observations that are responsible for the rejection of the KS test. Many dots are colored red in the interval of beam deflection ranging from -550 to -350 and few dots are colored red near the other end at 200. From the box plot (see panel (e) in Figure 2.2), it is seen that the first and the third quartiles are relatively close to the limits of the interval.

For illustrative purposes, we have also taken a small random subsample of 20 observations. We want to see whether the bimodal pattern would also be noticeable in such a small sample. The KS test for the subsample gives a value of $D_n = 0.249\sqrt{n}$ ($n = 20$) which is now not significant anymore (p=0.142). This result can also be derived from the corresponding detrended PP-plot in panel (d) of Figure 2.2. On the other hand, the histogram and the box plot in panels (b) and (f) respectively of Figure 2.2 still suggest non-uniformity. Finally, the summary statistics in Table 2.2 clearly indicate a deviation from uniformity in both the original sample and the subsample. This small experiment may suggest that the KS test is not powerful enough to detect the non-uniformity in such a small sample. In Chapter 5, we will see that the localised Pearson χ^2 test succeeds in producing significant results in both the original and the small subsample. More formal arguments will there be given too. We should of course be careful with this conclusion, because it involved only one subsample.

2.1.3 Chemical Concentration data

Thas (2001) and Rayner and Best (1989) cited data from a study on the effect of environmental pollutants on animals. These data was originally given by Risebrough (1972) and contains concentrations of various chemicals in the yolk lipids of pelican eggs. For 65 Anacapa birds the concentrations of polychlorinated biphenyl (PCB) are considered. These data will further be referred to as the PCB data. Here, the data-analytic question is whether the PCB concentrations are normally distributed. This assumption is needed when one is for example interested in calculating a parametric confidence interval for the population mean. If the assumption of normality is satisfied, then the parametric confidence interval is an optimal interval, in the sense that among all unbiased confidence intervals, it is the most narrow interval.



(a) (b)

(c) (d)

(e) (f)

Figure 2.2: Histograms (a) and (b), detrended PP-plots (c) and (d) and box plots (e) and (f) for the Lew data and a subsample of 20 observations of the Lew data. The region between (outside) the dotted lines in the detrended PP-plots, shows the acception (rejection) region of the KS test at the 5% significance level.

Table 2.2: Descriptive summary statistics for the Lew data and a subsample of size $n = 20$ together with the population characteristic for uniformity.

	Min	1st Quartile	Median	3rd Quartile	Max
sample	-579.0	-451.00	-162.0	93.00	300.0
subsample	-577.0	-535.00	-273.5	17.00	194.0
expected	-580.0	-359.75	-139.5	80.75	301.0

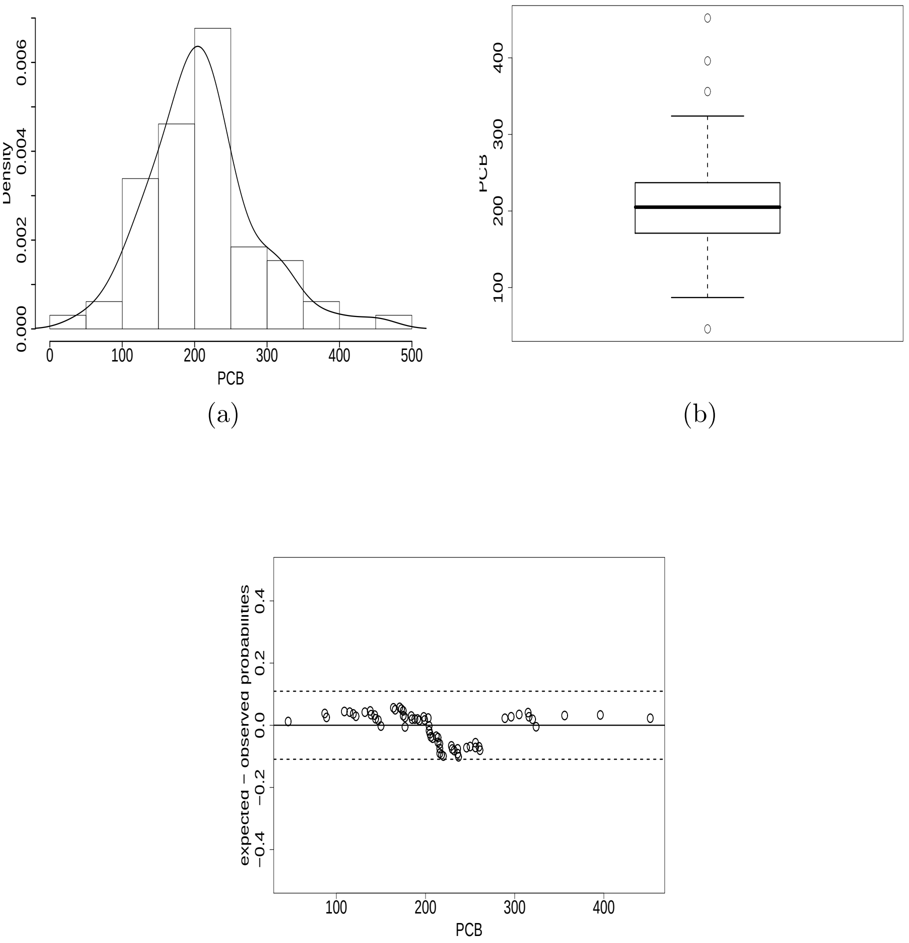
	Mean	Variance	Skewness	Kurtosis
sample	-177.4	76913.13	-0.051	-1.496
subsample	-230.0	84136.47	0.183	-1.686
expected	-139.5	64680.08	0.000	-1.200

Alternatively, the question of normality may be of interest in its own right. The answer to that question can again be given by GOF tests. However, the testing problem described here is a *composite* one-sample GOF problem, where *composite* refers to the hypothesised distribution, which is a member of a parametric family of distributions. In general, the formulation of the composite one-sample null hypothesis is as follows

H_0: the underlying distribution of the data is F_0, where F_0 can be any member of some parametrised family of distributions.

For the GOF problem here, the distribution under the null hypothesis belongs to the family of normal distributions characterised by the mean μ and the variance σ^2. The mean and the variance of the true distribution are not known to the researcher. Hence, only the form of the hypothesised distribution is known and therefore the null distribution actually covers an infinite number of distributions of the same form.

Usually, the composite GOF problem is handled by the same test statistic as for the simple null hypothesis, with the unknown parameters simply replaced by some appropriate estimates. The mean μ and the variance σ^2 could for example be replaced by their sample counterparts. By doing this the null distribution of the test statistic often becomes more difficult. While GOF test statistics for simple null hypotheses are usually distribution-free, the asymptotic null distribution of their counterparts for composite null hypotheses may depend on the parametric family and the unknown parameters. For example, the asymptotic null distribution of the KS test for simple null hypotheses does not depend on the distribution specified under the null hypothesis, while the KS test for testing normality with unknown mean and variance needs a correction (see Lilliefors (1967)).



(a)

(b)

(c)

Figure 2.3: Histogram (a), box plot (b) and detrended PP-plot (c) for the PCB data. The region between (outside) the dotted lines in the detrended PP-plots, shows the acception (rejection) region of the KS test at the 5% significance level.

Table 2.3: Descriptive summary statistics for the PCB data together with the population characteristic for normality. We assume that the population mean and variance are equal to the sample mean and variance, respectively.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
observed	46	171.00	205.000	237	452
expected	$-\infty$	160.88	210	259.121	$+\infty$
	Mean	Variance	Skewness	Kurtosis	
observed	210	5303.656	0.725	1.432	
expected	210	5303.656	0.000	0.000	

The KS test statistic for composite normality results in a value $D_n = 0.109\sqrt{n}$ ($n = 65$) with corresponding p-value equal to 0.052. This means that no strong evidence against normality is found at the 5% significance level. The histogram with density estimate (panel (a) of Figure 2.3), however, shows a small bump to the right of the main mode. Note that this impression strongly depends on the number of bars in the histogram and the chosen bandwidth for the kernel density estimator. The box plot, in panel (b) of Figure 2.3, shows three *outliers* for which the concentration is markedly high and one outlier with a very small concentration. These outliers are responsible for the rather big tails in the histogram. Apart from that, the box is fairly symmetric, i.e. the median is right in the middle of the box and the *whiskers*. The whiskers are the horizontal lines and indicate the largest (smallest) observation that is smaller (larger) the median plus (minus) 1.5 times the interquartile range. Observations outside that box are considered outliers. Finally, in the detrended PP-plot (panel (c) of Figure 2.3), all points fall nicely in the acception region for the KS test. From the PP-plot, we may conclude that the data is fairly normally distributed.

For completeness, we comment on the summary statistics for the PCB data in Table 2.3. Note that the expected quartiles are calculated under the assumption that the true mean and variance of the distribution are equal to their sample equivalents. This assumption will not exactly be satisfied, which complicates the comparison between observed and expected values. While a GOF test provides the data-analyst with objective criteria concerning the characteristics of a distribution, a comparison of the observed and expected summary statistics is very difficult to judge only by eye. Consider for example comparing observed and expected values for the skewness and kurtosis. The skewness and kurtosis for a normal distribution do not depend on the mean and variance and are expected to be zero for all normal distributions. The observed values are different from zero, but the question is how large the difference should be in order to be

Table 2.4: Descriptive summary statistics for the Fastfood data together with the population characteristic for normality. We assume that the population mean and variance are equal to the sample mean and variance, respectively.

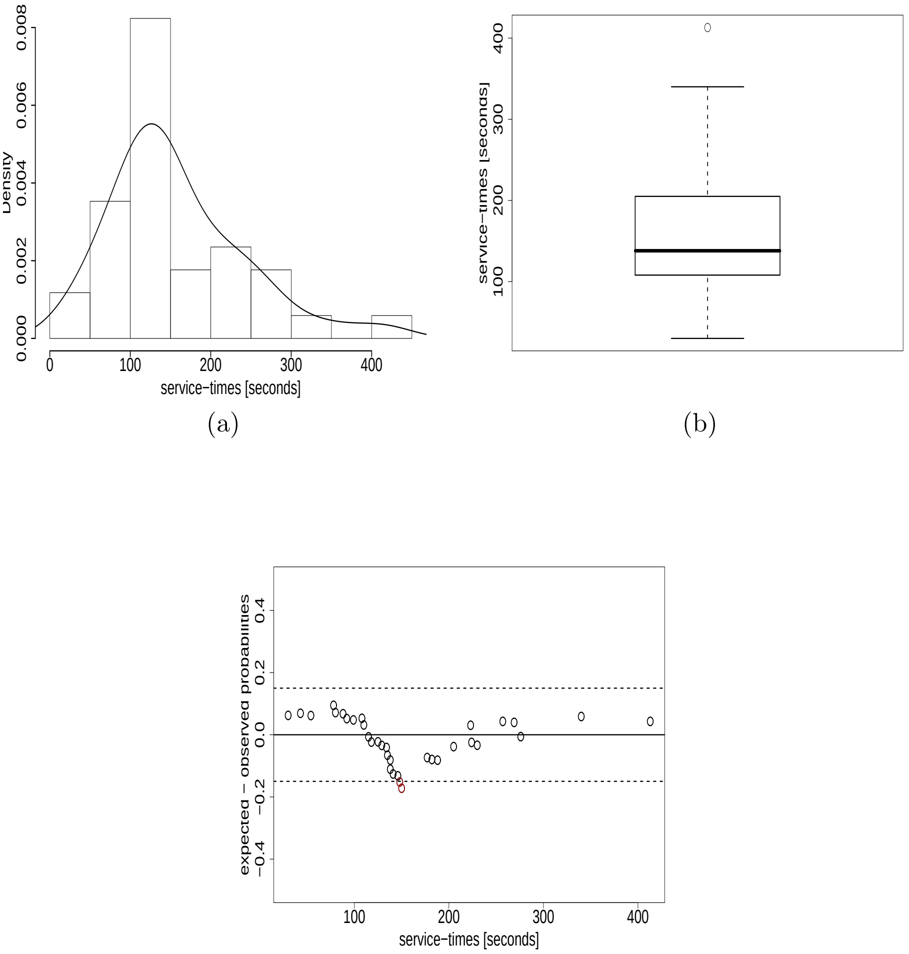
	Minimum	1st Quartile	Median	3rd Quartile	Maximum
observed	30	108.500	138.000	200.800	413
expected	$-\infty$	101.973	158.324	214.674	$+\infty$
	Mean	Variance	Skewness	Kurtosis	
observed	158.324	6979.801	1.126	1.575	
expected	158.324	6979.801	0.000	0.000	

significant. That is the reason why we need to combine the explorative glance at the data with the formal statistical GOF tests, which are described in the next chapters. Based on what we explored in this section, it seems possible that we find a significant difference in skewness and/or kurtosis.

2.1.4 Fastfood data

Hollander and Wolfe (1999) published data on service-times for a Tallahassee drive-through fastfood restaurant. The service-time is defined as the time (in seconds) from the moment the car pulled up to the speaker to order, to the moment the car left the window with the order. The data were obtained around dinner time on Thursday evening. We assume that the observed times are independent. In particular, it is reasonable to assume that the time needed to serve one customer does not have any influence on the time needed to serve the next or previous customer. Hence, the observations can be viewed as a random sample from an unknown distribution. The question could now be whether the underlying distribution is normal or whether it has characteristics that severely deviate from a normal distribution.

The KS test results in $D_n = 0.187\sqrt{n}$ ($n = 34$) with p-value equal to 0.004, which indicates a severe deviation from normality. On the detrended PP-plot in panel (c) of Figure 2.4, it is seen that the deviation corresponds to two observations, at service-time near 150 seconds, falling within the rejection region. From the histogram in panel (a) of Figure 2.4, we derive that this deviation from normality is thus located where most of the observations are located, i.e. at the mode of the distribution. Moreover, the kernel density estimate in that panel gives a positively skewed impression. This impression is confirmed by the box plot in panel (b) and by the summary statistics in Table 2.4.



(a) (b)

(c)

Figure 2.4: Histogram (a), box plot (b) and detrended PP-plot (c) for the Fastfood data. The region between (outside) the dotted lines in the detrended PP-plots, shows the acception (rejection) region of the KS test at the 5% significance level.

2.1.5 Old Faithful geyser data

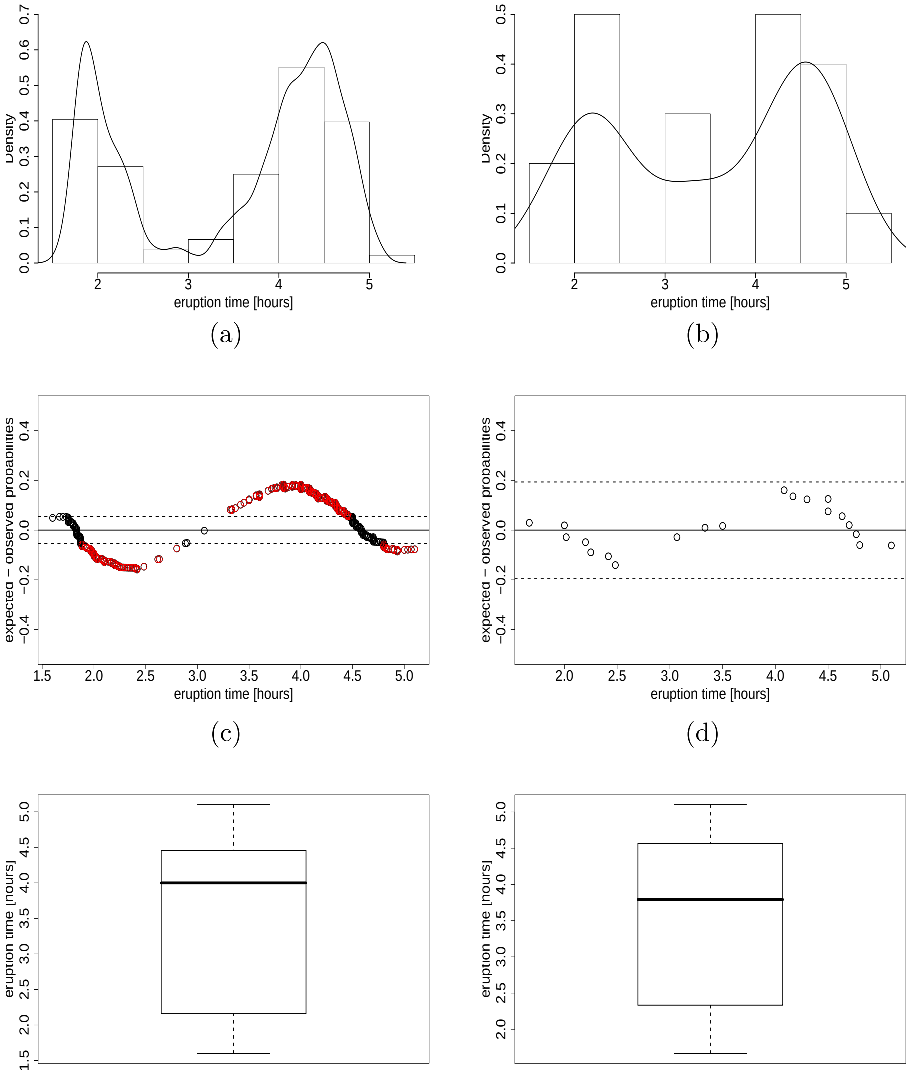
The Old Faithful in Yellowstone National Park is probably the most photographed geyser in the world. The data of the same name has been widely discussed as well (see e.g. Azzalini and Bowman (1990), Härdle (1991) or Scott (1992)). Several datasets are circulating. They all consists of two variables: the waiting time to an eruption and the time each eruption takes. The dataset we consider has 272 registered eruptions. We are only interested in the eruption time and would like to get more insight in its distributional characteristics. From the histogram in panel (a) of Figure 2.5, it is clear that the data has two modes. Therefore a normal distribution, which is unimodal, would be far from appropriate. Nevertheless, we here still perform tests for normality, in order to illustrate how various tests give insight in the deviation from the null distribution. The purpose of this analysis is more meant for getting knowledge concerning the deviations from normality.

The KS test has value $D_n = 0.181\sqrt{n}$ with $n = 272$ and a p-value smaller than 0.001, which shows a severe deviation from normality as was expected. The detrended PP-plot in panel (c) of Figure 2.5 shows that the deviations are located at the regions where the two modes are occurring. Almost all points are in the rejection region of the KS test. The box plot in panel (e) of Figure 2.5 shows that the distribution is clearly asymmetric, due to the second mode being much larger than the first mode. The summary statistics are given in Table 2.5 and are not in agreement with a symmetric unimodal distribution.

Suppose now we only had 20 observations from this dataset. The question is then whether it would still be as easy to recognise this bimodal pattern. The KS test for a random subsample gives a non-significant result ($D_n = 0.185\sqrt{n}$ with $n = 20$ and $p = 0.069$). Hence, the detrended PP-plot (panel (d) of Figure 2.5) has no points in the rejection region. The box plot and the histogram (panel (b) and (f), respectively) still indicate that the data is probably not normal, but the bimodal pattern is not that obvious anymore. In Chapters 5 and 6, we apply other tests and tools that are powerful enough to reveal the deviation from normality even in such a small dataset.

2.2 Circular data

Circular or *directional* data arise in many fields. For instance, in meteorology wind directions are often measured, biologists may be interested in the directions of migrating animals, or health scientists study the arrival times of patients at an intensive care unit. All these data have in common that they are measured on a circle, which represents either a compass or a clock. The measurement scale is characterised by invariance to the choice of the origin, and the distance



(a) (b)

(c) (d)

(e) (f)

Figure 2.5: Histograms (a) and (b), detrended PP-plots (c) and (d) and box plots (e) and (f) for the Old Faithful geyser (OFG) data and a subsample of 20 observations of the OFG data. The region between (outside) the dotted lines in the detrended PP-plots, shows the acception (rejection) region of the KS test at the 5% significance level.

Table 2.5: Descriptive summary statistics for the Old Faithful geyser data and a subsample of size $n = 20$ together with the population characteristic for normality. We assume that the population mean and variance are equal to the sample mean and variance, respectively.

	Min	1st Quartile	Median	3rd Quartile	Max
sample	1.600	2.163	4.000	4.454	5.1
expected	$-\infty$	2.718	3.488	4.258	∞
subsample	1.667	2.375	3.792	4.533	5.1
expected	$-\infty$	2.743	3.524	4.305	∞
	Mean	Variance	Skewness	Kurtosis	
sample	3.488	1.303	-0.418	-1.506	
expected	3.488	1.303	0.000	0.000	
subsample	3.524	1.341	-0.252	-1.612	
expected	3.524	1.341	0.000	0.000	

between two observations is given by the smallest arc instead of the numerical difference. In the next section some typical circular data examples are described and some explorative analysis is done. More elaborate analyses of these data is given in the next chapters. Again, all corresponding datasets can be found in Appendix C.

2.2.1 Birth time data

Rayner and Best (1989) quoted the following 37 times for consecutive births in a hospital, which were originally given by Mood et al. (1974). Although the birth times are recorded from 7.02pm on the first day till 4.31pm twelve days later, we will ignore this information. The reason is that we are only interested in the specific times of birth throughout the day. In particular, we would like to verify whether the time points occur uniformly at random or whether there is some “preferred” birth time. Since these time points can be interpreted as measurements on a scale with a period of one day, we are clearly dealing with circular data. The data can thus be presented on the circle as shown in Figure 2.6. The rotation direction is chosen to be clockwise and the origin is chosen to be at midnight. In statistical analysis on these data, it is important that choosing another origin or changing the rotation direction should not change the conclusions . Furthermore, it is clear that for times of birth throughout the day the distance between two observations is not given by the numerical difference. For instance, the difference between two births at 2.00am and 11.00pm (or

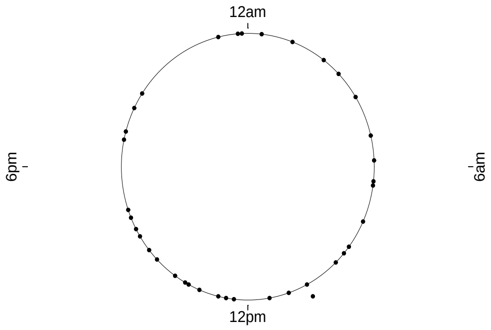


Figure 2.6: Birth time data

equivalently 2h00 and 23h00, on a 24 hours clock) is given by the smallest arc which is only 3 hours instead of $23 - 2 = 21$ hours. Note that Rayner and Best (1989) interpreted these data as linear data.

For such data the classical (linear) distributions are clearly not appropriate, but many circular distributions have been proposed (see e.g. Jammalamadaka and SenGupta (2001) and Fisher (1993)). The circular uniform (CU) distribution and the circular normal (CN) distribution, also called the von Mises (VM) distribution, are the two most important circular distributions and are formally defined in the next chapter. The uniform distribution on the circle is similar to the uniform distribution on a real interval since its density curve is constant, and it is therefore invariant under rotation. On the other hand, the density curve of the CN distribution is not rotation invariant. It is a symmetric unimodal distribution and the circular equivalent to the normal distribution on the real line.

Just as with linear data, an important aspect of circular data analysis is testing for GOF. Although for the former type of data many tests exist, these are often not suited for circular distributions. The question here is whether the birth times occur uniformly throughout the day. This is in fact the one-sample GOF problem for circular uniformity, where circular uniformity means that each time point on the 24h clock (or on the circle in Figure 2.6), has equal probability to be a birth time. This *simple circular GOF problem* is discussed in the next chapter. Here we confine the discussion to a formulation of the distributional characteristics of the data. Panel (a) of Figure 2.7 shows a classical histogram with kernel density estimate for the Birth time data, obtained after projecting the time points on the real line. This plot suggests that the preferred time of birth is around 12pm and less births occur at night.

The impression from such a histogram can be misleading because it is sensitive to the point at which the circle is cut off. A useful way to modify the linear histogram is to make a rose diagram, in which the bars of a histogram are replaced by circular sectors. The area of the sector is proportional to the frequency in the corresponding arc. Panel (c) in Figure 2.7 shows a rose diagram for the Birth time data, from which we see that the idea of a preferred direction at noon is somewhat less pronounced. The same holds for the circular kernel density estimate presented in panel (d). Here the kernel density estimate was based on a von Mises kernel for which the concentration parameter is chosen by means of cross-validation (see Hall, Watson, and Cabrera (1990)). In Chapter 3 we give some more information on circular kernel density estimation.

We now discuss some descriptive summary statistics that are suitable for circular data. Regarding estimation of the mean birth time it is clear that the average as it is defined for linear data is not appropriate here. The reason is that its value is not *origin-equivariant.* For instance, the average for the Birth time data is 12.08pm when the origin is at midnight, while taking the origin at 6pm, the average is 7.36am. Instead we proceed as follows. Let $z_j = e^{ix_j} = \cos x_j + i \sin x_j$, $j = 1, \dots, n$ represent the data points on the unit circle. The *circular mean direction* of $x_1, \dots, x_n$ is denoted by $\overline{X}_c$ and defined as the direction of the mean resultant vector of the sample unit vectors $(\cos x_j, \sin x_j)$, $j = 1, \dots, n$. Let

$$\overline{C} = \frac{1}{n} \sum_{j=1}^{n} \cos x_j \text{ and } \overline{S} = \frac{1}{n} \sum_{j=1}^{n} \sin x_j,$$

then $\overline{X}_c$ is the solution of the equations

$$\overline{C} = \overline{R} \cos \overline{X}_c \text{ and } \overline{S} = \overline{R} \sin \overline{X}_c, \tag{2.3}$$

where the *mean resultant length* $\overline{R}$ is given by

$$\overline{R} = (\overline{C}^2 + \overline{S}^2)^{1/2}. \tag{2.4}$$

Note that $\overline{X}_c$ is not defined when $\overline{R} = 0$. If $\overline{R}$ is close to 0, then there is no concentration about any particular direction, and the data approach a circular uniform distribution. Therefore, the larger the value of $\overline{R}$, the more evidence against circular uniformity. When $\overline{R} > 0$, $\overline{X}_c$ is explicitly given by

$$\overline{X}_c = \begin{cases} \arctan(\overline{S}/\overline{C}) & \text{if } \overline{C} \geq 0 \\ \arctan(\overline{S}/\overline{C}) + \pi & \text{if } \overline{C} < 0, \end{cases} \tag{2.5}$$

where arctan takes values in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. This definition is *equivariant* under rotation. In particular, adding a constant γ to the sample directions $x_1, \dots, x_n$ results in a mean direction $\overline{X}_c + \gamma$. The mean resultant length $\overline{R}$ is invariant under rotation.

The circular mean direction for the Birth time data is $\overline{X}_c$ = 12.12pm and the resultant length of the mean vector is $\overline{R}$ =0.19. Since all sample vectors are unit vectors, we have $0 \leq \overline{R} \leq 1$, where $\overline{R} = 1$ would mean that all observations are in the same direction. The *sample circular variance* is defined as

$$V = 1 - \overline{R},$$

where large and small values indicate widely dispersed and tightly clustered directions, respectively.

The sample variance for the Birth time data (V=0.81) is rather large which means that the time points are well spread around the circle. There exist other measures of dispersion for which we refer to e.g. Mardia and Jupp (2000).

It is useful to combine sample mean direction and sample circular variance into the *first sample trigonometric moment about the zero direction*,

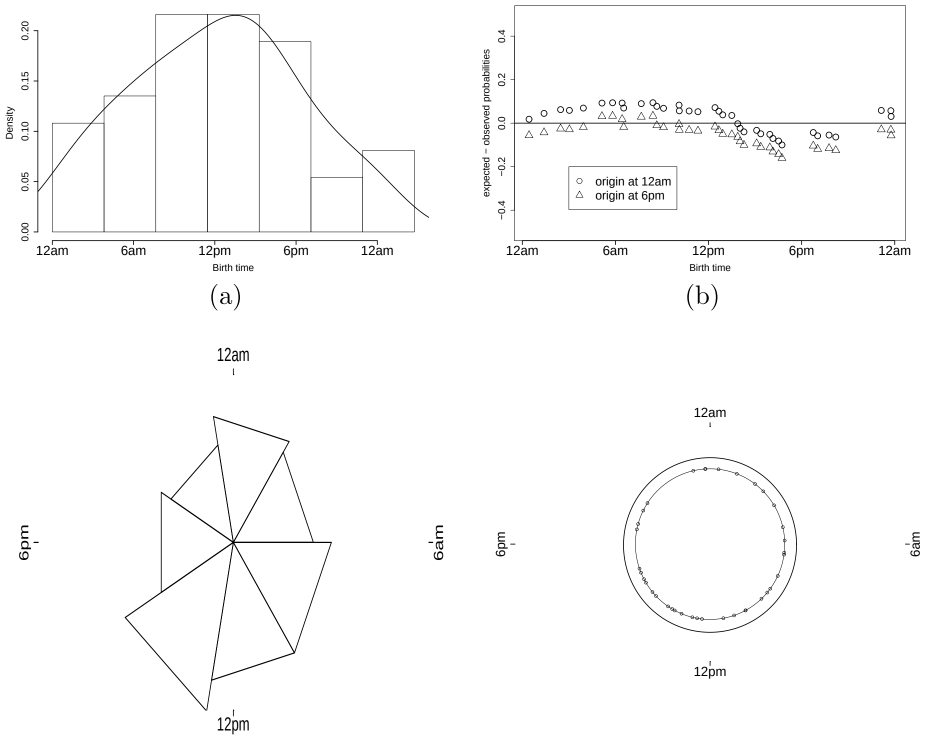
$$m_1 = \overline{C} + i\overline{S} = \overline{R}e^{i\overline{X}_c}. \tag{2.6}$$

For the Birth time data, the complex value $m_1 = -0.19 - 0.01i$ is the first order trigonometric moment.

Panel (b) in Figure 2.7 shows the detrended PP-plot for the linearised Birth time data. Different symbols are used to indicate where the data is cut off, or, in other words, which origin is chosen. It is clear from the figure that such a plot depends on the choice of the origin. We should draw a detrended PP-plot for every choice of the origin so as to obtain the complete picture of the locations where the data deviate from the uniform distribution. It is clear that the same is required for the KS test, which is related to the detrended PP-plot as explained in Section 2.1. In Chapter 6 we extend the detrended PP-plot to an interval-based PP-plot which is directly applicably to circular data and is related to the origin-invariant Kuiper test. Here, we already give the result of the Kuiper test, which is in fact the largest value of the KS test statistics among the set of KS statistic values obtained by taking all possible different origins. More details are given in Section 3.4.2. Applying the Kuiper test to the Birth time data we obtain $K_n = 1.218$ which corresponds to a non-significant p-value of 0.508. So, from our initial analysis of the Birth time data, we may conclude that there is no evidence that a specific time point is preferred to give birth.

2.2.2 Homing pigeons

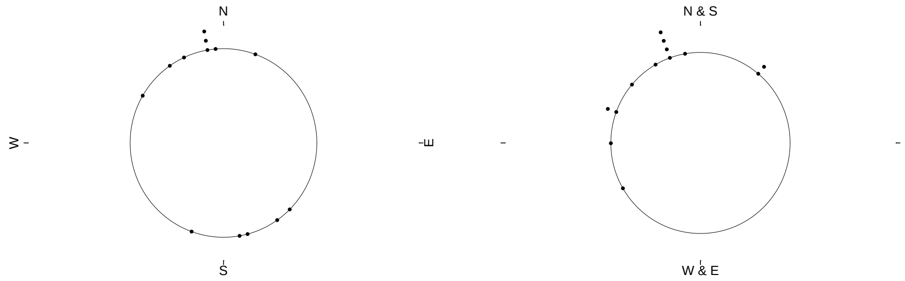
Batschelet (1981) reported data about homing pigeons that were released one by one in the Toggenburg Valley in Switzerland under sub-Alpine conditions. The birds did not adjust quickly to the homing direction but preferred to fly along the axis of the valley. The origin (0°) is taken at the North direction.



(a) (b)

(c) (d)

Figure 2.7: A classical histogram with linear density estimation for projected data (a) A detrended PP-plot for uniformity (b) A rose diagram (c) and a circular density estimation (d) for the Birth time data.

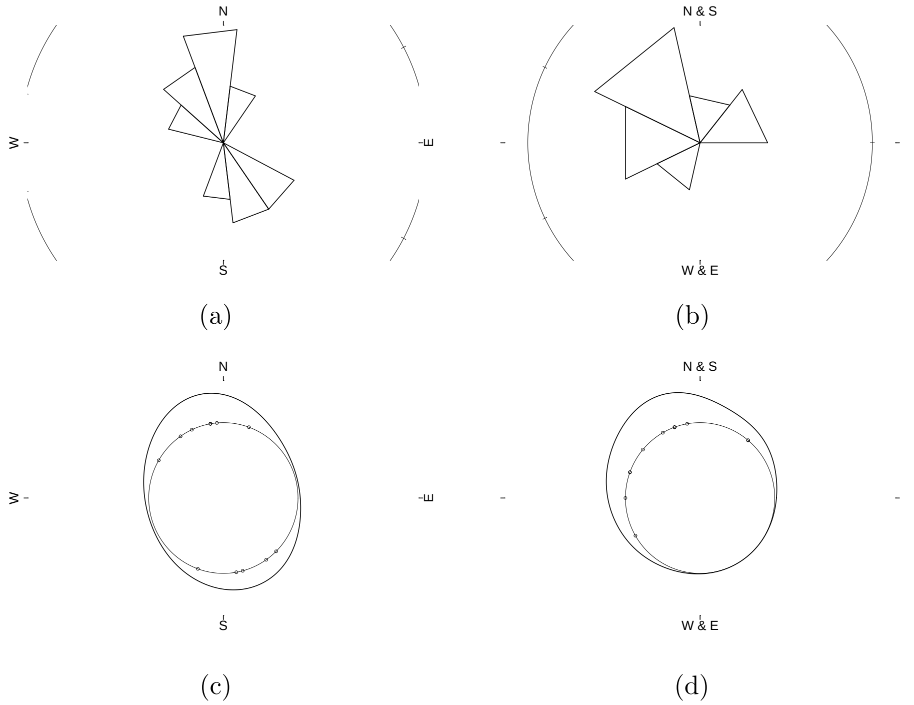


(a) (b)

Figure 2.8: (a) Original Homing pigeons data and (b) the doubled angles of the Homing pigeons data.

These observations, which are plotted in panel (a) of Figure 2.8, have also been studied by Jammalamadaka and SenGupta (2001). Numerous other experiments about homing pigeons are reported and analysed (see Mardia and Jupp (2000) and Fisher (1993)). The question of interest in these experiments is whether the birds have a preferred direction of flight. If no direction is preferred, the data would be uniformly distributed on the circle. Hence a similar GOF problem arises as with the Birth time data. The Kuiper test results in $K_n = 1.505$ with p-value 0.170, which means that no significant difference from circular uniformity can be recognised at the 5% significance level. However, from the rose diagram and the kernel density estimate in panels (a) and (c) of Figure 2.9, it is seen that the birds systematically choose one of two opposite directions. The data thus possibly come from a bimodal distribution, where the modes are diametrically opposed. The fact that the modes are opposed may be the reason why the Kuiper test could not find a difference from circular uniformity. We also see from the summary statistics in the upper part of Table 2.6 that the mean resultant length $\overline{R}$ is rather low. Hence, the values of this statistic also explains why the GOF test could not find a difference from uniformity.

We may argue that the high mountains on both sides of the valley force the pigeons to fly along the axis of the valley and that there is no preference for flying upstream or downstream. Under this assumption the distribution is *axial* and, therefore, the method of doubling the angles may be applied. Indeed, doubling the angles is equivalent to reducing all angles modulo 180^o. Panel (b) of Figure 2.8 shows the observed angles on a circle with circumference 180^o. Note that data which are perfectly uniformly distributed on the circle with circumference 360^o, will also be uniform on the circle with circumference 180^o



(a) (b)

(c) (d)

Figure 2.9: A rose diagram (a) and a circular density estimate (c) for the original Home pigeons data. A rose diagram (b) and a circular density estimate (d) for the doubled Home pigeons data.

after doubling. Applying the Kuiper test for uniformity to the doubled data convincingly rejects the null hypothesis ($K_n = 2.151, p = 0.003$). The second row in Table 2.6 contains the values for the first trigonometric moment of the doubled data. The sample circular mean is now clearly in the direction of the unique mode of the distribution and the length of the resultant vector is much larger than that of the original data.

The definition of the first trigonometric moment (2.6) is extended to the p^{th} *order trigonometric moment around the zero direction*, as

$$m_p = \overline{a}_p + i\overline{b}_p = \overline{R}_p e^{i\overline{X}_p}, \tag{2.7}$$

where

$$\overline{a}_p = \frac{1}{n}\sum_{j=1}^{n}\cos(pX_j) \text{ and } \overline{b}_p = \frac{1}{n}\sum_{j=1}^{n}\sin(pX_j) \tag{2.8}$$

and where $\overline{X}_p$ and $\overline{R}_p$ denote the sample mean direction and the sample mean resultant length of $pX_1, pX_2, \ldots, pX_n$. Furthermore, the p^{th} *central trigonomet-*

Table 2.6: Descriptive summary statistics for the Homing pigeons data.

	$\overline{X}_c$	$\overline{R}$	m_1		
original data	344.330°	0.216	0.208-0.058i		
doubled data	326.533°	0.735	0 .613-0.405i		
	$\overline{X}_2$	$\overline{R}_2$	m_2	q_1	q_2
original data	326.533°	0.735	0.613-0.405i	0.039	1.190
doubled data	297.111°	0.278	0.127-0.247i	-0.143	-0.206

ric moment is defined as

$$m_p^\star = \overline{a}_p^\star + i\overline{b}_p^\star, \tag{2.9}$$

where

$$\overline{a}_p^\star = \frac{1}{n}\sum_{j=1}^{n}\cos(p(X_j - \overline{X}_c) \text{ and } \overline{b}_p^\star = \frac{1}{n}\sum_{j=1}^{n}\sin(p(X_j - \overline{X}_c)). \tag{2.10}$$

The second order trigonometric moment is related to the *circular skewness* and *kurtosis* which are defined as

$$q_1 = \frac{\overline{R}_2 \sin(\overline{X}_2 - 2\overline{X}_c)}{(1-\overline{R})^{3/2}} \tag{2.11}$$

$$q_2 = \frac{\overline{R}_2 \cos(\overline{X}_2 - 2\overline{X}_c) - \overline{R}^4}{(1-\overline{R})^2}, \tag{2.12}$$

respectively. If the distribution is symmetric and unimodal, q_1 is nearly zero and if the distribution has a unimodal peak similar to the CN distribution, q_2 is close to zero as well. The population versions of the trigonometric moments are defined in Section 3.1.2 and are important to characterise a circular distribution.

For the Birth time data, we have $q_1 = -0.124$ and $q_2 = 0.081$. For the Homing pigeons data, the values for the second trigonometric moment together with the circular skewness and kurtosis are in the lower part of Table 2.6. Note that the values for the second order trigonometric moment of the original data are obviously equal to the values of the first trigonometric moment of the doubled data. Furthermore, for the doubled values skewness and kurtosis are both close to zero, while for the original data, kurtosis is much larger. Hence, the doubled data seem to agree with coming from a CN distribution. More GOF techniques are applied to these data in next chapters.

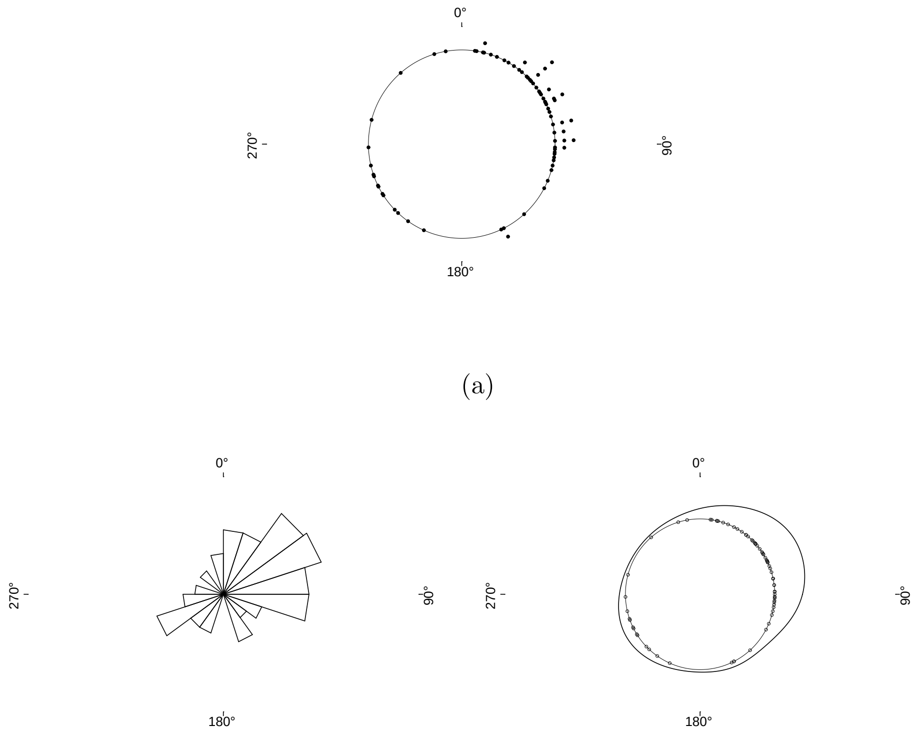
2.2.3 Turtles data

The Turtles data is another example on animal movement, which is quoted by Fisher (1993), who took it from Stephens (1969). The data consist of directions taken by turtles after some treatment. Interest lies in whether the turtles had a preferred direction to move away from the place they were released. Some of the turtles confused forward with backward. We assume that the turtles wanted to move away in a particular direction, but that due to the treatment some of the turtles were not able to orientate well. The most common distribution to model such data is a CN distribution. This distribution (see also Section 3.1.2) is symmetric, unimodal and the circular analogue of the normal distribution on the line. Since the distribution is characterised by two parameters, i.e. the location μ and the concentration κ, we encounter here a *composite circular GOF problem.* Recall from Section 2.1 that the KS test for linear uniformity could be extended to test linear normality. In the same way, the Kuiper test can be applied to test for circular normality. For more details we refer to Section 3.4.1. Here we only give the results of this classical test. For the Turtles data we find $K_n = 1.568$ with $p = 0.008$, so the null hypothesis of circular normality is rejected at the 5% significance level.

This dataset was also analysed by Fernández-Durán (2004) in the context of density estimation and Mardia and Jupp (2000) used it as an example of a bimodal distribution. A mixture model of two CN distributions was proposed to describe the data. Looking at the raw data, the histogram and the kernel density estimate in Figure 2.10, we see that indeed two modes are recognised in opposite directions. Also the summary statistics in Table 2.7 are comparable to those of the original Homing pigeons data. However, doubling the data is not a biologically motivated option since we would lose useful information about the confused turtles who choose to go backwards.

2.2.4 Ants data

The Ants data is discussed in Batschelet (1981), but originally this orientation experiment in biology was published by Jander (1957). The data refer to directions chosen by 100 ants. The ants were put in an evenly illuminated arena with a black target placed at 180°. The animals could see the target and were supposed to run towards it. We are interested in whether the data follow a CN distribution. As it was demonstrated by Fisher (1993), this distribution is not a suitable model. Indeed, the Kuiper test is highly significant ($K_n = 11.369$, $p < 0.001$). Fernández-Durán (2004) and Jammalamadaka and Kozubowski (2003) used the example to demonstrate the flexibility of their proposed family of distributions to skewed and / or peaked data. Here, we briefly present the explorative analysis of the data, while in the next chapters we give a more detailed



(a)

(b)

(c)

Figure 2.10: Raw data (a), rose diagram (b) and kernel density estimate (c) for the Turtles data.

Table 2.7: Descriptive summary statistics for the Turtles data, the Ants data, the Direzione data and the Arrival data.

	$\overline{X}_c$	$\overline{R}$	m_1		
Turtles	64.171°	0.497	0.217+0.447i		
Ants	183.139°	0.610	-0.609-0.033i		
Direzione	16.740°	0.656	0.628+0.189i		
Arrival	5.15pm	0.323	-0.062-0.317i		
	$\overline{X}_2$	$\overline{R}_2$	m_2	q_1	q_2
Turtles	124.874°	0.481	-0.275+0.395i	0.082	1.657
Ants	13.303°	0.459	0.446-0.106i	-0.230	2.083
Direzione	8.556°	0.474	0.469+0.071i	0.989	2.069
Arrival	2.12am	0.065	0.055+0.036i	0.097	-0.105

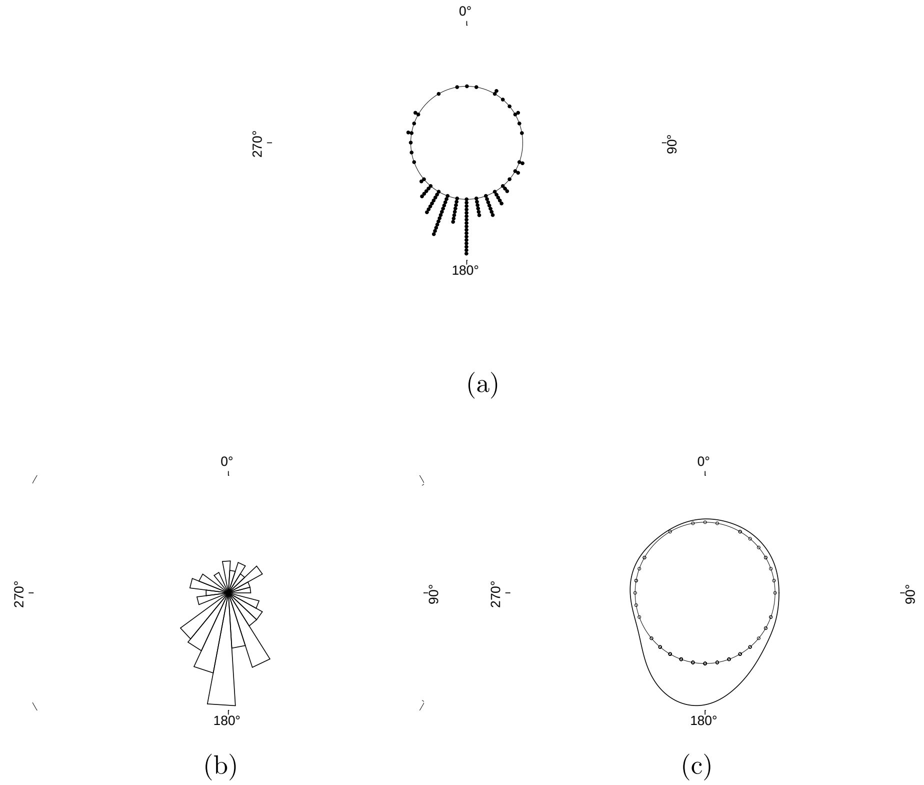
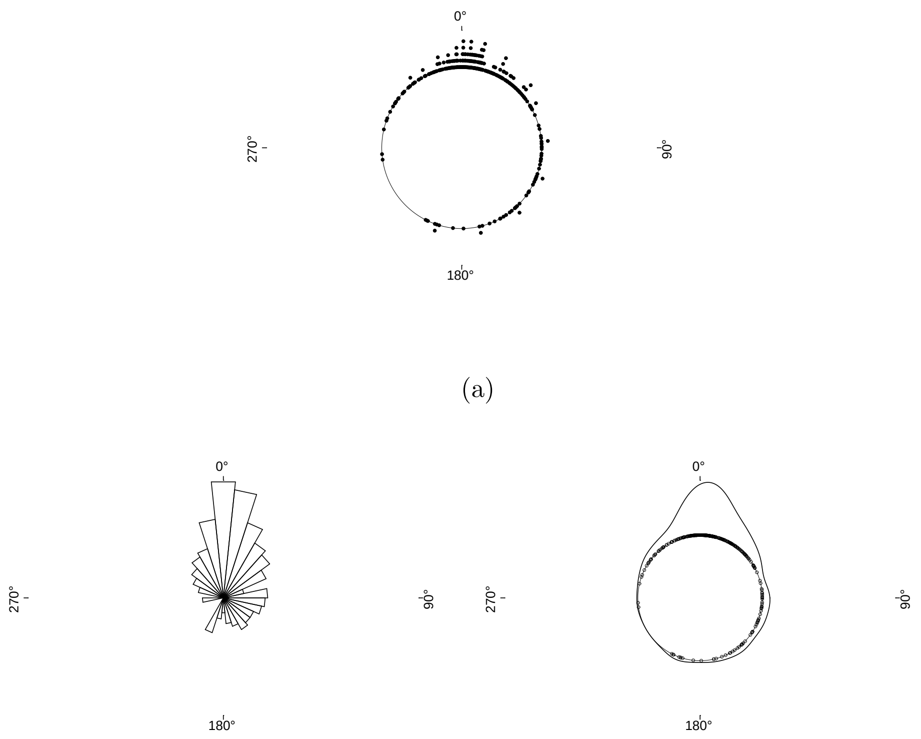


Figure 2.11: Raw data (a), rose diagram (b) and kernel density estimator (c) for the Ants data.

analysis to see in what way the true distribution deviates from the hypothesised normal.

Figure 2.11 shows the raw data (panel (a)), the rose diagram (panel (b)) and the kernel density estimate (panel (c)). They all give a unimodal impression, with mode at about 180^o (where the target was placed) and with a negative skewness. The circular sample mean and the circular skewness in Table 2.7 confirm this impression. The mean resultant length is quite large, which indicates strong evidence for a unimodal distribution. Note, that the kurtosis is very large as well. This suggests that the degree of peakedness is much higher than that of a CN distribution.



(a)

(b)

(c)

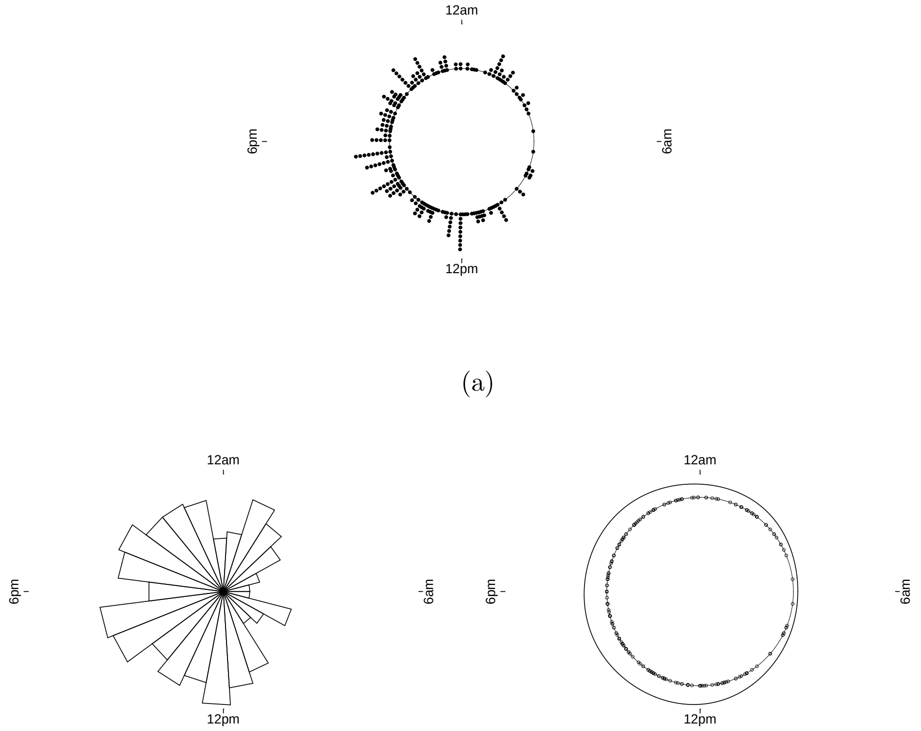
Figure 2.12: Raw data (a), rose diagram (b) and kernel density estimator (c) for the Direzione data.

2.2.5 Direzione data

The Direzione data, taken from Agostinelli (2006), contain 310 measurements of wind direction in the Italian Alpes. The raw data plot, the rose diagram and the kernel density estimate in Figure 2.12 suggest a unimodal distribution with its mode at the North direction (about 16°, see Table 2.7). The question is again whether the underlying distribution for this dataset is a CN distribution. The Kuiper test has a highly significant p-value ($p < 0.001$). From the summary statistics in Table 2.7 and the explorative graphs (Figure 2.12), we have the impression that the distribution is highly positively skewed with a very long tail on the positive side. Agostinelli (2006) argued that this large dataset possibly contains outliers and he provided robust estimation techniques. Another possibility is that this long tail is a second cluster of observations that is widespread along the first half of the circle.

2.2.6 Arrival data

Finally, the Arrival data (Fisher 1993) consists of the arrival times on a 24 hours clock of 254 patients at an intensive care unit, over a period of 12 months. The question of interest is the same as for the three previous examples. Looking at the summary statistics, we see that the CN distribution might not be a good fit. In particular, $\overline{R}$ is quite small. On the other hand, skewness and kurtosis are not as large as in the previous datasets. The data are shown in panel (a) of Figure 2.13, together with the rose diagram and kernel density estimate (panel (b) and (c), respectively). These plots suggest two large clusters of arrivals around 12pm and 5pm and three small clusters of arrivals around 2am, 7am and 10pm. Nevertheless, the Kuiper test ($K_n = 1.174$) does not yield a significant result (p=0.11). In Chapters 4 and 5, however, we obtain significant results with our new data-driven smooth test and with the data-driven circular SSPc test. Moreover, the graphical tool introduced in Chapter 6 localises the deviation from normality.



(a)

(b)

(c)

Figure 2.13: Raw data (a), rose diagram (b) and kernel density estimator (c) for the Arrival data.

CHAPTER 3

Some goodness-of-fit tests

In the previous chapter we illustrated that graphical methods to explore the distribution of the data are important to get a first impression of the distributional characteristics and can therefore not be neglected. On the other hand, the conclusions drawn from these visualisation techniques are often subjective. Moreover, the impression of e.g. a histogram or a kernel density estimator depends highly on the choice of the number of classes or the bandwidth, respectively. Therefore, they should not be used without formal statistical tools that give unambiguous answers to questions about the distribution of the data.

The major part of this chapter deals with GOF tests for linear and circular data. We give a historical overview of the most important linear and circular one-sample GOF tests based on the literature. The discussion will be limited though to tests that are simple, widely used and of practical importance. Additionally, we consider tests that are related to the new tests that will be proposed in the next chapters. Finally, some tests with good power characteristics are also included, which will serve as strong competitors to our new tests.

In the linear case as well as in the circular case, the GOF tests can be broadly classified into three categories. The first class of Pearson's χ^2 statistics is described in Section 3.2. Section 3.3 introduces the second class, which are the smooth tests, while Section 3.4 gives an overview of the techniques based on the EDF, comprising the third class. How these three classes of statistics are related to each other is explained in Section 3.5. Section 3.6 describes additional relevant tests that can not be classified under one of the three previous categories. For linear data, D' Agostino and Stephens (1986) give more details about many existing GOF techniques. For circular data, no such large spectrum of GOF tests is developed yet. However, most books about circular statistics contain one or more chapters where circular GOF tests are described (see e.g. Jammalamadaka and SenGupta (2001), Fisher (1993) and Mardia and Jupp (2000)).

Recall that in this thesis we are interested in one-sample GOF tests for different types of data. If such a GOF hypothesis test gives a significant result, we conclude that the probability density underlying the data does not follow the hypothesised density. A question that immediately arises from such a conclusion is in what way the data deviate from the hypothesised model. To be able to answer that question, the alternative hypothesis has to be informative or directional. Examples of directional deviations from the null hypothesis include a difference in location, variance or higher order moments. This information can be obtained by the data-driven smooth GOF test. Moreover, it leads naturally to an estimate of the density which can be used as a diagnostic tool to see how the true distribution deviates from the hypothesised. The resulting density estimator is known as an orthonormal series estimator and is briefly described in Section 3.7 together with two other popular non-parametric density estimation techniques.

3.1 GOF tests for linear and circular data

3.1.1 Statement of the one-sample GOF problem

Let $x_1, \ldots, x_n$ denote the data values observed either on the line or on the circle. To refer to one of these sample spaces we use the notation $\mathcal{S}$. The data values are assumed to be generated by $X_1, \ldots, X_n$, which are continuous or discrete i.i.d. random variables with unknown distribution function $F(x)$. Thus, $F(x)$ is an unspecified statistical model that generated the observations, which is often called the data generating mechanism.

Although all examples treated in this thesis contain continuous data, we introduce the GOF problem for discrete as well as for continuous data. The reason is that some test statistics for the continuous case can be formulated as a function of the Pearson's χ^2 statistic applied to grouped or categorised data. In particular, some of the tests that we develop in this thesis will make use of this formulation.

Suppose we are now interested in whether a prespecified model is valid for the given data. In particular, we wish to test if the data is generated by a certain distribution function $F_0(x, \boldsymbol{\beta})$, which is equivalent to testing the null hypothesis

$$H_0 : F(x) = F_0(x, \boldsymbol{\beta}), \text{for all } x \text{ and some } \boldsymbol{\beta} \in \boldsymbol{\Theta} \subset I\!R^p, \tag{3.1}$$

where $\boldsymbol{\beta}$ is either a known or an unknown p-dimensional parameter vector. The expression "for all" means that the equality in (3.1) should hold for all x in $\mathcal{S}$ on which both F and F_0 are defined. For notational comfort, but without loss of generality, we will further write (3.1) as

$$H_0 : F(x) = F_0(x, \boldsymbol{\beta}).$$

The null hypothesis is called *simple* if the distribution $F_0(x, \boldsymbol{\beta})$ is completely specified and unique. This is only the case when $\boldsymbol{\beta}$ is known. On the other hand, if $\boldsymbol{\beta}$ is unknown, the null distribution $F_0(x, \beta)$ represents a parametric family of distributions which contains more than one element and the null hypothesis is therefore called *composite*.

We can also state the null hypothesis of the GOF problem in terms of PDFs instead of CDFs, i.e.

$$H_0 : f(x) = f_0(x, \boldsymbol{\beta}) \tag{3.2}$$

for the continuous GOF problem, while for the discrete GOF problem we use the notation of the probability mass function, i.e.

$$H_0 : p(x) = p_0(x, \boldsymbol{\beta}). \tag{3.3}$$

Before we give more details about the statistical aspects of linear and circular GOF problems, we first note that the classical linear CDFs and corresponding PDFs are clearly not appropriate for circular data. Therefore, in the next section circular distributions are introduced (see also Jammalamadaka and SenGupta (2001)).

Note that in many books about circular statistics, the convention is to use θ as the notation for the direction on the circumference of a circle. In this thesis, we prefer to use x for both types of data and it will be clear from the context which type of data are referred to.

3.1.2 Circular distributions

We first introduce a circular distribution of a continuous random variable X through its PDF, its CDF and its characteristic function and we compare these definitions with the linear case. We consider circular distributions, without loss of generality, on the unit circle, i.e. a circle with unit radius, which is also denoted by $\text{arc}(0, 2\pi)$. Distributions on any other circle can be transformed to the unit circle. At the end of this section, we define the most common circular distributions.

Circular PDF

Let $f(x)$ be a PDF on the line. Then $f(x)$ is assumed to be non-negative and

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{3.4}$$

A circular PDF $f(x)$ is also defined to be non-negative but condition (3.4) becomes

$$\int_{0}^{2\pi} f(x)dx = 1.$$

This is because the total probability mass is concentrated on the circumference of the unit circle. Moreover, a circular f is periodic, i.e. $f(x) = f(x + 2k\pi)$ for any integer k.

Circular CDF

A distribution on the unit circle can also be characterised by its CDF $F(x)$, which clearly depends on the initial direction and the orientation of the circle. Suppose that an arbitrary origin and orientation have been chosen, then $F(x)$ can be defined as

$$F(x) = \int_{0}^{x} f(y)dy \quad \text{for } 0 \leq x \leq 2\pi.$$

By definition, we have $F(0) = 0$ and $F(2\pi) = 1$.

Circular characteristic function and trigonometric moments

A description of the distribution via its characteristic function is more useful in the circular case. As in the linear case, the characteristic function is defined as

$$\phi(t) = \mathrm{E}\left[e^{itX}\right] = \int_{0}^{2\pi} e^{itx} dF(x),$$

but since x and $x + 2\pi$ represent the same direction, we need

$$\mathrm{E}\left[e^{itx}\right] = \mathrm{E}\left[e^{it(x+2\pi)}\right].$$

Hence, the condition

$$e^{it2\pi} = 1 \tag{3.5}$$

must be satisfied for all t for which ϕ is defined. Since (3.5) only holds for integer valued t, the characteristic function $\phi(t)$ of a circular distribution is only defined for $t = 0, \pm 1, \pm 2, \ldots$.

The value of this function at an integer $t = p$ is also called the p^{th}-*trigonometric moment* and is often denoted by

$$\phi_p = \rho_p e^{i\mu_p} = \alpha_p + i\beta_p, \ \ p = 0, \pm 1, \pm 2, \ldots$$

where $0 \leq \rho_p \leq 1$,

$$\alpha_p = \mathrm{E}\left[\cos(pX)\right] = \int_0^{2\pi} \cos(px) dF(x),$$

and

$$\beta_p = \mathrm{E}\left[\sin(pX)\right] = \int_0^{2\pi} \sin(px) dF(x).$$

The first trigonometric moment, $\phi_1 = \rho_1 e^{i\mu_1}$, is simply written as $\rho e^{i\mu}$, where μ and ρ are the population measures for the *mean direction* and the *concentration* towards this mean direction, respectively. The concentration ρ is also referred to as *mean resultant length.* Note that the corresponding sample versions $\overline{X}_c$ and $\overline{R}$ are defined in (2.5) and (2.4), respectively.

The p^{th} *central trigonometric moment* is defined as

$$\mathrm{E}\left[e^{ip(X-\mu)}\right] = \alpha_p^\star + i\beta_p^\star,$$

which is essentially the trigonometric moment about the mean direction μ, where

$$\alpha_p^\star = \mathrm{E}\left[\cos(p(X-\mu))\right]$$

and

$$\beta_p^\star = \mathrm{E}\left[\sin(p(X-\mu))\right].$$

Any linear or circular probability distribution is completely determined by its characteristic function. Consequently, for the circular case, this result is particularly useful in the sense that any circular distribution is thus completely determined by its trigonometric moments. The relation with Fourier series explains this argument. In particular, next to the interpretation in terms of characteristic functions or trigonometric moments, the sequence of complex values $\{\phi_p\}, p = 0 \pm 1 \pm 2, \ldots$ can also be interpreted in terms of Fourier coefficients of $f(x)$. If the circular distribution $f(x)$ is square integrable on $[0, 2\pi]$, i.e.

$$\int_0^{2\pi} f(x)^2 dx < \infty$$

then its Fourier expansion is given by

$$f(x) = \frac{1}{2\pi} \sum_{p=-\infty}^{\infty} \phi_p e^{-ipx} = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{\infty} (\alpha_p \cos(px) + \beta_p \sin(px))\right]. \tag{3.6}$$

Fernández-Durán (2004) proposed a family of flexible distributions based on a truncated version of the expansion in (3.6). It turns out that this finite sum is a nonnegative density only under certain conditions.

It is interesting to note that symmetric distributions about the origin have real characteristic functions and therefore have a Fourier series expansion only in terms of cosine functions, i.e.

$$f(x) = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{\infty} \alpha_p \cos(px)\right].$$

For the linear case, the equivalent result of (3.6) is the inversion formula for characteristic functions of real-valued random variables and states that if $\phi(p)$ is absolutely integrable, then

$$f(x) = \int_{-\infty}^{+\infty} \phi(p) e^{-ipx} dp.$$

Note that here the density is not written in terms of easily interpretable moments.

The most common circular distributions are the circular uniform and the circular normal distribution. There exist many other useful and interesting families of circular models, which are however not discussed in this thesis. In fact, any linear distribution can be wrapped around or projected on the circle. For an overview of general techniques to obtain circular distributions from known linear distributions, we refer to Jammalamadaka and SenGupta (2001).

Circular uniform distribution

The circular analogue to the linear uniform distribution is called the *circular uniform distribution* (CU) and has PDF

$$f_{CU}(x) = \frac{1}{2\pi}, \quad 0 \leq x < 2\pi.$$

It is often called the isotropic distribution, since all directions are equally likely. Moreover, it is the unique distribution which is invariant under rotation.

It follows from (3.6) that for this density all trigonometric moments are zero, except ϕ_0 which is 1, the normalisation constant. The mean resultant length ρ is zero and the mean direction μ is undefined, which means that the density has no preferred direction.

The CDF is

$$F_{CU}(x) = \frac{x}{2\pi}, \quad 0 \leq x < 2\pi. \tag{3.7}$$

Circular normal or von Mises distribution

The most useful distribution for symmetric unimodal samples is the circular normal (CN) or the von Mises (VM) distribution with PDF

$$f_{CN}(x;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(x-\mu)}, \ \ 0 \leq x < 2\pi,$$

where $0 \leq \mu < 2\pi$ and $\kappa > 0$ are nuisance parameters and

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos x} dx \tag{3.8}$$

is the modified Bessel function of the first kind and order zero.

The density is symmetric around its mean direction μ, which is also the mode of the distribution. The antimode, which is located at the opposite direction of the mode, is $\mu + \pi$. The parameter κ gives an indication of the concentration around the mode. The reason for this is that the ratio of the density at the mode to the density at the antimode is uniquely determined by κ. In particular, $f(\mu)/f(\mu+\pi) = e^{2\kappa}$.

In Figure 3.1 the von Mises density with mean direction $\mu = \pi$ is shown for different values of the concentration parameter κ. In the first panel the PDFs are plotted on the circle while in the second panel the PDFs are shown unwrapped on the interval $[0, 2\pi]$.

It can be seen that higher values of κ correspond to higher concentrations towards the mean direction. If $\kappa = 0$, the distribution reduces to the CU distribution and if κ goes to infinity, the distribution approaches a degenerate point mass at μ.

Since the density is symmetric about μ, we have for all $p = 0, \pm 1, \ldots,$

$$\beta_p^\star = \mathrm{E}\left[\sin(p(X-\mu))\right] = 0.$$

Furthermore, we have

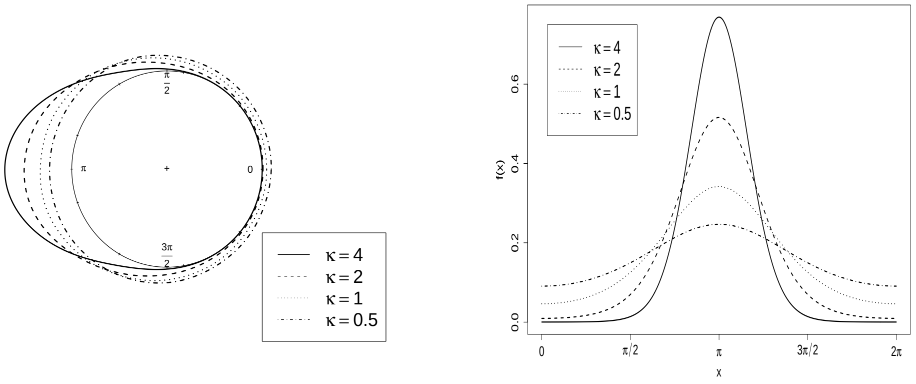
$$\begin{aligned} \alpha_p^\star &= \frac{1}{2\pi}\int_0^{2\pi} \cos(p(x-\mu)) e^{\kappa(\cos(x-\mu))} dx \\ &= \frac{I_p(\kappa)}{I_0(\kappa)}, \end{aligned}$$

where

$$I_p(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \cos(px) e^{\kappa\cos x} dx$$

is the modified Bessel function of the first kind and order p. Consequently, the p^{th}-trigonometric moments are

$$\phi_p = (\alpha_p^\star + i\beta_p^\star)e^{ip\mu} = A_p(\kappa)e^{ip\mu}, \ \ p = 0, \pm 1, \pm 2, \ldots, \tag{3.9}$$



(a) (b)

Figure 3.1: The von Mises density with mean direction $\mu = \pi$ and concentration $\kappa = 0.5, 1, 2$ and 4. Panel (a) shows the densities on the circle, while panel (b) shows the unwrapped versions of the densities on the interval $[0, 2\pi]$.

where $A_p(\kappa)$ denotes $\frac{I_p(\kappa)}{I_0(\kappa)}$. The direction of the first trigonometric moment is μ, as we expected for symmetry reasons, and its resultant length ρ is $A(\kappa) = A_1(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$.

The CDF is often not useful because it has no closed form. It can be obtained by integrating the PDF but is not easy to evaluate. In the remainder of this thesis we use the notation F_{CN} to refer to that CDF.

3.1.3 Additional issues for circular GOF testing

Principle of invariance

For observations measured on the unit circle, an origin and an orientation have to be chosen arbitrarily. This choice should not influence the results of the statistical methods applied to these data. In particular, the value of test statistics for circular data should be invariant under rotation and reflection on the circle so as to be of any practical use.

We consider GOF problems which are invariant under these types of transformations and we explain which restrictions should be imposed on the statistical procedures so that this invariance property is guaranteed. In Chapter 6 of Lehmann and Romano (2005) a general discussion about invariant problems under a group of transformations is given. We now formulate the most important results applied to the group of translations, or, equivalently, the group of rotations.

Suppose M is a function that maps elements of the sample space $\mathcal{S}^n$ onto some subset of $\mathbb{R}^n$. First, we define M as a *maximal origin-invariant* function of the random variables $X_1, \ldots, X_n$ if

(1) it is origin-invariant, and if

(2) $M(X_1, \ldots, X_n) = M(X_1^\star, \ldots, X_n^\star)$ implies $X_i = X_i^\star + c \quad (i = 1, \ldots, n)$ where c is some constant.

The latter represents a change in origin or a rotation. It can be shown that a test statistic $T(X_1, \ldots, X_n)$ is origin-invariant if and only if it depends on $X_1, \ldots, X_n$ only through some maximal origin-invariant function $M(X_1, \ldots, X_n)$, i.e. there exists some function g such that

$$T(X_1, \ldots, X_n) = g(M(X_1, \ldots, X_n)),$$

for all $(X, \ldots, X_n) \in \mathcal{S}^n$. The set of arc length differences

$$(X_1 - X_2, X_2 - X_3, \ldots, X_n - X_{n-1})$$

is an example of a maximal origin-invariant function. These differences are also called *spacings* and they are clearly invariant under rotations. Furthermore, suppose $X_i - X_{i+1} = X_i^\star - X_{i+1}^\star$ for $i = 1, \ldots, n-1$. Let $X_n^\star - X_n = c$, then we have directly that $X_i = X_i^\star + c$, for all $i = 1, \ldots, n$. Tests based on spacings are therefore directly applicable to circular data. For more details about this class of tests we refer to Section 3.6. Furthermore, each rotationally invariant statistic that is useful for circular data, can be expressed in terms of spacings. This is however not always an easy task. Moreover, doing this is not worth the effort because other maximal invariant functions can be used as well. For example the Kuiper and the Watson tests, which are described in Section 3.4, are based on another maximal origin-invariant function.

In Section 4.1 we use a particularly interesting maximal origin-invariant function given by the set of arc differences

$$(X_1 - \overline{X}_n^c, X_2 - \overline{X}_n^c, \ldots, X_n - \overline{X}_n^c), \tag{3.10}$$

where $\overline{X}_n^c$ is the estimator of the circular mean direction defined in Section 2.2. The proof of its maximal invariance is similar as for the set of spacings.

It is important to note that the property of origin-invariance is no guarantee for the test statistic to be meaningful. In particular, it is always possible to construct an origin-invariant statistic from a test statistic which is originally meant to solve linear GOF problems. For example, if we compute the original test statistic for the centered observations in (3.10) instead of applying it directly to the original observations, we end up with an origin-invariant test statistic, which is however not necessarily useful.

Probability integral transformation

As in the linear case, testing the GOF null hypothesis for a completely specified and continuous circular distribution can be translated into testing for circular uniformity via the probability integral transformation (PIT). Suppose that $X_1, \dots, X_n$ are i.i.d. from $F_0(x, \boldsymbol{\beta})$ and consider the simple GOF null hypothesis test as in (3.1), i.e. $H_0 : F(x) = F_0(x, \boldsymbol{\beta})$, where F_0 is continuous and $\boldsymbol{\beta}$ is known. Then this problem reduces to a test for uniformity on the unit circle, with null hypothesis

$$H_0 : F(u) = \frac{u}{2\pi}, \quad 0 \leq u < 2\pi \tag{3.11}$$

for which we use the transformed sample $u_i = 2\pi F_0(x_i, \boldsymbol{\beta})$, $i = 1, \dots, n$. Hence, any completely specified GOF problem can be transformed to the problem of testing for circular uniformity. This means that once the GOF problem for circular uniformity is solved, any other simple GOF problem can be solved as well. However, in most practical situations the distribution under the null hypothesis is not completely specified. This is often the case, for instance, when we are interested in testing the null hypothesis of circular normality, where the parameters μ and κ are unknown and have to be estimated from the data. In contrast to the simple GOF problem on the unit circle, only few GOF tests for composite null hypothesis on the unit circle have been described in the literature. We refer to Sections 3.2, 3.3, 3.4 and 3.6 for the description of the relevant tests on the unit circle for simple as well as for composite null hypotheses.

3.1.4 Additional issues for general GOF testing

Parameter estimation and nuisance parameters

In the case of testing for composite null hypothesis, we often need to estimate unknown parameters from certain families of parametric models. For example, applying a nonparametric test for circular normality $F_0(x; \mu, \kappa)$ involves estimation of the unknown parameters μ and κ within the family of CN distributions. In general, in order to identify the appropriate member of the parametric family $F_0(x, \boldsymbol{\beta})$ for the composite null hypothesis, the unknown parameter vector $\boldsymbol{\beta}$ has to be estimated from the data. However, these parameters are not of primary interest to the researcher and are therefore referred to as *nuisance parameters*. In the presence of nuisance parameters, the null distribution of the test statistic is often more difficult to find. One usually resorts to simulations to obtain the required critical values.

Omnibus and directional tests

Suppose that the sample size is sufficiently large. If the alternative hypothesis to the GOF problem in (3.1) is the complement of the null hypothesis, i.e.

$$H_1 : F(x) \neq F_0(x, \boldsymbol{\beta}),$$

for at least one $x \in \mathcal{S}$, the test is called an *omnibus* test if it is consistent for testing against H_1. For instance, take as null hypothesis $F_0 = F_{CU}$, as in (3.7), then an omnibus test should be sensitive to all distributions different from the CU distribution. On the other hand, if the alternative hypothesis is a smaller subset of the complement of the null hypothesis, e.g. $H_1 : F = F_{CN}$, then an appropriate test is called *directional* and will therefore only have high probability to reject the null hypothesis if the data follow that specific alternative.

Omnibus tests are consistent against any deviation from the null hypothesis. This seems to be an appealing property since this makes it likely that all possible deviations can be detected given that the sample size is sufficiently large. However, a drawback of this type of tests is that in case the null hypothesis is rejected, we may have no idea in what way the true distribution differs from the distribution specified in H_0. Moreover, directional tests usually have higher power towards the specific alternative for which they are constructed. On the other hand, if the true distribution is not that particular alternative, the directional test may have negligible power. Smooth tests, which are described in Section 3.3, can be seen as a compromise between these two types of tests. Essentially, the individual components of the smooth test statistic provide information on particular alternative directions and their sum results in a statistic which has omnibus features.

Goodness-of-fit versus lack-of-fit

Until now, we used the terms *statistical model* and *distribution* interchangeably. In the literature, fitting a statistical model has often a more general interpretation than fitting a density curve to a sample of observations generated by one random variable.

As an example of a statistical model, consider a traditional linear regression model, where the distribution of a response variable Y_i is assumed to depend on an explanatory variable X_i through the simple linear regression equation

$$Y_i = \mu + \beta X_i + \epsilon_i, i = 1, \ldots, n, \tag{3.12}$$

where ϵ_i are assumed to be independently normally distributed with mean 0 and constant variance, i.e. $\epsilon \sim N(0, \sigma^2)$ for some unknown σ.

From (3.12), the conditional mean of the response variable Y given the value of the predictor X is

$$\mathrm{E}\left[Y|X\right] = \mu + \beta X. \tag{3.13}$$

For more details about parametric linear regression models we refer to, for example, Neter, Kutner, Nachtsheim, and Wasserman (1996) or Draper and Smith (1998). In this context, it is important to know to what extent the proposed mean model in (3.13) fits the observed data well. Indeed, if the estimated parametric model turns out to be a poor fit, then inferences made using that model can be misleading. Statistical tools for these kind of model checks are called *lack-of-fit* (LOF) methods. Hart (1997) gives an extensive overview of nonparametric tests for LOF of a parametric regression model.

Returning to model representation (3.12), the LOF problem for statistical models poses essentially the same question as formulated in the GOF problems for distributions. Moreover, sometimes a statistical model can also be seen as a problem of fitting a distribution to the data. For example, the problem in (3.12) is equivalent to the distributional fit given by

$$Y_i \sim N(\mu + \beta X_i, \sigma^2). \tag{3.14}$$

Testing whether (3.14) is a good description of the data, conditional on the covariates $X_1, \ldots, X_n$, is essentially a GOF problem, but with a particular parameterisation of the mean, involving extraneous variables.

Despite this close relation between statistical models and distributions, the term LOF is usually preferred over GOF when assessing the quality of statistical models.

Because we look at distributions as statistical models, we use the term LOF next to GOF. However, we make some nuance when we use these terms. The term GOF is used when we want to emphasize the formal statistical decision making using hypothesis tests. On the other hand, tools for LOF refer in this thesis to more informative statistical analyses that aim at understanding the deviation from the null hypothesis. For example, a LOF method may be designed to locate in which subset of the sample the true distribution deviates strongly from the hypothesised distribution.

So far, we have discussed some general concepts concerning GOF testing procedures for linear and circular data. The next three sections are devoted to three important classes of nonparametric tests for GOF, namely Pearson's χ^2 tests, smooth tests and EDF tests.

3.2 Pearson's χ^2 GOF tests

More than a century ago, Pearson (1900) introduced the first GOF test, which has probably been one of the most frequently used statistical tests. Even though this test is in principle only applicable to discrete data, it has also frequently been used for continuous data. Since continuous data have to be discretised first, using Pearson χ^2 test results in loss of information. Therefore, other tests have since been developed for continuous data, which usually have better performance in terms of power. Nevertheless, in this section we thoroughly describe this test as we need it in the construction of our new class of tests in Chapter 5. The discussion will be restricted, however, to simple null hypotheses. For composite null hypotheses, the test is often referred to as the Pearson-Fisher test and was first proposed by Fisher (1924). After constructing Pearson's test for simple null hypotheses, we give some possibilities for applying Pearson's χ^2 tests to circular data.

3.2.1 An illustration of the original construction of Pearson's test through Mendel's data

Pearson's χ^2 test was one of the foundations of modern statistics. Moreover, it resulted from the aim to answer the much-discussed question about Mendel's inheritance theory, which is in turn the foundation of modern genetics. Mendel investigated these basic elements of genetics through a large observational study on peas. He stated in his *law of segregation* that each organism has two genes for each trait. The different forms of a gene are called alleles. The two alleles determine the genotype. When both alleles are present, one allele may mask or hide the other. The characteristic that is expressed is called the phenotype. The stronger allele is said to be dominant, and the weaker allele, which is masked, is said to be recessive. When expressing dominant and recessive alleles, the dominant allele is by convention written as a capital letter, and the recessive allele as the same letter, but lower case.

In one of his experiments, Mendel observed 556 peas, classified according to shape (either round (R) or angular (a)) and color (either yellow (Y) or green (g)). Round and yellow are dominant, meaning that if these alleles are present, they will be expressed. Hence, according to Mendel's law, which assumes also that all combinations of genes are equally likely, the following 16 genotypes have equal probability:

RRYY	RRYg	RRgY	RRgg
RaYY	RaYg	RagY	Ragg
aRYY	aRYg	aRgY	aRgg
aaYY	aaYg	aagY	aagg.

Table 3.1: Observed and expected counts from Mendel's experiment of 556 peas.

class	1	2	3	4
phenotype	RY	Rg	aY	ag
observed counts	315	108	101	32
expected counts	312.75	104.25	104.25	34.75

However, since phenotypes are determined by dominant genes, we expect the four phenotypes RY, Rg, aY and ag to occur with probabilities $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$ and $\frac{1}{16}$, respectively.

The observed number of peas for the four phenotypes obtained from Mendel's experiment are in Table 3.1. Using elementary probability theory, the expected frequencies under the null hypothesis that Mendel's law is correct are denoted as $E_i = n\pi_{0i}$, $i = 1, \ldots, 4$, where $n = 566$ is the total sample size and π_{0i} is the probability for the peas to be classified in class i. These values are presented in Table 3.1 as well. Now the question is how we should formally check whether or not these data fit Mendel's expectation. In other words, are the expected counts in agreement with the observed counts? In general, Pearson's statistic is formulated as a kind of distance measure between expected and observed counts,

$$X_n^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{3.15}$$

where k is the number of classes and O_i and E_i are the observed and expected frequencies in class i, respectively. To see that X_n^2 is essentially a GOF statistic we consider Pearson's original formulation of the problem. Observe that the frequencies for the four phenotypes, denoted by $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$, follow a multinomial distribution given by

$$P(\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\pi}) = \frac{n!}{\pi_1!\pi_2!\pi_3!\pi_4!} {\pi_1}^{x_1} {\pi_2}^{x_2} {\pi_3}^{x_3} {\pi_4}^{x_4}, \tag{3.16}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ is the vector of probabilities corresponding to the four categories. If Mendel's law is correct these probabilities are equal to $\boldsymbol{\pi_0} = (\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$. Thus, the discrete GOF problem with null hypothesis as in (3.3), is equivalent to testing the simple null hypothesis about the parameters $\boldsymbol{\pi}$ from the multinomial distribution in (3.16), i.e.

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0. \tag{3.17}$$

Pearson's statistic can be written as

$$X_n^2 = \sum_{i=1}^{k} \frac{(X_i - n\pi_{0i})^2}{n\pi_{0i}}. \tag{3.18}$$

Pearson proved that X_n^2 has asymptotically a chi-squared distribution with $k-1$ degrees of freedom under H_0, that is, as $n \to \infty$,

$$X_n^2 \xrightarrow{d} \chi_{k-1}^2.$$

For Mendel's data the observed value of the statistic is $X_n^2 = 0.74$. Using the asymptotic null distribution χ_3^2, the p-value becomes 0.9254, which means that the null hypothesis is accepted. It is interesting to note here that doubts about Mendel's results have been raised because of the fact that the p-value seems unrealistically large. In fact, there has been a lot of controversy regarding the reliability of Mendel's observations (see e.g. Magnello, 1998).

Note that the asymptotic distribution is generally regarded a good approximation as soon as the expected count E_i in each of the k classes is larger then 5 (see e.g. Lancaster, 1969).

As Pearson's test is originally constructed for categorical data or discrete data, an important question is how to deal with grouped continuous data. Greenwood and Nikulin (1996) thoroughly studied this issue, but nowadays these guidelines are of less practical importance since other more appealing techniques can be applied to continuous data as will be seen further in this thesis.

3.2.2 Pearson's test for circular data

In Section 3.1.3 we have explained that a test statistic should be independent of its arbitrarily chosen origin for being meaningful for circular data. One of the questions that arises is whether the Pearson χ^2 in its original form is useful for circular data.

To answer that question, consider first a discrete circular distribution. For example, suppose we want to describe the distribution of the stopping position of the ball on a roulette wheel. If the wheel is unbiased, the distribution of the stopping position of the ball is a discrete uniform distribution on 37 equally spaced points on a circle. In general, for the discrete GOF problem (3.3) on the circle, the Pearson χ^2 statistic is origin-invariant since observed and expected frequencies would not change if another origin is taken. Therefore, one can apply the test without modification.

Discrete data also arise from grouping continuous data. In the simplest case, the partition used for the grouping does not depend on the data, in the sense that it is fixed beforehand. For example, if dates of birth are considered, the data can be presented on a circle with a period of one year and partitioned into twelve classes, each corresponding to one month. Since the partition is fixed, the test statistic is rotation invariant, and thus it makes sense to apply Pearson's test to this type of data.

On the other hand, for continuous distributions in general, Pearson's χ^2 test clearly depends on how the cells are chosen when grouping the data. Since this choice in turn depends on the choice of the origin, the test is not directly applicable to circular data.

Modifications to tests for uniformity in order to make them origin-invariant include considering the maximum or the average statistic for all possible partitions of a certain form. Rao (1972) suggested an average-type χ^2 test. Suppose n observations on the circumference are partitioned into k classes of equal length. Note that this partition depends on the origin x_0 and has the form $C_{x_0} = \big\{[x_0, x_0 + \frac{2\pi}{k}[, [x_0 + \frac{2\pi}{k}, x_0 + 2\frac{2\pi}{k}[, \ldots, [x_0 + (k-1)\frac{2\pi}{k}, x_0 + 2\pi[\big\}$. The number of classes k in this partition is fixed for the construction of Rao's test. The observed cell counts $X_i(x_0)$, $i = 1, \ldots, k$ are then computed as the number of observations within the i^{th} arc in the partition C_{x_0}. Under the null hypothesis of circular uniformity, the expected frequencies are all equal to $\frac{n}{k}$. Pearson's statistic becomes

$$X_n^2(x_0) = \sum_{i=1}^{k} \frac{(X_i(x_0) - \frac{n}{k})^2}{\frac{n}{k}}, \tag{3.19}$$

which clearly depends on the origin x_0. Moreover, it also depends on the particular choice of grouping as is the case on the line. In order to make this statistic independent of x_0, Rao (1972) proposed to integrate x_0 out, i.e.

$$X_n^2 = \frac{1}{2\pi} \int_0^{2\pi} X_n^2(x_0) dx_0. \tag{3.20}$$

He determined the asymptotic null distribution of (3.20) and a computational formula based on arc lengths between observations. These arc lengths between observations on the circle are often called *spacings*. To be rotation or translation invariant, any test on the circle should have an expression in terms of those spacings (Lehmann & Romano, 2005) (see also Section 3.1.3). Taking $k = 2$, the integral statistic (3.20) reduces to Anje's statistic (Anje, 1968)

$$R_n = \frac{1}{2\pi n} \int_0^{2\pi} (N(x_0) - \frac{n}{2})^2 dx_0, \tag{3.21}$$

where $N(x_0)$ denotes the number of observations in the arc$(x_0, x_0 + \pi)$, which is the semi-circle that starts at x_0. The computational formula for Anje's statistic is given by

$$R_n = \frac{n}{4} - \frac{1}{2n\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} d_0(x_i, x_j),$$

where $d_0(x_i, x_j) = \min(|x_i - x_j|, 2\pi - |x_i - x_j|) = \pi - |\pi - |x_i - x_j||$ denotes the circular distance between two points on the circle. The asymptotic null distribution of (3.21) was obtained by Watson (1967). Additionally, Anje (1968) considered the maximum version of (3.21), which is given by $B_n = \max_{x_0} N(x_0)$ or equivalently $n - B_n = \min_{x_0} N(x_0)$. This test, for which Anje (1968) found the exact and asymptotic null distributions, is often referred to as the Hodges-Anje test. Batschelet (1981) lists critical levels of $n - B_n$ for sample sizes n from 5 to 40. Rothman (1972) proposed an extension to Anje's tests, replacing $N(x_0)$ by $N(x; x_0)$, which denotes the number of observations in the arc$(x_0, x_0 + 2\pi x)$ for $0 \leq x \leq 1$. The statistic becomes

$$R_n(x) = \frac{1}{2\pi n} \int_0^{2\pi} (N(x; x_0) - nx)^2 dx_0. \tag{3.22}$$

Even though these tests are invariant to the choice of the origin, they still depend on the choice of the partition. As mentioned above, these issues are similar to those discussed for the original Pearson's χ^2 test on the line in Greenwood and Nikulin (1996).

To compensate partially for this dependence, Rothman (1972) generalised his $R_n(x)$ test by integrating over all the possible partitions with two cells. The most general form of his statistic is then given by

$$R_n^H = \frac{1}{2\pi n} \int_0^{2\pi} \int_0^{2\pi} (N(x; x_0) - nx)^2 dx_0 \, dH(x), \tag{3.23}$$

where $H(x)$ is an arbitrary distribution function on $[0, 2\pi]$ that may be interpreted as weight function in a mixture of $R_n(x)$ statistics. Rothman's test is now consistent against all alternatives to circular uniformity and reduces to the Watson statistic (see Section 3.4.3 below) when $H(x)$ is the CU distribution, which corresponds to each partition receiving equal weight. This link is further explored in Section 3.5.3 and Section 5.8 and a similar link for GOF tests for linear distributions is in Section 5.2.

Example 3.2.1. We apply the Hodges-Anje test for uniformity to the Homing pigeons data from Section 2.2.2. This test can be quickly performed without any calculation. It is easy to derive the value of the statistic just by looking at the plotted data. In particular, rotating the diameter around the center of the circle (see Figure 2.8), it can readily be seen that five is the minimum number of observations lying on one side of the diameter. This value corresponds to a p-value of 0.873, which means that a significant deviation from uniformity can not be established. However, since the data can be interpreted as axial data, the test can be performed on the doubled data (as explained in Section 2.2.2), which then leads to significance (p=0.003).

3.3 Smooth GOF tests

While Pearson is considered the founder of GOF tests, Neyman is often cited as the founder of smooth GOF tests. Neyman (1937) introduced his smooth test for uniformity and argued that any other completely specified continuous distribution could be handled using the PIT (cf. Section 3.1.3). However, when applying that transformation, the interpretation of the test is in terms of the transformed distribution instead of the original one. Therefore, we prefer to use the construction of the smooth test as in Rayner and Best (1989), which is in terms of the original distribution. Moreover, in contrast to Neyman (1937), Rayner and Best (1989) derived the smooth test as a score test.

To obtain that score test, a k-dimensional *smooth* family of alternatives is constructed which contains the hypothesised distribution. The term "smooth" refers to the fact that the alternatives differ "smoothly" from the hypothesised distribution. Neyman's test is *optimal* for this type of alternatives, where *optimal* in Neyman's sense means asymptotically locally uniformly most powerful symmetric.

Section 3.3.1 is devoted to the construction of the family of smooth alternatives and the derivation of the corresponding smooth test in his most general form, i.e. without specifying whether nuisance parameters are known or unknown. In Sections 3.3.2 and 3.3.3, more details about the smooth tests are outlined for simple and composite null hypotheses, respectively. The order of the smooth family is important for the power of the test and can therefore better be estimated from the data. These data-driven procedures are discussed in Section 3.3.4. Since in this thesis we are also interested in GOF tests for circular data, we devote Section 3.3.5 to a smooth test for circular distributions.

3.3.1 General construction

Consider again the continuous GOF problem $H_0 : f(x) = f_0(x, \boldsymbol{\beta})$ and embed the hypothesised distribution $f_0(x, \boldsymbol{\beta})$ in an order k family of smooth alternatives given by the density

$$g_k(x; \boldsymbol{\theta}, \boldsymbol{\beta}) = C(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left[\sum_{j=1}^{k} \theta_j h_j(x; \boldsymbol{\beta})\right] f_0(x; \boldsymbol{\beta}), \tag{3.24}$$

where $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_k)$ denotes a k-dimensional real parameter vector, $C(\boldsymbol{\theta}, \boldsymbol{\beta}) = \left(\int_{-\infty}^{\infty} g_k(x; \boldsymbol{\theta}, \boldsymbol{\beta}) dx\right)^{-1}$ is the normalising constant, and $\{h_j; j = 1, \ldots, k\}$ represents a set of functions orthonormal on the real line with respect to the density $f_0(x;\boldsymbol{\beta})$, i.e.

$$\int_{-\infty}^{+\infty} h_l(x;\boldsymbol{\beta}) h_m(x;\boldsymbol{\beta}) f_0(x;\boldsymbol{\beta}) dx = \delta_{l,m}, \tag{3.25}$$

$(l, m = 1, \ldots, k)$ where $\delta_{l,m}$ is the Kronecker delta. If the set of functions $\{h_j; j = 1, \ldots\}$ forms a *complete orthonormal basis* of functions with respect to f_0, the family in (3.24) contains every possible continuous density function provided that the order k may grow infinitely large. Additionally, the orthonormality condition in (3.25) has the advantage that the resulting score statistic, which will be given soon, can often be written into a sum of asymptotically independent components which are easily interpretable. Moreover, each of the individual components is related to each of the individual parameters from the smooth model, which therefore can be uniquely identified.

Testing for the hypothesised distribution $f_0(x;\boldsymbol{\beta})$ now reduces to testing $H_0 : \boldsymbol{\theta} = 0$. Note that the latter is essentially a parametric null hypothesis. The reason is that the true density is assumed to be a member of (3.24), which is a parametric family of densities as the order k is finite. Hence, any test for that null hypothesis is strictly speaking a parametric test. This implies that if these parametric assumptions are not fulfilled, the test will lose power. Even in case the k-dimensional vector $\boldsymbol{\theta}$ truly equals zero, the true distribution may not be equal to the hypothesised $f_0(x, \boldsymbol{\beta})$. In particular, this happens if the true distribution is a member of an m-order smooth family where $m > k$, $(\theta_1, \ldots, \theta_k) = \mathbf{0}$ and $(\theta_{k+1}, \ldots, \theta_m) \neq \mathbf{0}$.

Suppose $X_1, \ldots, X_n$ denotes a sample of i.i.d. observations which have density f_0 under the null hypothesis. Denote $\boldsymbol{h}(x;\boldsymbol{\beta})^T = (h_1(x;\boldsymbol{\beta}), \ldots, h_k(x;\boldsymbol{\beta}))$, so that the score vector for $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta} = 0$, can be written as

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ln g_k(X_i; \boldsymbol{\theta}, \boldsymbol{\beta}) \bigg|_{\boldsymbol{\theta}=\mathbf{0}} = \sum_{i=1}^{n} \boldsymbol{h}(X_i;\boldsymbol{\beta}).$$

Then, the score test statistic for testing $H_0 : \boldsymbol{\theta} = 0$ versus $H_1 : \boldsymbol{\theta} \neq 0$ in (3.24) has the form

$$S_k = \frac{1}{n} \boldsymbol{V}_{\boldsymbol{\beta}}^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{V}_{\boldsymbol{\beta}}, \tag{3.26}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the asymptotic covariance matrix of $\frac{1}{\sqrt{n}} \boldsymbol{V}_{\boldsymbol{\beta}}^T$, evaluated at $\boldsymbol{\theta} = 0$. The score statistic in (3.26) is asymptotically χ^2 distributed under the null hypothesis with degrees of freedom depending on the number of unknown nuisance parameters. Similarly, the explicit computation of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ depends on the nature of the parameters in $\boldsymbol{\beta}$, and the estimation method. Further details about those computations are treated separately for the simple and the composite null hypothesis in the next sections.

Barton (1953) proposed a different smooth family of alternatives, defined as

$$g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = \left[1 + \sum_{j=1}^{k} \theta_j h_j(x;\boldsymbol{\beta})\right] f_0(x;\boldsymbol{\beta}), \tag{3.27}$$

which has the advantage that no normalisation constant is needed. This parameterization is sometimes preferred because the orthonormality properties in (3.25) can be directly applied to compute the moments of the distributional family (see Hamdam, 1962). In this perspective, an estimate of (3.27) is useful to approximate the true density. This is because, as we will see in Section 3.7, estimates of the θ-parameters are easily derived from the components in the smooth test statistic. On the other hand, the drawback is that the density is not always guaranteed to be positive. Methods for correcting non-positive densities are proposed in e.g. Gajek (1986) and Glad and Hjort (2003).

The two families of alternatives in (3.24) and (3.27) are referred to as the *Neyman* and the *Barton model*, respectively. The score statistic for the Barton model is equal to that of the Neyman model and thus also given by (3.26).

3.3.2 Simple null hypothesis

In this section, we discuss the smooth test based on (3.26) for the simple null hypothesis, and thus the distribution under the null hypothesis, $f_0(x;\boldsymbol{\beta})$ is completely specified. Since this means that the parameter $\boldsymbol{\beta}$ is known, we omit this parameter in the next part of this section.

Using the orthonormality properties of the h-functions, the covariance matrix evaluated at $\boldsymbol{\theta} = 0$ of the score vector $\boldsymbol{V}$ in (3.26) is easily computed to be nI_k, where I_k is the k-dimensional identity matrix. The score statistic in (3.26) for testing $H_0 : \boldsymbol{\theta} = 0$ thus simplifies to

$$S_k = \frac{1}{n} \sum_{j=1}^{k} V_j^2. \tag{3.28}$$

Under the null hypothesis, as $n \to \infty$,

$$S_k \xrightarrow{d} \chi_k^2. \tag{3.29}$$

Since the terms $\frac{1}{n}V_j^2$, $j = 1, \ldots, k$ in (3.28) are the k asymptotically independent components of the test statistic, each of them can be used as a *directional* test. The interpretation of those components depends on the choice of the orthonormal system $\{h_j\}$. For example, if f_0 is the uniform density on $[0, 1]$, an orthonormal system based on trigonometric functions can be used (e.g. $h_j(x) = \sqrt{2}\cos(j\pi x)$). For this basis, which is however not a complete basis, a large value of the component $\frac{1}{n}V_m^2$ for some m indicates a symmetric oscillating deviation from uniformity with period $\frac{2}{m}$. On the other hand, using orthonormal polynomials, the interpretation becomes one in terms of moment deviations. In particular, large values of $\frac{1}{n}V_m^2$ then indicate large deviations from f_0 in the m-th moment. For the example of uniformity, the set of normalised Legendre polynomials, which are listed in Appendix A.2, is the appropriate basis to obtain that interpretation.

Henze and Klar (1996), Henze (1997) and Klar (2000) argued that the directional interpretation of the component tests is only guaranteed when at most one $\theta_j \neq 0$ in (3.24) or (3.27). The reason is that $\mathbf{\Sigma}$ is calculated under the full parametric null hypothesis, which means that f and f_0 are equal in all moments. The latter is not necessarily true under the null hypothesis of $\boldsymbol{\theta} = 0$ corresponding to the smooth test. Note that even when the null hypothesis is not true, it is still possible to have more than one $\theta_j \neq 0$. To repair the directional property, they proposed to rescale the score statistic, not by $\mathbf{\Sigma}$, but by its empirical covariance matrix

$$\mathbf{\Sigma}_{emp} = \frac{1}{n^2}\sum_{i=1}^{n} \boldsymbol{h}(X_i)\boldsymbol{h}^T(X_i).$$

The test statistic now becomes

$$S_k^{emp} = \frac{1}{n}\mathbf{V}^T\mathbf{\Sigma}_{emp}^{-1}\mathbf{V}$$

and its asymptotic null distribution is as before. It should be noted that the convergence of S_k^{emp} is very slow as compared to the convergence of S_k. So in practice it is useful to obtain the null distribution of S_k^{emp} by simulation.

In this thesis we take Henze and Klar's argument into consideration, in the sense that when the traditional smooth test S_k shows a significant result, we are cautious about its interpretation. On the other hand, we also take into account that from empirical studies in Rayner and Best (1989) this traditional approach seems to be quite good and informative for many distributions.

Example 3.3.2. We choose to take an order four alternative to perform the Neyman test for uniformity using the Legendre polynomials on the Birth time data (see Section 2.2.1). Note that the choice of the order is subjective. Later, in Section 3.3.4, we mention how we can avoid making such a subjective choice. The value of the statistic is $S_4 = 3.87$, which is not significant using the asymptotic χ^2 critical points (p-value=0.42). In general, even if the result of the test is not significant, it may still be interesting to have a closer look at the individual components. After all, the non-significance could be due to a sample size which is too small, rather than to the null hypothesis actually being correct. The individual components can then inform us about what type of deviation from uniformity is most likely and should be looked for, if further analysis is planned. The individual components $\frac{1}{n}V_2^2 = 1.80$ and $\frac{1}{n}V_4^2 = 2.01$ sum up to 3.81, which amounts to 98% of the overall value of the statistic. This indicates that, if the true distribution deviates from uniformity, this might be due to the second and fourth moments, i.e. variance and kurtosis. On the other hand, we should be careful, since this is in fact circular data and the smooth test used here is not invariant to a change in the origin. We demonstrate this by computing the same test statistic again, but now taking the origin at 11pm instead of the original choice of 12am. Then the overall statistic $S_4 = 6.12$ (p-value=0.19) is considerably higher and the interpretation of the individual components $\frac{1}{n}V_1^2 = 0.50$, $\frac{1}{n}V_2^2 = 2.15$, $\frac{1}{n}V_3^2 = 3.03$ and $\frac{1}{n}V_4^2 = 0.44$ change as well. Here, the second and third moments are the most important ones. In Section 3.3.5 and Chapter 4, we explain how origin-invariant smooth tests can be constructed, which are suitable for circular data.

3.3.3 Composite null hypothesis

In the previous section, the p-dimensional nuisance parameter $\boldsymbol{\beta}$ in the GOF problem was assumed to be known. However, in most practical situations this index parameter vector is not known in advance. Usually the unknown parameters are estimated from the data using maximum likelihood or the method of moments. As we will explain later, especially for smooth tests, it is convenient to replace unknown parameters with their maximum likelihood estimators. Suppose $X_1, \ldots, X_n$ is a random sample from $f_0(x, \boldsymbol{\beta})$. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ of the p-dimensional vector $\boldsymbol{\beta}$ is a solution of the set of p estimation equations

$$\sum_{i=1}^{n} \boldsymbol{b}(X_i) = \sum_{i=1}^{n} \frac{\partial \ln f_0(X_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}, \tag{3.30}$$

where $\boldsymbol{b}$ is the score function evaluated under H_0. For testing a composite null hypothesis, the efficient score statistic is needed. Then the efficient score test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ in (3.24) is defined as

$$S_k = \frac{1}{n} \boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^T \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}^{-1} \boldsymbol{V}_{\hat{\boldsymbol{\beta}}}, \tag{3.31}$$

where $\boldsymbol{V}_{\boldsymbol{\beta}}$ is the efficient score vector, which is given by

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \sum_{i} \{\boldsymbol{h}(X_i; \boldsymbol{\beta}) - \mathrm{Cov}\,[\boldsymbol{h}, \boldsymbol{b}]\, \mathrm{Var}\,[\boldsymbol{b}]^{-1}\, \boldsymbol{b}(X_i)\}. \tag{3.32}$$

Since we are working with MLE, the efficient score vector is as before with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$, i.e. $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \sum_{j=1}^{n} \boldsymbol{h}(X_j; \hat{\boldsymbol{\beta}})$, but the covariance matrix of $\frac{1}{\sqrt{n}} V_{\hat{\boldsymbol{\beta}}}$ is no longer necessarily diagonal. In particular,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \left[I_k - \mathrm{Cov}_0\left[\boldsymbol{h}, \boldsymbol{b}\right] \mathrm{Var}_0\left[\boldsymbol{b}\right]^{-1} \mathrm{Cov}_0\left[\boldsymbol{b}, \boldsymbol{h}\right] \right], \tag{3.33}$$

where the index 0 in the Cov and Var operator refers to calculations under the null hypothesis. Under the null hypothesis, as $n \to \infty$, we have $S_k \xrightarrow{d} \chi^2_{k-p}$.

If the density f_0 belongs to an exponential family (e.g. the normal distribution), the covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ reduces to a diagonal form. In this case, the sufficient statistics for the parameters of the exponential density are polynomials in the observations. Klar (2000) showed that then the MLE is equal to the method of moments estimator (MME). The MME is obtained by expressing equality between theoretical and sample moments. It is particularly useful to have that the MLE coincides with the MME, because that means that the first p elements of the score vector are identically zero and the score test statistic reduces to a sum of $k-p$ asymptotically independent χ^2_1 distributed components, i.e.

$$S_k = \frac{1}{n} \sum_{j=p+1}^{k} V_{\hat{\boldsymbol{\beta}}_j}^2. \tag{3.34}$$

To obtain directional tests that are diagnostic for the deviation from the hypothesised distribution in its moments, the covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ can again be replaced by the corresponding empirical covariance matrix. However, for the composite null hypotheses, Klar (2000) demonstrates that the directional property only holds when the MLE equals the MME and when the sample size is large.

Since the null hypothesis includes all distributions of a p-dimensional family of distributions, an efficient statistic compares that null distribution with its orthogonal complement, which is a $(k-p)$-dimensional alternative. By imposing p constraints on the parameter space, the degrees of freedom of the test naturally reduce to $k-p$. Therefore, for any f_0, it is always possible to write the score test statistic in (3.31) as a sum of $k-p$ asymptotically independent components. This is done by diagonalising the non-singular covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$.

Example 3.3.3. We apply the smooth test for composite normality to the PCB data (see Section 2.1.3). Since for the normal distribution MLE and MME coincide, the smooth test statistic for a family of alternatives of order $k = 6$ reduces to the sum of $k - 2 = 4$ asymptotically independent χ^2_1 distributed components. Because the convergence to the asymptotic distribution is slow, we apply the parametric bootstrap with 10,000 samples to obtain the null distribution of S_6 and its components. Since the normal distribution is location-scale invariant, which is a distribution that satisfies $f(x;\mu,\sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}, 0, 1)$, the null distribution of S_6 does not depend on the values of the parameters μ and σ^2. This means that the observations may first be standardised as $Z_i = \frac{X_i - \overline{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}}$ so that the normalised Hermite polynomials may be used as a set of orthonormal functions to the standard normal distribution. The Hermite polynomials are given in Appendix A.1. Simulations should thus only be performed for the standard normal distribution. We obtain $S_6 = 10.18$ (p-value $= 0.030$), where the first two non-zero components are $\frac{1}{\sqrt{n}} V_3 = 5.44$ (p-value=0.019) and $\frac{1}{\sqrt{n}} V_4 = 4.12$ (p-value=0.025) and sum up to almost 95% of the total value of the test statistic. Hence, at the 5% significance level there is evidence for differences with respect to skewness and kurtosis but not to higher order moments. Note that we did not apply the rescaling method of Klar (2000), so we need to be careful with the diagnostic interpretation. However, relying on the empirical studies by Rayner and Best (1989), this interpretation seems justified.

Finally, we note that other smooth tests were proposed to deal with the GOF problem of composite normality. First, as mentioned before, it is possible to apply the PIT and use Legendre polynomials in a smooth test for uniformity on the transformed observations. The interpretation will then be in terms of the CDF instead of the PDF. Another contribution was due to Thomas and Pierce (1979), who did not use orthonormal polynomials but monomials to construct a smooth test for composite normality. The authors proved that the limiting distribution was simply a χ^2-distribution. However, the advantages of using the orthonormal polynomials formulated in Section 3.3.1 are lost with this approach.

3.3.4 Data-driven smooth tests

The problem is now how to choose the number of the components in Neyman's test. This choice is particularly important because the smooth test loses power if the order is either too small or too large.

Since Neyman's test is obtained as a score test for the parametric null hypothesis $\boldsymbol{\theta} = \mathbf{0}$ in the exponential family in (3.24), this test is asymptotically optimal against alternatives within this exponential family. However, if the true density f does not belong to that family of alternatives, in particular when the dimension of the proposed family is too small to capture the density f, the smooth test is not consistent anymore and loses power. Note that any "regular" density f can be represented by a log-linear expansion of infinite dimension. Hence, we may assume that there exists an order k which is large enough so that f belongs to this smooth exponential family.

On the other hand, the order can be chosen too large as well, resulting in a so called *dilution effect*. Suppose the true distribution differs from the hypothesised only with respect to its location, then a smooth test with order 1 is the optimal choice for the detection of the location shift. If instead a larger order for the smooth test is chosen, some sensitivity for the location effect is lost to each of the other extra components.

Consequently, it is clear that the *optimal* choice for the order is the smallest order for which the family of alternatives contains the true distribution function. Unfortunately, in practice, we often have no idea in what sense the true distribution deviates from the hypothesised, which implies that choosing the optimal order is both hard and subjective. To overcome that problem, the optimal order can be estimated from the data. The resulting smooth test, using the estimated order, is called the data-driven smooth test and was first introduced by Ledwina (1994).

Simple null hypothesis

For the simple null hypothesis, Ledwina (1994) considered choosing an appropriate dimension as a model selection problem and uses Schwarz' Bayesian Information Criterion (BIC) as the selection rule. An advantage of the BIC is that it asymptotically selects the right model with probability one. For a Neyman model of order k, the BIC is defined as

$$\mathrm{BIC}_n(k) = l(k) - \frac{1}{2}k\log(n),$$

where $l(k)$ is the log-likelihood of the order k Neyman model, maximised in all the model parameters. The last term accounts for the complexity of the model and has larger impact if the sample size is large. The optimal order according to the BIC is then

$$K = \min\{k : 1 \leq k \leq m, \mathrm{BIC}_n(k) \geq \mathrm{BIC}_n(j), j = 1, \ldots m\} \tag{3.35}$$

where m is the upper bound of the dimension, which we assume finite here. When the upper bound m is allowed to grow to infinity with the sample size n, the resulting data-driven test is omnibus consistent, see Kallenberg and Ledwina (1995), (1997) and Inglot, Kallenberg, and Ledwina (1997). The smooth test for the simple null hypothesis has the same form as before in (3.28) except that the arbitrary order k is replaced by the estimated order K, i.e.

$$S_K = \frac{1}{n}\boldsymbol{V}^t\boldsymbol{V} = \frac{1}{n}\sum_{j=1}^{K}\left(\sum_{i=1}^{n} h_j(X_i)\right)^2. \tag{3.36}$$

Composite null hypothesis

Kallenberg and Ledwina (1997) extended Ledwina's approach for simple null hypotheses to the composite case, and used a modification of the selection rule which makes the computations simpler. The modified selection rule directly uses the score statistic instead of the maximum log-likelihood. Since both selection rules are locally asymptotically equivalent, we use the same notation for the modified BIC, which is now given by

$$\mathrm{BIC}_n(k) = \frac{1}{n}\boldsymbol{V}_k^t\boldsymbol{V}_k - k\log(n),$$

where $\boldsymbol{V}_k = \sum_{i=1}^n \boldsymbol{h}(X_i;\hat{\boldsymbol{\beta}})$ is the k-dimensional score vector with the nuisance parameter $\boldsymbol{\beta}$ replaced by its MLE $\hat{\boldsymbol{\beta}}$. Note that any other $\sqrt{n}$-consistent estimator can be used as well, but in our discussion we restrict estimation to MLE. The optimal order K according to the BIC is as in (3.35) and the corresponding efficient score test for composite null hypothesis in (3.31) becomes

$$S_K = \frac{1}{n}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^T\left[I_K - \mathrm{Cov}_0\left[\boldsymbol{h},\boldsymbol{b}\right]\mathrm{Var}_0\left[\boldsymbol{b}\right]^{-1}\mathrm{Cov}_0\left[\boldsymbol{b},\boldsymbol{h}\right]\right]^{-1}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}. \tag{3.37}$$

Note that, for composite null hypotheses, it is better to replace the score statistic by the efficient score statistic for the computation of the BIC criteria as well (see Janic-Wróblewska, 2004). For each of the previous selection rules based on the BIC, it can be shown (see Ledwina, 1994, Kallenberg & Ledwina, 1997, Inglot et al., 1997 and Janic-Wróblewska, 2004) that under $H_0 : \boldsymbol{\theta} = 0$, it holds that

$$\lim_{n\to\infty} P(K = 1) = 1.$$

Hence for large sample sizes the selection rules always chooses the smallest order $K = 1$. This immediately implies that, under H_0, as $n \to \infty$,

$$S_K \xrightarrow{d} \chi_1^2.$$

This means that the limiting null distribution does not depend on m. It should be noted that convergence towards the null distribution is rather slow because the selection criterion does not always chooses $K = 1$ for finite sample sizes, so that it is better to use its simulated exact null distribution instead. However, simulations showed that the data-driven smooth test performs well against a wide range of alternatives.

In the previous discussion we only considered nested exponential models to derive the optimal order. Claeskens and Hjort (2004) considered any possible subset of the indices $\{1, \dots m\}$. Moreover, they also considered Akaike's Information criterion (AIC) as well, among others.

Example 3.3.4. To illustrate the data-driven procedure, we return to the PCB data we discussed in Example 3.3.3. We assume that the density of the data belongs to a family of alternatives to the normal distribution of order not larger than 10, so we believe it is appropriate to take $m = 10$. The selection criterion

(3.35) chooses $K = 3$ and for the corresponding smooth test we have $S_3 = 5.44$. The p-value which is again computed using 10,000 parametric bootstrap samples, equals 0.028 and is slightly smaller than before. We may conclude that the PCB data differs from normality with respect to its skewness.

The advantage of a data-driven smooth test is that, upon rejection of the null distribution, it provides useful information about the true distribution. In fact, as we will see in Section 3.7, it directly provides an estimate of the underlying density.

3.3.5 Smooth test for circular uniformity

In many fields in science, the question arises whether every time instant on a 24h clock, every direction or every angle occurs with the same probability. This GOF problem for circular uniformity (3.11) generally can not be solved adequately with the smooth test described in Section 3.3.2. In Example 3.3.2 we have illustrated on the Birth time data that the value of the test statistic is not origin-invariant and is therefore not useful for circular data. Nevertheless, Neyman's smooth test can be adapted for application to circular data. It is simply a matter of choosing an appropriate set of functions $\{h_j\}$ orthonormal on the CU distribution, so that an origin-invariant test statistic results. Bogdan et al. (2002) proposed to use the set of trigonometric functions $\{\sqrt{2}\cos(jx), \sqrt{2}\sin(jx);\quad j = 1\ldots k\}$, which is indeed orthonormal to the CU distribution. The proof that this choice does in fact result in an origin-invariant smooth test for circular uniformity will be given later in this section and is also in Bogdan et al. (2002). First we give some more details about the construction of the test statistic.

As before, the null density is embedded into a larger exponential family. The order k family of circular alternatives is given by

$$g_k(x,\boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} \Big(\theta_{2j-1}\sqrt{2}\cos(jx) + \theta_{2j}\sqrt{2}\sin(jx)\Big)\right] \quad 0 < x < 2\pi, \tag{3.38}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_{2k})$ denotes the parameter vector, $C(\boldsymbol{\theta})$ is a normalizing constant, and $\{\sqrt{2}\cos(jx), \sqrt{2}\sin(jx);\quad j = 1\ldots k\}$ is the set of orthonormal functions on the CU distribution, which is a complete set if $k \to \infty$.

Note that this family of alternatives, based on the system of trigonometric functions, is similar to the family of circular distributions proposed by Fernández-Durán (2004). He studied a linear expansion of the CU density, which is actually the Barton model but has the drawback that it does not necessarily result in a positive PDF. Taking a log-linear expansion of the density avoids this problem. In contrast to the family of alternatives discussed before, the order k of the family now does not refer to the number of parameters but rather to the largest order of the trigonometric functions used in the family.

The null hypothesis of circular uniformity is equivalent to testing the parametric hypothesis $H_0 : \boldsymbol{\theta} = 0$ against $H_1 : \boldsymbol{\theta} \neq 0$, and the score test statistic for this problem becomes

$$S_{2k} = \frac{1}{n}\boldsymbol{V}^t\boldsymbol{V} = \frac{1}{n}\sum_{j=1}^{2k} V_j^2, \tag{3.39}$$

where $\boldsymbol{V}^T = (V_1, \dots, V_{2k})$ is the score vector in which

$$V_{2j-1} = \sqrt{2}\sum_{l=1}^{n}\cos(jX_l) \text{ and } V_{2j} = \sqrt{2}\sum_{l=1}^{n}\sin(jX_l).$$

This statistic is indeed invariant under rotation since we can write

$$V_{2j-1}^2 + V_{2j}^2 = 2\left|\sum_{l=1}^{n}[\cos(jX_l) + i\sin(jX_l)]\right|^2 = 2\left|\sum_{l=1}^{n}\exp(ijX_l)\right|^2,$$

for each $j = 1, \dots k$. Changing the origin is equivalent to adding some constant γ to each of the X_l, which in turn leads to an extra factor $\exp(ij\gamma)$ with unit modulus.

As before, under the null hypothesis the smooth test statistic S_{2k} is asymptotically χ^2 distributed with $2k$ degrees of freedom and its individual components $\frac{1}{n}V_j^2, j = 1, \dots, 2k$ are asymptotically independently χ_1^2 distributed. Additionally, the jth component $\frac{1}{n}(V_{2j-1}^2 + V_{2j}^2)$ is proportional to the squared resultant length of the jth trigonometric moment. Hence, this two degrees of freedom component can be used as a directional test to detect differences in the jth trigonometric moment.

Later, in Section 3.5, we will describe the relation of this smooth test statistic to the Watson statistic, which is described in Section 3.4.3. For $k = 1$, the statistic S_2 reduces to probably the simplest test for CU, introduced by Rayleigh (1919). The Rayleigh test is introduced on the intuitive ground that the resultant length of the first trigonometric moment, say $\overline{R}$, is expected to be zero for uniformity, so large values of $2n\overline{R}$ indicate evidence against uniformity (see Section 2.2). From the theory of the smooth tests we know that the Rayleigh test only has optimal power against the alternatives

$$g_1(x, \boldsymbol{\theta}) = C(\boldsymbol{\theta})e^{\sqrt{2}(\theta_1\cos x + \theta_2\sin x)} = \frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(x-\mu)}, \quad 0 < x < 2\pi, \tag{3.40}$$

where the last identity is obtained using the trigonometric relations in which $\theta_1 = \kappa\sin\mu$, $\theta_2 = \kappa\cos\mu$ and $C(\boldsymbol{\theta}) = \frac{1}{2\pi I_0(\kappa)}$. The Rayleigh test is thus the score test for uniformity within the von Mises model.

To make a good choice on the number of the components k, Bogdan et al. (2002) proposed to use a modified version of the Schwarz selection rule following Inglot and Ledwina (1996), discussed in the previous section. In particular, let K be the smallest k for which $S_{2k} - 2k \log n$ is maximal. They proved that S_{2K} is consistent against every fixed alternative and asymptotically χ_1^2-distributed under the null hypothesis. However, simulations showed that convergence is relatively slow.

Example 3.3.5. We apply the origin-invariant smooth test to the Birth time data. Before we perform the data-driven version, we first give the results for order $k = 2$. This choice, which corresponds to a family of alternatives with four parameters, is made in order to compare the results with those obtained in Example 3.3.2. As before, we use asymptotic χ^2 critical points. The value of the test statistic is $S_4 = 3.57$, which is again not significant (p=0.470). The first component $\frac{1}{n}(V_1^2 + V_2^2) = 2.76$, accounts for the largest part of S_4 (77%) but also does not indicate a significant effect (p=0.250). Compared to the results from Example 3.3.2, which were based on a non-origin-invariant statistic, the p-values are slightly higher. This might indicate a small power loss. On the other hand, a conclusion can be formulated which does not change by taking another origin.

The p-values for the data-driven version of the smooth test are computed using 10,000 bootstrap samples. The selection criterion chooses $K = 1$ and the value of the test statistic $S_{2K} = 2.76$ again shows no significant deviation from circular uniformity (p=0.283).

This smooth test has the advantage that it is easy to compute. On the other hand it can only be applied to solve GOF problems for circular uniformity. If the null distribution is completely specified, one can apply the PIT, but then the interpretation in terms of trigonometric moments is on the transformed data. Moreover, for composite null hypotheses, which appear most often in practice, no smooth tests have been developed yet. Perhaps the reason is that it seems very difficult to find appropriate orthonormal functions. In Chapter 4 we develop a new general methodology for smooth tests for circular data. We shall see that the smooth test of Bogdan et al. (2002) is a special case of this new class of tests.

3.4 EDF GOF tests

In this section the large class of GOF statistics based on the empirical distribution function (EDF) is described. Each test of this class compares the hypothesised distribution function with the EDF, which is the most widely used estimator of the true distribution. The general construction of the EDF tests is discussed in Section 3.4.1.

The type of the EDF test depends on the distance measure used to make the comparison between the hypothesised distribution and the EDF. Sections 3.4.2 and 3.4.3 are devoted to the two most important types of EDF tests, which are referred to the *supremum* and the *integral* versions, respectively. The oldest, but still well known, supremum EDF test is the Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933 and Smirnov, 1939), while the Cramér-von Mises (CvM) test (Cramér, 1928 and von Mises, 1931, 1947) was the first integral EDF test.

Most EDF tests depend on the choice of the origin and in such are generally not applicable to circular distributions. However, the KS and the CvM tests both have their origin-invariant versions, which were introduced by Kuiper (1960) and Watson (1961), respectively. The Kuiper and the Watson tests are widely used for linear data as well and are discussed in Sections 3.4.2 and 3.4.3, respectively.

3.4.1 General Construction

In this section, the EDF is introduced and some of its properties are presented. Before we state the EDF statistic in its most general form, we first give some basic results on empirical processes. The reason is that a modern approach based on empirical processes provide concise expressions for both the EDF statistic and its asymptotic null distribution.

The empirical distribution function

Suppose $X_1, \ldots, X_n$ is a sample of n i.i.d. observations from an unknown distribution F. The empirical distribution function $\hat{F}_n$ is an estimator of the CDF $F(x) = P(X \leq x)$ and is given by

$$\hat{F}_n(x) \quad = \quad \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

where I is the indicator function. The EDF $\hat{F}_n(x)$ is a non-decreasing step function. If there are no ties in the sample, the EDF has steps of size $\frac{1}{n}$ at each observation X_i. The definition of an EDF for circular data is analogous but depends on both an arbitrarily chosen origin and orientation (clockwise or anti-clockwise).

Since for every x, $n\hat{F}_n(x)$ represents a number of observations in a total of n, it has a binomial distribution with parameters n and $F(x)$. Consequently, we know that for every x, $\hat{F}_n(x)$ is a consistent and unbiased estimator of $F(x)$. Furthermore, using the central limit theorem, we have that for every x, as $n \to \infty$,

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{d} N(0, F(x)(1 - F(x))). \tag{3.41}$$

Note that the latter result holds for each x separately. An improved result, which is known as the *Glivenko-Cantelli* theorem, states uniform convergence of $\hat{F}_n$, i.e.

$$P\left(\lim_{n\to\infty}\sup_x |\hat{F}_n(x) - F(x)| = 0\right) = 1, \tag{3.42}$$

which can also be written as

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0. \tag{3.43}$$

Empirical processes

The distribution of $\sqrt{n}\left(\hat{F}_n(x) - F(x)\right)$ in (3.41) has only a pointwise interpretation. Since we would like to compare the functions entirely instead of just evaluating the difference at individual points, the suggestion is to go from *pointwise* to *functional* properties of the *empirical process*. The latter is given by

$$\mathbb{B}_n(x) = \sqrt{n}\left(\hat{F}_n(x) - F(x)\right). \tag{3.44}$$

Then the *functional* central limit theorem says that the empirical process $\mathbb{B}_n$ converges weakly to a zero mean *Gaussian* process $\mathbb{B}(x)$ with covariance function

$$\operatorname{Cov}\left[\mathbb{B}(x), \mathbb{B}(y)\right] = F(x \wedge y) - F(x)F(y), \tag{3.45}$$

where $\wedge$ denotes the minimum operator. For more details about this derivation we refer to van der Vaart (1998). For uniform F, $\mathbb{B}_n$ and $\mathbb{B}$ are referred to as the uniform empirical process and the *Brownian brigde*, respectively.

Another important theorem is the *continuous mapping* theorem, which is frequently used in this context. It says that if $\mathbb{B}_n \xrightarrow{d} \mathbb{B}$, then $g(\mathbb{B}_n) \xrightarrow{d} g(\mathbb{B})$, for g a continuous function.

General construction of the EDF statistic

Suppose we are interested in testing the GOF null hypothesis $H_0 : F(x) = F_0(x, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is an either known or unknown p-dimensional parameter vector. A meaningful statistic to test H_0 compares the hypothesised distribution $F_0(x, \boldsymbol{\beta})$ with the EDF $\hat{F}_n(x)$ using some sensible *distance* measure. For example, it is clear from property (3.43) that a useful statistic can be constructed based on the *supremum* norm of the difference in distributions. Note that any *distance* between $F_0(x, \boldsymbol{\beta})$ and $\hat{F}_n(x)$ is indeed equivalent to a *norm* of $\mathbb{B}_n(x) = \mathbb{B}_n(x, \boldsymbol{\beta}) = \sqrt{n}\left(\hat{F}_n(x) - F_0(x, \boldsymbol{\beta})\right)$. Consequently, EDF statistics can generally be denoted by

$$T_n = g(\mathbb{B}_n), \tag{3.46}$$

where g is a function that involves an appropriate norm of the empirical process. If g is continuous, the asymptotic null distribution of the statistic T_n is given by $g(\mathbb{B})$ as easily derived by applying the continuous mapping theorem. Usually there is no analytical expression for the distribution function of $g(\mathbb{B})$, and therefore simulation from $g(\mathbb{B})$ may be used instead. It is however important to note that both the empirical and the limiting Gaussian processes depend on the parameter $\boldsymbol{\beta}$. Therefore we make this dependence more explicit in the notation. In particular, we write the Gaussian process $\mathbb{B}$ as $\mathbb{B}(x) = \mathbb{B}(x, \boldsymbol{\beta})$ and its covariance function as

$$\mathrm{Cov}\,[\mathbb{B}(x, \boldsymbol{\beta}), \mathbb{B}(y, \boldsymbol{\beta})] = F_0(x \wedge y, \boldsymbol{\beta}) - F_0(x, \boldsymbol{\beta}) F_0(y, \boldsymbol{\beta}). \tag{3.47}$$

The explicit expression of this covariance function depends on the nature of the parameter $\boldsymbol{\beta}$ and the estimation method.

When the parameter $\boldsymbol{\beta}$ is known, there is no problem in simulating the Gaussian process $\mathbb{B}$. The nature of the function g then determines whether it is convenient to simulate the limiting distribution $g(\mathbb{B})$. For example, if g is the supremum norm, simulating $g(\mathbb{B})$ is straightforward.

When the parameters are unknown, the same statistic can be used, but the nuisance parameters are replaced by their appropriate estimators. The resulting process $\mathbb{B}_n(., \hat{\boldsymbol{\beta}})$ is called the *estimated empirical process.* The covariance function of the limiting process becomes more involved, however. Durbin (1973) and van der Vaart (1998) provided proofs on these asymptotic results for a general class of efficient estimators while making some assumptions on the hypothesised distribution F_0. Unfortunately, the asymptotic law of the estimated empirical process depends on F_0. Therefore, in the composite null hypothesis case, one usually turns to the parametric bootstrap to obtain an approximation of the null distribution of the test statistic. When F_0 is a location-scale invariant distribution, the covariance function in (3.47) becomes independent of the parameter $\boldsymbol{\beta}$, but the dependence on the distribution F_0 remains. In this special situation percentage points may be approximated by one single series of simulations for each sample size, and with an arbitrary choice for $\boldsymbol{\beta}$. For F_0 equal to e.g. the normal distribution, the limiting null distribution of some EDF tests have been tabulated. In particular, Lilliefors (1967) tabulated critical values for the KS test (see Section 3.4.2) while Stephens (1976) provided percentage points and formulas to approximate p-values for the CvM and the AD tests (see Section 3.4.3). On the other hand, the von Mises distribution is location-invariant but not scale-invariant. The critical values for the Watson statistic (see Section 3.4.3) for different values of the concentration parameter κ are given in Lockhart and Stephens (1985). Since in this thesis the null distributions of the EDF tests are always simulated for the composite null hypothesis, we confine the further discussion of the EDF tests to the simple null hypothesis. In particular, for linear and circular data, we discuss the null hypothesis on the unit interval $[0, 1]$ and the unit circumference $\mathrm{arc}(0, 1)$, respectively.

In the next two sections, two concrete types of norm functions g are discussed, which are referred to as the *supremum* and the *integral* norm.

3.4.2 Supremum EDF tests

The EDF test statistics for which the function g in (3.46) is the supremum norm, are called supremum EDF test statistics. They are described in this section.

Kolmogorov-Smirnov test

The class of supremum EDF tests found its origin in the KS test (Kolmogorov, 1933 and Smirnov, 1939). For testing the simple null hypothesis $H_0 : F(x) = F_0(x)$, the KS test statistic is given by

$$D_n = \sqrt{n} \sup_x |\hat{F}_n(x) - F_0(x)| = \sup_x |\mathbb{B}_n(x)| \tag{3.48}$$

and can also be written as $\max(D_n^+, D_n^-)$, where

$$\begin{aligned} D_n^+ &= \sqrt{n} \sup_x (\hat{F}_n(x) - F_0(x)) = \sup_x \mathbb{B}_n \\ D_n^- &= \sqrt{n} \inf_x (\hat{F}_n(x) - F_0(x)) = \sup_x (-\mathbb{B}_n). \end{aligned}$$

The KS statistic D_n and its related statistics D_n^+ and D_n^- can be read from the PP-plot. Details about the relation between the graph and the formal tests are in Section 2.1.1.

Although it has always been a much-discussed topic to find user-friendly expressions and tabulations of the exact (see e.g. Massey, 1951, Stephens, 1970,Dallal, 1986, and Drew, Glen, & Leemis, 1998) and the asymptotic null distributions (see Kolmogorov, 1933) of the KS test, we will not pay much attention to it. The reason is that using the results on the empirical processes described above, we have that, as $n \to \infty$,

$$D_n \xrightarrow{d} D = \sup_x |\mathbb{B}(x)|. \tag{3.49}$$

Therefore the asymptotic null distribution can easily and accurately be simulated. Moreover, convergence of the null distribution is quite fast.

Here we just give the distribution function of the asymptotic null distribution found by Kolmogorov (1933), i.e.

$$F_D(d) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp(-2j^2 d^2).$$

This expression clearly does not depend on the distribution F_0, nor on the true distribution F, which implies that the KS test is nonparametric. Note that this was not clear yet from the right hand side of (3.49).

Kuiper test

The circular analogue of the KS test is the Kuiper test (Kuiper 1960). The test statistic is given by

$$K_n = \sup_{x,y} |\mathbb{B}_n(x) - \mathbb{B}_n(y)| = \sqrt{n} \sup_{x,y} |\hat{F}_n(x) - \hat{F}_n(y) - (F_0(x) - F_0(y))|. \quad (3.50)$$

This interpretation can be made more clear by rewriting the KS test statistic as $D_n = \sup_x |\mathbb{B}_n(x) - \mathbb{B}_n(0)|$, which is a supremum over intervals $[0, x]$ with lower bound fixed at the arbitrary origin 0. Hence, the Kuiper test is in fact the largest value of the KS statistics generated by choosing every possible starting point on the circle (see also Barr & Shudde, 1973). Consequently, it is intuitively clear that the statistic is origin-invariant. For a more formal proof of the origin-invariance property we refer to Mardia and Jupp (2000) or Jammalamadaka and SenGupta (2001). Also, in the sense of Section 3.1.3 it is directly clear that the Kuiper test is origin-invariant because the set of differences $\{\hat{F}_n(x_i) - \hat{F}_n(x_j) - (F_0(x_i) - F_0(x_j))\}$ is a maximal origin-invariant set.

The Kuiper statistic can alternatively be written as

$$K_n = D_n^+ + D_n^-. \quad (3.51)$$

Its asymptotic null distribution is given by Kuiper (1960) and again implies the distribution-free property. Note that the Kuiper test can also be applied to linear data and is often more powerful than the KS test (see e.g. Abrahamson, 1967).

In Chapter 6 we introduce a graphical diagnostic tool, which is related to the Kuiper test in a sense that it is based on the same interval indexed process

$$\mathbb{B}_n(x) - \mathbb{B}_n(y).$$

The interpretation of this process is intuitively clear when rewriting the process as

$$\sqrt{n}\{\hat{F}_n(x, y) - (F_0(x, y)),$$

where $\hat{F}_n(x, y) = \hat{F}_n(x) - \hat{F}_n(y)$ and $F_0(x, y) = F_0(x) - F_0(y)$. Hence, the process values can be interpreted as the differences between observed and expected probabilities of the random variable X falling into the interval $[x, y]$. From this argument it is intuitively clear that for a particular interval $[x, y]$, the larger the process value, the more the true distribution deviates from the null distribution within that interval.

3.4.3 Integral tests

This section is devoted to another important class of EDF statistics, where the function g in (3.46) is based on an integral and which is therefore called the integral class of EDF statistics.

The class of Anderson-Darling tests

Consider the simple GOF problem for testing $H_0 : F(x) = F_0(x)$. Anderson and Darling (1952) introduced a family of statistics defined as

$$T_n = \int_0^1 \mathbb{B}_n(x)^2 w(F_0(x)) dF_0(x), \tag{3.52}$$

where $w(u)$ is some nonnegative weight function ($0 \leq u \leq 1$). The two most common choices for this weight function are $w(u) = 1$ and $w(u) = \frac{1}{u(1-u)}$. For the former weight function, the family of statistics reduces to the CvM statistic. The statistic based on the latter weight function is more popular and is referred to as *the* Anderson-Darling (AD) statistic. This choice of weight function is particularly useful because the empirical process $\mathbb{A}_n(u) = \frac{\mathbb{B}_n(u)}{u(1-u)}$ has covariance function

$$\text{Cov}\left[\mathbb{A}_n(s), \mathbb{A}_n(t)\right] = \frac{s \wedge t - st}{\sqrt{s(1-s)t(1-t)}},$$

and thus has constant unit variance. We denote the CvM and the AD statistics as W_n and A_n, respectively. The computational formulae for the CvM and the AD statistics are

$$\begin{aligned} W_n &= \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{n} \right) + \frac{1}{12n} \\ A_n &= -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)(\ln U_{(i)} + \ln(1 - U_{(n+1-i)})), \end{aligned}$$

where $U_{(1)} \leq \ldots \leq U_{(n)}$ are the order statistics of the variable $U = F_0(X)$.

Similarly, as for the KS and the Kuiper tests, Stephens (1970) provided a simple method of approximating the exact null distribution for the CvM test. He fitted a polynomial in inverse powers of $\sqrt{n}$ to simulated critical values. Based on that empirical study he suggested using a modified test statistic, for which it turned out that the asymptotic critical values are fairly accurate. For the AD test, the null distribution is already accurately approximated by its asymptotic distribution when $n > 3$.

The asymptotic null distributions of W_n and A_n can again be found using the weak convergence of the empirical processes $\mathbb{B}_n$ and $\mathbb{A}_n$, respectively. Furthermore, applying the continuous mapping theorem we have that, as $n \to \infty$,

$$W_n \xrightarrow{d} W = \int_0^1 \mathbb{B}^2(u)du \tag{3.53}$$

and

$$A_n \xrightarrow{d} A = \int_0^1 \mathbb{A}^2(u)du, \tag{3.54}$$

where $\mathbb{A}(u) = \frac{\mathbb{B}(u)}{\sqrt{u(1-u)}}$. However, the integral norm is not a convenient function to simulate from. Another expression, which is more useful for simulations, is based on the *Kac and Siegert* or *principal components* decomposition of the Gaussian processes in (3.53) and (3.54).

In general, the Kac and Siegert (1947) decomposition of a Gaussian process $\mathbb{P}$ is given by

$$\mathbb{P}(u) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\psi_j(u)Z_j, \tag{3.55}$$

where $\{\lambda_j\}$ and $\{\psi_j\}$ are the eigenvalues and the eigenfunctions of the integral equation

$$\int_0^1 \psi(u)\text{Cov}\left[\mathbb{P}(u), \mathbb{P}(v)\right] du = \lambda\psi(v). \tag{3.56}$$

The components Z_j are i.i.d. standard normal random variables equivalent to

$$\frac{1}{\sqrt{\lambda_j}} \int_0^1 \mathbb{P}(u)\psi_j(u)du,$$

which are called the *principle components*. The solutions of (3.56) can only be found easily in special cases. For $\mathbb{P} = \mathbb{B}$, it can be shown that (j=1,2,...)

$$\lambda_j = \frac{1}{j^2\pi^2} \text{ and } \psi_j(u) = \sqrt{2}\sin(j\pi u),$$

while for $\mathbb{P} = \mathbb{A}$, it can be shown that (j=1,2,...)

$$\lambda_j = \frac{1}{j(j+1)} \text{ and } \psi_j(u) = 2\sqrt{\frac{1}{j(j+1)}}\sqrt{u(1-u)}\frac{d}{du}L_j(u),$$

where L_j are the orthonormal Legendre polynomials. The definition of those polynomials is in (A.11) of Appendix A.2.

By substituting the principle component decomposition (3.55) of the Gaussian processes $\mathbb{B}$ and $\mathbb{A}$ into the right hand sides of (3.53) and (3.54), respectively, the limiting null distributions can be rewritten as

$$W = \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2} Z_j^2 \tag{3.57}$$

and

$$A = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2 \tag{3.58}$$

respectively, where $Z_1, Z_2, \dots$ are i.i.d. standard normal random variables. This representation is an infinite sum of weighted χ^2 variables, for which the weights decrease with increasing order j. The downweighting is quite severe so that it is possible to restrict simulation to the first most important components. When the principal component decomposition is applied to the empirical processes $\mathbb{B}_n$ or $\mathbb{A}_n$, the interpretation of the components is related to that of the components in the smooth test statistics, which will be discussed in Section 3.5.

The Watson test

Watson (1961) proposed a modification of the CvM test which is useful for testing GOF of circular distributions. The Watson statistic is defined as

$$\begin{aligned} U_n &= \int_0^1 \left(\mathbb{B}_n(x) - \int_0^1 \mathbb{B}_n(y) dF_0(y)\right)^2 dF_0(x), & (3.59) \\ &= n \int_0^1 \left(\hat{F}_n(x) - F_0(x) - \int_0^1 \hat{F}_n(y) - F_0(y) dF_0(y)\right)^2 dF_0(x). & (3.60) \end{aligned}$$

Alternatively U_n can be written as

$$\begin{aligned} U_n &= \frac{1}{2} \int_0^1 \int_0^1 (\mathbb{B}_n(x) - \mathbb{B}_n(y))^2 dF_0(x) dF_0(y) & (3.61) \\ &= \frac{n}{2} \int_0^1 \int_0^1 \left(\hat{F}_n(x) - \hat{F}_n(y) - (F_0(x) - F_0(y))\right)^2 dF_0(y) dF_0(x), & (3.62) \end{aligned}$$

or as,

$$U_n = \inf_{x_0} \int_{x_0}^{1+x_0} \mathbb{B}_n(x; x_0)^2 dF_0(x), \tag{3.63}$$

where $\mathbb{B}_n(x; x_0)$ is the empirical process calculated with x_0 as starting point, i.e. $\mathbb{B}_n(x; x_0) = \sqrt{n}(\hat{F}_n(x; x_0) - F_0(x; x_0))$. Here, $\hat{F}_n(x; x_0)$ and $F_0(x; x_0)$ are defined as

$$\hat{F}_n(x; x_0) = \frac{\sum_{i=1}^n I(x_0 \leq X_i \leq x)}{n} \text{ and } F_0(x; x_0) = \int_{x_0}^{x} f_0(y) dy, \tag{3.64}$$

where x is assumed to have values between x_0 and $x_0 + 1$. Since x is periodic, the latter assumption can always be achieved by adding the appropriate integer constant. The last expression (3.63) implies that the Watson statistic can be interpreted as the smallest value of the CvM statistics generated by choosing every possible starting point on the circle. From this interpretation it is again intuitively clear that the Watson test is origin-invariant. Jammalamadaka and SenGupta (2001) provided a more formal proof. The origin-invariance property can also be derived from each of the expressions (3.59)-(3.62) using the argument that the integrand is a maximal origin-invariant function.

The computational form of the Watson statistic is given by

$$U_n = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n} \right) - n(\overline{U} - \frac{1}{2}) + \frac{1}{12n}, \tag{3.65}$$

where $U_{(1)} \leq \ldots \leq U_{(n)}$ are the order statistics of the variable $U = F_0(X)$ and $\overline{U}$ is the sample mean. The asymptotic null distribution can be found using empirical process theory. In particular, it can be shown that

$$U_n \xrightarrow{d} U = \int_0^1 \left(\mathbb{B}(u) - \int_0^1 \mathbb{B}(v) dv \right)^2 du. \tag{3.66}$$

Similarly as for the CvM and the AD statistics, this form is not often useful for simulation.

As an alternative we give the Kac and Siegert representation of the asymptotic null distribution (see Watson (1961) or Shorack and Wellner (1986)). Watson found the solutions for the integral equation (3.56) for the Gaussian process $\mathbb{P} = \mathbb{B} - \int_0^1 \mathbb{B}(v)dv$ with corresponding covariance function

$$\text{Cov}\left[\mathbb{P}(u), \mathbb{P}(v)\right] = u \wedge v - (u+v)/2 + (u-v)/2 + 1/12. \tag{3.67}$$

The eigenvalues are $\lambda_{2j-1} = \lambda_{2j} = \frac{1}{4\pi^2 j^2}$, and the eigenfunctions are

$$\psi_{2j-1} = \sqrt{2}\sin(2\pi j) \text{ and } \psi_{2j} = \sqrt{2}\cos(2\pi j) \;\; j = 1, 2, \ldots.$$

Consequently, the principal components decomposition is

$$U = \sum_{j=1}^{\infty} \frac{1}{4\pi^2 j^2} (Z_{2j-1}^2 + Z_{2j}^2) = \sum_{j=1}^{\infty} \frac{1}{4\pi^2 j^2} X_j \tag{3.68}$$

where the $X_j = Z_{2j-1}^2 + Z_{2j}^2$ are i.i.d. χ_2^2. Similarly as for the CvM and for the AD test statistics, the asymptotic null distribution can be simulated from the first few components, which have the largest weights. Note that from the Kac and Siegert representations, it can be seen that for the simple null hypothesis, each test from the integral class of EDF statistics considered here is distribution-free.

The general framework for the principal components decomposition of integral statistics was introduced by Anderson and Darling (1952). Later, Durbin and Knott (1972) provided the exact and asymptotic distribution of the principal components. Shorack and Wellner (1986) give a comprehensive overview of this theory. These principle components decompositions can also be obtained using the theory of U- or V-statistics. In particular, the classical statistics typically take the form of a degenerate U- or V-statistic. To obtain the limiting null distribution of those classes of statistics, we need the solutions of an integral equation of the form (3.56), where the covariance function is replaced by the kernel of the U- or V-statistic. For more details about those U- and V-statistics we refer to Lee (1990). In Chapter 5, we will rewrite one of our new GOF test statistics as a V-statistic in order to obtain a useful limiting distribution.

Usually, finding the Kac and Siegert representation is not an easy task. Ahmad (1993) avoided the problem by using a modification of the EDF, resulting in a test statistic with limiting normal distribution under the null and the alternative hypothesis. He did this for the CvM and the Watson statistics for simple null hypotheses. Janssen, Swanepoel and Veraverbeke (2005) extended this procedure to the composite null hypothesis case. This is worth noting here since their resulting tests are distribution-free.

3.5 Link between Pearson, EDF and smooth tests

3.5.1 Pearson χ^2 versus smooth tests

Barton (1955), Cox and Hinkley (1974) and Kopecky and Pierce (1979) each provided a link between the Pearson χ^2 test and a smooth test. In this section we briefly mention how this link naturally arises. In particular, as in Rayner and Best (1989), we derive the Pearson χ^2 statistic as a smooth statistic for categorised data. Although this link exists for composite null hypotheses as well, we restrict the discussion to the simple null hypothesis case. Consider again the discrete GOF problem from Section 3.2.1. We want to test

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0,$$

where $\boldsymbol{\pi}_0$ is the m-dimensional vector of expected cell probabilities, which completely specifies the discrete model. These null probabilities $\boldsymbol{\pi}_0 = (\pi_{01}, \dots, \pi_{0m})$ are embedded in the order k smooth family of alternatives ($k < m$)

$$\pi_{ki} = C(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} \theta_j h_{ij} \right) \pi_{0i}, \quad i = 1, \dots m, \tag{3.69}$$

where $C(\boldsymbol{\theta})$ is the normalising constant and $\{\boldsymbol{h}_j^T = (h_{ij}, \dots, h_{mj})\}$ is a set of m-dimensional vectors that satisfy the orthonormality condition

$$\sum_{i=1}^{m} h_{ij} h_{il} \pi_{0i} = \delta_{jl} \text{ with } h_{0j} = 1, \quad j, l = 1, \dots, k.$$

It can be shown that, given the observed cell frequencies $\boldsymbol{X}^T = (X_1, \ldots, X_m)$, the score test for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ in (3.69) is given by

$$S_k = \frac{1}{n} \sum_{j=1}^{k} \left(\sum_{i=1}^{m} X_i h_{ij} \right)^2 .$$

This score test is asymptotically χ^2_k distributed under the null hypothesis and reduces to the Pearson χ^2 statistic (3.18) if $k = m - 1$. The latter property is particularly useful because it demonstrates that the Pearson statistic can be decomposed into $m - 1$ asymptotically independent components. The interpretation of the components depends on the choice of the system $\{\boldsymbol{h}_j\}$.

3.5.2 EDF versus smooth tests

In this section we discuss an interesting relation between EDF tests and smooth tests. In particular, Durbin and Knott (1972) showed that the components in the orthogonal representation of the AD statistic are similar to those in the Neyman smooth test using the normalised Legendre polynomials as the orthonormal basis. In a similar way other integral EDF statistics can be related to smooth tests by choosing an appropriate orthonormal set of functions to describe the family of smooth alternatives. Here, we give the relation to smooth tests for the CvM, the AD and the Watson tests.

Since EDF tests are omnibus consistent it is generally not known to which alternatives they have high or low power. As we will show soon, the link with smooth tests is particularly useful to get information about the power characteristics. It turns out that from the corresponding smooth test one can directly observe to which alternatives the EDF statistic has high power.

Similarly as for the Gaussian processes, the Kac and Siegert decomposition can be obtained for the empirical processes. In particular, the expression of the Gaussian process $\mathbb{P}(u)$ in (3.55) is analogous to the one for the corresponding empirical process $\mathbb{P}_n(u)$ which is given by

$$\mathbb{P}_n(u) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \psi_j(u) Z_{nj}, \tag{3.70}$$

where the eigenvalues λ_j and eigenfunctions $\{\psi_j\}$ are as before, but the principal components Z_{nj} can be simplified for each of the EDF statistics.

The CvM, the AD and the Watson statistics can thus be rewritten as

$$W_n = \sum_{j=1}^{\infty} \frac{1}{j^2 \pi^2} Z_{nj}^2 \text{ where } Z_{nj} = \sqrt{\frac{2}{n}} \sum_{i=1}^{n} \cos(j \pi U_i), \tag{3.71}$$

$$A_n = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_{nj}^2 \text{ where } Z_{nj} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_j(U_i), \tag{3.72}$$

and

$$U_n = \sum_{j=1}^{\infty} \frac{1}{4\pi^2 j^2} (Z_{n,2j-1}^2 + Z_{n,2j}^2), \tag{3.73}$$

where

$$Z_{n,2j-1} = \sqrt{\frac{2}{n}} \sum_{i=1}^{n} \cos(2j\pi U_i) \text{ and } Z_{n,2j} = \sqrt{\frac{2}{n}} \sum_{i=1}^{n} \sin(2j\pi U_i). \tag{3.74}$$

Here, each of the test statistics is represented as an infinite weighted sum of asymptotically independent squared components. Note that the weights for each of the statistics decrease fast with the order j. This means that only the first components are important.

Moreover, it is clear that each of the alternative forms of the test statistics is similar to the decomposition of some smooth test statistic from Section 3.3. The corresponding smooth test statistic is the score statistic derived from the family of smooth alternatives using the set of eigenfunctions $\{\psi_j\}$ in (3.70) in the construction of the set of orthonormal functions $\{h_j\}$ in (3.24).

In particular, the CvM and the AD test statistics are related to the smooth test for uniformity using the orthonormal systems based on the trigonometric functions $\{\sqrt{2}\cos(j\pi u); j = 1 \ldots k\}$ and the set of normalised Legendre polynomials $\{L_j; j = 1 \ldots k\}$, respectively (see Section 3.3.2). The Watson test is related to the smooth test for circular uniformity based on the complete basis of trigonometric functions $\{\cos(2j\pi u), \sin(2j\pi u); j = 1 \ldots k\}$, which is given in Section 3.3.5.

From these relations, we derive an important distinction between EDF tests and smooth tests. Although the EDF tests are omnibus consistent, they have often low power against alternatives from a smooth family with order strictly larger than 2. This can be seen from the severe downweighting in the decomposition for the EDF statistics, which implies that only the first two components are important. On the other hand, smooth tests of arbitrary order k are not omnibus consistent, but they have higher power against all alternatives in the family up to the chosen kth order.

As mentioned in Section 3.3.2, there is a difference in interpretation of the components of the smooth test statistics depending on the choice of the orthonormal system. Hence, this difference in interpretation can now be applied to EDF tests as well. This means that the CvM is sensitive to slowly oscillating alternatives while the AD test has good power for differences in the first moments. Furthermore, the Watson test is particularly sensitive to deviations in the first order trigonometric moments.

Example 3.5.6. We use the `circular` package in R of Lund and Agostinelli (2005) to apply the Kuiper and the Watson tests for circular uniformity on the Birth time data. Here, the modified versions of the Kuiper statistic and the Watson statistic proposed by Stephens (1970) are used, which results in values $K_n = 1.218$ and $U_n = 0.085$, respectively. The p-value for the Kuiper test equals 0.508 and is calculated using the first term of the relevant asymptotic distribution of K_n. This is an approximation which is accurate to the first two decimals. On the other hand, the value of the Watson statistic is outside the region where such an approximation is appropriate. Therefore, we simulate the null distribution using 100,000 uniform bootstrap samples which results in a p-value of 0.632. Clearly, neither test indicates evidence against the null hypothesis of circular uniformity. Compared to the results for the smooth test in Section 3.3.5, the p-values are much larger. A possible explanation is that the first trigonometric moment of the true distribution is similar to that of the circular uniform distribution. Since the EDF tests put high weight on that first component, they give non-significant results. The lower p-value for the smooth test might be an indication that the higher order moments do not correspond well with those of the hypothesised distribution.

3.5.3 EDF versus Pearson χ^2 tests

As mentioned in Section 3.2.2, the Watson test is a special case of Rothman's test (see Rothman, 1972), which is based on the Pearson χ^2 statistic. Later in Chapter 5 we discuss the class of Sample Space Partition (SSP) tests for linear data which is based on Pearson's statistic in a similar way as Rothman's test. As we will see in that chapter, there is an analogous relation between the class of SSP tests and the class of EDF tests. Furthermore, in Section 5.8, we extend the class of SSP tests to the class of origin-invariant SSP tests which is applicable to circular data.

3.6 Other GOF tests

Spacings tests

In the previous sections we referred to *spacing* tests. Therefore, we here briefly discuss this class of test statistics, which is useful to test the GOF of both the linear and circular uniform density (see e.g. Jammalamadaka & SenGupta, 2001). For an extensive overview on the topic of spacings we refer to Pyke (1965). Compared to EDF tests for uniformity, the latter perform better in detecting differences between CDFs, while spacings tests are particularly effective in revealing differences between PDFs. Before applying a spacings test, the general

GOF problem is, without loss of generality, reduced to the GOF problem for uniformity on $[0,1]$ using the PIT. Assume that such a transformation has been made and that $U_1, \ldots, U_n$ is the transformed sample. Suppose $U_{(1)} \leq \ldots \leq U_{(n)}$ are the ordered observations. Define the *spacings* on the line as

$$D_i = U_{(i)} - U_{(i-1)}, \quad i = 1, \ldots, n+1$$

where $U_0 = 0$ and $U_{n+1} = 1$. On the other hand, the *spacings* on the circle with unit circumference, also called *arc lengths*, are defined as

$$D_i = U_{(i)} - U_{(i-1)}, \quad i = 1, \ldots, n-1$$

where $U_0 = U_{(n)} - 1$. Under the null hypothesis of uniformity we expect the spacings to be equal to $1/n$. The family of spacings test statistics is described as

$$T_n = \frac{1}{n} \sum_{i=1}^{n} g(nD_i), \tag{3.75}$$

where g is some meaningful function. Two common choices for that function result in the *circular range* test and *Rao's spacings* test. Both were introduced in the context of circular data by Rao (1969), who also found the corresponding exact null distribution. The circular range statistic is defined as

$$R_1 = 1 - \max_{1 \leq i \leq n} D_i, \tag{3.76}$$

which corresponds to the smallest arc containing all observations. Rao's spacings test, which is sometimes referred to as the test of equal spacings, is based on

$$R_2 = \frac{1}{2} \sum_{i=1}^{n} |D_i - \frac{1}{n}| = \sum_{i=1}^{n} \max(D_i - \frac{1}{n}, 0). \tag{3.77}$$

The statistic is zero only for equally spaced data and is large for clustered data. Furthermore, the statistic can be interpreted as the uncovered part of the circumference when placing the arcs $(U_i, U_i + \frac{1}{n})$ $i = 1, \ldots n$ on the circle. This test is known to have good power properties when testing for uniformity against multimodal densities.

Tests for circular normality

In most practical situations for testing GOF on the circle we are dealing with a composite null hypothesis. The Kuiper and the Watson tests can be used for composite null hypotheses after replacing the parameters by their appropriate estimators. However, the null distribution of the resulting statistic becomes more complicated and is therefore often obtained using the parametric bootstrap. For the von Mises distribution, which is the most common distribution on the circle, Lockhart and Stephens (1985) modified the Watson statistic and tabulated critical values of its asymptotic null distribution. A modification of the Kuiper test has not been published yet. In Chapter 4 we develop a new smooth test for circular normality. For comparison we discuss in this section two other tests for the von Mises distribution which are referred as to the *entropy* test and the *BarCox* test.

The entropy test for circular normality was proposed by Lund and Jammalamadaka (2000). It is basically an extension of the entropy test for normality on the line introduced by Vasicek (1976). The *entropy* of a circular distribution $f(x), 0 \leq x \leq 2\pi$, is defined as

$$H(f) = -\int_0^{2\pi} f(x) \log f(x) dx \tag{3.78}$$

and reaches its maximum for the von Mises density, subject to a given mean direction and concentration. The entropy of a von Mises distribution is then estimated as

$$H(\hat{f}_{CN}) = \log \left[\frac{2\pi I_0(\hat{\kappa})}{\exp(\hat{\kappa} A(\hat{\kappa}))} \right], \tag{3.79}$$

where $\hat{\kappa}$ is the MLE of κ. An intuitive, straightforward approach now is to compare this estimated maximum entropy value with the sample entropy, which is given by

$$H_{mn} = \frac{1}{n} \sum_{i=1}^{n} \log\{ \frac{n}{2m} (X_{(i+m)} - X_{(i-m)}) \}, \tag{3.80}$$

where $2m$ is the size of the steps for the spacings $X_{(i+m)} - X_{(i-m)}$, with the restriction that $2m < n$. Vasicek (1976) proved the consistency of this nonparametric estimator as $n \to \infty$, $m \to \infty$ and $m/n \to 0$. The entropy statistic is then given by the ratio of the two estimators for the entropy, H_{mn} and $H(\hat{f}_{CN})$. In particular,

$$K_{mn} = 2\pi \frac{\exp\{H_{mn}\}}{\exp\{H(\hat{f}_{CN})\}}. \tag{3.81}$$

Under the null hypothesis of circular normality, we have that

$$K_{mn} \xrightarrow{p} 2\pi \text{ as } n, m \to \infty \text{ and } m/n \to 0.$$

On the other hand, the statistic tends to have lower values if the data do not come from a von Mises distribution. The null distribution is obtained by simulation and the critical values have been tabulated for various sample sizes and values of κ (see Lund & Jammalamadaka, 2000). The choice of m corresponds to the value that maximizes the test statistic.

As a second test for circular normality, Barndoff-Nielsen and Cox (1979, Section 5.3) derived a score test from a family of distributions constructed by expanding the exponent of the von Mises null density by the second order trigonometric moment. The density then becomes

$$f(x) = \frac{1}{2\pi I_0(\kappa)} e^{\beta_1 \cos(x) + \beta_2 \sin(x) + \theta_1 \cos(2x) + \theta_2 \sin(2x)}$$

and the null hypothesis to be examined is $H_0 : \theta_1 = \theta_2 = 0$.

The authors used saddle-point approximation methods to prove that

$$\frac{1}{\sqrt{n}} B_c = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \cos 2(X_i - \hat{\mu}) - \sqrt{n} I_2(\hat{\kappa})/I_0(\hat{\kappa}) \text{ and } \frac{1}{\sqrt{n}} B_s = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sin 2(X_i - \hat{\mu})$$

are asymptotically independently normally distributed with

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}} B_c\right] = \left[\frac{I_0^2 + I_0 I_4 - 2I_2^2}{2I_0^2} - \frac{(I_0 I_3 + I_0 I_1 - 2I_1 I_2)^2}{2I_0^2(I_0^2 + I_0 I_2 - 2I_1^2)}\right]$$

and

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}} B_s\right] = \left[\frac{(I_0 - I_4)(I_0 - I_2) - (I_1 - I_3)^2}{2I_0(I_0 - I_2)}\right],$$

where we omit the argument κ for notational comfort. In this way a chi-squared statistic with two degrees of freedom can be derived. The resulting test statistic is referred to as the BarCox test statistic and its value is denoted by B. In Chapter 4, we develop a new framework for the construction of smooth tests for composite null hypotheses on circular data for which the BarCox test is a special case.

3.7 Non-parametric density estimation

In this section, we review some techniques for estimating the probability density function from observed data. We consider two important approaches, both of which are nonparametric. Hence, no rigid assumptions will be made about the true distribution of the observed data. For an excellent survey on nonparametric density estimation for linear data, we refer to Silverman (1986).

The first approach is the kernel density estimation, which we already applied in Chapter 2. It will be briefly explained in Section 3.7.1. The second density estimator is called the *orthonormal series estimator* and will be discussed in Section 3.7.2. We will see that, in the context of GOF, this estimate naturally arises from the data-driven smooth test. Both the kernel density and the orthonormal series estimator originate from the linear context, but the circular analogues are relatively straightforward and will also be explained in the respective sections. Note that in Chapter 2 we also applied the histogram and the rose diagram, which constitute a third approach for density estimation. Their construction is very simple, but an important drawback is that the choice of both the origin and the bin width is subjective.

3.7.1 Kernel density estimator

Rosenblatt (1956) introduced the kernel density estimator for linear data which is given by

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{3.82}$$

where K is the kernel function and h is the window width that serves as a smoothing parameter. The Gaussian kernel is a common choice for K, and the window width is usually chosen in a data-driven way. Here we will use cross-validation as a method for estimating the *integrated squared error loss* (ISE), which is defined as

$$\Lambda(\hat{f}) = \int_{\mathcal{S}} (\hat{f}(x) - f(x))^2 dx, \tag{3.83}$$

for any estimator $\hat{f}(x)$ of the true density $f(x)$. The basic principle of cross-validation (CV) is to construct an estimator $\hat{f}_i$ based on the reduced sample $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, and use X_i to validate $\hat{f}_i$. Let

$$\hat{f}_i(x;h) = \frac{1}{n-1}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right), \tag{3.84}$$

be the density estimator constructed from all observations except X_i. Furthermore, let

$$\mathrm{CV}(h) = \frac{2}{n}\sum_{i=1}^{n} \hat{f}_i(X_i;h) - \int_{\mathcal{S}} \hat{f}^2(x;h)dx. \tag{3.85}$$

If $f(x)$ were known, $-\mathrm{CV}(h) + \int_{\mathcal{S}} f^2(x)dx$ would be, for all h, an unbiased estimator of the *mean integrated squared error loss* $\mathrm{E}\left[\Lambda(\hat{f}(x;h))\right]$ (MISE). Since $\int_{\mathcal{S}} f^2(x)dx$ does not denpend on h, maximising $\mathrm{CV}(h)$ gives a good choice of the smoothing parameter h.

The same procedure can be followed to construct a good *circular kernel density estimator*. In particular, Hall et al. (1990) proposed to replace the euclidean difference $x - X_i$ in (3.82) with the cosine of $x - X_i$, which results in the estimator defined by

$$\hat{f}(x;\kappa) = \frac{1}{n}\sum_{i=1}^{n} L\left(\kappa \cos\left(x - X_i\right)\right), \tag{3.86}$$

where L is an arbitrary normalised kernel and κ is the smoothing parameter. We here take $L(t)$ proportional to e^t. The circular kernel density estimator (3.86) is then the average of n von Mises densities localised at the observations $X_1, \ldots, X_n$ and with concentration parameter κ. Minimising the MISE

$\mathrm{E}\left[\Lambda(\hat{f}(x;\kappa)\right]$ for the circular kernel density estimator now amounts to maximising

$$\mathrm{CV}(\kappa) = \frac{2}{n}\sum_{i=1}^{n}\hat{f}_i(X_i;\kappa) - \int_{\mathcal{S}}\hat{f}^2(x;\kappa)dx, \tag{3.87}$$

where $\hat{f}_i(X_i;\kappa)$ is the circular kernel density estimator constructed from leaving out sample value X_i.

The above procedure to choose a good smoothing parameter for the linear as well as for the circular kernel density estimator is referred to as the method of unbiased cross-validation (UCV). We applied these kernel density estimators to the examples in Chapter 2, the results of which will be presented in Chapter 6 and compared to those obtained by the orthogonal series density estimator (see next section).

3.7.2 Orthonormal series density estimator

The orthonormal series density estimator was introduced by Cencov (1962) and is essentially the order k smooth density (3.27) considered by Barton (1953). This estimator is widely discussed and applied in the context of nonparametric density estimation and is an expansion of the form

$$g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = \left[1 + \sum_{j=1}^{k}\theta_j h_j(x;\boldsymbol{\beta})\right] f_0(x;\boldsymbol{\beta}), \tag{3.88}$$

where $\boldsymbol{\beta}$ is assumed to be either known or replaced by its MLE $\hat{\boldsymbol{\beta}}$ and $\{h_j; j = 1,\dots k\}$ are orthonormal functions with respect to f_0. We now aim at finding appropriate estimates for $\boldsymbol{\theta}$. Note that for the choice of the smooth model in (3.88), the computation of the estimate for $\boldsymbol{\theta}$ is simple. Within model (3.88), we have that

$$\mathrm{E}_k\left[h_j(X;\boldsymbol{\beta})\right] = \int_{-\infty}^{\infty} h_j(x;\boldsymbol{\beta})g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = \theta_j, \tag{3.89}$$

where $\mathrm{E}_k\left[.\right]$ denotes the expected value with respect to the model g_k. Note that the above equation still holds when g_k is replaced by the true density f, for which it is assumed that it may be represented by g_∞. Therefore, an unbiased and consistent estimator of θ_j is obtained by

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} h_j(X_i;\boldsymbol{\beta}) = \frac{1}{n}V_j,$$

where V_j is the jth component in the score vector $\boldsymbol{V}_{\boldsymbol{\beta}}$ of the test statistic (3.26). The density estimator is now given by

$$g_k(x;\hat{\boldsymbol{\theta}},\boldsymbol{\beta}) = \left[1+\sum_{j=1}^{k}\hat{\theta}_j h_j(x;\boldsymbol{\beta})\right] f_0(x;\boldsymbol{\beta}), \tag{3.90}$$

where $\boldsymbol{\beta}$ is assumed to be either known or replaced by its MLE $\hat{\boldsymbol{\beta}}$. For notational comfort we will, from now on, omit the dependence on $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\beta}$ and refer to this estimator as $\hat{g}_k(x)$.

There exist many variants of this estimator (see e.g. Anderson & de Figueiredo, 1980, Buckland, 1992, Clutton-Brock, 1990 and Diggle & Hall, 1986). In particular, one can choose different sets of orthonormal functions $\{h_j, j=1,\ldots,k\}$. Also, the distribution $f_0(x;\boldsymbol{\beta})$, which serves as a *starting distribution*, can be altered, and most importantly the selection of the order k can be chosen.

Here, for the set of orthonormal functions, we use either the trigonometric functions or the polynomials orthonormal to the starting distribution, which is chosen here to be uniform or normal. Concerning the choice of the order k we use the selection rule considered by Tarter and Kronmal (1976), which is based on minimising the MISE. However, the authors only used series expansions orthonormal to the uniform distribution. In order to discuss the optimal order selection for series expansion orthonormal to a general f_0, we consider the criterion proposed by Anderson and de Figueiredo (1980). The criterion is called the *weighted* ISE and is given by

$$\Lambda(\hat{g}_k) = \int_{-\infty}^{\infty}\frac{(\hat{g}_k(x)-f(x))^2}{f_0(x;\boldsymbol{\beta})}dx. \tag{3.91}$$

for which the expected value reduces to the same expression as in Tarter and Kronmal (1976), i.e.

$$\mathrm{E}\left[\Lambda(\hat{g}_k)\right] = \frac{1}{n}\sum_{j=1}^{k}(d_j^2-\theta_j^2)+\sum_{j=k+1}^{\infty}\theta_j^2, \tag{3.92}$$

where $d_j^2 = \mathrm{E}\left[h_j^2(X)\right]$. The optimal choice of k is the value that minimises $\mathrm{E}\left[\Lambda(\hat{g}_k)\right]$. An unbiased estimator $\hat{\Lambda}(\hat{g}_k)$ of $\mathrm{E}\left[\Lambda(\hat{g}_k)\right]$ is obtained if d_j^2 and θ_j^2 in (3.92) are replaced by their respective unbiased estimators,

$$\hat{d_j^2} = \frac{1}{n}\sum_{i=1}^{n}h_j^2(X_i)$$

and

$$\hat{\theta_j^2} = \frac{1}{n-1}\left(n\hat{\theta}_j^2-\hat{d_j^2}\right).$$

Minimising $\hat{\Lambda}(\hat{g}_k)$ results in the following decision rule:

Include the jth term until it fails the test

$$\hat{\theta}_j^2 > \frac{2}{n+1}\hat{d}_j^2.$$

Another way now to obtain an appropriate order for the orthonormal series estimator is the following. The data-driven smooth test described in Section 3.3.4 uses the BIC criterion to determine the order in the family of smooth alternatives (3.88). The conclusion of this GOF test, which indicates whether the true distribution equals f_0 or not, can be accompanied by the density estimator that naturally arises as a member of the family of smooth alternatives. In particular, at the rejection of the null hypothesis, the density estimator in (3.90) where the order is determined by the BIC criterion provides additional information about how the true distribution deviates from the hypothesised f_0. In this way, model estimation and GOF testing combine to give the data-analyst more statistical information.

Since the density estimators described above are not always guaranteed to be positive, we apply the correction proposed by Glad and Hjort (2003) when such a situation occurs. Suppose that $\int_{\mathcal{S}} \max\{0, \hat{f}\}dx \geq 1$, then the modified estimator is simply given by

$$\tilde{f}(x) = \max\{0, \hat{f}(x) - \epsilon\}, \tag{3.93}$$

where ϵ is chosen such that $\int_{\mathcal{S}} \tilde{f}(x)dx = 1$. The authors showed that the corrected estimator $\tilde{f}$ has smaller MISE than the original $\hat{f}$.

Example 3.7.7. We give an illustration of the previously described approach for linear data. In particular, we compute four versions of the orthonormal series estimator (3.90) for the density of the Fastfood data from Section 2.1.4. We choose the starting distribution f_0 to be either uniform or normal, and we use either the MISE or the BIC criterion to select the order of the expansion. In case of the uniform starting distribution, the orthonormal series expansion is based on the Legendre polynomials, which are given in Appendix A.2. Note that the set of trigonometric functions would also be a good choice. Since these functions are related to circular distributions, we discuss this type of estimators in Chapter 4. On the other hand, in case of the normal starting distribution, the orthonormal series expansion is based on the Hermite polynomials, which are given in Appendix A.1. The corresponding estimate is referred to as the *Hermite series estimate*, while the estimate with uniform starting density is referred to as the *Legendre series estimate*.

The resulting density estimates are shown in Figure 3.2. Note that we only consider nested models for both series density estimators. The kernel density estimate for which the window width is determined by means of UCV is added

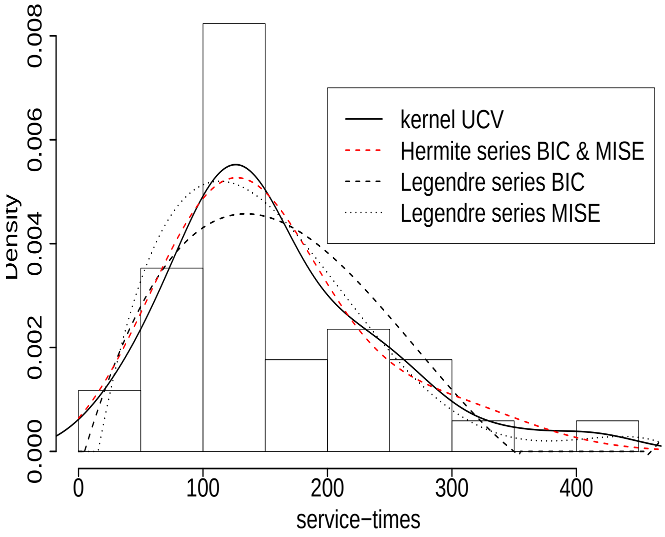


Figure 3.2: Density estimates for the Fastfood data. The kernel density estimate with window width determined by means of UCV and the Hermite and Legendre series density estimates based on BIC and MISE criteria are plotted.

to the plot as well. Let us first consider the Hermite series estimates. The data-driven smooth GOF test for normality is significant (p=0.022) at the 5% level. Since the BIC criterion chooses order 3, the deviation from normality is likely due to the skewness. The density estimate that results from this order selection in the family of smooth alternatives also confirms the positively skewed impression. Since the MISE criterion similarly selected only up to the third component for the construction of the Hermite series density estimate, both estimates coincide on the plot. This estimate is furthermore close to the kernel density estimate.

On the other hand, for the Legendre series estimate, the BIC criterion selected all components up to the third order while the MISE criterion selected all components up to the fourth order. Both Legendre series estimates have negative density values and are thus modified according to the algorithm of Glad and Hjort (2003) described before. They are more smooth and further away from the kernel density estimate than the Hermite series estimate. Since we have no knowledge about the true density we can not say which estimator performs best in this particular example. However, it has been argued several times that orthonormal series density estimates for which the starting distribution is already close to the true distribution are usually preferred (see also Hjort and Glad (1995) and Buckland (1992)). For this reason it may be assumed that the estimate based on the normal distribution will be the most accurate one. In Chapter 6, we use a new graphical tool to explore which density estimate is a good approximation to the true distribution.

For GOF purposes, the BIC criterion is a natural choice. However, this model selection criterion does not necessarily result in a good density estimate. For estimation purposes, we know that the MISE criterion is often preferred. In this example, apart from estimation purposes, we specifically would like to know in what sense the true distribution deviates from the null distribution. We see from Figure 3.2 that with respect to the normal density, the true density probably has a longer right tail. The difference between the true distribution and the uniform distribution is much larger. The Legendre series estimate based on the BIC criterion suggests that the largest difference is probably situated in the first moment. Note that this density estimate also suggests a larger right tail, however not very noticeable. The Legendre series density estimate based on the MISE criterion shows a more pronounced right tail and a more peaked density than the Legendre series estimate based on the BIC criterion. All these results confirm the skewed impression we had earlier in Section 2.1.4.

As we mentioned in Section 3.3.5, the Barton version of the family of alternatives to circular uniformity in (3.38) is proposed by Fernández-Durán (2004) to find an estimate of the density for circular data. This family of circular densities is written as

$$g_k(x,\boldsymbol{\theta}) = \frac{1}{2\pi}\left[1 + \sum_{j=1}^{k}\left(\theta_{2j-1}\sqrt{2}\cos(jx) + \theta_{2j}\sqrt{2}\sin(jx)\right)\right] \quad 0 < x < 2\pi. \tag{3.94}$$

Fernández-Durán (2004) used the AIC criterion to determine the optimal order of the estimate. Note that this family is constructed from functions orthonormal to the circular uniform distribution. In the next chapter we generalise the family of smooth alternatives in (3.94) to a family constructed from a general starting distribution. We refer to Section 4.5 for a more elaborate discussion on orthonormal series density estimates for circular data.

CHAPTER 4

Smooth goodness-of-fit tests on the circle

From the previous chapter we know that applying a smooth test to solve the GOF problem for a linear distribution has some important advantages. For example, the test statistic is easy to compute and the orthogonal decomposition of the statistic often leads to easily interpretable components. In particular, each of the components can be used as a directional test that reveals how the true distribution deviates for the hypothesised. Moreover, the components sum up to a test statistic with omnibus features by which we mean that most interesting deviations from the hypothesised distribution are detected, at least in large samples. Another interesting property of a smooth test is that from its construction an estimator of the true distribution function naturally arises.

In the context of circular distributions, the difficulty with the construction of smooth tests is to find appropriate orthonormal functions. The reason is that the class of orthonormal polynomials with respect to a circular distribution is only defined in the complex field. Until now, only a smooth test for circular uniformity has been proposed, based on basic trigonometric functions (see Section 3.3.5), which are real-valued functions. As in the linear case, any circular distribution can be tested for by first applying the PIT. However, performing smooth tests on transformed data has two important drawbacks. The first difficulty arises when interpreting the results, since it has to be done on the transformed data. The second disadvantage comes into play when p nuisance parameters are involved. For such a situation, which usually happens in practice, we explained in Section 3.3.3 that the first p components of the appropriate score statistic are identically zero when MLE and MME coincide. This interesting property is no longer valid for the score test based on the transformed data. These disadvantages motivate the search for smooth tests that are constructed explicitly for any arbitrary circular distribution.

In this chapter we propose a new class of smooth tests for circular distributions using the general theory of orthonormal polynomials on the unit circle (see e.g. Simon, 2005). In Section 4.1 we first introduce a "complex" framework for smooth tests on the unit circle, where the observations on the circle and the family of order k alternatives are defined on the field of complex numbers. The corresponding score test, however, is again a real-valued statistic which is asymptotically χ^2 distributed, and it has again an interpretation in terms of moment deviations. Since we apply the test to circular data, the origin-invariance property needs to be checked. We propose a straightforward adaptation of the statistic when it is not origin-invariant. Section 4.2 explains how this construction leads to the smooth test of Bogdan et al. (2002) in case of testing for circular uniformity. In Section 4.3, we apply the method to testing for a family of von Mises distributions. This test is a generalization of the test proposed by Barndoff-Nielsen and Cox (1979). Section 4.4 is devoted to the data-driven version of the new smooth test. The results of that data-driven smooth test are combined in Section 4.5 with a nonparametric estimate of the true circular density which arises naturally as a member from the proposed family of smooth models. All methods are illustrated on real data examples in Section 4.6. Some characteristics of the smooth test for circular normality are investigated in a simulation study in Section 4.7. Finally, in Section 4.8, we give a brief discussion of the proposed methodology.

4.1 General Construction

In this section the theory described in Rayner and Best (1989) is generalised in order to obtain the smooth tests for complex-valued circular data based on a complex smooth order k family of alternatives. Working with complex numbers allows a straightforward construction of smooth tests for any circular distribution. Specifically, a set of orthonormal polynomials with respect to the hypothesised circular distribution is required for the construction of the smooth family of alternatives, and such a set is in general only defined in the complex field. The orthonormal functions do not necessarily need to be polynomials. However, in our construction we choose to look for a set of polynomials with respect to a circular distribution since they are uniquely defined, form a complete basis and can easily be derived from recurrence relations.

4.1.1 Complex values on the circle

From now on we work with complex values for the data on the unit circle, instead of projecting the data points on the real interval $[0, 2\pi]$. Let $z_1, \ldots z_n$ denote the n observations on the unit circle expressed as complex numbers. Since the modulus of each observation is one, the values can be written as

$$z_j = e^{ix_j} = \cos x_j + i \sin x_j, \text{ for } j = 1, \ldots n, \tag{4.1}$$

where $i^2 = -1$ and $x_1, \ldots, x_n$ are the directions on the unit circle. The complex conjugate of z is denoted by $\overline{z}$ and equals $e^{-ix} = \cos x - i \sin x$. In what follows, we use the x and z notations interchangeably. This should not lead to confusion because Equation (4.1) states a one to one relation between z and x.

4.1.2 Construction of the smooth model

In this section we construct the smooth model based on a set of complex polynomials orthonormal to the hypothesised distribution $f_0(x, \boldsymbol{\beta})$. The model is referred to as the *complex smooth model*. In the next section we will derive the score statistic for the parameters in the complex smooth model, which is called the *complex score statistic*. Note that the asymptotic null distribution of the complex score statistic is not as usual because the score statistic is in terms of the complex-valued observations. However, it turns out that the asymptotic null distribution of the complex statistic is equal to the asymptotic distribution of an equivalent score statistic in terms of the real-valued directions on the circle. In particular, the equivalent score statistic, which is referred to as the *real score statistic*, is derived from a reparameterisation of the complex smooth model. This reparameterisation is given in this section and is referred to as the *real smooth model* since it is based on real-valued polynomials. The set of real-valued polynomials is however not necessarily orthonormal to f_0. Therefore we will later prefer to work with the complex smooth model. Note that by the equivalence described before, the general theory of Rayner and Best (1989) for real smooth tests can easily be be translated to complex smooth tests.

When testing the null hypothesis $H_0 : F(x) = F_0(x, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a known or unknown p-dimensional real-valued parameter, we consider the order k complex smooth family

$$g_k(x; \boldsymbol{\theta}, \boldsymbol{\beta}) = C(\boldsymbol{\theta}; \boldsymbol{\beta}) \exp \left[\sum_{j=1}^{k} \left(\theta_j h_j(z; \boldsymbol{\beta}) + \overline{\theta}_j \overline{h_j(z; \boldsymbol{\beta})} \right) \right] f_0(x, \boldsymbol{\beta}), \tag{4.2}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_k)$ denotes a complex parameter vector, $C(\boldsymbol{\theta}; \boldsymbol{\beta})$ is a normalising constant, and $\{h_1(z; \boldsymbol{\beta}), \ldots, h_k(z; \boldsymbol{\beta})\}$ is a set of polynomials orthonormal on the unit circle with respect to the density $f_0(x, \boldsymbol{\beta})$, i.e. they satisfy

$$\int_0^{2\pi} h_l(z; \boldsymbol{\beta})\overline{h_m(z; \boldsymbol{\beta})} f_0(x, \boldsymbol{\beta}) dx = \delta_{l,m}, \tag{4.3}$$

where $\delta_{l,m}$ is the Kronecker delta. The general theory of such polynomials is described by Simon (2005). The sequence of polynomials is unique if we agree on making the leading coefficient positive. This unique set of polynomials in z is then of the form

$$h_j(z; \boldsymbol{\beta}) = \kappa_{j,j} z^j + \kappa_{j,j-1} z^{j-1} + \ldots + \kappa_{j,0}, \ \kappa_{jj} > 0 \ \ j = 0, 1 \ldots, \tag{4.4}$$

where the coefficients $\kappa_{j,t}$ $(t = 0, \ldots, j)$ depend on the parameter vector $\boldsymbol{\beta}$ in $f_0(x; \boldsymbol{\beta})$. For the monic polynomials $q_j(z; \boldsymbol{\beta}) = \frac{h_j(z;\boldsymbol{\beta})}{\kappa_{jj}}$, the three term recurrence relation for orthogonal polynomials on the real line (see e.g. Szegö, 1975 or Chihara, 1978) is replaced by the Szegö recurrence

$$z q_j(z; \boldsymbol{\beta}) = q_{j+1}(z; \boldsymbol{\beta}) + \overline{\alpha_j} q_j^\star(z; \boldsymbol{\beta}), \ \ j = 0, 1, \ldots \tag{4.5}$$

where $q_j^\star(z; \boldsymbol{\beta}) = z^j \overline{q_j(z; \boldsymbol{\beta})}$ is the reversed polynomial and the α_j are called the *Verblunsky coefficients*. The latter satisfy $\alpha_j = -\overline{\kappa_{j+1,0}}$, $|\alpha_j| < 1$ for $j \geq 0$ and $\alpha_{-1} = -1$.

The expression of the family of smooth alternatives (4.2) is different from that for a smooth family of alternatives on the real line. In particular, the orthonormal functions $\{h_j\}$ are complex-valued functions in terms of $z = e^{ix}$ and the parameters $\theta_j, j = 1, \ldots, k$ are complex values for which both real and imaginary parts have to be considered. In order to make the density a meaningful and therefore real-valued function, the complex conjugate of $\sum_{j=1}^k \theta_j h_j(z; \boldsymbol{\beta})$ is added in the exponent of the expression. Note here that a polynomial h_j is not necessarily orthogonal to a polynomial $\overline{h_k}$ for $j \neq k$. The reason is that the orthonormality relation in (4.3) does not guarantee

$$\int_0^{2\pi} h_l(z; \boldsymbol{\beta}) h_m(z; \boldsymbol{\beta}) f_0(x, \boldsymbol{\beta}) dx = \delta_{l,m}, \tag{4.6}$$

or

$$\int_0^{2\pi} \overline{h_l(z; \boldsymbol{\beta}) h_m(z; \boldsymbol{\beta})} f_0(x, \boldsymbol{\beta}) dx = \delta_{l,m}. \tag{4.7}$$

As we will see in the next part of this chapter, the relations (4.6)-(4.7) are satisfied if f_0 is the CU distribution, while they are not satisfied if f_0 is e.g. the CN distribution.

Despite the difference in expression between smooth families on the real line and on the circle, there exists a one to one relation between the two families.

This relation is easily seen after reparameterising the expression in (4.2). This reparameterisation is used in the next section to develop the real score statistic, which turns out to be equal to the complex score statistic derived from the complex smooth family in (4.2). Suppose that the orthonormal polynomials $\{h_j\}$ in z have real coefficients $\kappa_{j,t}$ for $t \leq j = 0, \dots, k$. When f_0 is one of the circular distributions we consider in this thesis, namely the CU or the CN distribution, the orthonormal polynomials indeed have real coefficients. This means that $\overline{h_j(z;\boldsymbol{\beta})} = h_j(\overline{z};\boldsymbol{\beta})$, $j = 0, \dots, k$. Moreover, for every $j = 0, \dots, k$, the real and imaginary part of $h_j(z;\boldsymbol{\beta}) = h_j(\cos x + i \sin x;\boldsymbol{\beta})$ can be written as $h_j^c(x;\boldsymbol{\beta})$ and $h_j^s(x;\boldsymbol{\beta})$, respectively, where $h_j^c(x;\boldsymbol{\beta}) = \sum_{t=0}^{j} \kappa_{jt} \cos(tx)$ and $h_j^s(x;\boldsymbol{\beta}) = \sum_{t=0}^{j} \kappa_{jt} \sin(tx)$ for $j = 0, \dots, k$. Note that the polynomial in $\sin x$ has no constant term since $\sin 0 = 0$. This is in conformity with the imaginary part of the original polynomial $h_j(e^{ix};\boldsymbol{\beta})$, which has no constant part either. Let $\theta_j = \frac{1}{2}(\theta_{Rj} - i\theta_{Ij})$ for $j = 1, \dots, k$. This definition may seem odd, but it is chosen so that the family of alternatives (4.2) can then conveniently be rewritten as

$$g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = C(\boldsymbol{\theta};\boldsymbol{\beta}) \exp\left[\sum_{j=1}^{k} \left(\theta_{Rj} h_j^c(x;\boldsymbol{\beta}) + \theta_{Ij} h_j^s(x;\boldsymbol{\beta})\right)\right] f_0(x,\boldsymbol{\beta}), \tag{4.8}$$

where the set of real-valued polynomials $\{h_j^c(x), h_j^s(x)\}$ is not necessarily orthonormal with respect to the distribution $f_0(x)$. Here, we refer to the definition of orthonormality on the real line, as in Section 3.3.1. Again, as we will see in Sections 4.2 and 4.3, these orthonormality relations are satisfied for the CU distribution while they are not satisfied for the CN distribution.

Nevertheless, both parameterisations of the smooth family of alternatives are useful for the construction of a smooth statistic for testing $H_0 : F(x) = F_0(x;\boldsymbol{\beta})$. Moreover, the smooth statistics derived either from expression (4.2) or from (4.8) are equal. The complex representation in (4.2), however, uses explicitly the complex polynomials, which is convenient for most circular distributions.

4.1.3 Construction of the smooth test

Consider first the complex smooth family of alternatives in (4.2). Testing $H_0 : F(x) = F_0(x,\boldsymbol{\beta})$ is now equivalent to testing $H_0 : \boldsymbol{\theta} = \overline{\boldsymbol{\theta}} = \mathbf{0}$. The score statistic for the latter hypothesis is called the complex score statistic and has the form

$$S_{2k} = \frac{1}{n} \left(\boldsymbol{V}_{\boldsymbol{\beta}}^T, \overline{\boldsymbol{V}}_{\boldsymbol{\beta}}^T\right) \Sigma_{\boldsymbol{V}}^{-1} \begin{pmatrix} \overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \\ \boldsymbol{V}_{\boldsymbol{\beta}} \end{pmatrix}, \tag{4.9}$$

where $\Sigma_{\boldsymbol{V}}$ is the asymptotic complex covariance matrix of $\frac{1}{\sqrt{n}}\boldsymbol{V}^T = \frac{1}{\sqrt{n}}\left(\boldsymbol{V}_{\boldsymbol{\beta}}^T, \overline{\boldsymbol{V}}_{\boldsymbol{\beta}}^T\right)$, evaluated at $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}} = \mathbf{0}$, which is defined as,

$$\begin{aligned} n\Sigma_{\boldsymbol{V}} &= \operatorname{Cov}\left[\begin{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \\ \overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \end{pmatrix}\left(\overline{\boldsymbol{V}}_{\boldsymbol{\beta}}^T, \boldsymbol{V}_{\boldsymbol{\beta}}^T\right)\right] \\ &= \mathrm{E}\left[\begin{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \\ \overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \end{pmatrix}\left(\overline{\boldsymbol{V}}_{\boldsymbol{\beta}}^T, \boldsymbol{V}_{\boldsymbol{\beta}}^T\right)\right] - \mathrm{E}\left[\begin{array}{c} \boldsymbol{V}_{\boldsymbol{\beta}} \\ \overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \end{array}\right]\mathrm{E}\left[\overline{\boldsymbol{V}}_{\boldsymbol{\beta}}^T, \boldsymbol{V}_{\boldsymbol{\beta}}^T\right]. \end{aligned} \tag{4.10}$$

Note that because of the above definition, which is taken from Schreier and Scharf (2003), the covariance matrix is no longer equivalent to the usual Fisher information matrix. Nevertheless, the test statistic (4.9) is still asymptotically χ^2 distributed under the null hypothesis. The reason is that the complex score statistic (4.9) is equal to the score statistic for testing

$$H_0 : \boldsymbol{\theta}_R = \boldsymbol{\theta}_I = \mathbf{0},$$

where $\boldsymbol{\theta}_R = (\theta_{R1}, \ldots, \theta_{Rk})$ and $\boldsymbol{\theta}_I = (\theta_{I1}, \ldots, \theta_{Ik})$ in the real smooth model (4.8), which is the reparameterised family of alternatives. In particular, this real score statistic has the form

$$S_{2k}^{\star} = \frac{1}{n}\left(\boldsymbol{W}_{R,\boldsymbol{\beta}}^T, \boldsymbol{W}_{I,\boldsymbol{\beta}}^T\right)\Sigma_{\boldsymbol{W}}^{-1}\begin{pmatrix} \boldsymbol{W}_{R,\boldsymbol{\beta}} \\ \boldsymbol{W}_{I,\boldsymbol{\beta}} \end{pmatrix}, \tag{4.11}$$

where $\Sigma_{\boldsymbol{W}}$ is the asymptotic covariance matrix of the score vector $\frac{1}{\sqrt{n}}\boldsymbol{W}^T = \frac{1}{\sqrt{n}}\left(\boldsymbol{W}_{R,\boldsymbol{\beta}}^T, \boldsymbol{W}_{I,\boldsymbol{\beta}}^T\right)$, evaluated at $\boldsymbol{\theta}_R = \boldsymbol{\theta}_I = \mathbf{0}$, which is defined as usual,

$$\Sigma_{\boldsymbol{W}} = \operatorname{Cov}\left[\begin{pmatrix} \boldsymbol{W}_{R,\boldsymbol{\beta}} \\ \boldsymbol{W}_{\boldsymbol{I},\boldsymbol{\beta}} \end{pmatrix}\left(\boldsymbol{W}_{R,\boldsymbol{\beta}}^T, \boldsymbol{W}_{I,\boldsymbol{\beta}}^T\right)\right]. \tag{4.12}$$

Here, the covariance matrix is defined in the usual way and is therefore directly related to the usual Fisher information matrix. Thus, the real score statistic (4.11) is asymptotically χ^2 distributed under the null hypothesis with the degrees of freedom depending on how many nuisance parameters in $\boldsymbol{\beta}$ are to be estimated. The transformation to build a complex vector $\boldsymbol{V}$ from its real and imaginary parts (see e.g. Schreier & Scharf, 2003) is used to show the equality of S_{2k} and $S_{2k}^{\star}$. In particular, we have that

$$\begin{pmatrix} \overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \\ \boldsymbol{V}_{\boldsymbol{\beta}} \end{pmatrix} = \boldsymbol{T}\begin{pmatrix} \boldsymbol{W}_{R,\boldsymbol{\beta}} \\ \boldsymbol{W}_{I,\boldsymbol{\beta}} \end{pmatrix}, \tag{4.13}$$

where $\boldsymbol{T} = \begin{pmatrix} I_k & -iI_k \\ I_k & iI_k \end{pmatrix}$, and I_k is the $k \times k$ identity matrix. Thus

$$\Sigma_{\boldsymbol{V}} = \boldsymbol{T}\Sigma_{\boldsymbol{W}}\overline{\boldsymbol{T}}^T,$$

which implies that

$$\begin{aligned}
S_{2k} &= \frac{1}{n}\left(\boldsymbol{W}^T_{R,\boldsymbol{\beta}}, \boldsymbol{W}^T_{I,\boldsymbol{\beta}}\right)\overline{\boldsymbol{T}}^T\left(\boldsymbol{T}\Sigma_{\boldsymbol{W}}\overline{\boldsymbol{T}}^T\right)^{-1}\boldsymbol{T}\begin{pmatrix}\boldsymbol{W}_{R,\boldsymbol{\beta}}\\ \boldsymbol{W}_{I,\boldsymbol{\beta}}\end{pmatrix}\\
&= \frac{1}{n}\left(\boldsymbol{W}^T_{R,\boldsymbol{\beta}}, \boldsymbol{W}^T_{I,\boldsymbol{\beta}}\right)\overline{\boldsymbol{T}}^T\left(\overline{\boldsymbol{T}}^T\right)^{-1}\Sigma_{\boldsymbol{W}}^{-1}\boldsymbol{T}^{-1}\boldsymbol{T}\begin{pmatrix}\boldsymbol{W}_{R,\boldsymbol{\beta}}\\ \boldsymbol{W}_{I,\boldsymbol{\beta}}\end{pmatrix}\\
&= S^\star_{2k}.
\end{aligned}$$

Because of the orthonormality relations the computation of S_{2k} is easier than the computation of $S^\star_{2k}$. Therefore we prefer to use formula (4.9) in the next part of this chapter. For further details about the computation of the test statistic and its asymptotic null distribution, we refer to Sections 4.2 and 4.3 for treatment of testing for circular uniformity and circular normality, respectively.

4.1.4 Origin-invariance

The resulting test statistic S_{2k} in its general form is not necessarily origin-invariant. Assume, without loss of generality, that the covariance matrix $\Sigma_{\boldsymbol{V}}$ is the identity matrix I_{2k} and omit the index $\boldsymbol{\beta}$ for notational comfort. The test statistic can then be rewritten as

$$\begin{aligned}
S_{2k} &= \frac{2}{n}\boldsymbol{V}^T\overline{\boldsymbol{V}}\\
&= \frac{2}{n}\sum_{j=1}^{k}\left[\sum_{l=1}^{n}h_j(e^{iX_l})\right]\left[\sum_{l=1}^{n}h_j(e^{-iX_l})\right]. \qquad (4.14)
\end{aligned}$$

On the other hand, adding a constant γ to each X_l, results in

$$\frac{2}{n}\sum_{j=1}^{k}\left[\sum_{l=1}^{n}h_j(e^{i(X_l+\gamma)})\right]\left[\sum_{l=1}^{n}h_j(e^{-i(X_l+\gamma)})\right] \qquad (4.15)$$

which is not necessarily equal to (4.14) for a general orthonormal set of polynomials $\{h_j\}$. In fact, (4.14) and (4.15) are equal only if $h_j(e^{\pm i(X_l+\gamma)})$ can be rewritten as $e^{\pm ij\gamma}h_j(e^{\pm iX_l})$ for every j, which is only the case if the set of polynomials $\{h_j\}$ are monomials of order j. The set of monomials in e^{ix} corresponds to the set of orthonormal polynomials for the CU distribution, for which we refer to the next section.

The score statistic (4.9) can be made origin-invariant by replacing the random variables $X_1, X_2, \ldots, X_n$ with their centered counterparts

$$X_1 - \overline{X}^c_n, X_2 - \overline{X}^c_n, \ldots, X_n - \overline{X}^c_n,$$

which is a maximal origin-invariant function (see Section 3.1.3). Note that $\overline{X}^c_n$ is not the conventional sample mean of directions but the estimator of the circular mean direction as defined in Section 2.2. This estimator is origin equivariant. On could think of applying such a trick to any other statistic which is originally constructed for linear distributions, in order to make it origin-invariant and therefore applicable to circular data. Although the statistic is then origin-invariant, it is not always guaranteed to be a useful statistic. Nevertheless, in the case of testing for the von Mises distribution (see Section 4.3) this choice naturally arises as $\overline{X}_n^c$ is exactly the MLE of the location parameter of the distribution.

In the next two sections we calculate the complex score statistic in (4.9) and derive its asymptotic null distribution for the special cases of circular uniformity and circular normality.

4.2 Simple null hypothesis

We will first focus on the problem of testing for circular uniformity. As mentioned in Section 3.1.3, if $X_1, \ldots, X_n$ are assumed to be i.i.d. from $F_0(x)$ under the null hypothesis, the GOF test for $H_0 : F(x) = F_0(x)$, where F_0 is completely specified and continuous, can be reduced to a test for uniformity on the unit circle, i.e. $H_0 : F(u) = \frac{u}{2\pi}, 0 \leq u < 2\pi$ based on the transformed sample $u_i = 2\pi F_0(x_i)$, $i = 1 \ldots n$. Here we assume that such a transformation has been made, but, without loss of generality, we use the notation x instead of u.

The orthonormal polynomials with respect to the CU distribution on the unit circle are immediately found. First, note that the Verblunsky coefficients $\alpha_0, \alpha_1, \ldots, \alpha_k$ are zero, because the restrictions of the form

$$\frac{1}{2\pi}\int_0^{2\pi} h_j(z)dz = \frac{1}{2\pi}\int_0^{2\pi} \kappa_{j,j}e^{ijx} + \kappa_{j,j-1}e^{i(j-1)x} + \ldots + \kappa_{j,0}\, dx = 0$$

reduce to $\kappa_{j,0} = 0$ for $j = 1, 2, \ldots, k$. Furthermore, using the Szegö recurrence relation (4.5), we have $h_j(z) = q_j(z) = z^j = e^{jix}$, $j = 0, 1 \ldots, k$. Taking into account that $\frac{1}{\sqrt{n}}\boldsymbol{V}^T$ which equals

$$\frac{1}{\sqrt{n}}\left(\sum_{j=1}^n e^{iX_j}, \sum_{j=1}^n e^{2iX_j}, \ldots, \sum_{j=1}^n e^{kiX_j}, \sum_{j=1}^n e^{-iX_j}, \sum_{j=1}^n e^{-2iX_j}, \ldots, \sum_{j=1}^n e^{-kiX_j}\right) \tag{4.16}$$

has asymptotic covariance matrix $\Sigma_{\boldsymbol{V}} = E[\boldsymbol{V}^T\overline{\boldsymbol{V}}] = I_{2k}$, the score statistic (4.9) becomes

$$\begin{aligned}
S_{2k} &= \frac{1}{n}\boldsymbol{V}^T\overline{\boldsymbol{V}} \\
&= \frac{2}{n}\sum_{j=1}^{k} V_j\overline{V_j} \\
&= \frac{2}{n}\sum_{j=1}^{k}\left(\sum_{l=1}^{n} e^{ijX_l}\right)\left(\sum_{l=1}^{n} e^{-ijX_l}\right) \\
&= \frac{2}{n}\sum_{j=1}^{k}\left(\sum_{l=1}^{n}\cos(jX_l) + i\sin(jX_l)\right)\left(\sum_{l=1}^{n}\cos(jX_l) - i\sin(jX_l)\right) \\
&= \frac{2}{n}\sum_{j=1}^{k}\left[\left(\sum_{l=1}^{n}\cos(jX_l)\right)^2 + \left(\sum_{l=1}^{n}\sin(jX_l)\right)^2\right],
\end{aligned}$$

which is the same smooth test statistic as in Bogdan et al. (2002). However, they started their construction with a different order k family of alternatives,

$$g_k(x,\boldsymbol{\theta}) = C(\boldsymbol{\theta})\exp\left[\sum_{j=1}^{k}\Big(\theta_{2j-1}\sqrt{2}\cos(jx) + \theta_{2j}\sqrt{2}\sin(jx)\Big)\right] \quad 0 < x < 2\pi, \tag{4.17}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_{2k})$ denotes the parameter vector, $C(\boldsymbol{\theta})$ is a normalising constant, and $\{\sqrt{2}\cos(jx), \sqrt{2}\sin(jx)\}$ is a complete set of orthonormal functions on the CU distribution. Note that this family of distributions is equivalent to the proposed family in (4.8), taking $\theta_{Rj} = \sqrt{2}\theta_{2j-1}$ and $\theta_{Ij} = \sqrt{2}\theta_{2j}$. Moreover, this family is also similar to the one suggested by Fernández-Durán (2004) in the context of density estimation. Both families of alternatives (4.2) and (4.8) can thus be used as density estimators for the true circular distribution.

From Bogdan et al. (2002) we also know that, under the null hypothesis, the statistic S_{2k} is origin-invariant and asymptotically χ^2 distributed with $2k$ degrees of freedom. Additionally, the jth component, $\frac{2}{n}V_j\overline{V}_j$, is the squared resultant length of the jth trigonometric moment between the true and the hypothesised distribution. Hence, again the individual two degrees of freedom component can be used as a directional test to detect differences in the jth trigonometric moment. The choice of the order k is discussed in Section 4.4.

4.3 Composite null hypothesis of circular normality

Since many important examples involve the CN distribution, we confine our further discussion to testing the composite null hypothesis of circular normality, i.e.

$$H_0 : f(x) = f_0(x; \mu, \kappa), \tag{4.18}$$

where f_0 can be rewritten as

$$f_0(x; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa cos(x-\mu)} = \frac{1}{2\pi I_0(\kappa)} e^{\frac{1}{2}\kappa(z+\frac{1}{z})}, \quad 0 \leq x < 2\pi,$$

where $z = e^{i(x-\mu)}$, $0 \leq \mu < 2\pi$ and $\kappa > 0$ are nuisance parameters, and $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero which is defined in 3.8. Note that the relation between x and z is slightly different from the general setting in Section 4.1. In particular, the mean direction μ is first subtracted from x before the exponent is taken. The choice for this relation between x and z arises naturally because the von Mises distribution is a location family with μ as "location" parameter. To emphasise the dependence of z on μ we write z_μ in what follows.

4.3.1 MLE and MME

The parameters μ and κ are replaced by their MLE $\hat{\mu}$ and $\hat{\kappa}$, respectively, which are the solutions to

$$\kappa \sum_{i=1}^{n} \sin(X_i - \mu) = 0 \tag{4.19}$$

and

$$-nA(\kappa) + \sum_{i=1}^{n} \cos(X_i - \mu) = 0, \tag{4.20}$$

where $A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$ for which $I_1(\kappa) = \frac{dI_0(\kappa)}{d\kappa}$ is the modified Bessel function of order 1. The set of estimation equations (4.19)-(4.20) reduces to

$$\tan(\mu) = \frac{S}{C} \tag{4.21}$$

and

$$A(\kappa) = \frac{R}{n}, \tag{4.22}$$

where $S = \sum_{i=1}^{n} \sin(X_i)$, $C = \sum_{i=1}^{n} \cos(X_i)$ and $R = \sum_{i=1}^{n} \cos(X_i - \mu)$. These equations give unique solutions for μ and κ. From the first equation it is clear that the MLE of μ is equal to the circular mean direction, which is defined in Section 2.2. That is, we have $\hat{\mu} = \overline{X}_n^c$. The MLE of κ has no explicit expression but can be found by evaluating $A(\kappa)$ for different values of κ. Since A is a non-linear function, the MLE $\hat{\kappa}$ is a biased estimator of κ. Best and Fisher (1981) therefore proposed the estimator

$$\hat{\kappa}^\star = \begin{cases} \max(\hat{\kappa} - 2/(n\hat{\kappa}), 0), & \hat{\kappa} < 2, \\ (n-1)^3\hat{\kappa}/(n^3 + n), & \hat{\kappa} \geq 2, \end{cases} \tag{4.23}$$

which is approximately unbiased unless both n and κ are small. This estimator is asymptotically equivalent to the MLE and is used in what follows. The MMEs of μ and κ involve equating the theoretical and the sample trigonometric moments. The first theoretical trigonometric moment can be derived from (3.9) and is given by

$$E(e^{iX}) = A(\kappa)e^{i\mu},$$

while the first sample trigonometric moment is as in (2.7) and given by

$$\frac{1}{n}\sum_{j=1}^{n} e^{iX_j}.$$

Consequently, we have the set of equations

$$A(\kappa)\cos(\mu) = \frac{1}{n}\sum_{j=1}^{n}\cos(X_j) = \frac{1}{n}C$$

and

$$A(\kappa)\sin(\mu) = \frac{1}{n}\sum_{j=1}^{n}\sin(X_j) = \frac{1}{n}S$$

which results in the same equations as (4.21) and (4.22). This means that for the von Mises distribution, the MLE and MME coincide. Consequently, as we will see soon, the first and the $(k+1)$th elements of the score vector are exactly equal to zero.

4.3.2 The orthonormal polynomials

Recall that the set of orthonormal polynomials $\{h_j(z_{\hat{\mu}}; \hat{\kappa}), j = 0, \ldots, k\}$ with respect to the von Mises distribution on the unit circle can be found using the Szegö recurrence relation (see (4.5)). The Verblunsky coefficients $\alpha_1, \alpha_2, \ldots, \alpha_k$ in that recurrence relation are now found using the non-linear recurrence relation (see Periwal & Shewitz, 1990)

$$-\frac{\kappa}{2}(1-\alpha_j^2)(\alpha_{j+1}+\alpha_{j-1}) = (j+1)\alpha_j \quad \text{for } \alpha_j \neq 0.$$

The initial values are $\alpha_{-1} = -1$ and $\alpha_0 = A(\hat{\kappa})$. Consequently, the Verblunsky coefficients are real and therefore the coefficients κ_{jt}, $t \leq j = 0, \ldots, k$ for the polynomials $\{h_j; j = 0, \ldots, k\}$ in (4.4) are real as well. The first polynomials are

$$h_0(z_{\hat{\mu}}; \hat{\kappa}) = 1 \text{ and } h_1(z_{\hat{\mu}}; \hat{\kappa}) = \frac{e^{i(x-\hat{\mu})} - A(\hat{\kappa})}{\sqrt{1 - A^2(\hat{\kappa})}}.$$

The subsequent polynomials can be found using the Szegö recurrence relation (4.5), which simplifies to

$$z_{\hat{\mu}} q_j(z_{\hat{\mu}}; \hat{\kappa}) = q_{j+1}(z_{\hat{\mu}}; \hat{\kappa}) + \alpha_j z_{\hat{\mu}}^j q_j(1/z_{\hat{\mu}}; \hat{\kappa}).$$

For example, the second order orthogonal polynomial is given by

$$q_2(z_{\hat{\mu}}; \hat{\kappa}) = z_{\hat{\mu}}^2 + z_{\hat{\mu}}(1 - \alpha_1)\alpha_0 - \alpha_1, \tag{4.24}$$

where $\alpha_1 = 1 - \left(\frac{\alpha_0}{1-\alpha_0^2} \frac{2}{\hat{\kappa}} \right)$. In general, the normalised polynomials are $h_j(z_{\hat{\mu}}; \hat{\kappa}) = \frac{q_j(z_{\hat{\mu}}; \hat{\kappa})}{N_j}$, $j = 1, \ldots k$, where

$$N_j = \int_0^{2\pi} q_j(z_{\hat{\mu}}; \hat{\kappa}) q_j(\overline{z_{\hat{\mu}}}; \hat{\kappa}) f_0(x; \hat{\mu}, \hat{\kappa}) dx$$

is the normalising constant.

4.3.3 The efficient score test

Let $\boldsymbol{h}^T(z_{\hat{\mu}}; \hat{\kappa}) = (h_1(z_{\hat{\mu}}; \hat{\kappa}), \ldots, h_k(z_{\hat{\mu}}; \hat{\kappa}))$ so that the score vectors for $\boldsymbol{\theta}$ and $\overline{\boldsymbol{\theta}}$ can be written as $\boldsymbol{V}_{\hat{\mu},\hat{\kappa}}^T = \sum_{t=1}^n \boldsymbol{h}^T(z_{t\hat{\mu}}; \hat{\kappa})$ and $\overline{\boldsymbol{V}}_{\hat{\mu},\hat{\kappa}}^T = \sum_{t=1}^n \overline{\boldsymbol{h}}^T(z_{t\hat{\mu}}; \hat{\kappa})$, respectively. As mentioned before, the first and $(k+1)$th elements of the score vector

$$\boldsymbol{V}^T = \left(\boldsymbol{V}_{\hat{\mu},\hat{\kappa}}^T, \overline{\boldsymbol{V}}_{\hat{\mu},\hat{\kappa}}^T \right)$$

can be removed since they are identically zero. The reason is that the MLE and MME of the parameters μ and κ coincide. In particular, the real and imaginary parts of the first element are equivalent to the left hand side of the estimation equations (4.19) and (4.20), respectively. The same holds for the $(k+1)$th element since it is the complex conjugate of the first element. We therefore also remove the first element h_1 from the vector $\boldsymbol{h}$ and from the family of alternatives (4.2) so that only $k-1$ complex parameters are considered in the vector $\boldsymbol{\theta}$. Note that in practice we use the approximately unbiased estimator $\hat{\kappa}^\star$ defined in (4.23). Since this estimator is asymptotically equal to the MLE $\hat{\kappa}$, the asymptotic results described in this section remain valid when using $\hat{\kappa}^\star$ instead of $\hat{\kappa}$.

For easy computation of the efficient score statistic, we use an equivalent complex expression for the estimation equations (4.19)-(4.20), for which the real and imaginary parts refer to the original estimation equations. This complex estimation equation can for example be written as

$$\sum_{t=1}^n h_1(Z_{t\hat{\mu}}; \hat{\kappa}) = 0$$

or by its complex conjugate

$$\sum_{t=1}^{n} \overline{h}_1(Z_{t\hat{\mu}}; \hat{\kappa}) = 0,$$

where $Z_{t\hat{\mu}} = e^{i(X_t - \hat{\mu})}$. The efficient score test statistic for testing $H_0 : \boldsymbol{\theta} = \overline{\boldsymbol{\theta}} = \mathbf{0}$ becomes

$$S_{2k} = \frac{1}{n} \boldsymbol{V}^T \boldsymbol{\Sigma}_{\boldsymbol{V}}^{-1} \overline{\boldsymbol{V}},$$

where $\boldsymbol{V}$ is the $2(k-1)$-dimensional efficient score vector. The asymptotic covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{V}}$ of $\frac{1}{\sqrt{n}}\boldsymbol{V}$ is no longer diagonal. In particular,

$$\boldsymbol{\Sigma}_{\boldsymbol{V}} = \left[\Sigma_{\boldsymbol{hh}} - \Sigma_{hh_1} \Sigma_{h_1 h_1}^{-1} \Sigma_{hh_1}^T\right],$$

where

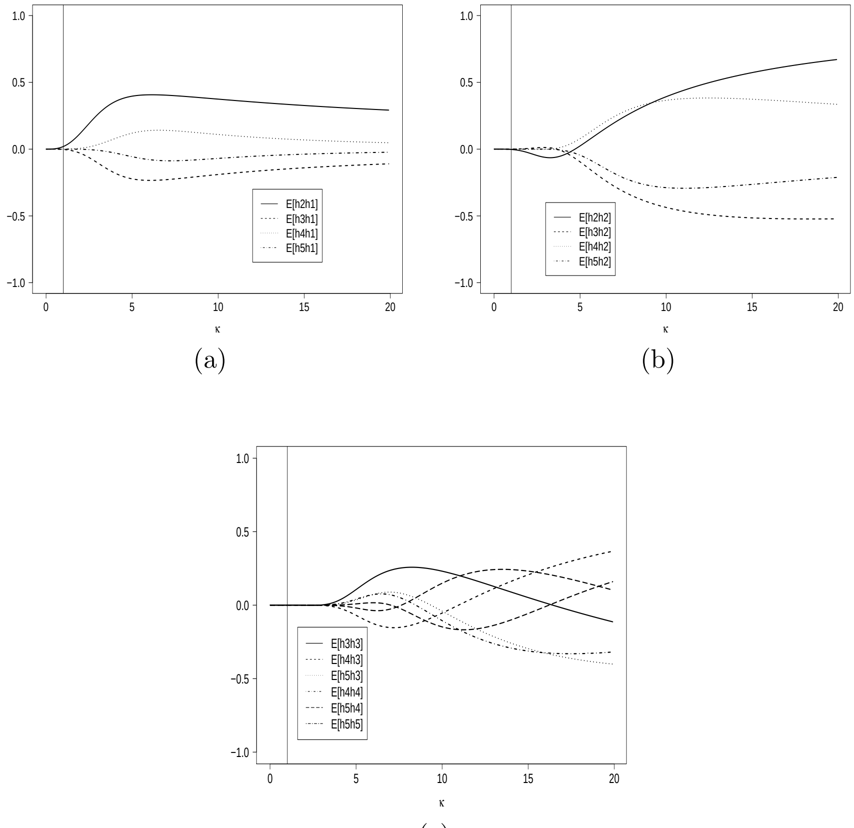
$$\begin{aligned} \Sigma_{\boldsymbol{hh}} &= \begin{pmatrix} I_{k-1} & \mathrm{E}\left[\boldsymbol{h}\boldsymbol{h}^T\right] \\ \mathrm{E}\left[\overline{\boldsymbol{h}}\overline{\boldsymbol{h}}^T\right] & I_{k-1} \end{pmatrix}, \\ \Sigma_{hh_1} &= \begin{pmatrix} \mathbf{0} & \mathrm{E}\left[\boldsymbol{h}h_1\right] \\ \mathrm{E}\left[\overline{\boldsymbol{h}}\overline{h}_1\right] & \mathbf{0} \end{pmatrix} \text{ and} \\ \Sigma_{h_1 h_1} &= \begin{pmatrix} 1 & \mathrm{E}\left[h_1 h_1\right] \\ \mathrm{E}\left[\overline{h}_1 \overline{h}_1\right] & 1 \end{pmatrix}. \end{aligned}$$

Although f_0 belongs to the exponential family, the asymptotic covariance matrix does not reduce to a diagonal form. The reason is that while the individual sets of polynomials $\{h_j; j = 1, \ldots, k\}$ and $\{\overline{h}_j; j = 1, \ldots, k\}$ are orthonormal to f_0, the complete set of polynomials $\{h_j, \overline{h}_j; j = 1, \ldots, k\}$ is not orthonormal to f_0. In particular, from (4.6) and (4.7) we know that $\mathrm{E}\left[h_l h_m\right]$ and $\mathrm{E}\left[\overline{h}_l \overline{h}_m\right]$ are equal but not necessarily zero for $l \neq m$. However, for values of κ smaller than 1, $\mathrm{E}\left[h_l h_m\right] = 0$ holds for every $l \neq m$. This is illustrated in Figure 4.1, where $\mathrm{E}\left[h_l h_m\right]$, for $l \leq m = 1, \ldots, 5$ is plotted versus κ ranging from 0.1 to 20 and where the vertical line indicates $\kappa = 1$. Note that these values are always real because the imaginary part of $\mathrm{E}\left[h_l h_m\right]$ is zero for $l \neq m$. In particular, the imaginary part is equal to $\mathrm{E}\left[h_l^c h_m^s\right] + \mathrm{E}\left[h_m^c h_l^s\right]$, in which both terms are zero. In general, any term $\mathrm{E}\left[h_l^c h_m^s\right]$ can be written as

$$\int_0^{2\pi} \left[(\kappa_{l,l} \cos(lx) + \ldots + \kappa_{l,1} \cos x) h_m^s(x) f_0(x) + \kappa_{l,0} h_m^s(x) f_0(x)\right] dx \qquad (4.25)$$

where we omit dependence on $\hat{\mu}$ and $\hat{\kappa}$ for notational comfort. The last term of the integrand in (4.25) is zero since it is essentially the imaginary part of the orthogonality relation $\mathrm{E}\left[h_m\right] = 0$. The other terms in (4.25) are

$$\kappa_{l,t} \kappa_{m,s} \int_0^{2\pi} \cos(tx) \sin(sx) f_0(x) dx, \quad t = 1, \ldots, l \text{ and } s = 1, \ldots, m$$



(a) (b)

(c)

Figure 4.1: The values for $\mathrm{E}\left[h_l h_m\right] = \mathrm{E}\left[\overline{h}_l \overline{h}_m\right], l \leq m = 1, \ldots, 5$ are plotted versus κ. In Panel (a) m=1, Panel (b) m=2 and Panel (c) m=3,4,5.

where the integrand is an odd function over the unit circle, for which the integral vanishes.

As $\mathrm{E}\left[h_l h_m\right] = 0$ for $l \neq m$ holds for $\kappa < 1$, $\Sigma_{\boldsymbol{hh}}$ reduces to the identity matrix and $\Sigma_{\boldsymbol{h}h_1}$ reduces to the null matrix. Hence, for $\kappa < 1$, $\boldsymbol{\Sigma_V}$ reduces to the diagonal matrix $I_{2(k-1)}$.

Let $X^\star = X - \hat{\mu} = X - \overline{X}_n^c$, then from Section 4.1 we know that $(X_1^\star, \ldots, X_n^\star)$ is a maximal invariant function. Hence, for $\kappa < 1$, the statistic is equal to

$$\begin{aligned} S_{2k} &= \frac{1}{n} \sum_{j=2}^{k} V_{(\hat{\mu},\hat{\kappa})j} \overline{V}_{(\hat{\mu},\hat{\kappa})j} \\ &= \frac{2}{n} \sum_{j=2}^{k} \left(\sum_{l=1}^{n} h_j \left[e^{iX_l^\star} \right] \right) \left(\sum_{l=1}^{n} h_j \left[e^{-iX_l^\star} \right] \right) \\ &= \frac{2}{n} \sum_{j=2}^{k} \left[\left(\sum_{l=1}^{n} h_j^c(X_l^\star) \right)^2 + \left(\sum_{l=1}^{n} h_j^s(X_l^\star) \right)^2 \right] \end{aligned} \tag{4.26}$$

where $h_j^c(x)$ and $h_j^s(x)$ for $j = 2, \ldots, k$ are as in Section 4.1. Since the statistic is a function of the maximal invariant $(X_1^\star, \ldots, X_n^\star)$, it is an origin-invariant test statistic. The asymptotic results are the same as for the simple null hypothesis. In particular, the null distribution of the smooth test S_{2k} is asymptotically $\chi^2_{2(k-1)}$ distributed. Moreover, the jth component is approximately equal to $\frac{2}{n} V_{(\hat{\mu},\hat{\kappa})j} \overline{V}_{(\hat{\mu},\hat{\kappa})j}$, which is the squared resultant length of the jth trigonometric moment and can therefore be used as a directional test to detect differences in the jth trigonometric moment. Note that the decomposition in (4.26), and therefore also the directional interpretation, only holds if $\kappa < 1$. As in the linear case (see Section 3.3.2), we should be careful with the directional interpretation of the components tests. Following Henze and Klar (1996), Henze (1997) and Klar (2000), we also rescale the score statistic by its empirical covariance matrix Σ_{emp} in order to obtain the proper diagnostic interpretation. The test statistic then becomes

$$S^{emp} = \frac{1}{n} \boldsymbol{V}^T \Sigma_{emp}^{-1} \overline{\boldsymbol{V}}$$

and its asymptotic null distribution is as before. Note that here convergence is very slow, similarly as in the linear case.

For $\kappa \geq 1$, such a decomposition is not possible. Therefore, we write the inverse asymptotic covariance matrix as $\Sigma_{\boldsymbol{V}}^{-1} = \Gamma = \begin{pmatrix} \Gamma_{\boldsymbol{V}\overline{\boldsymbol{V}}} & \Gamma_{\boldsymbol{V}\boldsymbol{V}} \\ \Gamma_{\overline{\boldsymbol{V}}\overline{\boldsymbol{V}}} & \Gamma_{\overline{\boldsymbol{V}}\boldsymbol{V}} \end{pmatrix}$, where $\Gamma_{\boldsymbol{V}\overline{\boldsymbol{V}}} = \Gamma_{\overline{\boldsymbol{V}}\boldsymbol{V}}$ and $\Gamma_{\boldsymbol{V}\boldsymbol{V}} = \Gamma_{\overline{\boldsymbol{V}}\overline{\boldsymbol{V}}}$. The efficient score statistic can now be written as

$$S_{2k} = \frac{2}{n} \boldsymbol{V}_{\hat{\mu},\hat{\kappa}}^T \Gamma_{\boldsymbol{V}\overline{\boldsymbol{V}}} \overline{\boldsymbol{V}}_{\hat{\mu},\hat{\kappa}} + \frac{2}{n} \boldsymbol{V}_{\hat{\mu},\hat{\kappa}}^T \Gamma_{\boldsymbol{V}\boldsymbol{V}} \boldsymbol{V}_{\hat{\mu},\hat{\kappa}} \tag{4.27}$$

and is asymptotically $\chi^2_{2(k-1)}$ distributed.

4.3.4 Relation to the BarCox test

In this section we show the equivalence between the BarCox test described in Section 3.6 and our smooth test of order $k = 2$. The score vector of the smooth test of order 2 is of the form

$$\boldsymbol{V} = \begin{pmatrix} \sum_{t=1}^{n} e^{i2X_t^\star} + e^{iX_t^\star}(1-\alpha_1)\alpha_0 - \alpha_1 \\ \sum_{t=1}^{n} e^{-i2X_t^\star} + e^{-iX_t^\star}(1-\alpha_1)\alpha_0 - \alpha_1 \end{pmatrix}.$$

From Section 4.1, we know that this vector is equivalent to the vector $\boldsymbol{W}$ containing the real and imaginary parts of the first component of $\boldsymbol{V}$. In particular, using the transformation $\boldsymbol{T}$ defined in that section, the relation between the two score vectors is $\boldsymbol{V} = \boldsymbol{T}\boldsymbol{W}$, where

$$\boldsymbol{W} = \begin{pmatrix} \sum_{t=1}^{n} \cos(2X_t^\star) + \cos(X_t^\star)(1-\alpha_1)\alpha_0 - \alpha_1 \\ \sum_{t=1}^{n} \sin(2X_t^\star) + \sin(X_t^\star)(1-\alpha_1)\alpha_0 \end{pmatrix}.$$

This vector is equivalent to the vector $(B_c, B_s)^T$ on which the BarCox test is based. Indeed, the first component of $\boldsymbol{W}$ is based on the score function $h_2^c(x^\star)$ which can be written as

$$h_2^c(x^\star) = (\cos(2x^\star) - \mathrm{E}[\cos 2X^\star]) + \alpha_0(1-\alpha_1)(\cos x^\star - \mathrm{E}[\cos X^\star])$$

where we have that $\mathrm{E}[\cos X^\star] = \alpha_0 = \frac{I_1}{I_0}$ and $\mathrm{E}[\cos 2X^\star] = \frac{I_2(\hat{\kappa})}{I_0(\hat{\kappa})} = \alpha_0^2 - \alpha_0^2\alpha_1 + \alpha_1$. The latter equality is due to the recurrence relation for the modified Bessel functions of the first kind of natural order m, which results in $I_2(\hat{\kappa}) = -\frac{2I_1(\hat{\kappa})}{\hat{\kappa}} + I_0(\hat{\kappa})$ for $m = 2$. Similarly, the second component of $\boldsymbol{W}$ is based on the score function $h_2^s(x^\star)$. This score function is trivially rewritten as

$$h_2^s(x^\star) = (\sin(2x^\star) - \mathrm{E}[\sin 2X^\star]) + \alpha_0(1-\alpha_1)(\sin x^\star - \mathrm{E}[\sin X^\star]),$$

since $\mathrm{E}[\sin 2X^\star] = \mathrm{E}[\sin X^\star] = 0$. On the other hand, we have that B_c and B_s are based on the functions $(\cos(2x^\star) - \mathrm{E}[\cos 2X^\star])$ and $(\sin(2x^\star) - \mathrm{E}[\sin 2X^\star])$, respectively. Hence the equivalence between the two tests is established.

4.4 Data-driven smooth tests

We propose to choose the order k of the family of alternatives in a similar way as proposed by Bogdan et al. (2002) (see Section 3.3.5). This choice is essentially the order k for which the BIC criterion is maximised. Reconsider the general form for the complex score statistic S_{2k} in (4.9) for testing $H_0 : \boldsymbol{\theta} = \overline{\boldsymbol{\theta}} = \boldsymbol{0}$ in the order k family of complex smooth alternatives (4.2). The BIC criterion is defined as

$$\mathrm{BIC}_n(k, \theta) = S_{2k} - p\log(n),$$

where the last term is the penalty term that accounts for the model complexity in which p is the degrees of freedom. Let m denote a finite positive integer. The order selection rule can then be written as

$$K_{BIC} = \inf\{k : 1 \leq k \leq m, \mathrm{BIC}_n(k, \theta) \geq \mathrm{BIC}_n(j, \theta), j = 1 \ldots, m\},$$

and the resulting data-driven smooth test statistic is $S_{2K_{BIC}}$. Although Bogdan et al. (2002) allowed the maximal order m to grow to infinity with increasing sample size n, we prefer to follow the convention of Claeskens and Hjort (2004) by considering the maximal order m as fixed and finite. This reflects real situations where the order is always limited to a certain finite maximum. Following Claeskens and Hjort (2004), we also consider the AIC criterion for the order selection. The AIC is given by

$$\mathrm{AIC}_n(k, \theta) = S_{2k} - 2p,$$

and the corresponding selection rule is given by

$$K_{AIC} = \inf\{k : 1 \leq k \leq m, \mathrm{AIC}_n(k, \theta) \geq \mathrm{AIC}_n(j, \theta), j = 1 \ldots, m\}.$$

The corresponding data-driven smooth test statistic is given by $S_{2K_{\mathrm{AIC}}}$. Henceforth, we denote the data-driven smooth test statistics by S_{AIC} and S_{BIC}, where the lower index indicates which criterion is used to determine the order of the family of alternatives in Equation (4.2). The null distributions for the data-driven statistics based on the AIC and the BIC order selectors are both obtained by simulation. We refer to Claeskens and Hjort (2004) for more details.

4.5 Nonparametric density estimation

Similarly as in the linear case, the results of the data-driven smooth test for a circular distribution described in Section 4.4 can be associated with an orthonormal series density estimator. The orthonormal series expansion that is used to find the density estimate, is the Barton version of the order k complex smooth family of alternatives proposed in (4.2), i.e.

$$g_k(x; \boldsymbol{\theta}, \boldsymbol{\beta}) = \left[1 + \sum_{j=1}^{k} \left(\theta_j h_j(z; \boldsymbol{\beta}) + \overline{\theta}_j \overline{h_j(z; \boldsymbol{\beta})}\right)\right] f_0(x, \boldsymbol{\beta}), \tag{4.28}$$

where $\boldsymbol{\beta}$ is assumed to be either known or replaced by its MLE. Since, in general, the orthonormality relation for the polynomials $\{h_j\}$ on the unit circle is limited to the relation (4.3), and does not guarantee the relations (4.6)-(4.7) to be satisfied, we are not able to find simple expressions for θ_j or at least not as simple as those that were found in the linear case (see (3.89)). The expressions for θ_j are now

$$\mathrm{E}_k\left[\overline{h_j(Z;\boldsymbol{\beta})}\right] = \theta_j + \sum_{l=1}^{k} \overline{\theta}_l \int_{-\infty}^{\infty} \overline{h_j(z;\boldsymbol{\beta})h_l(z;\boldsymbol{\beta})} f_0(x,\boldsymbol{\beta})dx, \quad j = 1,\ldots,k \tag{4.29}$$

which is a system of equations for which the solution can be found assuming that the integrals $\int_{-\infty}^{\infty} \overline{h_j(z;\boldsymbol{\beta})h_l(z;\boldsymbol{\beta})} f_0(x,\boldsymbol{\beta})dx = \mathrm{E}_0\left[\overline{h_j(z;\boldsymbol{\beta})h_l(z;\boldsymbol{\beta})}\right]$ exist. Unbiased estimators of θ_j are then obtained by solving the set of equations

$$\frac{1}{n}\sum_{m=1}^{n} \overline{h_j(Z_m;\hat{\boldsymbol{\beta}})} = \hat{\theta}_j + \sum_{l=1}^{k} \overline{\hat{\theta}}_l \mathrm{E}_0\left[\overline{h_j(Z;\hat{\boldsymbol{\beta}})h_l(Z;\hat{\boldsymbol{\beta}})}\right], \quad j = 1,\ldots,k. \tag{4.30}$$

Note that when f_0 is the CU distribution, the family (4.28) reduces to the one proposed by Fernández-Durán (2004). The author used the AIC criterion to determine the optimal order of the model. Within this model we have that $\mathrm{E}_k\left[e^{ijX}\right] = \theta_j$ and hence θ_j can be estimated by

$$\hat{\theta}_j = \frac{1}{n}\sum_{l=1}^{n} e^{-ijX_l}, \tag{4.31}$$

where $\sum_{l=1}^{n} e^{-ijX_l}$ is the $(k+j)$th element of the score vector in (4.16). The estimate for the complex conjugate of θ_j immediately follows since $\hat{\overline{\theta}}_j = \overline{\hat{\theta}}_j$.

If f_0 is the CN distribution $f_0(z;\mu,\kappa)$, the terms $\mathrm{E}_0\left[\overline{h_j(Z;\hat{\mu},\hat{\kappa})h_l(Z;\hat{\mu},\hat{\kappa})}\right]$ vanish when $\kappa < 1$ (see Section 4.3.3) and in that case the parameter estimates are easily found as well. When $\kappa > 1$, estimating θ_j is more complicated. To simplify the computations, we assume from now on that $\mathrm{E}_0\left[\overline{h_j(Z;\hat{\mu},\hat{\kappa})h_l(Z;\hat{\mu},\hat{\kappa})}\right] = \mathrm{E}_0\left[h_j(Z;\hat{\mu},\hat{\kappa})h_l(Z;\hat{\mu},\hat{\kappa})\right] = 0$ for all j.

When in (4.28) we replace the vector $\boldsymbol{\theta}$ by its unbiased estimator $\hat{\boldsymbol{\theta}}$, we obtain a density estimator given by

$$g_k(x;\hat{\boldsymbol{\theta}},\boldsymbol{\beta}) = \left[1 + \sum_{j=1}^{k}\left(\hat{\theta}_j h_j(z;\boldsymbol{\beta}) + \overline{\hat{\theta}}_j \overline{h_j(z;\boldsymbol{\beta})}\right)\right] f_0(x,\boldsymbol{\beta}). \tag{4.32}$$

In our discussion we choose either the CU or the CN distribution (with $\kappa < 1$) as starting distribution f_0 and refer to these estimates as the *CU series* and the *CN series* density estimates, respectively. We propose three different selection criteria to determine the optimal order. First, similar to the linear case in Section 3.7, we can optimise the order by minimising the weighted ISE criterion. The circular analogue is given by

$$\Lambda(\hat{g}_k) = \int_{-\infty}^{\infty} \frac{(\hat{g}_k(x) - f(x))\overline{(\hat{g}_k(x) - f(x))}}{f_0(x;\boldsymbol{\beta})}dx. \tag{4.33}$$

The expected value of this loss function can be written as

$$\mathrm{E}\left[\Lambda(\hat{g}_k)\right] = \frac{2}{n}\sum_{j=1}^{k}(d^\star_j - \theta_j\overline{\theta_j}) + 2\sum_{j=k+1}^{\infty}\theta_j\overline{\theta_j}, \tag{4.34}$$

where $d^\star_j = \mathrm{E}\left[h_j(Z)\overline{h_j(Z)}\right]$, with $Z = e^{iX}$. An unbiased estimator $\hat{\Lambda}(\hat{g}_k)$ of $\mathrm{E}\left[\Lambda(\hat{g}_k)\right]$ is obtained if $d^\star_j$ and $\theta_j\overline{\theta_j}$ in (4.34) are replaced by their respective unbiased estimators,

$$\hat{d^\star_j} = \frac{1}{n}\sum_{l=1}^{n}h_j(Z_l)\overline{h_j(Z_l)}$$

and

$$\theta_i\hat{\overline{\theta}}_i = \frac{1}{n-1}\left(n\hat{\theta}_i\hat{\overline{\theta}}_i - \hat{d^\star_j}\right).$$

Minimising $\hat{\Lambda}(\hat{g}_k)$ results in the decision rule:

Include the j*th term until it fails the test*

$$\hat{\theta}_j\hat{\overline{\theta}}_j > \frac{2}{n+1}\hat{d^\star_j}.$$

The other two possibilities considered here are the AIC and BIC criteria used in the definitions of the data-driven smooth tests.

When the resulting density estimates are non-positive or not appropriately normalised, we use the correction of Glad and Hjort (2003), as described in Section 3.7.

We illustrate the usefulness of these density estimators in combination with the data-driven smooth test on the Arrival example in the next section.

4.6 Examples

In this section we consider four examples of circular data. Each of the examples has been introduced in Chapter 2. An interesting question is whether the underlying distribution for these datasets is CN. To assess the validity of the CN distribution, we apply the data-driven smooth test as well as the Watson, the Kuiper and the Entropy tests for circular normality. We also consider the BarCox test, which is equivalent to the circular smooth test of order $k = 2$. Parametric bootstrap with 100,000 resamples is used to calculate the p-values for all tests. For the Watson test, tables with some critical values are available (Stephens 1985), but bootstrap is preferred here as well, as it allows more precise p-values.

4.6.1 Turtles Data

As discussed in Chapter 2, the Turtles dataset has a bimodal impression, where the two modes are diametrically opposed. Note that the mode at about 60 degrees is much larger than the alleged other mode. If this second mode is really present, we expect that the null hypothesis of the unimodal CN would be rejected. The Watson statistic U_n=0.16 (p=0.001), the Kuiper statistic K_n=1.57 (p=0.008) and the Entropy statistic K_{mn}=5.07 (p=0.002) are all highly indicative against circular normality. The BarCox test, with $B = 24.91$ (p <0.001), shows even more evidence against symmetric unimodality. Here we used the asymptotic p-value since none of the 100,000 bootstrap statistics was larger than the observed statistic. Similarly, the score tests $S_{\text{AIC}} = 33.64$ (p <0.001) and $S_{\text{BIC}} = 24.46$ (p <0.001) are highly significant. The AIC criterion selected components up to the fourth order while the BIC criterion only selected the second order components.

Note that the interpretation in terms of trigonometric moments is probably not strictly valid here because the decomposition in (4.26) only holds for $\kappa < 1$, and κ is estimated as $\hat{\kappa}^\star = 1.12$. On the other hand, in Figure 4.1 we saw that the values of $\mathrm{E}\left[h_l h_m\right]$ for, $l \leq m = 1, \ldots, 5$ and for $\kappa = \hat{\kappa}^\star$ are still very close to zero. The corresponding covariance matrix $\mathbf{\Sigma_V}$ is in fact the identity matrix up to two decimals. Hence a careful interpretation in terms of trigonometric moments is still possible.

The rescaled version $S^{emp}_{\text{AIC}} = 16.32$ (p=0.089) is not significant at the 5% level. Similarly $S^{emp}_{\text{BIC}} = 16.32$ (p=0.019) has a higher p-value as compared to its non-corrected counterpart. Both criteria selected only the second order component. The reason for the higher p-values is that the empirical variance is much higher than the asymptotic variance. These results indicate that scaling the score vectors properly before interpreting the different components, may result in different outcomes. Nevertheless, it is not unnatural that for a relatively small sample size (n=76) one needs to be more careful with the interpretation of the different components. It is of course possible that more data would result in evidence against the von Mises distribution.

Since the properly scaled score test is less powerful for small sample sizes, we formulate a conclusion based on the original smooth test. From these results, we conclude that the data deviate from the CN distribution primarily with respect to the second order trigonometric moment, which is in accordance with the bimodal impression of the data. See also the conclusions from the simulation study in Section 4.7.

As the two modes are opposite to each other, it is of interest whether the movements of the turtles are distributed around one single axis. In fact, it is possible that the movements are drawn from the same distribution apart from being diametrically opposed. To check this, we double the data values and carry out the same tests. More details about this doubling procedure are in Section 2.2. Now, none of the tests shows evidence against the null hypothesis anymore, which suggests that the doubled data indeed may originate from a von Mises distribution.

4.6.2 Ants Data

Regarding the Ants data, we have seen before that the ants clearly prefer the direction of about 180°, which corresponds to the direction where the black target was placed. A von Mises distribution could therefore be a good fit. However, the Watson statistic $U_n = 0.3196$, the Kuiper statistic $K_n = 11.34$ and the Entropy statistic $K_{mn} = 4.29$ all have p-values smaller than 0.001. The BarCox statistic $B = 20.86$ (p <0.001) is also highly significant. From these tests it is clear that the von Mises distribution is not a good representation. How does the true distribution differ from circular normality? Our smooth test gives an answer to that question.

Here the estimated value of κ is $\hat{\kappa}^\star$ =1.538. For this value, the covariance matrix $\boldsymbol{\Sigma_V}$ is the identity matrix up to one decimal. Hence, we can apply the interpretation of the components in terms of trigonometric moments, but again we need to be careful.

The score statistics S_{AIC} and S_{BIC} are both equal to 20.53 and their p-values are p=0.005 and p=0.001, respectively. Both selection criteria chose the order $k = 2$. This suggests that the difference from circular normality is due to the second order trigonometric moment.

In order to give a proper diagnostic interpretation we look at the properly scaled score statistics. In particular, we have $S^{emp}_{\text{AIC}} = 27.59$ (p =0.036) and $S^{emp}_{\text{BIC}} = 18.99$ (p=0.029). Similarly as for the Turtles data, the p-values are much larger than for the usual score tests. However, they are still significant at the 5% significance level. Since the sample size is relatively small as well (n=100), the same comment on the scaling procedure as for the Turtles data is useful here.

For the Ants data, we conclude that there is much evidence against circular normality. The second order trigonometric moment, which is related to the circular skewness and kurtosis (see Section 2.2) is responsible for this deviation.

4.6.3 Direzione Data

For the Direzione data, the raw data plot suggests a unimodal distribution with its mode at the North direction. The Kuiper test ($K_n = 3.781$), the Watson test ($U_n = 1.223$), the Entropy test ($K_{mn} = 5.076$) and the BarCox test ($B =$

83.842) have highly significant p-values ($p < 0.001$). Similarly, the score tests $S_{\text{AIC}} = 112.037$ and $S_{\text{BIC}} = 83.152$ ($p < 0.001$) clearly indicate that the data does not follow a CN distribution. The AIC criterion selected components up to the fifth order, while the BIC criterion selected only the second order component. From these results we may argue from a data analytical point of view that the empirical distribution shows inconsistencies with the CN distribution in at most the fifth order moment. Moreover, the BIC based test suggests that the second order component already indicates a severe deviation from circular normality.

The estimation of the concentration parameter yields $\hat{\kappa}^\star = 1.76$, which again results in a covariance matrix $\boldsymbol{\Sigma_V}$ which is the identity matrix up to one decimal. Therefore, applying the interpretation of the components in terms of trigonometric moments is again possible.

The properly scaled score statistics $S^{emp}_{\text{AIC}} = S^{emp}_{\text{BIC}} = 163.717$ ($p < 0.001$) also indicate severe deviation from circular normality for the moments up to the fifth order. The second order component of S^{emp}_{AIC} and S^{emp}_{BIC} has the largest contribution (82.314). This result indicates that there is a difference from circular normality with respect to the second order trigonometric moment. Since the second order trigonometric moment is related to the circular skewness and kurtosis, this is consistent with the skewed unimodal impression of the data.

4.6.4 Arrival Data

In the explorative analysis of the Arrival data in Section 2.2.6 two large clusters and three small clusters of arrivals are recognised. The Kuiper statistic $K_n = 1.174$ (p=0.122), the Watson statistic $U_n = 0.057$ (p=0.116) and the Entropy statistic $K_{mn} = 5.852$ (p=0.097) indicate that there is no significant deviation from circular normality at the 5% level. However, from the BarCox test $B = 6.364$ (p=0.042) and the score tests $S_{\text{AIC}} = 21.698$ (p=0.003) and $S_{\text{BIC}} = 6.341$ (p=0.045) we conclude at the 5% significance level that the von Mises distribution is not an appropriate distribution to describe this data. The choice for the order of the family of alternatives based on the AIC and the BIC criterion is five and two, respectively.

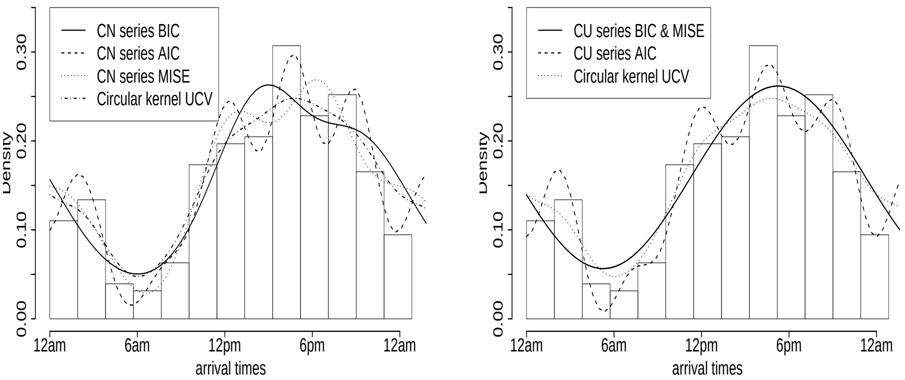
Since for this example the estimation of the concentration $\hat{\kappa}^\star$ is 0.67, the decomposition in (4.26) holds and the interpretation in terms of trigonometric moments is justified. The properly scaled score tests result in $S^{emp}_{\text{AIC}} = 35.433$ ($p <0.001$) and $S^{emp}_{\text{BIC}} = 6.073$ ($p = 0.053$). The latter score test indicates no significant result at the 5% level while the test based on the AIC criterion finds large evidence against circular normality. The AIC criterion selected the components up to the fifth order. From the simulation study in Section 4.7, it will become clear that the smooth test based on the AIC criterion has better power against higher order alternatives (order four and five) than the smooth test based on the BIC criterion.

Hence, these results suggest that the LOF might be related to the fifth order trigonometric moment. The fifth term in the family of alternatives describes a multimodal distribution with five modes. This fifth order departure is thus consistent with the five clusters of arrival times described above.

To visualise this deviation from normality, we use the family of smooth alternatives to find an appropriate density estimate. In particular, as described in Section 4.5, we choose the order of the family according to the AIC, BIC and MISE criteria and plug in the corresponding estimates. Since the concentration parameter has an estimated value smaller than 1, we may use the simplified formulae in (4.31) for the parameter $\boldsymbol{\theta}$. The BIC, AIC and MISE criteria select components up to order two, five and three, respectively. The corresponding CN series estimates are plotted in panel (a) of Figure 4.2 together with the kernel density estimate, for which the window width is chosen via UCV (see Section 3.7). In these plots we projected the data on the real line to make it easier to compare the density estimates. We thus have to keep in mind that begin and end points of the considered interval coincide.

The kernel density estimate is the most smooth estimate, followed in respective order by the CN series based on the BIC, the MISE and the AIC criterion. The latter estimate shows the locations of the five modes, which were recognised by the data-driven smooth test. These modes are also called *bumps* or clusters of observations. Note that the locations of the five clusters are exactly where we expected them (Section 2.2.6). The other density estimates have fewer modes, which are on slightly different locations. To get a better idea which estimate is the most appropriate we refer to the application of our new explorative tool in Chapter 6.

Panel (b) of Figure 4.2 shows the CU series density estimates based on the same order selection criteria. They are included in our study since they were shown to be useful for general circular densities by Fernández-Durán (2004), who used the AIC criterion. According to this criterion, the density estimate again has five modes and here the locations are once more exactly as we expected. On the other hand, the BIC and the MISE criterion both choose only the first term and give therefore a unimodal impression. Again, it is very difficult to say which density estimate is the most appropriate since we obviously have no information about the true distribution. In Chapter 6, we apply an explorative tool to get more insight.



(a) (b)

Figure 4.2: Density estimates for the arrival data. The kernel density estimate (both panels) with window width determined by means of unbiased cross-validation and the CN (panel (a)) and CU (panel (b)) series density estimates based on BIC, AIC and MISE criteria are plotted.

k	K_{AIC}	K_{BIC}
2	81433 (81.433%)	97959 (97.959%)
3	11361 (11.361%)	1880 (1.88%)
4	4657 (4.657%)	139 (0.139%)
5	2549 (2.549%)	22 (0.022%)

Table 4.1: Counts of the selected order under the null hypothesis, based on $100,000$ samples of size 50.

4.7 Simulation Study

In this section we present the results of a simulation study in which the diagnostic characteristics of the smooth tests for the composite null hypothesis of circular normality are investigated. We consider the smooth tests S_{BIC} and S_{AIC} and compare their powers with the Watson, the Kuiper, the Entropy and the BarCox tests for circular normality. Since data on the unit circle are directions that take values in the interval as small as $[0, 2\pi]$, we assume that in practical situations no more than five modes will appear in that interval. Therefore, we restricted the choice for the order to maximum five.

First we study the behavior of the order selection rule under the null hypothesis. We have performed 100,000 Monte Carlo simulation runs. In each run, a sample of 50 observations is randomly selected from a CN distribution. The number of times each order is selected by the AIC and BIC criterion is presented in Table 4.1. These results indicate that the BIC criterion selects almost always order two. For the AIC criterion we see that the fifth order is selected for 2549 samples, which is 2.549%. However, for most of the simulated samples (81%), the AIC criterium again selected order two. For sample sizes $n = 30$ and $n = 100$, similar results are obtained. Since for relatively small sample sizes (n=30, 50 and 100), the properly scaled smooth tests s_{AIC}^{emp} and S_{BIC}^{emp} introduce too much variability, we expect that its power results will be too low compared to the original smooth tests. Therefore, it is not useful to include these statistics in this simulation study.

For the empirical power study, the critical points at the 5% significance level for the smooth tests as well as for the Kuiper and the Entropy tests are obtained using $100,000$ Monte Carlo simulations. For the Watson test we use the tabulated critical points (see Lockhart & Stephens, 1985), while for the BarCox test we use asymptotic critical points. All powers are estimated based on $10,000$ simulated samples of size $n = 30$ and $n = 50$. As an alternative to the von Mises distribution, we consider mixtures of the von Mises distributions with densities given by

$$h(x) = \sum_{j=1}^{k} p_j f_0(x, \mu_j, \kappa_j),$$

where $0 \leq p_j \leq 1$ for $j = 1, \dots k$ and $p_1 + \dots p_k = 1$. This family obviously includes the von Mises distribution for $k = p_1 = 1$. Table 4.2 shows the alternatives of this family used in our study. They are described as mixtures of CN distributions. The alternatives are chosen in such a way that k equals the number of modes in the distribution. As a second alternative, we use a family of distributions proposed by Bogdan et al. (2002), with density

$$g_j(x) = 1 + \rho \cos(jx),$$

for $1 \leq j \leq 5$ and $-1 \leq \rho \leq 1$. The latter restriction is needed so as to prevent the density function from being negative. Note that this family of alternatives is an approximation to the CN distribution if $j = 1$. On the other hand if $\rho = 0$ the family reduces to the CU distribution. All powers are estimated based on 10,000 simulation runs. First we report on the simulations for the mixtures of the CN distributions. The simulated powers are presented in Table 4.3. For all bimodal alternatives, the BarCox test, which is essentially based on the first two components of our smooth test, has clearly the best power results. Watson performs well for symmetric bimodal alternatives (MVM$_1$ and

Alternative	k	$p_1,\ldots,p_k$	$\mu_1,\ldots,\mu_k$	$\kappa_1,\ldots,\kappa_k$
MVM_1	2	0.5,0.5	$0,\frac{\pi}{2}$	4,4
MVM_2	2	0.25,0.75	$0,\frac{\pi}{2}$	4,4
MVM_3	2	0.5,0.5	$0,\pi$	2,2
MVM_4	2	0.25,0.75	$0,\pi$	4,2
MVM_5	3	0.5,0.2,0.3	$0,\frac{2\pi}{3},\frac{4\pi}{3}$	6,6,6
MVM_6	3	$\frac{1}{3},\frac{1}{3},\frac{1}{3}$	$0,\frac{2\pi}{3},\frac{4\pi}{3}$	6,6,6
MVM_7	4	0.4,0.2,0.25,0.15	$0,\frac{\pi}{4},\pi,\frac{7\pi}{4}$	4,2,4,2
MVM_8	4	0.4,0.2,0.25,0.15	$0,\frac{\pi}{4},\pi,\frac{7\pi}{4}$	7,7,7,7
MVM_9	4	0.25,0.25,0.25,0.25	$0,\frac{\pi}{2},\pi,\frac{3\pi}{2}$	9,9,9,9
MVM_{10}	5	0.2,0.2,0.2,0.2,0.2	$0,\frac{2\pi}{5},\frac{4\pi}{5},\frac{6\pi}{5},\frac{8\pi}{5}$	12,12,12,12,12

Table 4.2: Mixture alternatives considered in the simulation study.

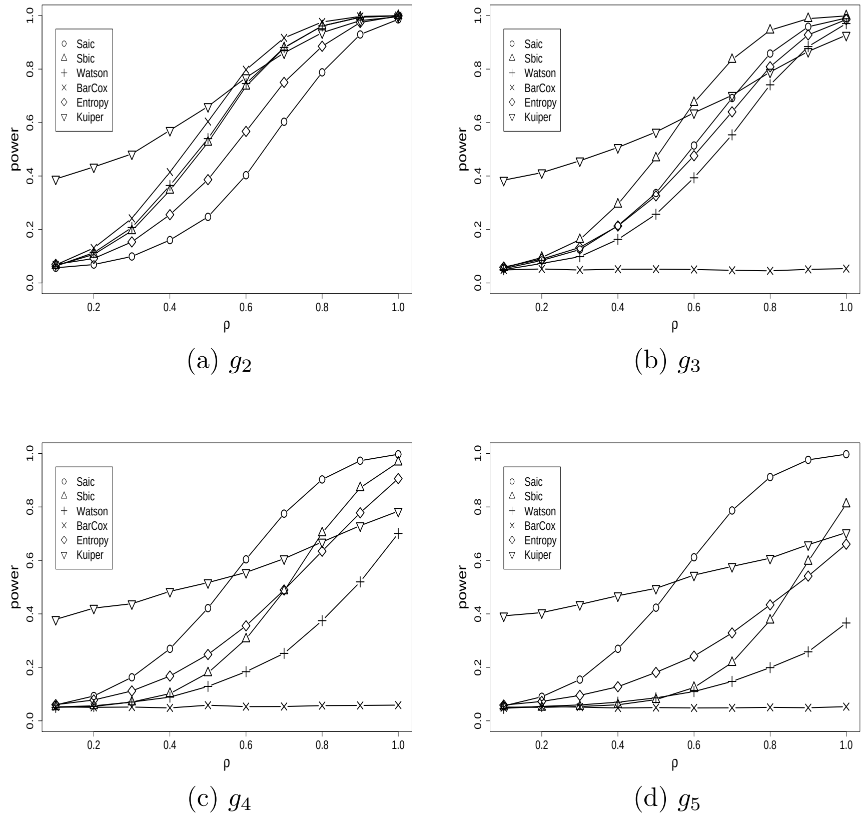
Alternative	S_{BIC}	S_{AIC}	K_n	U_n	K_{mn}	B
MVM_1	0.730	0.354	0.681	0.761	0.748	0.801
MVM_2	0.608	0.285	0.445	0.551	0.581	0.670
MVM_3	0.744	0.430	0.618	0.762	0.567	0.805
MVM_4	0.937	0.802	0.681	0.872	0.851	0.949
MVM_5	0.975	0.949	0.879	0.948	0.916	0.210
MVM_6	0.978	0.914	0.525	0.773	0.855	0.003
MVM_7	0.859	0.590	0.565	0.782	0.706	0.897
MVM_8	0.998	0.981	0.774	0.959	0.990	0.999
MVM_9	0.600	0.830	0.069	0.109	0.517	0.002
MVM_{10}	0.133	0.720	0.006	0.003	0.232	<0.001

Table 4.3: Simulated powers of our data-driven smooth tests (S_{BIC} and S_{AIC}), the Kuiper test (K_n), the Watson (U_n), the Entropy (K_{mn}) and the BarCox test (B) for the mixtures of the von Mises distribution, based on 10^6 samples.

MVM_3), whereas for asymmetric alternatives (MVM_2 and MVM_4) our smooth test based on the BIC criterion has higher power. The Entropy and the Kuiper tests do not have good powers for bimodal alternatives. Moreover, Kuiper has considerably lower powers for all mixtures we considered. For the three-modal alternatives ($k = 3$), the BarCox test shows a power breakdown whereas our data-driven smooth test has the best power results (both with the BIC and the AIC criterion), followed by the Watson and the Entropy tests. This is even more pronounced for symmetric modes. The alternatives MVM_7 and MVM_8 have four modes which do not receive equal weight. For these alternatives all tests show a similar performance as in the bimodal case. For the symmetric four-modal alternative MVM_9, only our smooth tests are doing well. Finally, the five-modal alternative (MVM_{10}) is only satisfactorily detected by the smooth test based on the AIC selection criterion.

Before we present the simulation results for the other family of alternatives, we first verify whether the selection criteria AIC and BIC adapt well towards the alternative. In particular, we expect the criteria to choose order k for a k-modal alternative. We consider alternatives MVM_3, MVM_6, MVM_9, MVM_{10} as examples of distributions with $k = 2, 3, 4$ and 5 modes. Table 4.4 shows the number of times each order is selected in $10,000$ samples of size 50. These results indicate that the AIC criterion most often chooses the right number of modes, while the BIC criterion only does well if the order of the alternative is two or three. For higher order alternatives the BIC criterion selects the right order less frequently. For these reasons the power of the BIC selection criterion is best for alternatives with two or three modes, while the AIC criterion is a better selection criterion for alternatives of higher order. Consequently, we conclude that the data-driven smooth test generally performs well against many different types of alternatives.

Figure 4.3 shows the estimated power curves for simulations from the alternatives g_j, $j = 2, \ldots, 5$, which have j evenly distributed modes. The power is plotted as a function of the parameter ρ. The data-driven smooth test based on the BIC selection criterion has good overall power results. It is a good compromise between the classical test for bimodal distributions and the data-driven smooth test based on the AIC criterion for higher order alternatives. As a general guideline, we recommend to apply both S_{AIC} and S_{BIC}. In case both yield a significant result, but select a different order, then it is likely that the BIC selected a lower order than the AIC. In that case, it is probably safest to use the BIC result for interpreting the deviation, unless there are specific sign that an important higher order deviation is present (cf. Section 4.6). In case only one of S_{AIC} or S_{BIC} is significant, we suggest relying on the statistic which produced the significant result, which is then likely S_{AIC} in case of higher order deviation and S_{BIC} in case of lower order deviations.



(a) g_2 (b) g_3

(c) g_4 (d) g_5

Figure 4.3: Estimated power functions for the alternatives g_j, $j = 2, \ldots, 5$ with parameter ρ, $0.1 \leq \rho \leq 1$, $n = 50$ and $\alpha = 0.05$.

n^o modes	2	3	4	5
alternative	MVM$_3$	MVM$_6$	MVM$_9$	MVM$_{10}$
k		BIC		
2	9861	219	4001	8672
3	118	9724	6	6
4	19	49	5933	0
5	2	8	60	1322
k		AIC		
2	8420	18	467	2534
3	857	8816	12	62
4	462	723	8581	11
5	261	443	940	7393

Table 4.4: Counts of the selected order under the alternatives, based on $10,000$ samples of size 50.

4.8 Discussion

In this chapter we have proposed a class of smooth GOF tests for circular distributions which are called the *complex smooth tests*. The construction of the order k complex smooth model is crucial for the development of the complex score statistic. We showed that for certain circular distributions the complex smooth model can be rewritten as a real smooth model for which the score statistic is equal to the former complex score statistic. Hence, in some sense this class of tests generalises the framework of Rayner and Best (1989) for smooth tests on the real line. For circular uniformity and circular normality we gave the explicit form of the smooth test statistics and their asymptotic distributions. The complex smooth test for circular uniformity reduces to the smooth test of Bogdan et al. (2002). The smooth test for circular normality for an order two complex smooth model reduces to the score test of Barndoff-Nielsen and Cox (1979). The AIC and BIC selection rules are applied to make an appropriate choice of the order of the complex smooth model. This results in two versions of the data-driven smooth test, S_{AIC} and S_{BIC}, for circular distributions against a general class of order k smooth alternatives. The complex data-driven smooth test for the CN distribution has been applied on real data examples and a simulation study showed that they have good power against many alternatives. Moreover, it is illustrated by means of an example that the interpretation of the data-driven smooth test can be visualised by means of the directly related density estimate of the true circular density.

CHAPTER 5

Localised Pearson χ^2 Test

In this chapter we present some new results on the class of GOF tests closely related to the sample space partition tests (SSP) originally proposed by Thas (2001) (see also Thas and Ottoy (2003b)). The tests are constructed by integrating out the Pearson χ^2 statistic over all possible partitions of the sample space in c cells. This is essentially a similar generalisation to the one Rothman (1972) proposed for circular data, which is described in Section 3.2.2. However, Rothman only considered partitions of two cells. Where Rothman's statistic is a generalisation of the Watson statistic for circular data, the new tests are generalisations of the AD family of GOF tests for linear data and are indexed by the so called sample space partition size c. The resulting tests are therefore called the *linear* SSPc tests. Clearly, Rothman's test can be generalised by considering partitions of general size c. This type of tests will be referred to as the *circular* SSPc tests.

In the next section we give a formal introduction to the class of SSPc tests. In particular, the construction of the test statistics and their asymptotic null distributions are given. Special attention is given to the simplest case $c = 3$ for which the corresponding test statistic can be seen as a V-statistic and for which a decomposition in terms of Legendre polynomials can be found. From these results the limiting distribution under contiguous alternatives is an immediate consequence. In Section 5.3 we intuitively describe the behaviour of the SSPc test statistics. More specifically, we explore the limiting expected value of the

SSPc tests under a particular family of local alternatives. In Section 5.4 a data-driven version of the SSPc test is constructed so that an appropriate subset size c is chosen from the data at hand. All tests are applied to real data examples in Section 5.5. In Section 5.6 we present a power study in which our new linear SSPc tests are compared to some of the classical GOF tests described in Chapter 3. Further extensions to composite null hypotheses and to testing for circular distributions are described in Sections 5.7 and 5.8, respectively. Finally, a brief discussion of the proposed class of tests is given in Section 5.9.

5.1 Introduction

We are concerned with testing the simple null hypothesis of GOF,

$$H_0 : F(x) = F_0(x) \text{ for all } x \in \mathcal{S}, \tag{5.1}$$

where $F(x)$ and $F_0(x)$ are the true and hypothesised distribution function of the continuous univariate random variable X, and $\mathcal{S}$ represents the common sample space on which $F(x)$ and $F_0(x)$ are defined. Since we will consider *omnibus* tests, the alternative hypothesis is

$$H_1 : F(x) \neq F_0(x) \text{ for some } x \in \mathcal{S}. \tag{5.2}$$

A very popular type of GOF tests is the class of EDF tests which are based on the EDF $\hat{F}_n(x)$ and are described in Section 3.4. Within that class, the AD test is often referred to as one of the most powerful omnibus tests (see e.g. D' Agostino and Stephens, 1986).

Recently, some modifications and generalisations of the AD test have been proposed.

Both Zhang (2002) and Einmahl and McKeague (2003) considered integral statistics of the form

$$T_n^w = \int_{\mathcal{S}} P_n(x) dw(x), \tag{5.3}$$

where $w(x)$ is some weight function and $P_n(x)$ is a *localised* statistic for testing GOF in a binomial distribution which is induced by discretising the sample space $\mathcal{S}$ at location x. The AD statistic is obtained with $w(x) = F_0(x)$ and

$$P_n(x) = \frac{n(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))}, \tag{5.4}$$

which is the score test statistic for testing $F(x) = F_0(x)$ in the induced binomial distribution. Note that (5.4) is exactly equal to the Pearson χ^2 statistic X_n^2 of

(3.18) with $k = 2$, $\boldsymbol{\pi} = (F_0(x), 1 - F_0(x))$ and $\boldsymbol{X}^T = (n\hat{F}_n(x), n - n\hat{F}_n(x))$. Indeed, we have

$$\begin{aligned} X_n^2 &= \frac{(n\hat{F}_n(x) - nF_0(x))^2}{nF_0(x)} + \frac{((n - n\hat{F}_n(x)) - (n - nF_0(x)))^2}{n - nF_0(x)} \\ &= \frac{n(\hat{F}_n(x) - F_0(x))^2}{F_0(x)} + \frac{n((\hat{F}_n(x)) - (F_0(x)))^2}{1 - F_0(x)} \\ &= P_n(x). \end{aligned}$$

Thas (2001) and Thas and Ottoy (2003b) proposed a similar extension of the AD statistic as in (5.3) by taking $w(x) = \hat{F}_n(x)$, which results in the average of Pearson statistics, i.e.

$$\begin{aligned} T_n^{\hat{F}} &= \int_{\mathcal{S}} P_n(x) d\hat{F}_n(x) \\ &= \frac{1}{n} \sum_{i=1}^{n} P_n(X_i) \end{aligned}$$

Zhang (2002) considered several choices for the weight function $w(x)$ and took $P_n(x)$ equal to the Cressie-Read family of divergence statistics (Cressie and Read, 1984), which makes T_n^w a family of statistics indexed by the same parameter as in the Cressie-Read family. The same extension was independently proposed by Thas and Ottoy (2003b). Einmahl and McKeague (2003) choose the log-likelihood ratio statistic instead of $P_n(x)$ and set $w(x) = F_0(x)$. The term *localised*, in this context, is due to these authors.

Similarly, Rothman's test (1972) for circular uniformity in (3.23) can be rewritten as a localised test. In particular, we write Rothman's statistic as

$$R_n^w = \frac{1}{2\pi} \int_0^{2\pi} \int_{x_0}^{2\pi + x_0} P_n(x; x_0) dw(x; x_0) dx_0, \tag{5.5}$$

where $dw(x; x_0) = F_0(x; x_0)(1 - F_0(x; x_0))dH(x; x_0)$ and

$$P_n(x; x_0) = \frac{n(\hat{F}_n(x; x_0) - F_0(x; x_0))^2}{F_0(x; x_0)(1 - F_0(x; x_0))}. \tag{5.6}$$

The dependence on the starting point x_0 is now explicitly present in $P_n(x; x_0)$ and $w(x; x_0)$ in which $F_0(x; x_0)$ and $\hat{F}_n(x; x_0)$ are defined as in (3.64). We again have that $P_n(x; x_0)$ is exactly equal to the Pearson χ^2 statistic X_n^2, localised at x, with the origin at x_0.

All the integrals of the form (5.3) have in common that they are localised at exactly one point x. In the next part of this chapter we extend these tests by localising at a finite number of distinct elements of $\mathcal{S}$. This results in a family of tests, indexed by the subset size. Although this type of generalisation applies to several choices of $P_n(x)$ and $w(x)$, we will focus here on the class of tests with $P_n(x)$ the Pearson statistic and $w(x) = F_0(x)$. A similar generalisation of the Rothman test (5.5) for circular distributions is explored in Section 5.8.

5.2 The linear SSPc test

5.2.1 Construction of the test statistics

Let $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ denote a sample of i.i.d. observations with linear distribution function $F(x)$. Let $D_c = \{x_1, \ldots, x_{c-1}\} \in \mathcal{S}$ with $c = \#D_c + 1 \geq 2$, but finite. Suppose further that $x_{(1)} \leq \ldots \leq x_{(c-1)}$ are the ordered elements of D_c. Every D_c induces a multinomial distribution with probabilities

$$\pi_1 = F(x_{(1)}); \pi_2 = F(x_{(2)}) - F(x_{(1)}); \ldots; \pi_c = 1 - F(x_{(c-1)}). \tag{5.7}$$

Thus c has the interpretation of the number of cells of the D_c-induced table of counts. In Thas (2001) and Thas and Ottoy (2003a,b), c is referred to as the sample space partition (SSP) size.

As mentioned before, we will only consider the Pearson χ^2 statistic for multinomial GOF, though other choices may be interesting as well. Let $P_{c,n}(D_c)$ now denote the Pearson χ^2 statistic, i.e.

$$\begin{aligned} P_{c,n}(D_c) &= P_{c,n}(x_1, \ldots, x_{c-1}) \\ &= n \sum_{i=1}^{c} \frac{(\hat{F}_n(x_{(i)}) - \hat{F}_n(x_{(i-1)}) - (F_0(x_{(i)}) - F_0(x_{(i-1)})))^2}{F_0(x_{(i)}) - F_0(x_{(i-1)})}, \end{aligned}$$

where $x_{(0)} \equiv 0$ and $x_{(c)} \equiv 1$. We propose the test statistic

$$T_{c,n} = \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} P_{c,n}(x_1, \ldots, x_{c-1}) dF_0(x_1) \ldots dF_0(x_{c-1}). \tag{5.8}$$

We omit the superscript w which is from now on always equal to F_0. The statistical tests proposed in Thas (2001) and Thas and Ottoy (2003b) are related to $T_{c,n}$ in (5.8) in the sense that F_0 is replaced by $\hat{F}_n$, which is essentially an average instead of an integral. For $c \geq 2$, $T_{c,n}$ represents a class of statistics indexed by c, which is also referred to as SSP size. The tests based on $T_{c,n}$ are therefore also called the SSPc tests. When $c = 2$ the statistic reduces to the AD statistic, and thus in this sense the SSPc tests are a generalisation of the AD test. Without loss of generality we will further suppose that $F_0(x) = x$, i.e. the hypothesised distribution is the uniform distribution over $\mathcal{S} = [0, 1]$. This situation can be obtained for any simple null hypothesis by applying the PIT.

For further study, it is interesting to rewrite $T_{c,n}$ in a more attractive way. Let $\mathbb{B}_n(x) = \sqrt{n}\left(\hat{F}_n(x) - x\right)$, which is the empirical process defined in Section 3.4.1. Then,

$$\begin{aligned} T_{c,n} &= \int_0^1 \cdots \int_0^1 \left[\frac{\mathbb{B}_n^2(x_{(1)})}{x_{(1)}} + \frac{\mathbb{B}_n^2(x_{(c-1)})}{1-x_{(c-1)}}\right] dx_1 \dots dx_{c-1} + \end{aligned}$$
$$\int_0^1 \cdots \int_0^1 \left[\frac{(\mathbb{B}_n(x_{(2)}) - \mathbb{B}_n(x_{(1)}))^2}{x_{(2)} - x_{(1)}} + \dots \frac{(\mathbb{B}_n(x_{(c-1)}) - \mathbb{B}_n(x_{(c-2)}))^2}{x_{(c-1)} - x_{(c-2)}}\right] dx_1 \dots dx_{c-1}. \tag{5.9}$$

Let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the first and the second term of (5.9). Note that $\mathcal{I}_1$ results from the first and the last term in (5.8), while $\mathcal{I}_2$ results from the other terms in (5.8). Further computation of the integrals leads to

$$\begin{aligned} \mathcal{I}_1 &= (c-1)\int_0^1 \int_{x_1}^1 \cdots \int_{x_1}^1 \frac{\mathbb{B}_n^2(x_1)}{x_1} dx_{c-1} \dots dx_2 dx_1 \\ &+ (c-1)\int_0^1 \int_0^{x_{c-1}} \cdots \int_0^{x_{c-1}} \frac{\mathbb{B}_n^2(x_{c-1})}{1-x_{c-1}} dx_1 dx_2 \dots dx_{c-1}, \end{aligned} \tag{5.10}$$

while each term of $\mathcal{I}_2$ can be rewritten as

$$\begin{aligned} &\tbinom{c-1}{2} \int_0^1 \int_0^1 \int_{x_i \vee x_{i-1}}^1 \cdots \int_{x_i \vee x_{i-1}}^1 \int_0^{x_i \wedge x_{i-1}} \cdots \int_0^{x_i \wedge x_{i-1}} \\ &\frac{(\mathbb{B}_n(x_i) - \mathbb{B}_n(x_{i-1}))^2}{x_i - x_{i-1}} dx_1 \dots dx_{i-2} dx_{i+1} \dots dx_{c-1} dx_{i-1} dx_i \end{aligned} \tag{5.11}$$

where $\wedge$ and $\vee$ denote the minimum and maximum operator. Hence, $\mathcal{I}_2$ becomes

$$\begin{aligned} \mathcal{I}_2 &= \binom{c-1}{2} \int_0^1 \int_0^1 \frac{(\mathbb{B}_n(x) - \mathbb{B}_n(y))^2}{x-y} \\ &\left((1-x \vee y)^{c-3} + (1-x \vee y)^{c-4}(x \wedge y) + \dots (x \wedge y)^{c-3}\right) dxdy \end{aligned} \tag{5.12}$$

Then, the more attractive formula for the test statistic becomes

$$\begin{aligned} T_{c,n} &= \mathcal{I}_1 + \mathcal{I}_2 \\ &= (c-1)\int_0^1 ((1-x)^{c-1} + x^{c-1}) \frac{\mathbb{B}_n^2(x)}{x(1-x)} dx \\ &+ \binom{c-1}{2} \int_0^1 \int_0^1 \frac{(1-(x \vee y))^{c-2} - (x \wedge y)^{c-2}}{(1-(x \vee y)) - (x \wedge y)} \frac{(\mathbb{B}_n(x) - \mathbb{B}_n(y))^2}{|x-y|} dxdy. \end{aligned} \tag{5.13}$$

Let $A_{c,n}$ and $U_{c,n}$ denote the first and the second integral of (5.13), respectively. The statistic $A_{c,n}$ is basically a weighted CvM statistic with weight function

$$a_c(x) = \frac{(1-x)^{c-1} + x^{c-1}}{x(1-x)}, \tag{5.14}$$

which suggests that $A_{c,n}$, like the AD statistic, is sensitive to deviations in the tails of the distribution. The other term, $U_{c,n}$, may be seen as a weighted Watson statistic with weight function

$$w_c(x,y) = \frac{(1-(x \vee y))^{c-2} - (x \wedge y)^{c-2}}{(1-(x \vee y)-(x \wedge y))|x-y|}. \tag{5.15}$$

The weight function is best interpreted by expanding it into (suppose $x < y$)

$$w_c(x,y) = \frac{1}{|x-y|}[x^{c-3} + x^{c-4}(1-y) + x^{c-5}(1-y)^2 + \ldots + (1-y)^{c-3}]. \tag{5.16}$$

This expansion shows that $U_{c,n}$ consists basically of $c-1$ terms which are all sensitive to deviations from F_0 in small intervals $[x,y]$ and it also implies that this term gets more important than $A_{c,n}$ as c increases. Hence, with increasing c, we suspect that the statistic $T_{c,n}$ has an increasing sensitivity for deviations from F_0 in small intervals. In Section 5.6 this will be further empirically investigated in a simulation study.

5.2.2 Computational formulae

The representation in (5.13) is also useful for obtaining computational formulae. In particular, computational forms are easily obtained by integration. We give here explicit solutions for $c = 2, 3$ and 4. Let $X_{(i)}$ denote the i-th order statistic ($i = 1, \ldots, n$).

$c = 2$:

$$T_{2,n} = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(X_{(i)}) + \ln(1-X_{(n+1-i)}))$$

$c = 3$:

$$T_{3,n} = 2A_n - 4W_n + K_n, \tag{5.17}$$

where A_n and W_n represent the AD and CvM statistics, respectively, and

$$\begin{aligned} K_n &= \int_0^1\int_0^1 \frac{(\mathbb{B}_n(x)-\mathbb{B}_n(y))^2}{|x-y|}dxdy \\ &= -\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[X_{(i\vee j)}\ln(X_{(i\vee j)}) \\ &\quad +(1-X_{(i\wedge j)})\ln(1-X_{(i\wedge j)}) \\ &\quad |X_{(j)}-X_{(i)}|\ln|X_{(j)}-X_{(i)}| + X_{(i)}(1-X_{(i)}) + X_{(j)}(1-X_{(j)}) - \frac{1}{6}\Big] \end{aligned}$$

$c = 4$:

$$T_{4,n} = 3A_n - 10.5W_n + 3K_n + 1.5n\left(\bar{X} - \frac{1}{2}\right)^2$$

5.2.3 Asymptotic theory

The next theorem gives the limiting null distribution of $T_{c,n}$. A proof is given in Appendix B.1.

Theorem 5.1 *Let* $\{\mathbb{B}(x), x \in [0,1]\}$ *denote a Brownian bridge. Suppose* $c \geq 2$ *is given, then, under the simple null hypothesis, as* $n \to \infty$*,*

$$\begin{aligned} T_{c,n} \quad \xrightarrow{d} \quad T_{c,\infty} &= (c-1)\int_0^1 a_c(x)\mathbb{B}^2(x)dx \\ &+ \binom{c-1}{2}\int_0^1\int_0^1 w_c(x,y)(\mathbb{B}(x)-\mathbb{B}(y))^2 dxdy. \end{aligned} \tag{5.18}$$

The proof of the following theorem is in Appendix B.2.

Theorem 5.2 $T_{c,n}$ *is consistent against any fixed alternative.*

5.2.4 Empirical levels

To assess the usefulness of the asymptotic null distribution $T_{c,\infty}$ in small samples, we have performed a simulation study to obtain the empirical levels of the SSP3 and SSP4 tests when the corresponding quantiles of the asymptotic null distribution are used. Since the asymptotic null distribution is expressed in terms of integrals of Brownian bridges, it is typically approximated by simulation. In particular, we considered 50,000 simulation runs and in each run the integrals are approximated using a simulated Brownian bridge on a 10,000 points grid. From the resulting simulated asymptotic null distribution, a nominal level α quantile is derived. For $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.10$ the quantiles are given in Table 5.1. For these levels, and for sample sizes $n = 5$, $n = 20$, $n = 50$ and $n = 100$, the empirical levels are then calculated by comparing the simulated exact null distribution of the test statistics (10,000 simulation runs) with the approximated asymptotic quantiles. The results are presented in Table 5.1. They suggest that the SSP3 test may satisfactorily be applied using the asymptotic null distribution, even for sample sizes as small as $n = 5$. For the SSP4 test, on the other hand, a bias is observed, in particular for small levels. For larger levels, however, one may find the bias sufficiently small.

5.2.5 Limiting distribution of SSP3 statistic under contiguous alternatives

In this section, we focus on the statistic $T_{c,n}$ for $c = 3$ and derive its limiting distribution under contiguous alternatives. This is done by rewriting the statistic as a V-statistic and determining its decomposition using Legendre polynomials.

Table 5.1: The asymptotic quantiles (at $n = \infty$, obtained from expression (5.18) using simulated Brownian bridges) and the empirical levels for $n = 5$, $n = 20$, $n = 50$ and $n = 100$ based on 10,000 simulation runs.

n	SSP3			SSP4		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.013	0.053	0.099	0.021	0.067	0.112
20	0.011	0.051	0.099	0.020	0.069	0.119
50	0.011	0.051	0.100	0.020	0.069	0.119
100	0.011	0.051	0.100	0.019	0.068	0.117
∞	6.190	4.286	3.500	8.108	6.140	5.301

Recall that the AD statistic has an orthogonal decomposition in terms of Legendre polynomials (see Section 3.5.2). Note that we continue to consider the simple null hypothesis of uniformity.

SSP3 rewritten as a V-statistic

Instead of localising the Pearson statistic at one point x, as is done for the AD statistic ($c = 2$), the statistic (5.8) for $c = 3$ is localised at two points x_1 and x_2. The Pearson statistic for testing GOF in a multinomial distribution with 3 cells then reduces to

$$\begin{aligned} P_n(x_1, x_2) &= n\left[\frac{(\hat{F}_n(x_{(1)}) - x_{(1)})^2}{x_{(1)}} + \frac{(\hat{F}_n(x_{(2)}) - \hat{F}_n(x_{(1)}) - (x_{(2)} - x_{(1)}))^2}{x_{(2)} - x_{(1)}}\right. \\ &\qquad \left. + \frac{(\hat{F}_n(x_{(2)}) - x_{(2)}))^2}{1 - x_{(2)}}\right]. \end{aligned}$$

In particular, for each localisation at some $(x_{(1)}, x_{(2)})$, the hypothesis of uniformity induces a null hypothesis in terms of probability parameters of a multinomial distribution

$$\pi_1 = x_{(1)}; \pi_2 = x_{(2)} - x_{(1)}; \pi_3 = 1 - x_{(2)}. \qquad (5.19)$$

As with the AD statistic, the SSP3 statistic for testing the null hypothesis of uniformity is the integral of $P_n(x_1, x_2)$ w.r.t. the hypothesised distribution, i.e.

$$\begin{aligned} T_{3,n} &= \int_0^1 \int_0^1 P_n(x_1, x_2) dx_1 dx_2 \\ &= 2\int_0^1 ((1-x)^2 + x^2)\frac{\mathbb{B}_n^2(x)}{x(1-x)}dx + \int_0^1 \int_0^1 \frac{(\mathbb{B}_n(x) - \mathbb{B}_n(y))^2}{|x-y|}dxdy \end{aligned}$$

The latter equality follows from the attractive expression of the statistic in (5.13).

Both the AD and the SSP3 test statistics may be seen as V-statistics. Let X_i $(i = 1, \dots, n)$ be a sample of i.i.d. uniform variates on $[0, 1]$, and let $\mathbb{B}_n(u, X_i) = \sqrt{n}(I(X_i \leq u) - u)$, so that $\mathbb{B}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{B}_n(x, X_i)$. The general form of a V-statistic of degree 2 is

$$V_n(\Psi) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \Psi(X_i, X_j), \tag{5.20}$$

where Ψ denotes a *symmetric* nonzero kernel for which $\int_0^1 \int_0^1 \Psi(x, y)^2 dxdy < \infty$. The kernel Ψ is *symmetric* if $\Psi(x, y) = \Psi(y, x)$ for all $x, y \in [0, 1]$. Furthermore, the kernel is assumed to be *degenerate* throughout this section, i.e.

$$\int_0^1 \Psi(x, y)dy = 0 \text{ for every } 0 \leq x \leq 1. \tag{5.21}$$

From Gregory (1977) we know that the CvM statistic is obtained by taking the kernel

$$\Psi_{\text{CvM}}(x, y) = \frac{1}{n}\int_0^1 \mathbb{B}_n(u, x)\mathbb{B}_n(u, y)du, \tag{5.22}$$

while the AD statistic is obtained by taking the kernel

$$\Psi_{\text{AD}}(x, y) = \frac{1}{n}\int_0^1 \frac{\mathbb{B}_n(u, x)\mathbb{B}_n(u, y)}{u(1 - u)}du. \tag{5.23}$$

Our SSP3 statistic now corresponds to the kernel

$$\begin{aligned}
\Phi_{\text{SSP3}}(x, y) &= \frac{1}{n}\int_0^1 \int_0^1 \left[\frac{\mathbb{B}(u \wedge v, x)\mathbb{B}_n(u \wedge v, y)}{u \wedge v(1 - u \wedge v)} \frac{u \vee v(1 - u \wedge v)}{|u - v|}\right. \\
&+ \frac{\mathbb{B}_n(u \vee v, x)\mathbb{B}_n(u \vee v, y)}{u \vee v(1 - u \vee v)} \frac{u \vee v(1 - u \wedge v)}{|u - v|} \\
&- \left. 2\mathbb{B}_n(u \wedge v, x)\mathbb{B}_n(u \vee v, y)\frac{1}{|u - v|}\right] dudv.
\end{aligned}$$

Indeed, taking the sum in (5.20) for the kernel Φ_{SSP3} we have

$$\begin{aligned}
& V_n(\Phi_{\text{SSP3}}) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \Phi_{\text{SSP3}}(X_i, X_j) \\
= & \int_0^1 \int_0^1 \frac{(\mathbb{B}_n(u \wedge v))^2 - 2(\mathbb{B}_n(u \wedge v))(\mathbb{B}_n(u \vee v)) + (\mathbb{B}_n(u \vee v))^2}{|u - v|}dudv \\
+ & \int_0^1 \int_0^1 \frac{(\mathbb{B}_n(u \wedge v))^2}{|u - v|}\left(\frac{u \vee v}{u \wedge v} - 1\right) + \frac{(\mathbb{B}_n(u \vee v))^2}{|u - v|}\left(\frac{1 - u \wedge v}{1 - u \vee v} - 1\right) dudv \\
= & \int_0^1 \int_0^1 \frac{(\mathbb{B}_n(u) - \mathbb{B}_n(v))^2}{|u - v|}dudv + \int_0^1 \int_0^1 \frac{\mathbb{B}_n(u \wedge v)}{u \wedge v} + \frac{\mathbb{B}_n(u \vee v)}{1 - u \vee v}dudv \\
= & T_{3,n}.
\end{aligned}$$

The kernels of the CvM and the AD statistics can be written in a more attractive form. In particular,

$$\begin{aligned}\Psi_{\mathrm{CvM}} &= -x \vee y + \frac{1}{2}(x^2 + y^2) + \frac{1}{3} && (5.24)\\ \Psi_{\mathrm{AD}} &= -1 - \ln(x \vee y - xy). && (5.25)\end{aligned}$$

Note that both kernels Ψ_{CvM} and Ψ_{AD} are symmetric and degenerate. On the other hand, we see that Φ_{SSP3} is a non-symmetric function in its arguments, i.e. $\Phi_{\mathrm{SSP3}}(x,y) \neq \Phi_{\mathrm{SSP3}}(y,x)$ if $x \neq y$. Further note that

$$V_n(\Phi_{\mathrm{SSP3}}) = V_n(\Psi_{\mathrm{SSP3}}), \tag{5.26}$$

where

$$\begin{aligned}\Psi_{\mathrm{SSP3}}(x,y) = \frac{1}{n}\int_0^1\int_0^1 &\left[\frac{\mathbb{B}_n(u\wedge v,x)\mathbb{B}_n(u\wedge v,y)}{u\wedge v(1-u\wedge v)}\frac{u\vee v(1-u\wedge v)}{|u-v|}\right.\\ &+\frac{\mathbb{B}_n(u\vee v,x)\mathbb{B}_n(u\vee v,y)}{u\vee v(1-u\vee v)}\frac{u\vee v(1-u\wedge v)}{|u-v|} \qquad (5.27)\\ &\left.-\left(\mathbb{B}_n(u\wedge v,x)\mathbb{B}_n(u\vee v,y)+\mathbb{B}_n(u\vee v,x)\mathbb{B}_n(u\wedge v,y)\right)\frac{1}{|u-v|}\right]dudv.\end{aligned}$$

The kernel Ψ_{SSP3} is the symmetrised version of the kernel Φ_{SSP3}. In the following, without loss of generality, we will use Ψ_{SSP3} instead of Φ_{SSP3}. Furthermore, this kernel has a more attractive formula which is given by the next lemma. The lemma is proved in Appendix B.3.

Lemma 5.1 *The kernel in (5.27) is equal to*

$$\begin{aligned}\Psi_{\mathrm{SSP3}}(x,y) &= 2\left(|x-y|\ln|x-y| - (x\vee y)\ln(x\vee y)\right.\\ &\quad -((1-x)\vee(1-y))\ln((1-x)\vee(1-y))\\ &\quad +x\vee y + (1-x)\vee(1-y)\\ &\quad \left.-\ln(x\vee y) - \ln((1-x)\vee(1-y))\right) - 5. && (5.28)\end{aligned}$$

In addition,

$$\begin{aligned}\Psi_{\mathrm{SSP3}}(x,y) &= \Psi_{\mathrm{SSP3}}(y,x),\\ \Psi_{\mathrm{SSP3}}(x,y) &= \Psi_{\mathrm{SSP3}}(1-x,1-y),\\ \int_0^1 \Psi_{\mathrm{SSP3}}(x,y)dy &= 0,\ \ 0\le x\le 1. && (5.29)\end{aligned}$$

We also have the following two results.

Corollary 5.1 *The kernel of the SSP3 statistic can be expressed as a linear combination of the kernels of the CvM and the AD statistics, i.e.*

$$\Psi_{\mathrm{SSP3}}(x,y) = 2\Psi_{\mathrm{AD}}(x,y) - 4\Psi_{\mathrm{CvM}}(x,y) + \Omega(x,y), \tag{5.30}$$

where $\Omega(x,y)$ is given in (B.9).

Proof. This results from combining Equations (5.25)-(5.24) and (B.5)-(B.6). ☐

Corollary 5.2 *The SSP3 statistic can be expressed as*

$$T_{3,n} = V_n(\Psi_{\text{SSP3}}) = 2V_n(\Psi_{\text{AD}}) - 4V_n(\Psi_{\text{CvM}}) + V_n(\Omega). \qquad (5.31)$$

In addition $V_n(\Omega) = K_n \geq 0$ with probability 1.

Proof. The expression of $T_{3,n}$ follows immediately from (5.30). ☐

Asymptotic properties of V-statistics

We first give a result from Gregory (1977) on the limiting distribution of V-statistics under contiguous alternatives. We will see that Gregory's theorem is not directly applicable to our V-statistic $T_{3,n}$ since the system of eigenvalues and eigenfunctions for the kernel Ψ_{SSP3} is not known. Nevertheless, we can find its limiting distribution using an arbitrary system of orthonormal functions for which the coefficients in the resulting expansion have to be computed explicitly.

Let $\{\lambda_k, k \geq 0\}$ and $\{\psi_k, k \geq 0\}$ denote the eigenvalues and the system of orthonormal eigenfunctions of Ψ. Thus $\{\psi_k, k \geq 0\}$ and $\{\lambda_k, k \geq 0\}$ are solutions of the integral equation

$$\lambda\psi(x) = \int_0^1 \Psi(x,y)\psi(y)dy \qquad (5.32)$$

and for every $l, m = 0, \ldots$, ψ_l and ψ_m satisfy the orthonormality relation

$$\int_0^1 \psi_l(x)\psi_m(x)dx = \delta_{lm}, \qquad (5.33)$$

where δ_{lm} is the Kronecker delta.

This means that Ψ can be written as

$$\Psi(x,y) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(y). \qquad (5.34)$$

Hence, the test statistic $V_n(\Psi)$ can be orthogonally decomposed, resulting in the expansion

$$\begin{aligned} V_n(\Psi) &= \sum_{k=1}^{\infty} \lambda_k \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_k(X_i) \right)^2 & (5.35) \\ &= \frac{1}{n} \sum_{k=1}^{\infty} \lambda_k \left(\sum_{i=1}^{n} \psi_k(X_i) \right)^2 . & (5.36) \end{aligned}$$

Theorem 5.3 *(Gregory, 1977)*
Consider the sequence of alternative distributions G_n to the uniform null distribution given by

$$\frac{dG_n}{x} = 1 + \frac{1}{\sqrt{n}} q_n, \tag{5.37}$$

where $\{q_n\}$ is some sequence of functions in $L^2[0,1]$ converging to some function $q \in L^2[0,1]$, in which $L^2[0,1]$ denotes the set of squared integrable functions on $[0,1]$. If in addition $\sum_{k=1}^{\infty} \lambda_k < \infty$ then for $n \to \infty$

$$V_n(\Psi) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (Z_k + \upsilon_k)^2, \tag{5.38}$$

where $\upsilon_k = \int_0^1 q(x)\psi_k(x)dx$ and $Z_1, Z_2, \ldots$ are i.i.d. standard normal variables. Hence, under the null hypothesis we have, as $n \to \infty$

$$V_n(\Psi) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k Z_k^2. \tag{5.39}$$

Unfortunately, the eigenvalues λ_k and the eigenfunctions ψ_k of the kernel Ψ_{SSP3} are very hard to determine. Nevertheless, applying the following result will provide an alternative expression of the limiting distribution of $V_n(\Psi_{\text{SSP3}})$. The limiting null distribution is given in Theorem 5.4 and the limiting distribution under the contiguous alternatives described in Theorem 5.3 is given in Theorem 5.5.

Any kernel $\Psi \in L^2[0,1]^2$ has an L^2-expansion in terms of Legendre polynomials, i.e.

$$\Psi(x,y) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} (2k+1)(2l+1)\Psi_{kl} P_k(x) P_l(y), \tag{5.40}$$

with coefficients

$$\Psi_{kl} = \int_0^1 \int_0^1 \Psi(x,y) P_k(x) P_l(y) dx dy, \tag{5.41}$$

where $\sqrt{2k+1}P_k(x)$ are the orthonormal Legendre polynomials on $[0,1]$. The Legendre polynomials are defined in Appendix A, which also contains properties that will be used in the proof of Lemma 5.2.

Let

$$V_\infty = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \sqrt{(2k+1)(2l+1)}\Psi_{kl} Z_k Z_l, \tag{5.42}$$

where $Z_1, Z_2, \ldots$ is a sequence of i.i.d. standard normal random variables.

Theorem 5.4 *For a V-statistic based on a kernel as in (5.40), we have under the null hypothesis, as $n \to \infty$,*

$$V_n(\Psi) \xrightarrow{d} V_\infty. \tag{5.43}$$

Moreover, as $n \to \infty$

$$\sup_x |P(V_n(\Psi) \leq x) - P(V_\infty \leq x)| \to 0. \tag{5.44}$$

Proof. From (5.20) and (5.40) we have

$$V_n = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}(2k+1)(2l+1)\Psi_{kl}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} P_k(X_i)\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} P_l(X_i)\right). \tag{5.45}$$

According to the CLT, the real-valued random variables of the form $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} P_k(X_i)$ in (5.45) weakly converge to normal variates, i.e. as $n \to \infty$,

$$\frac{\sqrt{2k+1}}{\sqrt{n}}\sum_{i=1}^{n} P_k(X_i) \xrightarrow{d} Z_k, \tag{5.46}$$

for all $k \geq 1$, where the standard normal random variables $Z_1, Z_2 \ldots,$ are independent by virtue of the orthogonality conditions of the Legendre polynomials. From (5.45) and (5.46) and with the help of standard arguments we get (5.44).
□

Compared to the expansion in (5.35), the expansion in (5.45) is not orthonormal. This is the price we have to pay for not pursuing the solutions of the integral equation in (5.32). However, we will see in the next section that by applying the expansion in (5.45) to our SSP3 test statistic, all the coefficients except Ψ_{kk} and $\Psi_{k,k+2}$ $k = 1, 2, \ldots$ vanish, resulting in an almost orthogonal representation.

The next theorem gives the limiting distribution under contiguous alternatives.

Theorem 5.5 *Consider the sequence of alternative distributions described in Theorem 5.3. If in addition* $\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\Psi_{kl} < \infty$ *then for* $n \to \infty$

$$V_n(\Psi) \xrightarrow{d} \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sqrt{(2k+1)(2l+1)}\Psi_{kl}(Z_k + \upsilon_k)(Z_l + \upsilon_l), \tag{5.47}$$

where $\upsilon_k = \int_0^1 q(x)\sqrt{2k+1}P_k(x)dx$ *and* $Z_1, Z_2, \ldots$ *are i.i.d. standard normal variables. Under the null hypothesis, this expression reduces to the first part of Theorem 5.4.*

Proof. The proof is similar to the proof of Theorem 5.3 (see Gregory, 1977). □

The asymptotic distribution of the SSP3 statistic

We use Theorem 5.4 and Theorem 5.5 to find the limiting distribution of $T_{3,n} = V_n(\Psi_{\mathrm{SSP3}})$. In particular, we compute the coefficients Ψ_{kl} in (5.41) for $\Psi =$

Ψ_{SSP3} in (5.28). The results are summarised in the following lemma. The proof is given in Appendix B.4.

Lemma 5.2 *Consider the series*

$$\sum_{k=1}^{m} a_k P_k(x) P_k(y) + b_k (P_k(x) P_{k+2}(y) + P_{k+2}(x) P_k(y)), \tag{5.48}$$

where

$$\begin{aligned}
a_1 &= 2.1 \\
a_k &= \frac{(4k+7)(2k+1)}{k(k+1)(2k+3)} + \frac{2(2k+1)}{(2k-1)(2k+3)} \sigma_k, \quad k \geq 2 \\
b_k &= -\frac{1}{2k+3} \sigma_{k+2}, \\
\sigma_k &= \sum_{p=2}^{k-1} \frac{1}{p}, \ \textit{and } \sigma_1 = \sigma_2 = 0.
\end{aligned}$$

The series (5.48) converges in $L^2([0,1]^2)$ to $\Psi_{SSP3}(x,y)$. Moreover, this series is majorised by a function of the type

$$(x(1-x)y(1-y))^{-1/4} \in L^2([0,1]^2),$$

for all $0 < x, y < 1$.

Theorem 5.6 *The limiting distribution of the SSP3 statistic under contiguous alternatives (5.37) is*

$$V_\infty = \sum_{k=1}^{\infty} \left(\alpha_k (Z_k + \upsilon_k)^2 + \beta_k (Z_k + \upsilon_k)(Z_{k+2} + \upsilon_{k+2}) \right), \tag{5.49}$$

where

$$\begin{aligned}
\alpha_1 &= 0.7 \\
\alpha_k &= \frac{4k+7}{k(k+1)(2k+3)} + \frac{2\sigma_k}{(2k-1)(2k+3)}, \quad k \geq 2 \\
\beta_k &= -\frac{2\sigma_{k+2}}{(2k+3)\sqrt{(2k+1)(2k+5)}}
\end{aligned}$$

Under the null hypothesis, the limiting distribution is obtained by setting $\upsilon_k = 0, \ k = 1, 2, \ldots$ in (5.49).

Proof. The proof is given by applying Theorem 5.5. □

From now on we denote the limiting null distribution of the SSP3 statistic by $V_\infty^{H_0}$.

Table 5.2: The approximate asymptotic quantiles for the SSP3 statistic (at $n = \infty$, obtained from the expansion in terms of Legendre polynomials (5.49)) and the empirical levels for $n = 5$, $n = 20$, $n = 50$ and $n = 100$. In the left and right part of the table, the expansion (5.49)) is calculated up to the 100th (K=100) and the 10,000th term (K=10,000).

n	$K = 100$			$K = 10,000$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.014	0.060	0.114	0.012	0.052	0.104
20	0.013	0.056	0.114	0.010	0.049	0.100
50	0.011	0.057	0.111	0.009	0.050	0.098
100	0.012	0.056	0.115	0.010	0.050	0.103
∞	6.039	4.171	3.345	6.301	4.322	3.482

Empirical levels and convergence rate

In Section 5.2.4 we concluded that the asymptotic null distribution $T_{3,\infty}$ of the SSP3 statistic is useful in small samples. The alternative expression for that limiting distribution, $V_\infty^{H_0}$, may also by useful to determine asymptotic critical values. Since this expression is a sum with infinitely many terms, it is impossible to simulate from that distribution. However, note that the weights α_k and β_k in the expansion are decreasing in absolute value with increasing k. This means that we can get an approximation of $V_\infty^{H_0}$ by truncating the expansion at a certain order K. Here we will investigate which order is large enough to obtain good approximations of $V_\infty^{H_0}$. We calculate the quantiles of the distribution simulated for the expansion with terms up to order $K = 100$ and $K = 10,000$ and compare them to the quantiles obtained from the simulated $T_{3,\infty}$ based on Brownian bridges (see Table 5.1). In particular, we considered 10,000 simulation runs and the quantiles for $\alpha = 0.01, 0.05$ and 0.1 are given in Table 5.2. The quantiles for $K = 100$ are slightly smaller than the quantiles obtained from the simulated $T_{3,\infty}$. Hence, we expect that order $K = 100$ is still too small to obtain a good approximation of the limiting distribution. On the other hand, the quantiles for $K = 10,000$ are slightly larger for $\alpha = 0.01$ and 0.05, but about equal for $\alpha = 0.1$. Meanwhile, we assess whether using those newly generated quantiles for the asymptotic null distribution is useful in small samples. Therefore, the empirical levels are calculated for $n = 5, 20, 50$ and 100 (based on 10,000 simulation runs). It is clear that the quantiles for the simulated $V_\infty^{H_0}$ with $K = 100$ are not appropriate since the empirical levels are biased. On the other hand, the empirical values for $K = 10,000$ are comparable to the empirical levels obtained from the simulated $T_{3,\infty}$ in Table 5.1 and can therefore reliably be used in small samples.

5.3 Limiting behaviour of the SSPc test under contiguous alternatives

In this section we construct a particular family of contiguous alternatives and explore the behaviour of the SSP3 and SSP4 statistics under those alternatives. Usually, the behaviour of a test statistic is expressed in terms of the power characteristics. Instead of focusing directly on the power, we here examine the expected value of the statistic, which is related to the power as explained by Mudholkar et al. (1991). In particular, the authors reason that the expected value of the p-value of a test statistic is 0.5 under the null hypothesis and will decrease to zero as the distance of the true distribution from the null distributions increases. Since the expected value of the test statistic increases as the p-value decreases, we may use the former as an indication of the power.

Suppose the observations $x_1, \ldots, x_n$ are measured on the unit interval $[0, 1]$ and are generated by the family of contiguous alternatives to uniformity

$$f_n(x) = 1 + \frac{1}{\sqrt{n}}\delta(x), \tag{5.50}$$

where $\delta(x)$ is an arbitrary drift function that satisfies $\int_0^1 \delta(x)dx = 0$, and for which $f_n(x)$ is a proper density function for all $n \geq 1$. The corresponding distribution function of this family is given by

$$F_n(x) = x + \frac{1}{\sqrt{n}}\Delta(x), \tag{5.51}$$

where $\Delta(x) = \int_0^x \delta(u)du$. Under this family of contiguous alternatives, we know from Janssen (1995) that the empirical process

$$\mathbb{B}_n(x) = \sqrt{n}(\hat{F}_n(x) - x),$$

which can be written as

$$\sqrt{n}(\hat{F}_n(x) - F_n(x)) + \sqrt{n}(F_n(x) - x),$$

converges weakly to

$$\mathbb{B}(x) + \int_0^x \delta(u)du = \mathbb{B}(x) + \Delta(x),$$

where $\mathbb{B}(x)$ is a Brownian Bridge. Furthermore, we obtain that the process $\mathbb{B}_n(x) - \mathbb{B}_n(y)$ converges weakly under the family of alternatives (5.50) to $\mathbb{B}(x) - \mathbb{B}(y) + \int_y^x \delta(u)du = \mathbb{B}(x) - \mathbb{B}(y) + \Delta(y) - \Delta(x)$. This implies that the limiting distribution for the SSPc test statistics under this family of contiguous alternatives becomes

$$\begin{aligned} T_{c,n} \quad &\xrightarrow{d} \quad (c-1)\int_0^1 a_c(x)(\mathbb{B}(x)+\Delta(x))^2 dx \\ &+\binom{c-1}{2}\int_0^1\int_0^1 w_c(x,y)(\mathbb{B}(x)-\mathbb{B}(y)+\Delta(x)-\Delta(y))^2 dxdy. \end{aligned}$$

Hence, the limiting expected value of $T_{c,n}$ is given by

$$\begin{aligned} \mathrm{E}\left[T_{c,n}\right] \quad &= \quad (c-1)\mathrm{E}\left[A_{c,n}\right]+\binom{c-1}{2}\mathrm{E}\left[U_{c,n}\right] & (5.52) \\ &\longrightarrow \quad (c-1)\int_0^1 a_c(x)\Delta^2(x)dx \\ &+\binom{c-1}{2}\int_0^1\int_0^1 w_c(x,y)(\Delta(x)-\Delta(y))^2 dxdy. & (5.53) \end{aligned}$$

We similarly obtain the limiting expected values under this family of contiguous alternatives for the CvM, AD and Watson statistics, as

$$\mathrm{E}\left[W_n\right] \longrightarrow \int_0^1 \Delta^2(x)dx \tag{5.54}$$

$$\mathrm{E}\left[A_n\right] \longrightarrow \int_0^1 \frac{\Delta^2(x)}{x(1-x)}dx \tag{5.55}$$

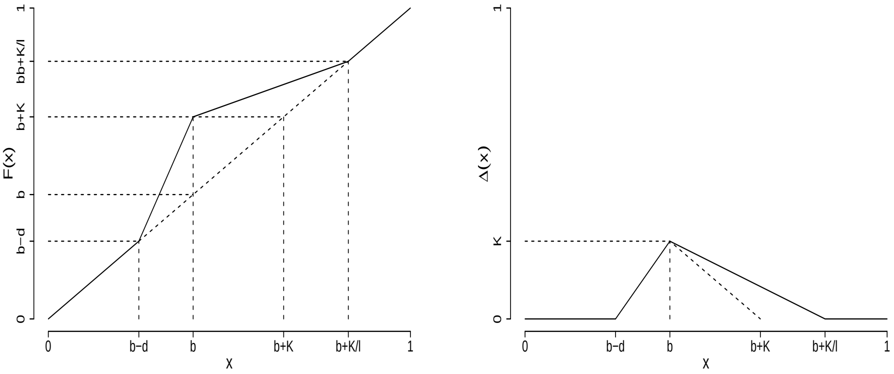
$$\mathrm{E}\left[U_n\right] \longrightarrow \int_0^1\int_0^1 (\Delta(x)-\Delta(y))^2 dxdy. \tag{5.56}$$

We are now interested to see for which functions $\Delta(x)$ the limiting value in (5.53) is large or in other words for which alternatives the SSPc test is most powerful. Similarly we will examine the behaviour of the classical statistics through their limiting values (5.54)-(5.56). Note that since we here only consider limiting expected values of the statistics and not the complete limiting distributions, this does not serve as a basis for comparing the power of the different statistics. For such a comparison we refer to the simulation study in Section 5.6.

As an example we suppose the data come from the family of contiguous alternatives (5.50) using a piecewise continuous function for $\Delta(x)$, which is characterised by four parameters as

$$\Delta(x)=\begin{cases} 0 & 0\le x\le b-d \\ \frac{K}{d}(x-(b-d)) & b-d\le x\le b \\ K-l(x-b) & b\le x\le b+K/l \\ 0 & b+K/l\le x\le 1. \end{cases} \tag{5.57}$$

To observe how the distribution $F_n(x)$ of (5.51) with Δ given by (5.57) depends on the four parameters, we show a graph of $F_n(x)$ in panel (a) of Figure 5.1,



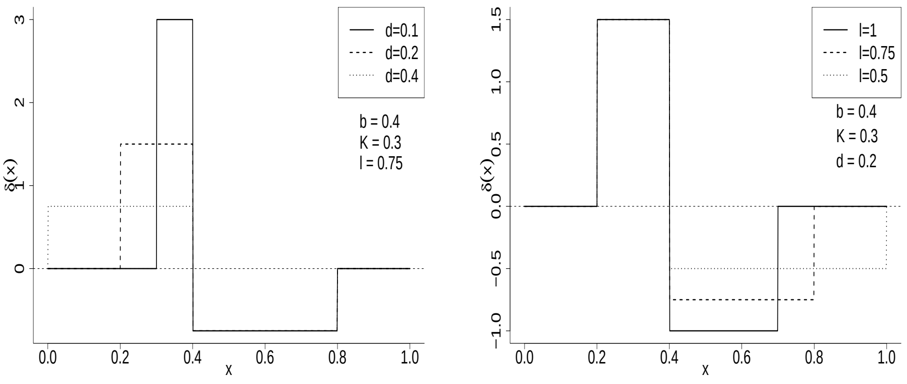
(a) (b)

Figure 5.1: The family of contiguous alternatives $F_n(x)$ with $n = 1$ (panel (a)) and the corresponding functions $\Delta(x)$ (panel (b)).

for the case of $n = 1$. When a sample is drawn from $F_n(x)$, the deviation from uniformity gets smaller for increasing sample size.

The deviation from uniformity starts at $b - d$ and increases linearly to reach its maximum in b. It then linearly decreases again until it vanishes at $b + K/l$. The parameter d is called the *run up* since large values of d indicate a slow increase towards the farthest deviated point. The parameter b is the *location* of the largest deviation, while l is the *decrease rate*. The function $\Delta(x)$ is shown in panel (b) of Figure 5.1, from which we see that K is the maximum of $\Delta(x)$. Hence, K serves as the maximal deviation from uniformity. In fact, for this family of alternatives, the KS test $D_n = \sup_x |\mathbb{B}_n(x)|$ converges weakly to $\sup_x |\mathbb{B}(x) + \Delta(x)|$ (Janssen, 1995) so that the limiting value of the KS test is large when $\sup_x |\Delta(x)| = K$ is large. Also, the Kuiper test statistic for such alternatives has large limiting expected value when $\sup_x \sup_y |\Delta(x) - \Delta(y)| = K$ is large. In the following, we keep the value of K fixed, so that the KS and the Kuiper test have constant power throughout. While the power of the KS and Kuiper test is fixed, we are interested in what the power properties are for the SSP3 and SSP4 tests for varying b, d and l. We also look at the properties of the classical CvM, AD and Watson.

In Figure 5.3 we show the functions Δ which will be considered. In particular, we set $K = 0.3$, and take the location b to be either 0.2, 0.4 or 0.6, corresponding to panels (a), (b) and (c), respectively. The run up d and the decrease rate l are varied. There is a limited range of possible values for d and l in order for $F_n(x)$ in (5.53) to be a valid distribution function on $[0, 1]$. The plots in Figure 5.3



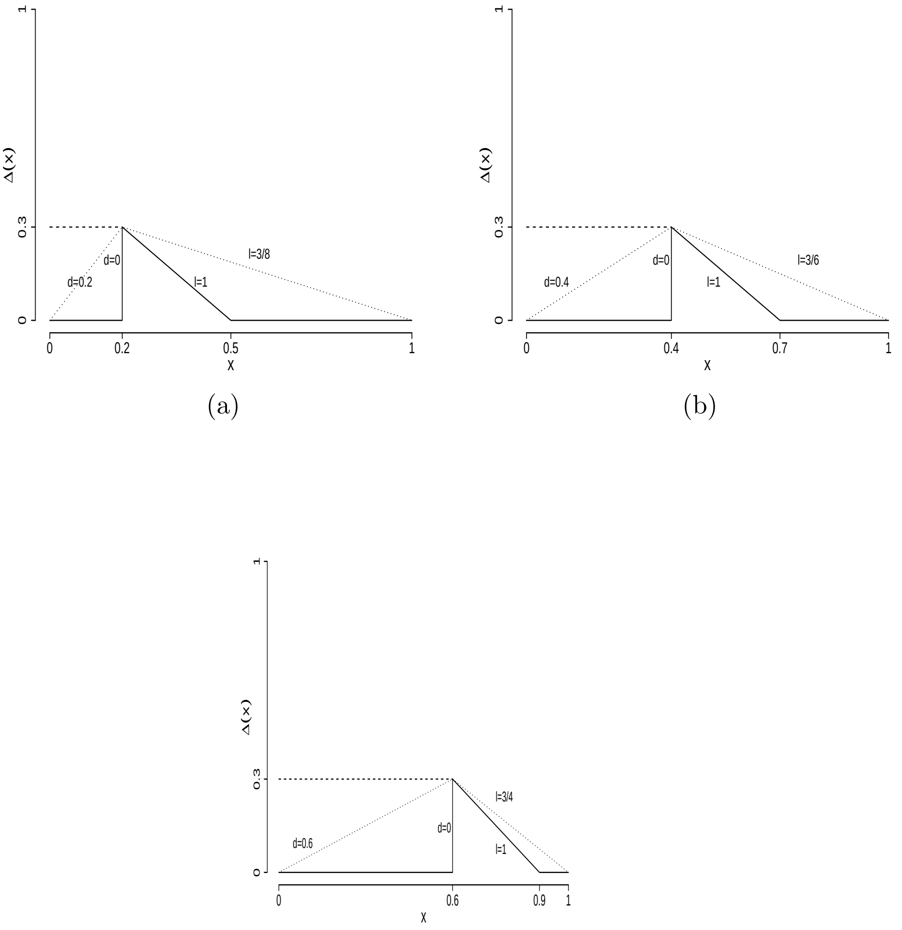
(a) (b)

Figure 5.2: Some examples of the functions $\delta(x)$ where $K = 0.3$ and $b = 0.4$. In (a) l is fixed and d takes different values. In (b) d is fixed and l takes different values.

show the two functions (dotted and full) corresponding to the end points of the range. In particular, when in panel (a) the location is equal to 0.2, the value of d ranges between 0 and 0.2. A larger run up indicates an earlier start of the LOF and the largest run up ($d = 0.2$) corresponds to a start at the begin point of the interval. On the other hand, a zero run up corresponds to a start of the LOF in the location $b = 0.2$. The decrease rate has always its largest possible value at 1, which corresponds to the fastest possible decrease (see panel (a) of Figure 5.1). The slowest decrease rate equals $l = \frac{K}{1-b}$, corresponding to a LOF with end point in 1.

To get a better idea which of the previously described alternatives can be interpreted as local alternatives, we show the function $\delta(x)$ for several values of d and l in Figure 5.2. Since Δ is piecewise linear, δ is piecewise constant. In these plots we set $K = 0.3$ and $b = 0.4$. The left panel then shows $\delta(x)$ with l fixed at 0.75 for three different values of d (0.1, 0.2, and 0.4). Clearly, smaller values of d correspond to more localised deviations in density. The right panel in Figure 5.2 shows how $\delta(x)$ varies with l, as we keep d fixed at 0.2. We see that the parameter l has no large influence on the "local impression"of the deviation. Therefore, for the family contiguous alternatives $F_n(x)$ of (5.51) with Δ given by (5.57), we will refer to a local or global LOF when d is small or large, respectively.

The expected values under these specific alternatives are computed by piecewise integration. Consider first the alternatives with LOF location at 0.2. The



(a) (b)

(c)

Figure 5.3: The functions $\Delta(x)$ in which, $K = 0.3$ and $b = 0.2, 0.4$ and 0.6 are plotted in the panels (a), (b) and (c), respectively. The values of the parameters d and l are specified in the panels.

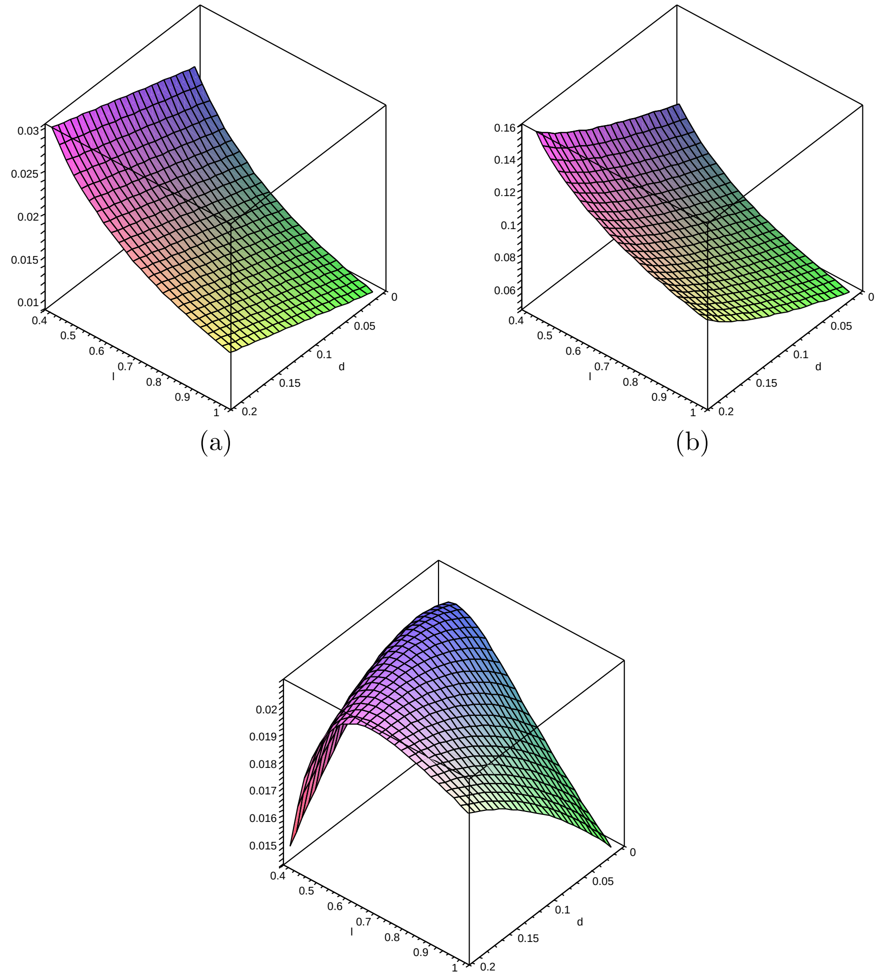
three panels of Figure 5.4 give the surface plots that represent the expected values of the CvM, the AD and the Watson statistics, respectively, versus the parameters d and l. The values for the CvM and AD statistics are high for small values of l and large values of d, which generally corresponds to a global LOF over the whole range of the interval. If the run up d decreases and the decrease rate l increases, which corresponds to more localised LOFs, lower values of both classical tests are expected. For the Watson test, in panel (c), a different situation is noticed. In particular, high values of U_n are expected if l and d are either both small or both large. On the other hand, we expect smaller values of U_n for the other two extreme combinations of the parameters d and l. Hence the Watson test has high expected values for some local as well as for some global alternatives.

The surfaces that represent the expected values of $T_{3,n}$ and $T_{4,n}$ versus d and l are given in panels (e) and (f) of Figure 5.5, respectively. We also plot the expected values of the terms $A_{c,n}$ and $U_{c,n}$ in (5.52) versus d and l. For $c = 3$, they are in panels (a) and (c), respectively and for $c = 4$, they are in panels (b) and (d), respectively. The surfaces for the terms $A_{3,n}$ and $A_{4,n}$ are similar to the surface for the AD statistic. However, the former surfaces give a slightly more flat impression, indicating that there is no large difference in expected values over the range of the parameters. For global LOF, both $U_{3,n}$ and $U_{4,n}$ have low expected values. From these results, we may conclude that for $c = 3$ and 4, the terms $U_{c,n}$ are more sensitive for local LOF, while the terms $A_{c,n}$ are most sensitive to global LOF.

The contribution of the individual terms to the values of the statistic $T_{c,n}$ differs between $c = 3$ and $c = 4$. In particular, for $c = 3$ the term $A_{c,n}$ dominates, while for $c = 4$ the term $U_{c,n}$ dominates. However, for both statistics we may conclude that combining $A_{c,n}$ and $U_{c,n}$ results in increased sensitivity for local alternatives as compared to the AD statistic.

The analogous plots for the locations $b = 0.4$ and $b = 0.6$ are presented in Figures 5.6-5.9. In particular, the limiting expected values for the classical tests under the specific alternatives in which $b = 0.4$ and $b = 0.6$ are shown in Figure 5.6 and 5.8, respectively, while those for the localised tests are in Figures 5.7 and 5.9, respectively.

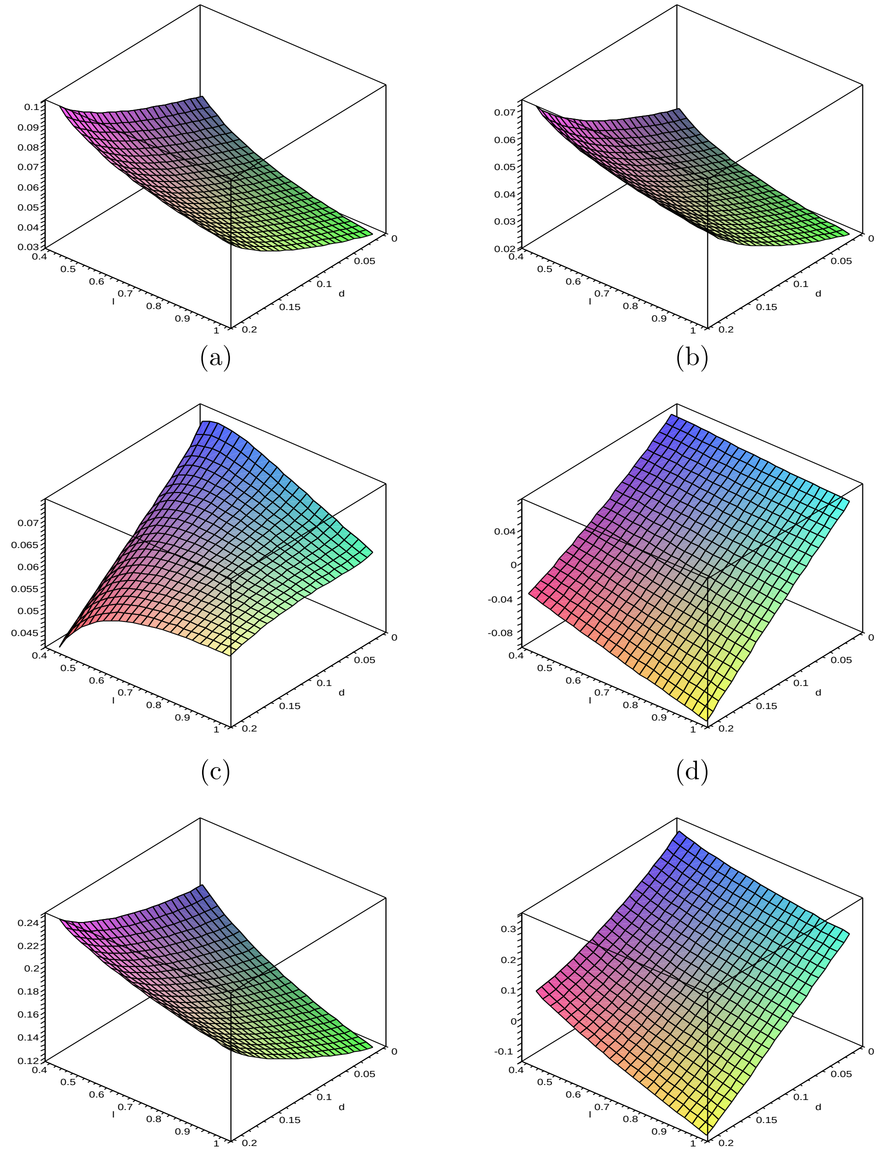
While the surface for the Watson statistic remains fairly similar when the location of the LOF is changed (see panels (c) of Figures 5.4, 5.6 and 5.8), the surfaces of the CvM and the AD exhibit larger differences between values for local and global LOF as the location increases (see panels (a) and (b) of Figures 5.4, 5.6 and 5.8). In particular, if the location is at the end of the interval, only large values of d yield large expected values of the statistics. The CvM and AD tests are indeed less powerful for deviations that occur near the end of the interval, as can immediately be seen from the construction of their statistics.



(a) (b)

(c)

Figure 5.4: The values for the W_n (a), A_n (b), and U_n (c) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.2$ as a function of the parameters d and l.



(a) (b)

(c) (d)

(e) (f)

Figure 5.5: The values for $A_{3,n}$ (a), $U_{3,n}$ (c), $T_{3,n}$ (e), $A_{4,n}$ (b), $U_{4,n}$ (d) and $T_{4,n}$ (f) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.2$ as a function of the parameters d and l.

For the $A_{c,n}$ terms of the localised statistics, similar steeper surfaces are observed (see panels (c) and (d) of Figures 5.5, 5.7 and 5.9). On the other hand, the surfaces of the $U_{c,n}$ terms give more or less the same impression for the three different locations (see panel (c) and (d) of Figures 5.5, 5.7 and 5.9). Hence these terms do not depend on the location of the LOF. The resulting localised statistics show similar behaviour as well, indicating higher sensitivity towards local alternatives. These results will be confirmed in the simulation study in Section 5.6.

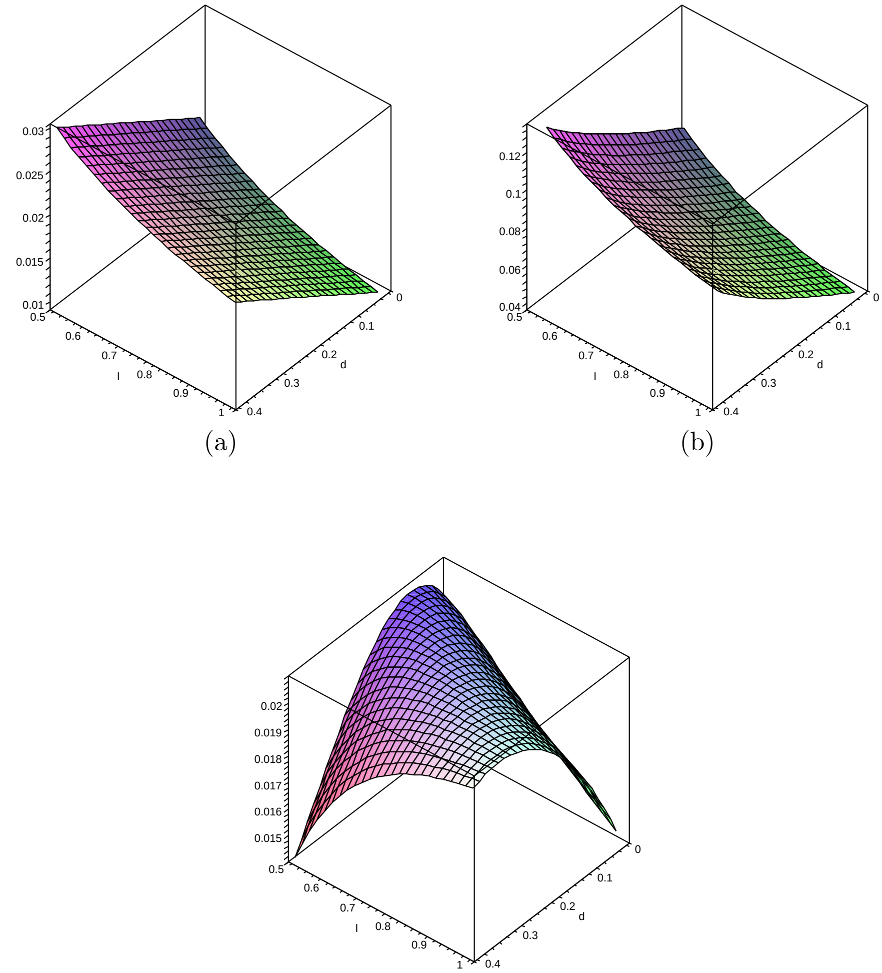
5.4 Data-driven Test

5.4.1 Construction of the test statistic

In Section 5.2.3 it has been shown that for every finite $c \geq 2$ the SSPc test is omnibus consistent. On the other hand, for finite sample sizes, it is expected that the power depends on the choice of c. In particular, for some alternatives a small c may result in a high power, while for other alternatives an SSPc test with a large c may perform better. Since the true distribution function F is unknown to the user, the choice of c is arbitrary, at the risk of selecting a low power test. For this reason, we propose a method to select an appropriate value for c based on the observations in the sample. By means of a particular selection rule, a SSP size C_n is obtained. The data-driven test statistic is then defined as $T_{C_n,n}$.

Before we continue with the specification of the selection rule, we remark that our data-driven test differs conceptually from many other data-driven GOF tests in the literature. Ledwina (1994) introduced the data-driven methodology for selecting the number of components in Neyman's smooth GOF statistics, which is described in Section 3.3.4. Until then, smooth tests were always based on a finitely truncated series of components, resulting in two drawbacks. First, if too many components are considered, there is the risk of power loss under low order alternatives (dilution effect). Second, a finite number of components does not result in an omnibus consistent test. Ledwina (1994) made the data-driven test omnibus consistent by allowing the maximal selectable order to converge to infinity as the sample size goes to infinity. This is exactly the major difference between her and our approach: we do not need the extension to the data-driven version so as to make the SSPc test omnibus consistent, because in Theorem 5.2 it has been shown that the SSPc test is omnibus consistent for any $c \geq 2$. Moreover, as we will show below, we may restrict the sample space of C_n to some finite space Γ of permissible SSP sizes.

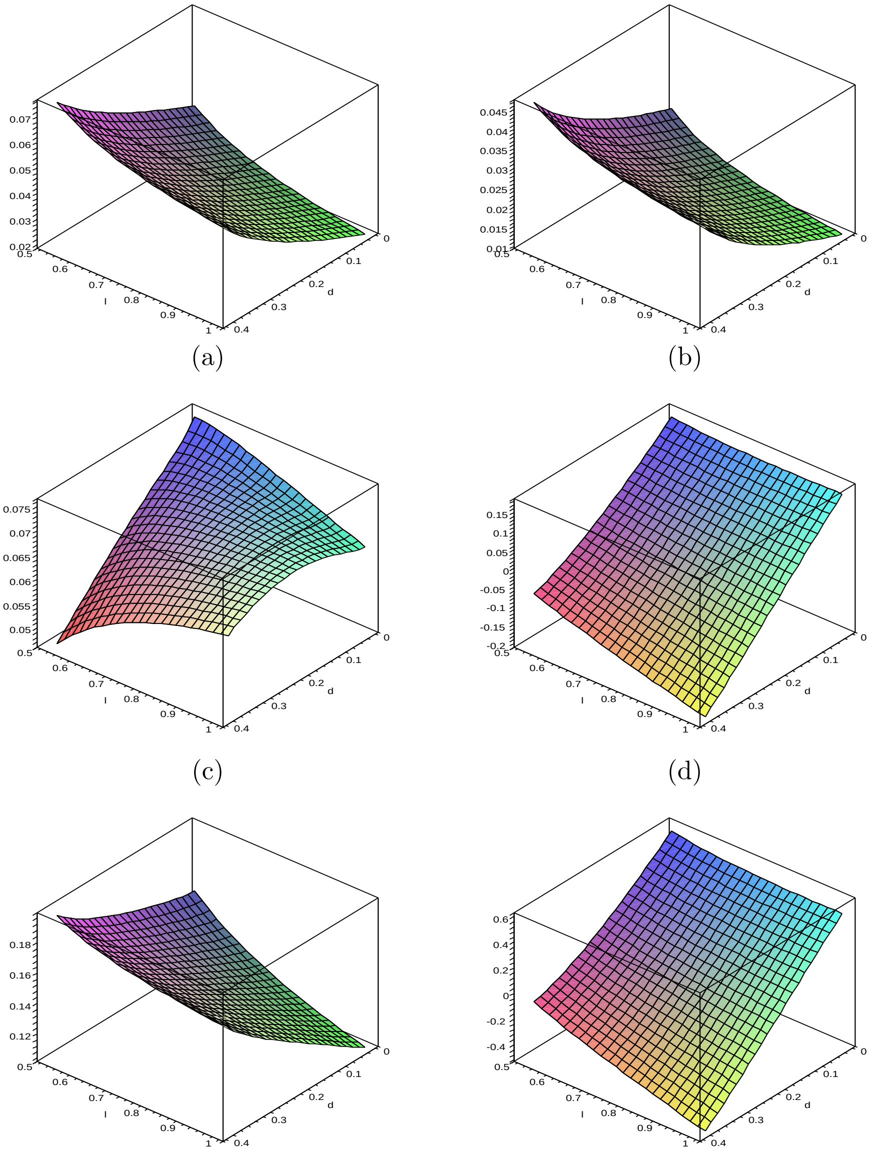
Originally, Ledwina (1994) used Schwarz's selection rule (Schwarz, 1978). Later, a computationally simpler rule was proposed (Kallenberg & Ledwina,



(a) (b)

(c)

Figure 5.6: The values for the W_n (a), A_n (b) and U_n (c) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.4$ as a function of the parameters d and l.

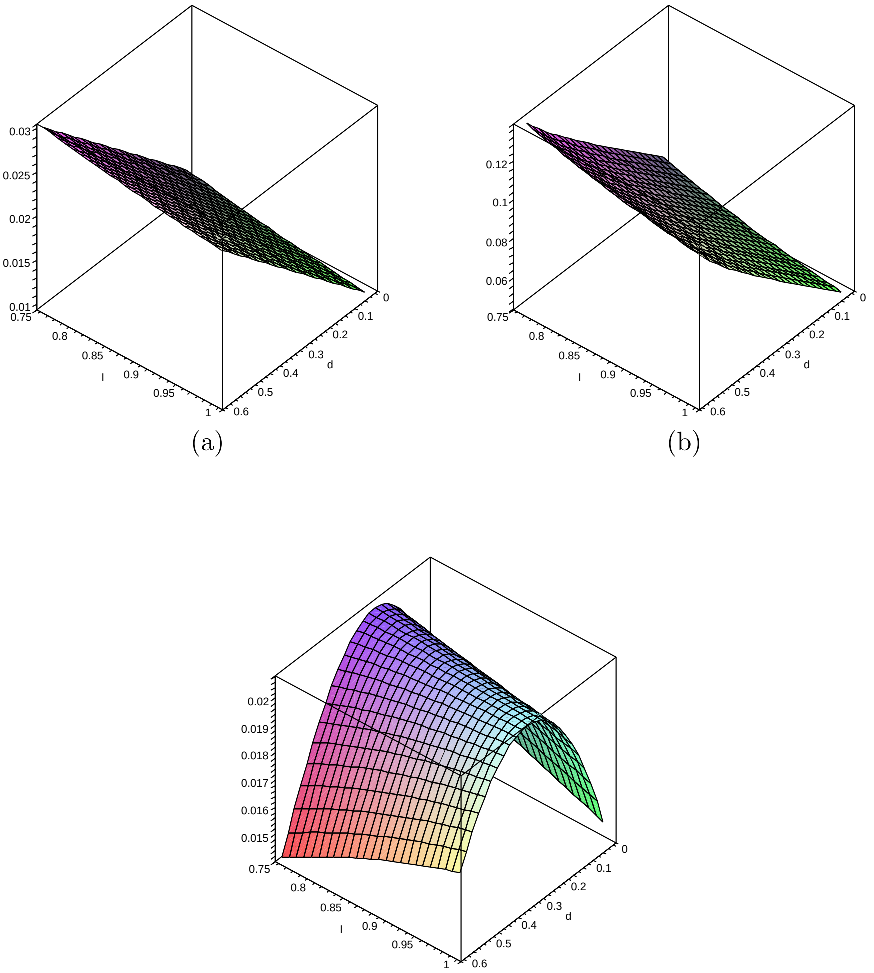


(a) (b)

(c) (d)

(e) (f)

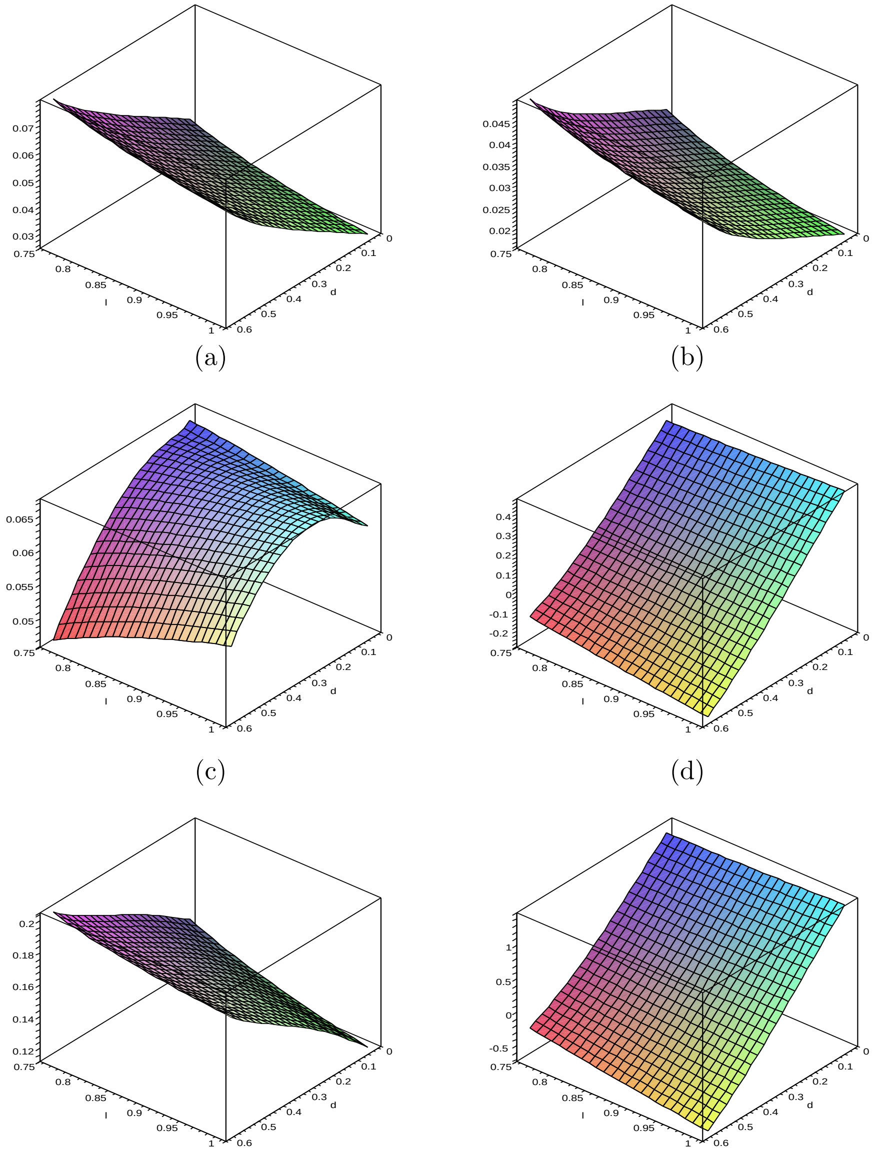
Figure 5.7: The values for $A_{3,n}$ (a), $U_{3,n}$ (c), $T_{3,n}$ (e), $A_{4,n}$ (b), $U_{4,n}$ (d) and $T_{4,n}$ (f) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.4$ as a function of the parameters d and l.



(a) (b)

(c)

Figure 5.8: The values for the W_n (a), A_n (b) and U_n (c) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.6$ as a function of the parameters d and l.



(a) (b)

(c) (d)

(e) (f)

Figure 5.9: The values for $A_{3,n}$ (a), $U_{3,n}$ (c), $T_{3,n}$ (e), $A_{4,n}$ (b), $U_{4,n}$ (d) and $T_{4,n}$ (f) for contiguous alternatives $F_n(x)$ in which $K = 0.3$ and $b = 0.6$ as a function of the parameters d and l.

1997), which is though still referred to as Schwarz's Bayesian Information Criterion (BIC). We propose a class of selection rules, indexed by the penalty a_n which is a non-decreasing sequence of real positive values. Let Γ denote the set of permissible SSP sizes, i.e. the selection rule will choose an appropriate SSP size among the elements of Γ. The class of selection rules which selects the SSP size C_n is given by

$$C_n = \mathrm{ArgMax}_{c\in\Gamma}\{T_{c,n} - 2(c-1)\ln a_n\}. \tag{5.58}$$

Although the form of this selection rule resembles the BIC ($a_n = n^{1/2}$) and the AIC ($a_n = e$) very closely, it has no sound theoretical justification, for $T_{c,n}$ is not a log-likelihood, as it is in AIC and BIC, nor a score statistic as it is in the modified BIC of Kallenberg and Ledwina (1997). Apart from the choices $a_n = n^{1/2}$ and $a_n = e$, we also consider a double logarithmic penalty term (LL), $a_n = \ln n$. We will refer to the data-driven versions as SSP-AIC, SSP-BIC and SSP-LL, depending on the penalty used.

From the simulation study (see below) and from personal experience, we propose to take $\Gamma = \{2,3,4\}$, or at most $\Gamma = \{2,3,4,5\}$. With these choices good powers have been observed.

5.4.2 Asymptotic theory

In this section we present some asymptotic distribution theory. Proofs are presented in Appendix B.5.

Theorem 5.7 *Let c_m denote the minimal SSP size, i.e. $c_m = \min_c \Gamma$. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0,*

$$P[C_n = c_m] \to 1$$

as $n \to \infty$.

Based on this result, the asymptotic null distribution of the data-driven test statistic $T_{C_n,n}$ is easily obtained.

Theorem 5.8 *Let $c_m = \min_c \Gamma$. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, the asymptotic null distribution of $T_{C_n,n}$ is given by*

$$\begin{aligned} T_{C_n,n} \quad \xrightarrow{d} \quad & (c_m - 1)\int_0^1 a_{c_m}(x)\mathbb{B}^2(x)dx + \\ & \binom{c_m-1}{2}\int_0^1\int_0^1 w_{c_m}(x,y)\left(\mathbb{B}(x)-\mathbb{B}(y)\right)^2 dxdy. \end{aligned}$$

Theorem 5.9 *$T_{C_n,n}$ is consistent against any fixed alternative.*

Proof. By the omnibus consistency of the SSPc test for every $c \in \Gamma$ (Theorem 5.2), the omnibus consistency of the data-driven test based on $T_{C_n,n}$ follows immediately. $\square$

5.5 Examples

In Chapter 2 we have introduced the simple one-sample GOF problem through some real data examples. In this section, we apply the SSPc tests for $c = 2, 3$ and 4 and the data-driven SSP tests using the AIC, BIC and LL criteria to the linear data examples introduced in Section 2.1. We will denote the order chosen by the AIC, BIC and LL criteria as C^{AIC}, C^{BIC} and C^{LL}, respectively.

For comparison reasons, the classical KS, CvM and data-driven smooth tests (see Chapter 3) are applied as well. The data-driven smooth test is denoted by S_K, where K is the order chosen by the BIC criterion. Finally, the ZA test described in Section 5.6 below is also included. We determine all p-values using 100,000 simulation runs, so as to make them comparable.

5.5.1 Lottery data

The p-values for all GOF tests for linear uniformity applied to the Lottery data are presented in the first column of Table 5.3. None of the tests give evidence against the null hypothesis. Note that the BIC selection criterion of the data-driven smooth test selects order $K = 2$. The smallest p-value (0.304) is registered for the data-driven smooth test. All other p-values are larger than 0.5. Applying the SSP tests is particularly useful for this data since a deviation from uniformity would give some interesting information about how the selection of the lottery numbers deviates from uniformity. In particular, a global deviation from uniformity for the Lottery data would imply that a large part of the possible lottery numbers are not selected from a uniform distribution, while a local deviation would refer to only some of the lottery numbers that cause the non-uniformity. The data-driven SSPc test chooses partition size $c = 2$ for each of the different selection criteria. The largest p-values are those for the Zhang test, the SSP3 and the SSP4 tests. Hence, these results suggest that no local deviations from uniformity are present. From this discussion we accept the null hypothesis of uniformity and conclude that the lottery numbers are distributed according to a uniform distribution.

Suppose now the numbers were systematically changed in the sense that 400 is subtracted from all numbers between 800 and 875. The GOF test results for the changed Lottery data are shown in the second column of Table 5.3. This induced deviation from uniformity is not detected by the KS, CvM, AD (which

Table 5.3: GOF test results (p-values and order selections) for the Lottery and the Lew data.

	Lottery	changed Lottery	original Lew	subsample Lew
KS	0.689	0.120	0.011	0.142
CvM	0.509	0.134	0.009	0.123
ZA	0.896	0.561	<0.001	0.045
S_K	0.304	0.472	<0.001	0.016
K	2	1	3	2
SSP2	0.624	0.186	<0.001	0.040
SSP3	0.781	0.057	<0.001	0.033
SSP4	0.863	0.02	<0.001	0.036
SSP-AIC	0.624	0.02	<0.001	0.034
C^{AIC}	2	4	4	4
SSP-BIC	0.624	0.186	<0.001	0.042
C^{BIC}	2	4	4	2
SSP-LL	0.624	0.005	<0.001	0.029
C^{LL}	2	4	4	4

is the SSP2 test), the data-driven smooth test and ZA test. The SSP3 test shows a borderline result, while the SSP4 test is highly significant at the 5% level. The data-driven SSP-AIC and SSP-LL also show significant results. Since the corresponding selection criteria choose $c = 4$, this clearly indicates a local deviation from uniformity.

5.5.2 Lew data

Regarding the Lew data, the results of the GOF tests for testing whether the beam deflection observations come from a uniform distribution on [-580,301] are in the second column of Table 5.3. Here, all tests give clear evidence against the null hypothesis. The KS test and the CvM test have the largest p-values. The data-driven smooth test selected the components in the score test up to the third order, meaning that the deviation from uniformity is related to the first three order moments. Looking at the individual components ($V_1^2 = 4.450$, $V_2^2 = 10.551$ and $V_3^2 = 35.062$), it is seen that the first and the second order component form only a small part of the total value of the statistic. From this we may conclude that the true distribution for the beam deflection points deviates from uniformity only with respect to the third moment, which in turn is related to the skewness.

Every data-driven SSPc test chooses SSP size four. This means that the deviation from uniformity is more detectable in small subintervals than in more global intervals. Note that from the individual SSP2 test a global LOF is concluded to be present as well.

Consider now the small ($n = 20$) randomly selected subsample of the Lew data. The last column of Table 5.3 lists the corresponding p-values, which expectedly are higher that those of the full sample. In fact, the KS and the CvM tests are not significant anymore at the 5% level. The Zhang test and the data-driven SSP-BIC test have p-values which are just below the significance level of 5%. Here, the BIC criterion chooses SSP size two, which indicates a global LOF. The data-driven smooth test has again the smallest p-value, followed by the data-driven SSP-LL test. For the smooth test, components up to the second order are selected. Hence, only deviations in the first two moments are responsible for the LOF. The SSP-LL test selected SSP size four indicating again a local LOF. In Chapter 6 we will demonstrate with our new graphical tool where these deviations from uniformity are located.

For this example we knew that the original beam deflection data was not at all uniformly distributed. On the contrary it had a bimodal pattern. Here, we have shown that this local deviation from uniformity can already be seen from the data-driven smooth test and the data-driven localised SSPc test in a small subsample of size 20.

We here also include the results for the circular Birth time data. This is done to illustrate the usefulness of appropriate circular SSPc tests, which are proposed in Section 5.8.

5.5.3 Birth time data

None of the classical tests indicated a significant difference from uniformity for the Birth time data. The p-values range from 0.283 for the data-driven smooth test of Bogdan to 0.632 for the Watson statistic (see Examples 3.3.5 and 3.5.2). In Example 3.3.2, we have applied Neyman's smooth test to the Birth time data and demonstrated that the conclusions are not invariant to the chosen origin. For the linear SSPc tests, the results are not origin-invariant either. Nevertheless, we here state the results for the origin chosen at midnight in order to compare them with the results for the appropriate circular SSPc in Section 5.8. The p-values for all the data-driven linear SSPc tests where $\Gamma = \{2, 3, 4\}$ equal 0.686, where all selection criteria (AIC, BIC and LL) chose $c = 2$. The p-value is equal for all data-driven SSP tests since most (± 85%) of the bootstrap samples choose the same SSP size. This p-value is the largest among all tests. The p-values for the separate SSPc tests are 0.686, 0.648 and 0.367 for $c = 2, 3$ and 4, respectively. Note that the p-values decrease as c

increases, meaning that over the whole range ($c = 2$) the fit is accepted to be uniform. In smaller intervals ($c = 4$), the conclusion is similar but the p-value is only half as large. Nevertheless, all the data-driven versions choose $c = 2$. This indicates no deviation from linear uniformity if the origin is at 12am. However, taking any other origin leads to different p-values and may lead to different conclusions. In Section 5.8, we will present the results for the circular version of the SSPc tests. The conclusions for the circular SSP tests will not depend on the choice of the origin.

5.6 Simulation study

In the previous sections it has been shown that the SSPc test and its data-driven versions are omnibus consistent (Theorems 5.2 and 5.9), which is an asymptotic property. In practice, however, the finite power characteristics are of more importance. In this section we give the results of a Monte Carlo study in which we investigate the power of the SSPc tests ($c = 2, 3, 4$) and their data-driven versions ($\Gamma = \{2, 3, 4\}$) for sample sizes 20 and 50. The computational formulae of Section 5.2.2 are used. For comparison purposes we also include some traditional GOF tests that are described in Chapter 3. In particular, we consider the CvM test, the KS test and the data-driven Neyman smooth test. Furthermore, Zhang's (2002) test based on his Z_A statistic (ZA) is included. The ZA test statistic is an integral statistic of the form (5.3) with $P_n(x)$ the localised likelihood ratio statistic and with $dw(x) = \hat{F}_n^{-1}(x)(1 - \hat{F}_n(x))^{-1} d\hat{F}_n(x)$. The computational form is given by

$$Z_A = -\sum_{i=1}^{n} \left[\frac{\ln(X_{(i)})}{n - i + 0.5} + \frac{\ln(1 - X_{(i)})}{i - 0.5} \right].$$

Zhang reports in his simulation study that ZA has generally good power characteristics. Finally, note that since the AD test is a special case of the family of the SSPc tests ($c = 2$), it is already included in the study.

We have performed simulations under two different types of alternatives to the null hypothesis of standard normality. The first type is a normal distribution with mean μ and variance σ^2. This simple alternative is mainly included to assess the sensitivity of the SSPc tests to changes in mean μ when the variance is kept constant at its correct value $\sigma^2 = 1$, and the sensitivity to changes in σ^2 when the mean is kept at its hypothesised value $\mu = 0$. For the former series of experiments, μ is varied between 0 and 1. In the latter series, σ is varied from 1 to 2.2.

In Section 5.2 we have argued that the form of the SSPc statistic suggests that with increasing SSP size c, it may become more and more sensitive to deviations from F_0 in small intervals. As an example of such alternatives, we have included a family of mixtures of normal distributions. In particular, this mixture has density

$$f_{\delta,\gamma}(x) = (1-\gamma)\phi(x;0,1) + \gamma\phi(x;\delta,0.01), \tag{5.59}$$

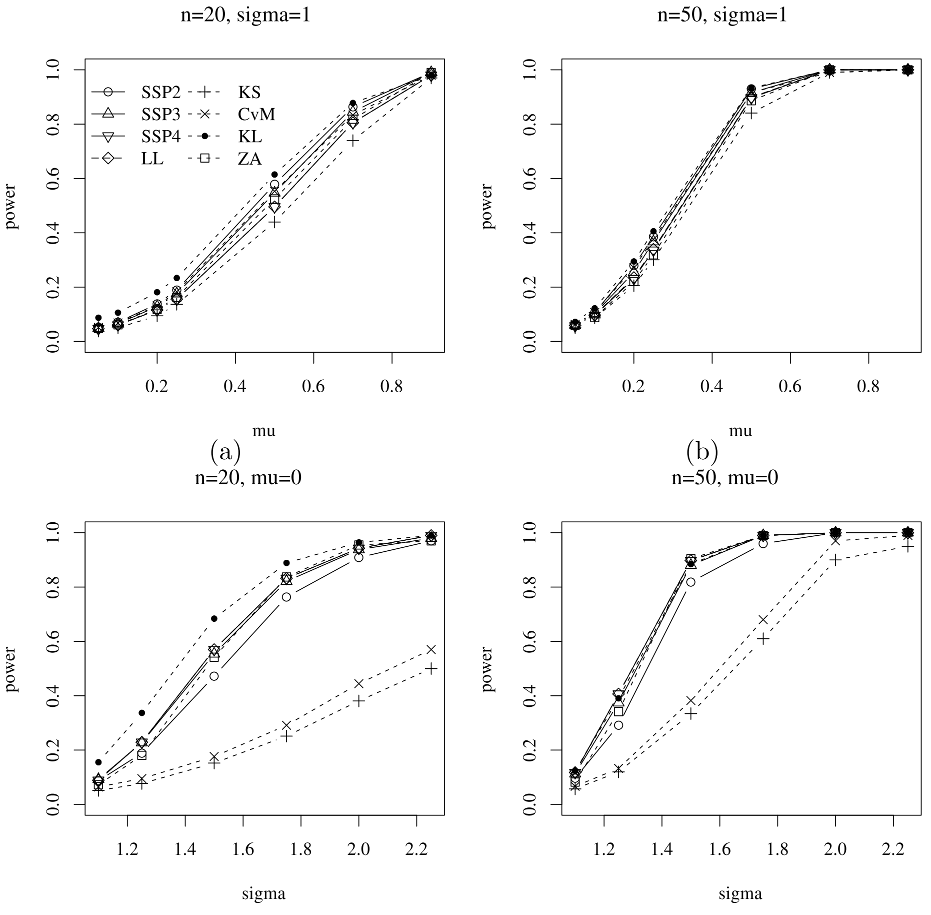
where $\phi(x,\mu,\sigma)$ is the density of a normal distribution with mean μ and standard deviation σ. Note that $(\delta,\gamma) = (0,0)$ results in the hypothesised standard normal distribution. This mixture may be interpreted as a standard normal distribution which is contaminated with another normal distribution with small standard deviation, which is therefore clearly "localised"around its mean δ. We will refer to this alternative as the contaminated normal distribution.

All powers are estimated based on 10,000 Monte Carlo simulation runs. Tests are performed at the 5% level of significance. To make all powers comparable, we have used simulated critical points for all tests (based on 50,000 simulation runs). The results are shown in Figures 5.10 (normal) and 5.11 (contaminated normal). To avoid the figures to become too messy, we have limited the presentation of the results of the data-driven SSP test to the one with the largest power.

The results for the normal alternatives with constant variance are in the upper panels of Figure 5.10 (the left panel corresponds to $n = 20$, the right to $n = 50$). It can be seen that all tests have quite similar powers. However, it is important to note that there is a small loss in power when the SSP size c is increased from $c = 2$ to $c = 3$ and further to $c = 4$. The lower panels represent the situation where the mean is fixed at $\mu = 0$. We see that now the opposite is observed: the power clearly increases with increasing SSP size. These two series of simple normal alternatives demonstrate the importance of the choice of the SSP size. From the three data-driven SSP tests, it is the SSP-LL test that outperforms the other two. In the fixed σ^2 series, this test has powers in between those of the SSP2 and SSP3 test, and in the fixed μ simulations, the data-driven test is almost indistinguishable from the SSP4 test. Further, it is interesting to note that overall the best powers are obtained with the data-driven smooth test (referred to as KL in the legend of Figures 5.10 and 5.11) when $n = 20$, but at the larger sample size ($n = 50$) the SSP4 test sometimes performs better. Zhang's ZA test always has powers in between those of the SSP2 and SSP4 tests. Finally, note that the traditional KS and CvM tests have considerably lower powers for detecting variance misspecifications.

Since all tests are very sensitive to changes in the mean, we have shifted the simulated data from the contaminated normal alternatives by subtracting the true mean, which is equal to $\gamma\delta$.

The results of the simulation study for the contaminated normal distribution are presented in Figure 5.11 for $\delta = 1$ (top), $\delta = 1.5$ (middle) and $\delta = 2$ (bottom). From these plots, which show the power as a function of γ, we generally



(a) (b) (c) (d)

Figure 5.10: Estimated power curves for the normal distribution alternative. The legend is only shown in the first plot

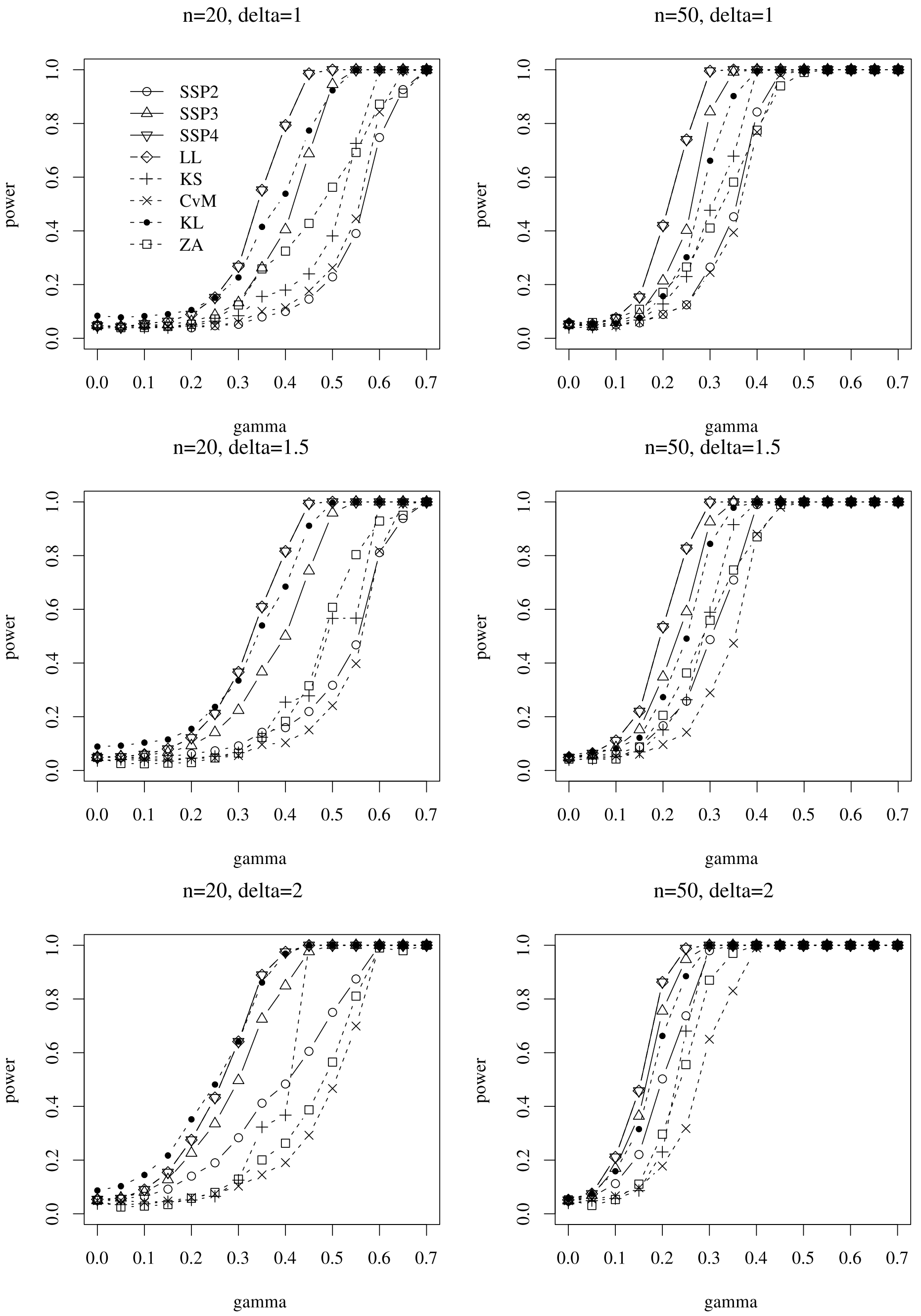


Figure 5.11: Estimated power curves for the contaminated normal distribution alternative. The legend is only shown in the first plot

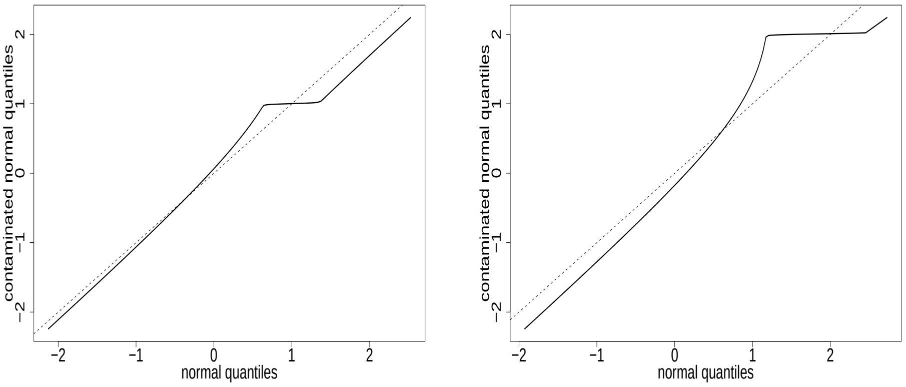


Figure 5.12: QQ-plots of the contaminated normal distributions $(\delta, \gamma) = (1, 0.2)$ (left) and $(\delta, \gamma) = (2, 0.2)$ (right) versus the standard normal distribution.

conclude that the power of the SSPc tests increases as the SSP size c increases. Furthermore, it is seen that the data-driven version with the LL penalty succeeds to select an appropriate value for c. In particular, the behaviour of the SSP-LL test is almost exactly equal to that of the SSP4 test. The powers of the KS, CvM and ZA tests are never as large as the powers of the new SSP3 and SSP4 tests. Among all tests included, the data-driven smooth test (KL) shows the highest powers when $\delta = 2$ for sample size $n = 20$, but it is outperformed by the SSPc tests for the larger sample size $n = 50$. The same sample size effect on the smooth test is also seen for $\delta = 1.5$ and $\delta = 1$. Furthermore, as δ decreases from 2 over 1.5 to 1, the smooth test loses power compared to the SSPc tests. A small absolute value of δ means that the contamination is better "hidden" in the probability mass of the standard normal compound of the mixture. This is illustrated in Figure 5.12, where we show normal QQ-plots for the contaminated normal alternatives with $(\delta, \gamma) = (1, 0.2)$ (left) and with $(\delta, \gamma) = (2, 0.2)$ (right). The QQ-plot in case of $(\delta, \gamma) = (2, 0.2)$ shows a more "local" departure from standard normality in the sense that the deviation is more concentrated in a small interval of the support (the peak at the location of the LOF is larger compared to the global deviation from the straight line). With $(\delta, \gamma) = (1, 0.2)$ the departure is seen in a large interval (the peak at the LOF is relatively smaller).

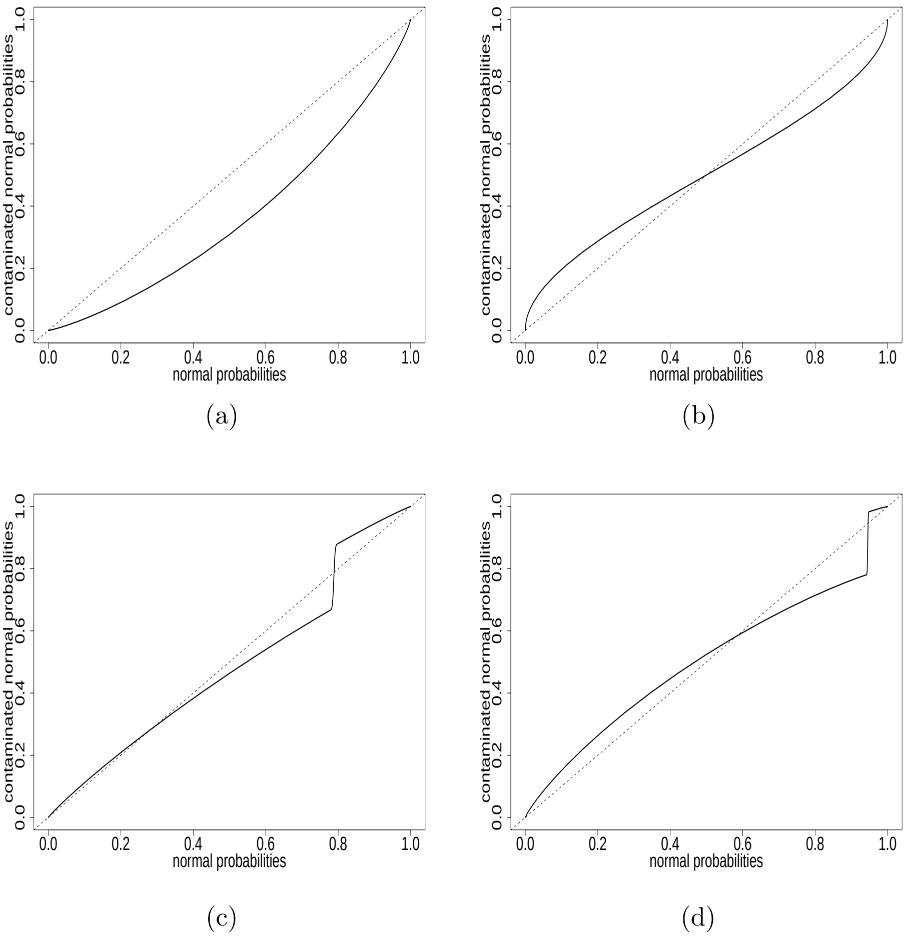
These observations confirm our intuitive interpretation of the weight functions in the SSPc statistic (Section 5.2).

In order to see whether the conclusions from the discussion in Section 5.3

about the limiting expected values of the SSPc tests are in accordance with the power results above, we compare the two families of alternatives considered in both discussions. In particular, we draw PP-plots comparing some alternative distributions to the null distribution as considered in the above simulation study. We then compare these plots to the plot of $F_n(x)$ in (5.51) with Δ given by (5.57) in panel (a) of Figure 5.1. Note that the latter plot can analogously be interpreted as a PP-plot since the null distribution is considered to be uniform on $[0,1]$. Recall that small values of d in δ correspond to local deviations from uniformity. These local deviations are recognised in the PP-plot as a steep, almost vertical line at location b. The PP-plots now for the distributions $\phi(x;0,1.5)$, $\phi(x;0.5,1)$, $f_{1,0.2}(x)$ and $f_{2,0.2}(x)$ are shown in Figure 5.13, panels (a), (b), (c) and (d), respectively. The probabilities of the distributions $\phi(x;0.5,1)$ and $\phi(x;0,1.5)$ are plotted versus the standard normal probabilities, while those of the contaminated normal distributions are plotted versus the normal probabilities with mean $\delta\gamma$ and standard deviation 1. In the PP-plots of the first two alternatives, which represent a shift in location and scale respectively, no steep vertical line is recognised. These alternatives are examples of a global LOF. The PP-plot of a "shift in mean" alternative is comparable to that of an alternative $F_n(x)$ to uniformity where d is large and l is small, even though the curve on the former PP-plot is smoother and below the diagonal. A PP-plot for a normal alternative with mean at -0.5 would yield the same curve but reflected around the diagonal line. For such alternatives, all tests had similar power. The alternative which represents the shift in scale in panel (b) of Figure 5.13 does not correspond to a member of the family $F_n(x)$. The PP-plots of the two contaminated normal alternatives correspond to local alternatives in the family $F_n(x)$. Indeed, at the location of the LOF, the PP-plot shows a vertical line. For such alternatives, the expected values of the statistics were relatively high for the SSP4 test, while the SSP3 test also had a reasonably high expected value due to the $U_{c,n}$ term. Generally, we may conclude that the intuition based on the asymptotic results in Section 5.3 is confirmed by the simulation study, in the sense that the SSPc tests are most effective in case of local deviation.

5.7 SSPc test for composite null hypothesis

An interesting direction of research is to investigate the properties of the tests in the case of a composite null hypothesis. In particular, we want to test whether the true F belongs to some hypothesised family of distributions with unknown parameters. Similarly to all other tests, the same SSPc statistic that was used for the simple null hypothesis, is also used for testing the composite null hypothesis. The only adaptation is the replacement of $F_0(x)$ by $F_0(x;\hat{\boldsymbol{\beta}})$, where



(a) (b)

(c) (d)

Figure 5.13: PP-plots for the distributions $\phi(x; 0.5, 1)$ (a), $\phi(x; 0, 1.5)$ (b) and $f_{\delta,\gamma}(x)$, with (δ, γ) equal to $(1, 0.2)$ (c) and $(2, 0.2)$(d).

$\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. An important consequence is that the (asymptotic) null distribution changes and often becomes more complicated. We have briefly mentioned this problem for general EDF GOF tests in Section 3.4. Here we will give a more elaborate discussion on EDF tests in the presence of nuisance parameters. Note that the SSPc family of test statistics belongs to the EDF class since its expression is also in terms of the empirical process $\mathbb{B}_n$. Therefore, it is important to see how $\mathbb{B}_n$ behaves under nuisance parameter estimation.

We again make the dependence on the p-dimensional parameter $\boldsymbol{\beta}$ more explicit by using the notations $\mathbb{B}_n(x;\boldsymbol{\beta})$ and $\mathbb{B}(x,\boldsymbol{\beta})$ for the empirical and Gaussian processes, respectively. The covariance function of the Gaussian process is as before given by

$$\text{Cov}\left[\mathbb{B}(x,\boldsymbol{\beta}),\mathbb{B}(y,\boldsymbol{\beta})\right] = F_0(x\wedge y,\boldsymbol{\beta}) - F_0(x,\boldsymbol{\beta})F_0(y,\boldsymbol{\beta}). \tag{5.60}$$

When the nuisance parameters are estimated, the estimators are plugged in into the empirical process, resulting in the *estimated empirical process*, say

$$\hat{\mathbb{B}}_n(x) = \mathbb{B}_n(x,\hat{\boldsymbol{\beta}}). \tag{5.61}$$

To find the asymptotic behaviour of $\hat{\mathbb{B}}_n(x)$ some assumptions on the distribution F_0 and on the estimation method are required. The estimator $\hat{\boldsymbol{\beta}}$ has to be a *locally asymptotically linear* estimator, denoted by $\hat{\boldsymbol{\beta}}_n$, which means that the following expansion holds,

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Upsilon}(X_i,\boldsymbol{\beta}) + o_P(1), \tag{5.62}$$

where $\boldsymbol{\Upsilon}^T = (\Upsilon_1,\ldots,\Upsilon_p)$ is a continuously differentiable vector function $I\!R^p \to I\!R^p$ and has $\text{E}\left[\dot{\Upsilon}_j^2\right] < \infty$ and $\text{E}\left[\Upsilon_j^2\right] < \infty$ ($\dot{\Upsilon}$ denotes the first derivative of Υ). This property holds for many well known estimators, e.g. maximum likelihood estimators and method of moment estimators. The following theorem, and its proof, can be found in e.g. van der Vaart (1998).

Theorem 5.10 *Given a locally asymptotically linear estimator $\hat{\boldsymbol{\beta}}_n$, the estimated empirical process $\hat{\mathbb{B}}_n$ converges weakly to a zero mean Gaussian process $\hat{\mathbb{B}}$ with covariance function*

$$\begin{aligned} Cov\left[\hat{\mathbb{B}}(x),\hat{\mathbb{B}}(y)\right] &= F_0(x\wedge y,\boldsymbol{\beta}) - F_0(x,\boldsymbol{\beta})F_0(y,\boldsymbol{\beta}) \\ &\quad -\boldsymbol{\Lambda}^T(x,\boldsymbol{\beta})\boldsymbol{g}(y,\boldsymbol{\beta}) - \boldsymbol{\Lambda}^T(y,\boldsymbol{\beta})\boldsymbol{g}(x,\boldsymbol{\beta}) \\ &\quad +\boldsymbol{g}^T(x,\boldsymbol{\beta})\Sigma_{\boldsymbol{\Upsilon}}\boldsymbol{g}(y,\boldsymbol{\beta}), \end{aligned} \tag{5.63}$$

where $\boldsymbol{\Lambda}(x,\boldsymbol{\beta}) = \int_{-\infty}^{x}\boldsymbol{\Upsilon}(z,\boldsymbol{\beta})dF_0(z,\boldsymbol{\beta})$, $\boldsymbol{g}(x,\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}F_0(x,\boldsymbol{\beta})$, and $\boldsymbol{\Sigma}_{\boldsymbol{\Upsilon}} = Var[\boldsymbol{\Upsilon}(X,\boldsymbol{\beta})]$.

Let $\hat{T}_{c,n}$ denote the statistic given by (5.13) where the empirical process $\mathbb{B}_n(x)$ is replaced by $\hat{\mathbb{B}}_n(x)$. Then the asymptotic null distribution of this SSPc statistic for composite null hypothesis follows from the preceding discussion and from (5.18), i.e.

$$\begin{aligned}\hat{T}_{c,n} \quad \xrightarrow{d} \quad \hat{T}_c &= (c-1)\int_0^1 a_c(x)\hat{\mathbb{B}}^2(x)dx \\ &+ \binom{c-1}{2}\int_0^1\int_0^1 w_c(x,y)(\hat{\mathbb{B}}(x)-\hat{\mathbb{B}}(y))^2dxdy. \qquad (5.64)\end{aligned}$$

Note that the SSPc test for a composite null hypothesis is not distribution-free since its limit distribution (5.64) depends on the unknown $\boldsymbol{\beta}$ and on the hypothesised distribution F_0. We therefore prefer to use the parametric bootstrap to approximate the null distribution of $\hat{T}_{c,n}$. Although the parametric bootstrap procedure is already used in a similar context in Section 3.4, we here explicitly mention the results concerning the asymptotic equivalence between the empirical process and its bootstrapped analogue.

The parametric bootstrap procedure consists in generating i.i.d. random variables from $F(x,\hat{\boldsymbol{\beta}})$, say $X_1^*,\ldots,X_n^*$. Let $\hat{\boldsymbol{\beta}}^*$ be the estimator of the parameter vector $\boldsymbol{\beta}$ in the bootstrap sample. For each bootstrap sample, the empirical process is given by

$$\mathbb{B}_n(x;\hat{\boldsymbol{\beta}}^*) = \sqrt{n}(\hat{F}_n^*(x) - F(x;\hat{\boldsymbol{\beta}}^*)), \qquad (5.65)$$

where $\hat{F}_n^*(x)$ denotes the EDF of $X_1^*,\ldots,X_n^*$. Many bootstrap samples are generated, resulting in many empirical process values. These simulated process values then serve as an approximation to the distribution of $\mathbb{B}_n(x;\hat{\boldsymbol{\beta}})$ under the null hypothesis. This procedure is considered sensible since it is proved by Babu and Rao (2004) and Stute, Gonzáles-Manteiga, and Presedo-Quindimil (1993) that the distribution of the bootstrapped empirical process $\mathbb{B}_n(x;\hat{\boldsymbol{\beta}}^*)$, weakly converges to the same Gaussian process $\hat{\mathbb{B}}(x)$ as the original empirical process $\mathbb{B}_n(x;\hat{\boldsymbol{\beta}})$. As mentioned in Section 3.4.1, the process $\hat{\mathbb{B}}$ becomes independent of the parameter $\boldsymbol{\beta}$ when F_0 is a location-scale invariant distribution. Since the dependence on F_0 remains, one single series of simulations for each sample size with arbitrary chosen $\boldsymbol{\beta}$ is sufficient.

Similarly as in the simple case, a data-driven version of the SSP test can be constructed and its distribution under the null hypothesis can be obtained using the parametric bootstrap.

5.7.1 Examples

In Chapter 2 we also considered three linear data examples to introduce the composite one-sample GOF problem. We now apply the SSPc tests for $c = 2, 3, 4$, the SSP-AIC, SSP-BIC and SSP-LL to those examples to test for normality.

The classical KS, the CvM, the data-driven smooth tests for composite normality (see Chapter 3) and the ZA test are considered as well. Note that the latter can be extended to a test for composite normality, similarly to the SSPc test. We determine all p-values using 100,000 parametric bootstrap samples. Since the normal distribution belongs to a location-scale family, only the sample size is crucial in the bootstrap simulations. In particular, the parameters of the hypothesised normal distribution have no effect on the bootstrapped null distribution and can therefore be chosen arbitrarily. We choose to generate bootstrap samples from the standard normal distribution to obtain the null distribution for each of the test statistics.

Chemical concentration data

For the PCB data, the different GOF tests for normality give different results, as can be seen in the first column of Table 5.4. First, the data-driven smooth test (S_K) and the CvM test are significant at the 5% level. Further, the KS test and the SSP-BIC test give borderline p-values. Finally, the Zhang, the SSP-LL and the SSP-AIC are clearly not significant at the 5% level. The BIC criterion for the data-driven smooth test chooses order $K = 3$, which implies that the significant result is due to the third order moment. This is consistent with the skewed impression of the data as was apparent in Section 2.1.3. Each criterion for the data-driven SSPc test selected SSP size $c = 2$. Form their non-significant p-values, we may conclude that there is neither a global nor a local deviation from normality.

Fastfood data

For the Fastfood data, we again use the same GOF tests to check whether the service-times are normally distributed. The results are in the second column of Table 5.4. Here, the Kuiper test has the smallest p-value, followed by the Zhang test. At the 5% level, the data-driven smooth test is significant as well and informs us that the deviation is possibly due to the skewness. This again confirms the skewed impression from the explorative plots in Section 2.1.4.

For this example it is seen that the choice of the SSP size is important to obtain significant results. Indeed, the SSP4 test is not significant. All selection criteria choose SSP size $c = 2$, which seems to indicate that the LOF is located over the whole range of the sample size. However, from the simulation study in Section 5.6 we know that the choice of $c = 2$ also tends to occur in case of local deviations situated close to the mode (while higher values of c occur for local deviations away from the mode). It would be reasonable to conclude that such a situation is present here, since the detrended PP-plot also localised the deviation from normality in the mode of the distribution (see Section 2.1.4).

Table 5.4: GOF test results (p-values and order selections) for the PCB, the Fastfood and the Old Faithful Geyser (OFG) data.

	PCB	Fastfood	original OFG	subsample OFG
KS	0.052	0.004	<0.001	0.069
CvM	0.036	0.03	<0.001	0.028
ZA	0.106	0.019	<0.001	0.07
S_K	0.035	0.022	0.001	0.03
K	3	3	10	3
SSP2	0.05	0.036	<0.001	0.024
SSP3	0.055	0.044	<0.001	0.022
SSP4	0.058	0.053	<0.001	0.019
SSP-AIC	0.209	0.037	<0.001	0.024
C^{AIC}	2	2	3	2
SSP-BIC	0.055	0.036	<0.001	0.024
C^{BIC}	2	2	3	2
SSP-LL	0.124	0.036	<0.001	0.024
C^{LL}	2	2	3	2

Old faithful geyser data

Finally, consider the Old Faithful geyser data. For the original eruption times, all GOF tests give extremely significant p-values (<0.001, see Table 5.4, third column), which is an indication that the normal distribution is not appropriate. Let us see how these results carry over to the small random subsample of size 20. The results in the last column in Table 5.4, are obviously less striking. Moreover, the Zhang and the KS test show no difference from normality at the 5% significance level. The CvM test does reject the null hypothesis, but here we have no information about how the distribution deviates from normality. The data-driven smooth and the data-driven SSP tests, on the other hand, reveal interesting information about the true distribution. In particular, the data-driven smooth test selected order $K = 3$, which means that in the subsample a deviation from normality with respect to the skewness is present. The selection criteria for the data-driven SSP tests each chose $c = 2$, which implies that the LOF is over the whole range. Note that for these data-driven GOF tests the selected order is smaller for the subset than for the original data.

5.8 The circular SSPc test

5.8.1 Construction of the test statistics

In this section we extend the Rothman test statistic of (5.5) with $w(x;x_0) = F_0(x;x_0)$, i.e.

$$R_n = \frac{1}{2\pi}\int_0^{2\pi}\int_{x_0}^{2\pi+x_0} P_n(x;x_0)dF_0(x;x_0)dx_0, \tag{5.66}$$

to the class of *localised circular SSPc tests* by considering partitions of general size c. Note that, as in the linear case, the superscript w is omitted since w is always chosen equal to F_0. The class of circular SSPc statistics can also be derived from the class of linear SSPc statistics. In fact, we make the class of linear SSPc statistics origin-invariant by taking the integral over all possible origins.

We consider $X_1, \ldots, X_n$ to be a sample of i.i.d. observations with circular distribution $F_0(x)$, $0 \leq x \leq 2\pi$. Note that here the origin is at zero. When we choose to take the origin at x_0, we make this explicit in the notation. The circle is denoted by $\mathcal{S}_{x_0} = \mathrm{arc}(x_0, x_0 + 2\pi)$ and the set of $c-1$ dividing points on that circle is given by $D_c^{x_0} = \{x_1, \ldots, x_{c-1}\} \in \mathcal{S}_{x_0}$. The ordered elements of $D_c^{x_0}$ are $x_0 < x_{(1)} \leq \ldots \leq x_{(c-1)} < x_0 + 2\pi$. Similarly as in the linear case, every $D_c^{x_0}$ induces a multinomial distribution with probabilities

$$\pi_1 = F_0(x_{(1)};x_0); \pi_2 = F_0(x_{(2)};x_0) - F_0(x_{(1)};x_0); \ldots; \pi_c = 1 - F_0(x_{(c-1)};x_0). \tag{5.67}$$

Let $P_{c,n}^{x_0}(D_c^{x_0})$ denote the Pearson χ^2 statistic for testing for a multinomial distribution with probabilities (5.67) induced by the partition $D_c^{x_0}$ of size c, i.e.

$$\begin{aligned} P_{c,n}^{x_0}(D_c^{x_0}) &= P_{c,n}^{x_0}(x_1, \ldots, x_{c-1}) \\ &= n\sum_{i=1}^{c} \frac{(\hat{F}_n(x_{(i)};x_0) - \hat{F}_n(x_{(i-1)};x_0) - (F_0(x_{(i)};x_0) - F_0(x_{(i-1)};x_0)))^2}{F_0(x_{(i)};x_0) - F_0(x_{(i-1)};x_0)}, \end{aligned}$$

where $x_{(0)} \equiv x_0$ and $x_{(c)} \equiv 2\pi + x_0$ coincide on the circle. From (3.64), we have that $\hat{F}_n(x_{(0)};x_0) = F_0(x_{(0)};x_0) = 0$ and $\hat{F}_n(x_{(c)};x_0) = F_0(x_{(c)};x_0) = 1$. We propose the class of test statistics

$$R_{c,n} = \int_0^{2\pi}\int_{\mathcal{S}_{x_0}} \cdots \int_{\mathcal{S}_{x_0}} P_{c,n}^{x_0}(x_1, \ldots, x_{c-1})dF_0(x_1;x_0)\ldots dF_0(x_{c-1};x_0)dx_0. \tag{5.68}$$

for which the Rothman test statistic is a special case. In particular, Rothman's statistic $R_n = R_{2,n}$ considers partitions of two cells. Alternatively, the statistic

(5.68) can be viewed as an extension of the linear SSPc statistic to an origin-invariant SSPc statistic. This can be seen by rewriting $R_{c,n}$ as a function of

$$T_{c,n}^{x_0} = \int_{\mathcal{S}_{x_0}} \dots \int_{\mathcal{S}_{x_0}} P_{c,n}^{x_0}(x_1, \dots, x_{c-1}) dF_0(x_1; x_0) \dots dF_0(x_{c-1}; x_0),$$

where the superscript x_0 indicates the dependence on the origin. With this notation we get

$$R_{c,n} = \int_0^{2\pi} T_{c,n}^{x_0} dx_0. \tag{5.69}$$

This relation is particularly helpful for rewriting $R_{c,n}$ in a more attractive form. We assume without loss of generality that the hypothesised distribution is the CU distribution on the circle $\mathcal{S}_{x_0}$. Because we would like to use the properties of the linear SSPc tests, we choose to test for circular uniformity on the circle with circumference equal to one instead of 2π. Then the CDF of the CU distribution is given by

$$F_0(x; x_0) = x - x_0 \tag{5.70}$$

where x is forced to be between x_0 and $1 + x_0$. This means that a multiple of the period 1 is added if necessary. Let $\mathbb{B}_n(x; x_0) = \sqrt{n}\left(\hat{F}_n(x; x_0) - (x - x_0)\right)$ denote the empirical process starting at x_0. Then, we can write

$$\begin{aligned} R_{c,n} = {} & (c-1) \int_0^1 \int_{x_0}^{1+x_0} a_{c,n}(x; x_0) \mathbb{B}_n^2(x; x_0) dx dx_0 \qquad (5.71) \\ & + \binom{c-1}{2} \int_0^1 \int_{x_0}^{1+x_0} \int_{x_0}^{1+x_0} w_{c,n}(x, y; x_0) (\mathbb{B}_n(x; x_0) - \mathbb{B}_n(y; x_0))^2 dx dy dx_0, \end{aligned}$$

where

$$\begin{aligned} a_c(x; x_0) &= \frac{((1 - (x - x_0))^{c-1} + (x - x_0)^{c-1})}{(x - x_0)(1 - (x - x_0))} \text{ and} \\ w_c(x, y; x_0) &= \frac{(1 - (x \vee y) + x_0)^{c-2} - ((x \wedge y) + x_0)^{c-2}}{(1 - (x \vee y) - (x \wedge y) + 2x_0)|x - y|}. \end{aligned}$$

5.8.2 Asymptotic theory

In this section, we give a theorem about the limiting null distribution of $R_{c,n}$. The proof is similar as in the linear case, and is omitted here.

Theorem 5.11 *Let $\{\mathbb{B}(x; x_0), x \in arc[x_0, x_0 + 1]\}$ denote a Brownian bridge which starts at x_0 and ends at $x_0 + 1$ with mean zero and covariance function given by*

$$Cov[\mathbb{B}(x; x_0), \mathbb{B}(y; x_0)] = F_0(x \wedge y; x_0) - F_0(x; x_0) F_0(y; x_0). \tag{5.72}$$

Suppose $c \geq 2$ is given, then, under the null hypothesis, as $n \to \infty$,

$$\begin{aligned} R_{c,n} \xrightarrow{d} R_{c,\infty} &= (c-1)\int_0^1 \int_{x_0}^{1+x_0} a_c(x;x_0)\mathbb{B}^2(x;x_0)dxdx_0 \\ &+ \binom{c-1}{2}\int_0^1 \int_{x_0}^{1+x_0}\int_{x_0}^{1+x_0} w_c(x,y;x_0)(\mathbb{B}(x;x_0)-\mathbb{B}(y;x_0))^2 dxdydx_0. \end{aligned} \tag{5.73}$$

Omnibus consistency is established similarly as in the linear case.

5.8.3 Computational formulae

We give explicit computational formulae for $c = 2$, $c = 3$ and $c = 4$. Recall that $T_{2,n}$ is the AD statistic and $T_{3,n}$ and $T_{4,n}$ can be written as linear combination of the AD (A_n), the CvM (W_n) and the weighted Watson (K_n) statistics. Furthermore, by (5.69), (5.20) and (5.22)-(5.23), we have that

$$R_{2,n} = \int_0^1 V_n(\Psi^{x_0}_{\text{AD}})dx_0, \tag{5.74}$$

$$R_{3,n} = \int_0^1 2V_n(\Psi^{x_0}_{\text{AD}}) - 4V_n(\Psi^{x_0}_{\text{CvM}}) + V_n(\Omega^{x_0})dx_0 \tag{5.75}$$

$$R_{4,n} = \int_0^1 3V_n(\Psi^{x_0}_{\text{AD}}) - 10.5V_n(\Psi^{x_0}_{\text{CvM}}) \tag{5.76}$$

$$+3V_n(\Omega^{x_0}) + 1.5V_n(\Xi^{x_0})dx_0. \tag{5.77}$$

where

$$\begin{aligned} \Psi^{x_0}_{\text{AD}}(x,y) &= \frac{1}{n}\int_{x_0}^{1+x_0} \frac{\mathbb{B}_n(u,x;x_0)\mathbb{B}_n(u,y;x_0)}{(u-x_0)(1-u+x_0)}du \\ \Psi^{x_0}_{\text{CvM}}(x,y) &= \frac{1}{n}\int_{x_0}^{1+x_0} \mathbb{B}_n(u,x;x_0)\mathbb{B}_n(u,y;x_0)du \\ \Omega^{x_0}(x,y) &= \frac{1}{n}\int_{x_0}^{1+x_0}\int_{x_0}^{1+x_0} (\mathbb{B}_n(u\wedge v,x;x_0) - \mathbb{B}_n(u \vee v,x;x_0)) \\ &\quad (\mathbb{B}_n(u\wedge v,y;x_0) - \mathbb{B}_n(u\vee v,y;x_0))\frac{1}{|u-v|}dudv \\ \Xi^{x_0}(x,y) &= ((x-x_0)\text{mod}1 - 0.5)((y-x_0)\text{mod}1 - 0.5), \end{aligned}$$

where mod is the modulo operator. We can rewrite (5.74)-(5.77) as

$$R_{2,n} = V_n\left(\int_0^1 \Psi_{\mathrm{AD}}^{x_0} dx_0\right), \tag{5.78}$$

$$R_{3,n} = 2V_n\left(\int_0^1 \Psi_{\mathrm{AD}}^{x_0} dx_0\right) - 4V_n\left(\int_0^1 \Psi_{\mathrm{CvM}}^{x_0} dx_0\right) + V_n\left(\int_0^1 \Omega^{x_0} dx_0\right) \tag{5.79}$$

$$R_{4,n} = 3V_n\left(\int_0^1 \Psi_{\mathrm{AD}}^{x_0} dx_0\right) - 10.5V_n\left(\int_0^1 \Psi_{\mathrm{CvM}}^{x_0} dx_0\right) \tag{5.80}$$

$$+ 3V_n\left(\int_0^1 \Omega^{x_0} dx_0\right) + 1.5V_n\left(\int_0^1 \Xi^{x_0} dx_0\right), \tag{5.81}$$

so that the derivation of the computational formulae is reduced to integrating out x_0 in the kernels of A_n, W_n and K_n. Hence, the origin-invariant versions of the SSP2, SSP3 and SSP4 statistics reduce to

$$R_{2,n} = V_n(\Psi_{\mathrm{AD}}^{\mathrm{oi}}), \tag{5.82}$$

$$R_{3,n} = 2V_n(\Psi_{\mathrm{AD}}^{\mathrm{oi}}) - 4V_n(\Psi_{\mathrm{CvM}}^{\mathrm{oi}}) + V_n(\Omega^{\mathrm{oi}}) \tag{5.83}$$

$$R_{4,n} = 3V_n(\Psi_{\mathrm{AD}}^{\mathrm{oi}}) - 10.5V_n(\Psi_{\mathrm{CvM}}^{\mathrm{oi}}) + 3V_n(\Omega^{\mathrm{oi}}) + 1.5V_n(\Xi^{\mathrm{oi}}), \tag{5.84}$$

respectively. Here $\Psi_{\mathrm{AD}}^{\mathrm{oi}}$, $\Psi_{\mathrm{CvM}}^{\mathrm{oi}}$, Ω^{oi} and Ξ^{oi} are the origin-invariant analogues (obtained by integrating out the origin) of the kernels $\Psi_{\mathrm{AD}}, \Psi_{\mathrm{CvM}}$ and Ω, which are (after some simple but lengthy algebra)

$$\begin{aligned}
\Psi_{\mathrm{AD}}^{\mathrm{oi}} &= 2|x-y|\ln|x-y| + 2(1-|x-y|)\ln(1-|x-y|) + 1, \\
\Psi_{\mathrm{CvM}}^{\mathrm{oi}} &= \frac{7}{6} - xy - (1 + x \wedge y)|x-y| - (1 + x \vee y)(1 - x \vee y), \\
\Omega^{\mathrm{oi}} &= 2(1-|x-y|)\ln(1-|x-y|) - 2(x-y)^2 + 2|y-x|(\ln(|y-x|) + 1) + \frac{2}{3} \\
\Xi^{\mathrm{oi}} &= -0.5(x \vee y)^3 + 0.5(x \wedge y)^3 + x^2 + y^2 - 0.5|x-y| \\
&\quad - 0.5xy|x-y| - 0.5(x+y)^2 + 0.5(x+y)((x \vee y)^2 - (x \wedge y)^2) + \frac{5}{60}.
\end{aligned}$$

5.8.4 Data-driven Test

Similarly as for the linear SSPc tests, a data-driven version of the circular SSPc test can be considered. The choice of the partition size is given by (5.58) where $T_{c,n}$ is replaced by $R_{c,n}$, i.e.

$$C_n = \mathrm{ArgMax}_{c \in \Gamma}\{R_{c,n} - 2(c-1)\ln a_n\}. \tag{5.85}$$

The data-driven test statistic is defined as $R_{C_n,n}$. We consider the same choices for the penalty term a_n in (5.85) as in the linear case and refer to the corresponding data-driven versions as CSSP-AIC, CSSP-BIC and CSSP-LL. We further propose to take the set of permissible SSP sizes as in the linear case, i.e. $\Gamma = \{2, 3, 4\}$. We refer to the simulation study in Section 5.8.6 below, in which it is demonstrated that this choice of Γ yields good powers.

Similarly as in the linear case, we can find the asymptotic limiting distribution of the data-driven test statistic $R_{C_n,n}$. We state the results without proof since the proofs are analogous to those of Theorem 5.7 and 5.8.

Theorem 5.12 *Let c_m denote the minimal CSSP size, i.e. $c_m = \min_c \Gamma$. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0,*

$$P[C_n = c_m] \to 1$$

as $n \to \infty$.

Theorem 5.13 *Let $c_m = \min_c \Gamma$. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, the asymptotic null distribution of $R_{C_n,n}$ is given by*

$$\begin{aligned} R_{c,n} \xrightarrow{d} R_{c,\infty} &= (c_m - 1)\int_0^1 \int_{x_0}^{1+x_0} a_{c_m}(x; x_0)\mathbb{B}^2(x; x_0)dxdx_0 \\ &+ \binom{c_m - 1}{2} \int_0^1 \int_{x_0}^{1+x_0} \int_{x_0}^{1+x_0} w_{c_m}(x, y; x_0)(\mathbb{B}(x; x_0) - \mathbb{B}(y; x_0))^2 dxdydx_0. \end{aligned}$$

where $\{\mathbb{B}(x; x_0), x \in arc[x_0, x_0 + 1]\}$ is a Brownian bridge which starts at x_0 and ends at $x_0 + 1$ with mean zero and covariance function given by

$$Cov[\mathbb{B}(x; x_0), \mathbb{B}(y; x_0)] = F_0(x \wedge y; x_0) - F_0(x; x_0)F_0(y; x_0). \qquad (5.86)$$

By the omnibus consistency of the CSSPc test for every $c \in \Gamma$, the omnibus consistency of the data-driven test based on $R_{C_n,n}$ follows immediately.

5.8.5 Examples

In this section, we briefly present the results for the origin-invariant SSP*c* tests for $c = 2$ and $c = 3$ applied on two examples. We also perform the three data-driven versions of the circular SSP*c* test. The p-values are computed using 10,000 bootstrap simulations.

Birth time data

Recall that for the Birth time data, no evidence against uniformity had been found based on the Kuiper and Watson test in Example 3.5.2 and based on the smooth test with $k = 2$ in Example 3.3.5. The first column of Table 5.5 shows the p-values of the classical origin-invariant GOF tests. For explanation of

Table 5.5: The p-values of the origin-invariant GOF tests for the Birth time (first column) and the Homing pigeons data (second column).

GOF test	Birth time	Homing pigeons
Kuiper	0.508	0.170
Watson	0.632	0.128
Rayleigh	0.255	0.571
Bogdan	0.283	0.001
Rao	0.975	0.091
Anje	0.288	0.654
CSSP-AIC	0.451	0.002
CSSP-BIC	0.437	0.001
CSSP-LL	0.439	0.106

these classical tests, we refer to Chapter 3. The p-values obtained for the in fact inappropriate SSP-AIC, SSP-BIC and SSP-LL tests were obtained in Section 5.5.3 as all equal to 0.686, where each variant selected $c = 2$. The individual CSSP2 test has a p-value of 0.442, while CSSP3 and CSSP4 results in slightly larger p-values (0.493 and 0.526, respectively). The data-driven versions CSSP-AIC, CSSP-BIC and CSSP-LL are non-significant as well (p=0.451, p=0.437 and p=0.439, respectively) and each of the corresponding selection criteria chooses $c = 2$. These results are given to see the relatively large difference in p-values between the linear and the circular versions. In other situations the results of the linear and circular tests may lead to different conclusions. Once more we stress that the circular version is applicable to both types of data, while the linear versions are not appropriate to circular data.

Homing pigeons data

For the bimodal Homing pigeons data introduced in Chapter 2, the second column of Table 5.5 lists the p-values for the origin-invariant GOF tests for uniformity. The Rayleigh and Anje test have the highest p-values, indicating no significant difference from uniformity. Recall that the Rayleigh test is in fact the first component of the smooth test for circular uniformity introduced by Bogdan et al. (2002). The high p-value for the Rayleigh test is due to the fact that the test is only sensitive for unimodal alternatives. Anje's test statistic is the integral version of the Hodges-Anje which has a non-significant result as well (p=0.873, see Example 1 in Section 3.2.2). The classical statistics of the EDF type (Watson and the Kuiper test) are also non-significant, although the corresponding p-values are much lower than those of the Rayleigh and Anje tests. The data-driven smooth test of Bogdan is the only GOF test that recognises the difference from circular uniformity at the 5% level of significance. Moreover, this test informs us about the bimodality, since the BIC selection criterion chooses order $K = 2$. The circular data-driven SSP tests CSSP-AIC and CSSP-BIC also have significant p-values (p=0.002 and p=0.001). Both selection criteria choose SSP size $c = 2$. This choice gives us the extra information that the significance is due to a global LOF. On the other hand, the individual CSSP2 test has a non-significant p-value (0.092), while the p-values for the CSSP3 and CSSP4 are much larger (0.398 and 0.313, respectively). The data-driven version CSSP-LL, which selected $c = 2$, is non-significant as well (p=0.106). This is probably because the sample size (n=13) is too small. We may conclude that our new CSSP test is useful to detect bimodalty. Moreover, it gives the additional information that the LOF is located over the whole circumference.

5.8.6 Simulation Study

In this section we present a small power study to investigate the small sample properties of the CSSPc test. Similarly as for the linear SSPc test, we give results of a Monte Carlo study in which we investigate the power of the CSSPc tests ($c = 2, 3, 4$) and their three data-driven versions ($\Gamma = \{2, 3, 4\}$) for sample sizes 20 and 50. The computational formulae of Section 5.8.3 are used. We compare the new tests with classical GOF tests described in Chapter 3. In particular, we consider the Watson test, the Kuiper test, the Rao spacing test, the Rayleigh test, the data-driven smooth test and the Anje test. We perform simulations under the same alternatives as for the linear SSPc test, but we replaced the normal distribution on the real line with the CN distribution on the circle with unit circumference. The hypothesised distribution is the CN distribution with parameters $\mu = 0$ and $\kappa = 1$. The first type of alternative considered is a CN distribution where either the location parameter μ is changed while keeping $\kappa = 1$, or the concentration parameter κ is changed while fixing $\mu = 0$. For the former series, we vary μ between 0 and 0.5. Note that because of the origin-invariance properties of the test statistics, the powers should be the same for μ varying between 1 and 0.5. In the latter series, κ is varied from 1 to 2.2. Note that larger values for the concentration parameter induce a more peaked density at $\mu = 0$.

As the CSSPc test is the origin-invariant version of the linear SSPc test, we expect the test to be more sensitive to deviations from the null hypothesis in small intervals. Such alternatives are represented in this study by the circular analogues of the contaminated normals in (5.59), given by

$$f_{\delta,\gamma}(x) = (1 - \gamma) f_{CN}(x; 0, 1) + \gamma f_{CN}(x; \delta, 100),$$

where f_{CN} is the density of the CN distribution. This family of mixture distributions reduces to the hypothesised distribution if $\gamma = 0$. Note that the concentration parameter κ of the second component in the mixture is very large, which implies a second mode highly concentrated at δ. We refer to this "localised "alternative as the contaminated CN distribution.

We have performed 1,000 Monte Carlo simulations to estimate the powers. The tests are performed at the 5% significance level and we again use simulated critical points (based on 50,000 simulation runs). The results are shown in Figures 5.14 and 5.16 for the CN and the contaminated CN distribution, respectively. Since the power curves are sometimes too close to one another to differentiate them, we also provide plots where we zoom in on the curves. These plots are in Figures 5.15 and 5.17, for the CN and the contaminated CN distribution respectively. Note that, the power curves for the data-driven CSSP-AIC tests are very close to the curves of the CSSP4, while those for the CSSP-BIC and the CSSP-LL are close to the power curves of the CSSP2.

For CN alternatives with constant concentration, in panels (a) and (b) of Figure 5.14 we see that all tests, except the data-driven smooth test of Bogdan and the Rao spacings test, have similar powers. The latter two classical tests have lower power, with the Rao spacings test as the worst. From the corresponding detail plots in Figure 5.15, we see that our CSSP tests, represented by the solid curves, have in fact the largest powers. Recall that for the linear SSPc tests, a small loss in power was seen when the SSP size is increased from $c = 2$ to 3 and 4. That power loss is also present here, although it is much smaller than in the linear case.

When μ is fixed at zero, in panels (c) and (d), the opposite is observed as for the linear SSPc test, the power increases with the increasing SSP size, although only slightly. Note that the Anje and the Rayleigh tests have now slightly larger power than the CSSP tests. The choice of the SSP size depends on the criterion used. If the AIC criterion is used, the power is almost indistinguishable from the CSSP4 test. On the other hand, when one of the other two criteria is used, the power curves approach more the curve of the CSSP2 test.

As in the linear case, we have shifted the simulated data from the contaminated CN alternatives by subtracting the true mean, which is again equal to $\delta\gamma$. This correction is made so as to investigate the powers for localised alternatives where the change in location is less important.

The results of the simulation study for the contaminated CN distribution are presented in Figure 5.16 for $\delta = 1$ (top), $\delta = 1.5$ (middle) and $\delta = 2$ (bottom). Corresponding plots where we zoomed in on the middle part of the curves, are in Figure 5.17. From all these plots, we conclude that the various versions of the CSSP test outperform all other tests. The detail plots show that the power of the CSSPc test increases as the SSP size increases. The data-driven CSSP-AIC test succeeds well in selecting an appropriate value for c. In fact, its behaviour is almost exactly equal to that of the CSSP4 test.

5.8.7 Circular SSPc test for composite null hypothesis

Similarly as for the linear SSPc test, the circular SSPc test can be used for composite null hypotheses. Let $\hat{R}_{c,n}$ denote the statistic given by (5.71) where the empirical process $\mathbb{B}_n(x;x_0)$ is replaced by $\hat{\mathbb{B}}_n(x;x_0) = \mathbb{B}_n(x;x_0,\hat{\boldsymbol{\beta}})$. Then the asymptotic null distribution is stated here without proof.

Proposition 5.1 *Let* $\{\hat{\mathbb{B}}(x;x_0), x \in arc[x_0, x_0+1]\}$ *denote a Gaussian process which starts at* x_0 *and ends in* x_0+1 *and has covariance function as in (5.63) where* $F_0(x,\boldsymbol{\beta})$ *and* $\boldsymbol{\Lambda}(x,\boldsymbol{\beta})$ *are replaced by* $F_0(x;x_0,\boldsymbol{\beta}) = \int_{x_0}^{x} f_0(y,\boldsymbol{\beta})dy$ *and* $\boldsymbol{\Lambda}(x;x_0,\boldsymbol{\beta}) = \int_{x_0}^{x} \boldsymbol{\Upsilon}(z;x_0,\boldsymbol{\beta})dF_0(z;x_0,\boldsymbol{\beta})$. *Suppose* $c \geq 2$ *is given, then, under the null hypothesis, as* $n \to \infty$,

$$\hat{R}_{c,n} \xrightarrow{d} \hat{R}_{c,\infty} = (c-1)\int_0^1 \int_{x_0}^{1+x_0} a_c(x;x_0)\hat{\mathbb{B}}^2(x;x_0)dxdx_0 + \binom{c-1}{2}\int_0^1 \int_{x_0}^{1+x_0}\int_{x_0}^{1+x_0} w_c(x,y;x_0)(\hat{\mathbb{B}}(x;x_0) - \hat{\mathbb{B}}(y;x_0))^2 dxdydx_0. \quad (5.87)$$

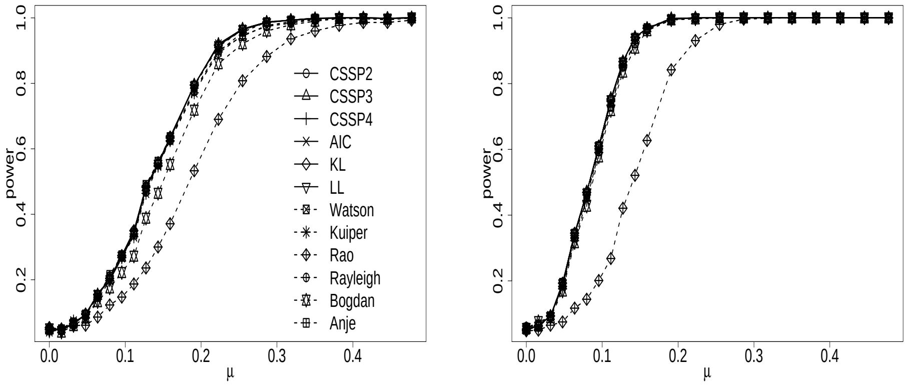
Also for the circular SSPc test statistic for composite distributions, we use parametric bootstrap to approximate the null distribution.

5.9 Discussion

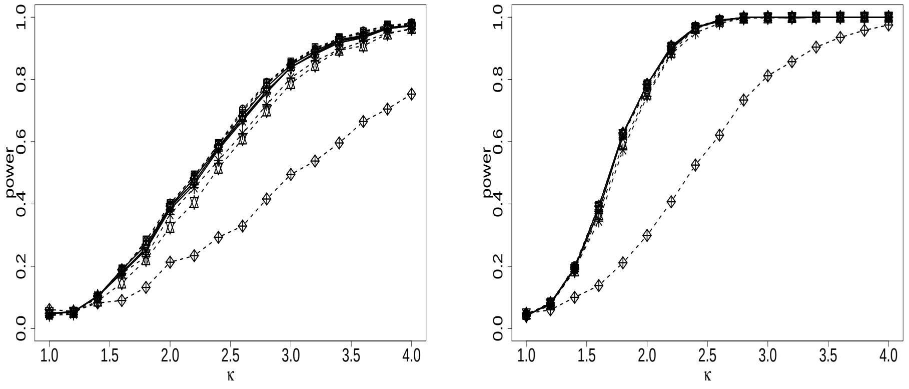
In this chapter we have first presented a new class of GOF tests for the simple one-sample problem for linear data. The test statistic is constructed as an average of localised Pearson χ^2-statistics with arbitrary degrees of freedom. The degrees of freedom of the Pearson's statistics are directly related to the indexing parameter (SSP size c) of our new class. The tests are generalisations of the AD test, which is included in the class by taking $c = 2$. The methodology presented in this chapter may be used to obtain similar extensions to the tests proposed by Zhang (2002) and Einmahl and McKeague (2003).

The simulation study that we have presented, indicates that a substantial power gain may result from choosing some $c > 2$. On the other hand, the study also showed that for some alternatives, the highest power is obtained with $c = 2$. To avoid the problem of choosing the right value for the indexing parameter c, we have proposed a data-driven version of the test. The simulations confirmed that the selection rule succeeds quite well in selecting a good choice for c.

The weight functions that are involved in the test statistic, as well as the simulation results, suggest that the new tests are very sensitive to deviations

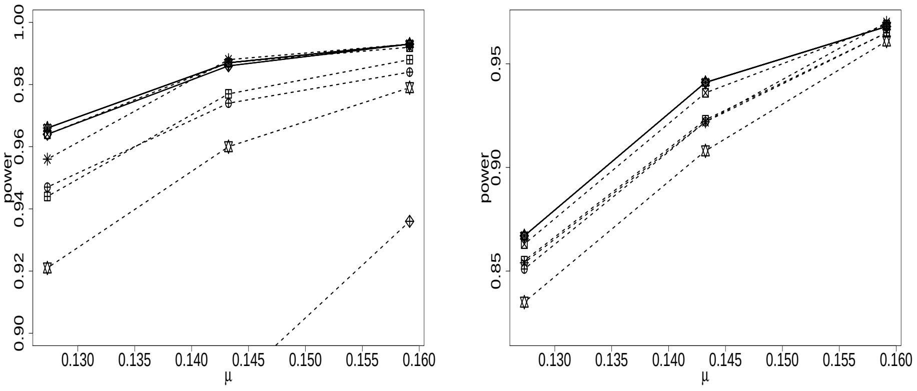


(a) $n = 20$, $\kappa = 1$ (b) $n = 50$, $\kappa = 1$

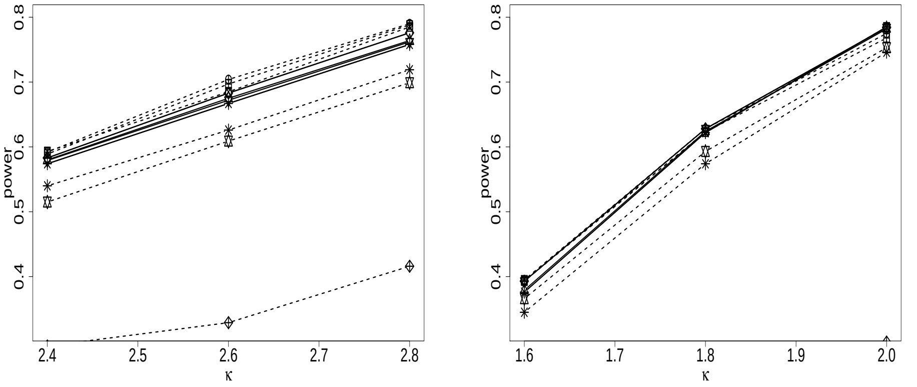


(c) $n = 20$, $\mu = 0$ (d) $n = 50$, $\mu = 0$

Figure 5.14: Estimated power curves for the CN alternative. The legend is only shown in the first plot.

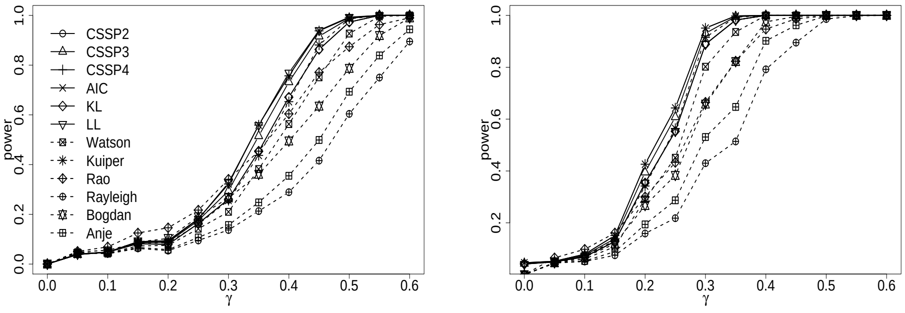


(a) $n = 20$, $\kappa = 1$ (b) $n = 50$, $\kappa = 1$



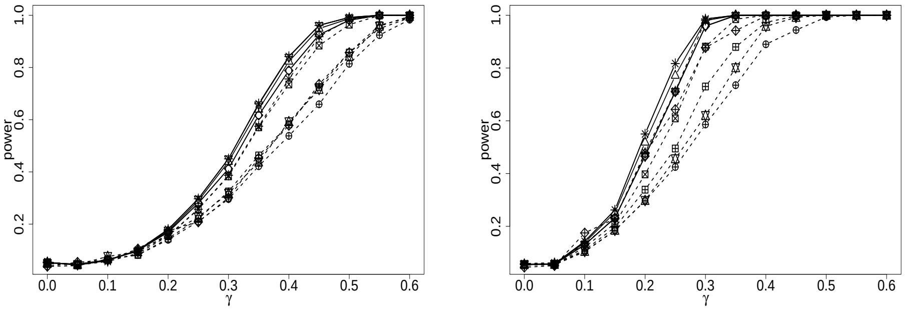
(c) $n = 20$, $\mu = 0$ (d) $n = 50$, $\mu = 0$

Figure 5.15: Estimated power curves for the CN alternative. The legend is the same as in Figure 5.14.



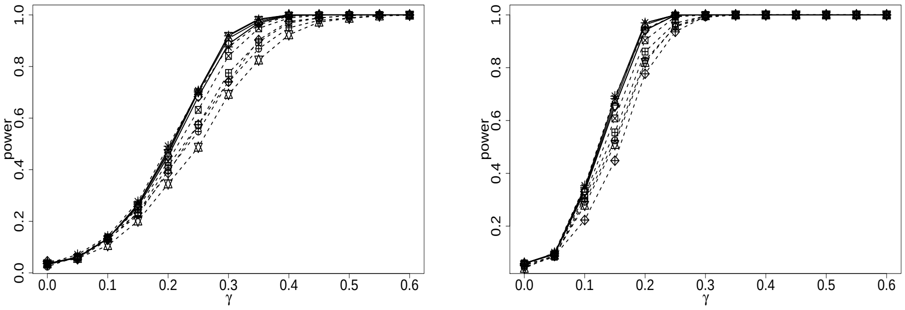
(a) $n = 20$, $\delta = 0.15$

(b) $n = 50$, $\delta = 0.15$



(c) $n = 20$, $\delta = 0.3$

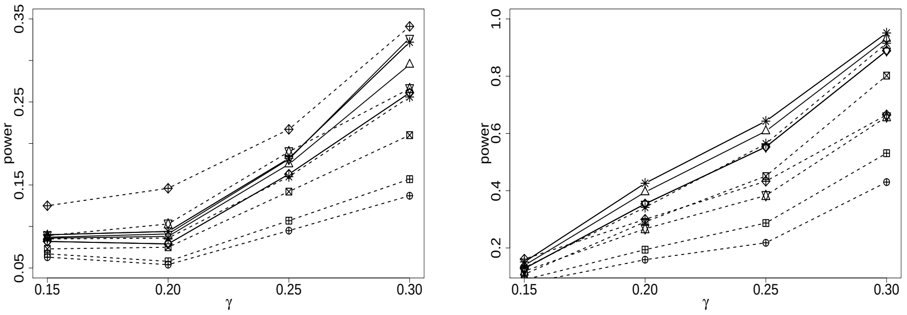
(d) $n = 50$, $\delta = 0.3$



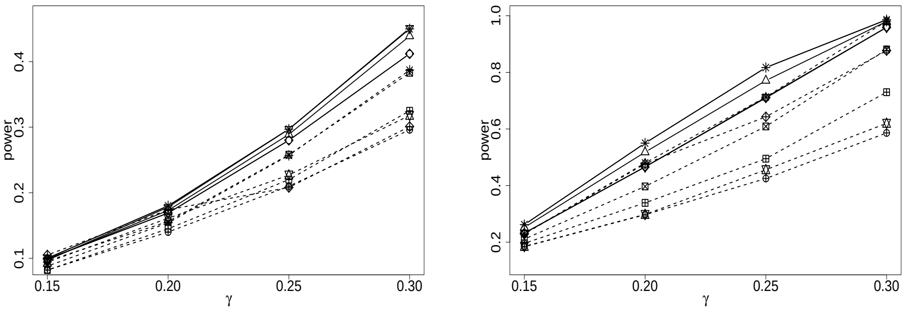
(e) $n = 20$, $\delta = 0.5$

(f) $n = 50$, $\delta = 0.5$

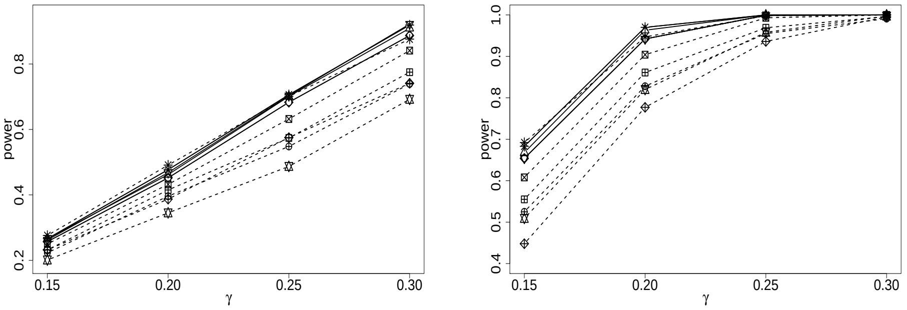
Figure 5.16: Estimated power curves for the contaminated CN distribution alternative. The legend is only shown in the first plot.



(a) $n = 20$, $\delta = 0.15$ (b) $n = 50$, $\delta = 0.15$



(c) $n = 20$, $\delta = 0.3$ (d) $n = 50$, $\delta = 0.3$



(e) $n = 20$, $\delta = 0.5$ (f) $n = 50$, $\delta = 0.5$

Figure 5.17: Estimated power curves for the contaminated CN distribution alternative. The legend is the same as in Figure 5.16.

from the hypothesised distribution F_0 in small intervals of the support of F_0. Furthermore, this sensitivity increases with the increasing SSP size c.

We have also presented the asymptotic null distributions of the new GOF tests, and we proved that all tests are omnibus consistent.

In the special case of SSP size $c = 3$, we have written the statistic as a V-statistic so as to find an appropriate decomposition in terms of Legendre polynomials. This decomposition led to the limiting distribution of the statistic under contiguous alternatives.

Furthermore, we have extended the use of the new class of GOF tests to composite null hypotheses using the estimated empirical process. If the nuisance parameters are estimated by asymptotically linear estimators, this estimated empirical process converges weakly to a Gaussian process with known covariance function. However, that limiting Gaussian process is difficult to compute and to simulate from. Therefore, we suggest using the parametric bootstrap to obtain the approximate null distribution.

Finally, the new class of GOF tests for linear data has been extended to a new class of GOF tests for circular data. This has been done by making the class of statistics origin-invariant. We simply integrated out all possible origins to obtain the origin-invariant class of statistic. The limiting null distribution was given and computational formulae for SSP size $c = 2, 3$ and 4 were found. The methodology to obtain those formulae can easily be extended to any SSP size. The data-driven version and its asymptotic theory are similar as in the linear case. A simulation study has indicated that the circular SSP tests have the same power characteristics as the linear SSPc tests, although, the differences between their powers are less pronounced than in the linear case. All methods presented in this chapter are extensively applied to real data examples.

CHAPTER 6

The Interval-based PP-plot

This chapter is devoted to a new graphical diagnostic tool for the detection of LOF for circular distributions. Observations measured on the circle are characterised by the invariance to the choice of the origin, and the distance between observations is given by the smallest arc on the circle. Circular data thus differ substantially from linear data. These important differences motivate the search for statistical tools for which the conclusions do not depend on the chosen origin nor on the measurement direction. Classical GOF tests for circular distributions, such as e.g. the Kuiper test, were described in Chapter 3. In each of the Chapters 4 and 5, we have proposed a new class of circular GOF tests. When such a circular GOF test results in rejecting the null hypothesis, it is important to know in what way the true distribution deviates from the hypothesised. Our graphical tool enables us to localise those regions of LOF. Moreover, the graph is constructed from the empirical process which is also the basis for the origin-invariant Kuiper test statistic. This relation to the Kuiper test is particularly useful for constructing a formal version of the graph, so that conclusions for the Kuiper test can immediately be read from the graph.

The construction of the new plot is given in Section 6.1. We will argue that the location of the LOF is easily recognised by the new plot. However, the plot in its original form does not reveal significance. Therefore a formal version of the plot, which contributes to the need for an objective conclusion, is proposed in Section 6.2. Characteristics of the new plot are investigated through the limiting expected value of the process under a particular family of alternatives in Section 6.3. The tools developed in this chapter are illustrated on real data examples in Section 6.4. Since these plots are also informative graphical tools for linear data, we include two linear data examples as well. A small simulation study in Section 6.5 confirms what we intuitively expected from the applications in Section 6.4 and from the limiting expected value of the considered process in Section 6.3. In Section 6.6, we suggest using the new plot as a diagnostic tool to evaluate the appropriateness of a nonparametric density estimator as described in Sections 3.7.1 and 4.5. Finally, in Section 6.7, a brief discussion is given.

6.1 Construction of the IBPP-plot

In this section two types of diagnostic plots for GOF of circular data are proposed. These plots aim at "localising" the deviation from the hypothesised distribution in the sense that they visualise in which arc(s) on the circle the true and the hypothesised distribution are different. Although these plots are also informative graphical tools for linear data, we restrict the construction of these plots to circular data. Later, in Section 6.4, we return to the Old Faithful geyser data example to illustrate that the tool is also useful to localise *bumps* in a linear distribution. The term *bump* is used to indicate a location where a small cluster of observations is found which is unusual according to the null hypothesis. Moreover, in Section 6.5 we investigate the behaviour of the new plot under linear local alternatives through a simulation study.

Suppose X is a random variable on the circle with unit circumference, i.e. $\mathcal{S}$ =arc$(0, 1)$. Note that we take the origin at 0, but, as we will see soon, the construction of the new plot is invariant to that choice. The PP-plot is one of the most popular graphical tools used for diagnosing whether the true distribution $F(x)$ underlying the data agrees with a hypothesised distribution $F_0(x, \boldsymbol{\beta})$. Here $\boldsymbol{\beta}$ is a p-dimensional nuisance parameter, which is initially assumed to be known. As mentioned in Section 2.1.1, the detrended PP-plot, which is closely related to the PP-plot, is also useful to assess GOF for linear data. Moreover, in the same section we have explained how the KS test is related to the detrended PP-plot. However, the empirical process $\mathbb{B}_n(x) = \sqrt{n}(\hat{F}_n(x) - F_0(x, \boldsymbol{\beta}))$, which is the basis of that plot, is not origin-invariant. Therefore, the detrended PP-plot is not origin-invariant either and conclusions from graphical assessment for circular data could be different if another origin or rotation direction was chosen. This means that the plot is not fully appropriate for circular data. This has already been demonstrated on the Birth time data in Section 2.2.1.

As the detrended PP-plot is related to the KS statistic, we propose to construct an origin-invariant PP-plot based on the origin-invariant Kuiper statistic.

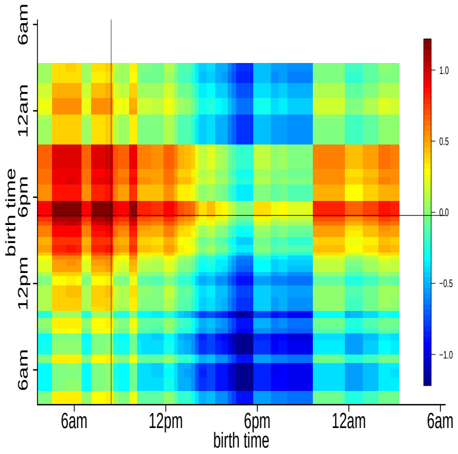


Figure 6.1: IBPP-plot for the Birth time data.

In particular, we extend the PP-plot to an origin-invariant PP-plot by using the interval indexed process

$$\mathbb{Z}_n(x, y) = \mathbb{B}_n(x) - \mathbb{B}_n(y). \tag{6.1}$$

The supremum of the absolute value of that process over all possible intervals results in the Kuiper test statistic, which is defined in (3.50) and can be interpreted as the circular analogue of the KS test statistic. For each arc(x, y) the corresponding value for the process $\mathbb{Z}_n(x, y)$ can be interpreted as the difference between the empirical probability and the expected probability of the random variable X falling into the arc(x, y). This interpretation is also mentioned in Section 3.4.2 and is obvious from the alternative expression of the process, i.e.

$$\mathbb{Z}_n(x, y) = \sqrt{n}\{\hat{F}_n(x, y) - F_0(x, y)\},$$

where $\hat{F}_n(x, y) = \hat{F}_n(x) - \hat{F}_n(y)$ and $F_0(x, y) = F_0(x, \boldsymbol{\beta}) - F_0(y, \boldsymbol{\beta})$. From this argument it is intuitively clear that for a particular arc(x, y), the larger the process value, the more the true distribution deviates from the null distribution within that arc. We suggest to plot the process values $\mathbb{Z}_n(x, y)$ versus x and y as a heat map, using a colour legend. The resulting plot is called the interval-based PP-plot (IBPP-plot) for which the name refers to the interval-based interpretation of the process. The IBPP-plot for the Birth time data is drawn in Figure 6.1. The horizontal and vertical axes of the heat map correspond to the beginning and end points of the arcs, respectively. Red and blue regions indicate large positive and large negative process values, respectively. On the other hand, yellow and green regions indicate values of $\mathbb{Z}_n$ which are small in absolute value. As a convention we plot the largest value of the process in the middle of the horizontal or the vertical axis. Additionally, the beginning and end points of $\mathrm{arc}(x, y)$ with $(x < y)$ where the absolute maximum process value is observed, are indicated with a vertical and a horizontal line, respectively. Note that the IBPP-plot is *antisymmetric*. Indeed, the process values $\mathbb{Z}_n(x, y)$ for which x is smaller than y are opposite in sign to the process values $\mathbb{Z}_n(y, x)$. Hence, we only have to look at the half plane above the line through the origin with slope one, which is referred to as the *diagonal* of the IBPP-plot. From Figure 6.1, we see that for the Birth time data the maximum process value $\mathbb{Z}_n$ is obtained in about arc(8.30am,5pm). The red region around that maximum process value indicates that process values for intervals on the circle including arc(8.30am,5pm) are large as well. This could possibly indicate that in these arcs relatively more births occur than expected under uniformity. The smaller blue region observed for intervals on the circle including arc(6pm,12am) indicate that for these regions probably less births are occur than expected under uniformity. It is now of our interest whether the results for this graph are significant. To obtain that information we construct a formal version of the IBPP-plot, which will be described in the next section.

We propose another type of diagnostic plot for GOF of circular data, which is closely related to the IBPP-plot. Therefore, consider the process

$$\mathbb{Y}_n(x) = \sup_y |\mathbb{B}_n(x) - \mathbb{B}_n(y)|, \tag{6.2}$$

which is in fact the supremum of the process $\mathbb{Z}_n(x, y)$ over one of the end points of the $\mathrm{arc}(x, y)$. We suggest to plot the process values $\mathbb{Y}_n(x)$ on a circle and we call it the *circular* PP-plot (CiPP-plot). Figure 6.2 shows the CiPP-plot for the Birth time data. The maximum of the absolute value of the process is indicated by a straight line through $x_{\max}$ and the origin, where $x_{\max}$ is the direction for which the process in 6.2 or equivalently the process in 6.1 obtains its maximum value. Note that two such lines through the origin are drawn since the maximum process value is obtained twice. Moreover, these two lines basically correspond to the beginning and end points of the $\mathrm{arc}(x, y)$ where the maximum process value of $\mathbb{Z}_n$ is obtained. A full line circle is drawn, with radius equal to that maximum process value. The interpretation of the CiPP-plot is similar to that of the IBPP-plot, but part of the information about the interval is lost since the supremum of the absolute values is taken over one of the dimensions. For the Birth time data relatively large and small process values are obtained for acrs with end points within arc(8.30am,5pm) and arc(6pm,5am), respectively.

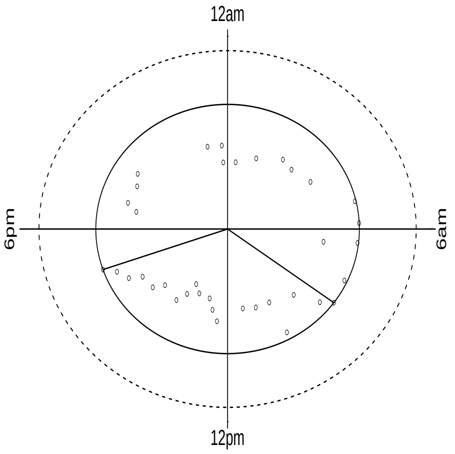


Figure 6.2: CiBPP-plot for the Birth time data.

6.2 The formal IBPP-plot

From the plots described in the previous section, interesting information regarding the GOF problem can be derived. This information is subjective, however. How large should a process value $\mathbb{Z}_n(x, y)$ be to conclude that there is a LOF in the arc(x, y)? We suggest two solutions. The first is based on the pointwise asymptotic null distribution of $\mathbb{Z}_n(x, y)$. Suppose that F_0 is completely specified and that we performed the PIT such that the GOF problem is reduced to a uniform null hypothesis on $[0, 1]$. Using the central limit theorem (CLT) we have that, under the null hypothesis,

$$\mathbb{Z}_n(x, y) \xrightarrow{d} N(0, |F_0(x) - F_0(y)|(1 - |F_0(x) - F_0(y)|))$$

for every $(x, y) \in \text{arc}(0, 1)^2 \backslash \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Hence, for each (x, y) the statistic in (6.1) can be compared with the e.g. 5% critical value of the corresponding null distribution. A disadvantage of this approach is that on the IBPP-plot many of these process values are plotted and interpreted simultaneously, and therefore a multiplicity problem arises. Our second solution overcomes this problem by taking the supremum of the process values over all possible arcs (x, y), which results in the Kuiper test statistic as defined in (3.50). Using the asymptotic theory of Shorack and Wellner (1986), we know that the asymptotic null distribution of the Kuiper test statistic is given by

$$\sup_{0 \le x, y \le 1} |\mathbb{Z}_n(x, y)| = \sup_{0 \le x, y \le 1} |\mathbb{B}_n(x) - \mathbb{B}_n(y)| \xrightarrow{d} \sup_{0 \le x, y \le 1} |\mathbb{B}(x) - \mathbb{B}(y)|, \quad (6.3)$$

where $\mathbb{B}$ is a Brownian bridge. Since a Brownian bridge $\mathbb{B}(x)$ has mean function zero and covariance function $\mathrm{Cov}\left[\mathbb{B}(x), \mathbb{B}(y)\right] = F_0(x) \wedge F_0(y) - F_0(x)F_0(y)$, the process $\mathbb{B}(x) - \mathbb{B}(y)$ has mean zero too, and variance function

$$\begin{aligned}\mathrm{var}(\mathbb{B}(x) - \mathbb{B}(y)) &= F_0(x) - F_0(x)^2 - 2(F_0(x) \wedge F_0(y) \\ &\quad -F_0(x)F_0(y)) + F_0(y) - F_0(y)^2 \\ &= |F_0(x) - F_0(y)|(1 - |F_0(x) - F_0(y)|).\end{aligned}$$

If the p-dimensional parameter vector $\boldsymbol{\beta}$ in $F_0(x, \boldsymbol{\beta})$ is unknown, we suggest replacing it by its MLE $\hat{\boldsymbol{\beta}}$ and the empirical process $\mathbb{B}_n(x)$ is then replaced by the estimated empirical processes $\hat{\mathbb{B}}_n(x)$ defined in (5.61). The corresponding estimated versions of the proces $\mathbb{Z}_n$ is similarly denoted by $\hat{\mathbb{Z}}_n$. The asymptotic distribution of the Kuiper statistic is then given by

$$\sup_{0 \leq x,y \leq 1} |\hat{\mathbb{Z}}_n(x, y)| = \sup_{0 \leq x,y \leq 1} |\hat{\mathbb{B}}_n(x) - \hat{\mathbb{B}}_n(y)| \xrightarrow{d} \sup_{0 \leq x,y \leq 1} |\hat{\mathbb{B}}(x) - \hat{\mathbb{B}}(y)|, \quad (6.4)$$

where $\hat{\mathbb{B}}$ is given in Theorem 5.10. Hence, when an unknown parameter vector $\boldsymbol{\beta}$ is present in F_0, the variance function $\mathrm{var}(\hat{\mathbb{B}}(x) - \hat{\mathbb{B}}(y))$ becomes more complicated. In particular, for a locally asymptotically linear estimator of $\boldsymbol{\beta}$ the $\mathrm{Cov}\left[\hat{\mathbb{B}}(x), \hat{\mathbb{B}}(y)\right]$ is given in (5.63) of Theorem 5.10. Since the computation of such a covariance function is usually difficult, the critical values for the Kuiper test for composite null hypotheses are obtained using the parametric bootstrap.

A formal version of the plot is then constructed by indicating only the process values that exceed the α-level critical value of the corresponding Kuiper test. In this way the new plot is directly linked to a formal statistical test, as is the case with the detrended PP-plot and the KS test.

We refer to this plot as the *formal IBPP-plot*, whereas the first version of the graph, which is still informative but which does not reveal significance, is referred to as the *exploratory IBPP-plot*. On the formal version of the IBPP-plot, black and white regions indicate significant and non-significant process values, respectively. As the Kuiper test is origin and rotation invariant, the construction of the two plots is guaranteed to be invariant too.

Since for each $\mathrm{arc}(x, y)$ the process value $\mathbb{Z}_n(x, y)$ has a different variance, it may be argued that the process values should be standardised for making the process values comparable among one another. Therefore, we here also consider the *standardised* IBPP-plot, which is based on the process values

$$\mathbb{Q}_n(x, y) = \frac{\mathbb{B}_n(x) - \mathbb{B}_n(y)}{\sqrt{|F_0(x) - F_0(y)|(1 - |F_0(x) - F_0(y)|)}} \quad (6.5)$$

instead of $\mathbb{Z}_n(x, y)$. Note that this standardisation is only appropriate for the simple null hypothesis case. For composite null hypotheses the variance becomes more complicated as explained above. A standardised IBPP-plot for composite null hypotheses is out of the scope of this thesis. Similarly as for the original IBPP-plot, we can use a formal version by indicating only the process values that exceed an α-level critical value. This critical value is determined by simulating the null distribution of the supremum of the standardised process $\mathbb{Q}_n(x, y)$ over all possible intervals. As the original IBPP-plot, the standardised IBPP-plot is thus also related to a formal test statistic given by

$$K_n^s = \sup_{0 \leq x,y \leq 1} |\mathbb{Q}_n(x, y)|, \tag{6.6}$$

which is in fact a weighted or standardised version of the Kuiper test statistic. Some properties of the original and standardised IBPP-plots are investigated in the next section. In particular, we investigate the limiting expected value of the corresponding processes under a family of contiguous alternatives. From that investigation we will see that the IBPP-plots are expected to localise the LOF well. A simulation study in Section 6.5 confirms these expectations.

6.3 Limiting behaviour of IBPP-plots under contiguous alternatives

In this section we illustrate the ability of the formal and exploratory IBPP-plots to locate the LOF, by displaying its expected behaviour for large sample sizes. We examine both the original and the standardised IBPP-plots. Since the (standardised) IBPP-plot is related to (a weighted version) of the Kuiper test, we use contiguous alternatives to circular uniformity for which the Kuiper test is known to be powerful. We limit the discussion to the simple null hypothesis, so that without loss of generality the null hypothesis is chosen as circular uniformity.

We expect the IBPP-plots to be able to localise in which arc(s) the true distribution deviates from the hypothesised. Moreover, we expect that the IBPP-plots excel in recognising alternatives that only deviate from the null distribution in some interval. We refer to such alternatives as “local alternatives”. To see how the IBPP-plots asymptotically behaves under such alternatives, we focus on the limiting expected values of the corresponding statistics $\mathbb{Q}_n(x, y)$ and $\mathbb{Z}_n(x, y)$. As explained in Section 5.3, the limiting expected values give an indication of the power of the test statistic.

Suppose the observations x are measured on the circle with unit circumference $\mathcal{S}$ =arc$(0, 1)$ and are generated by the family of contiguous alternatives to uniformity,

$$f_n(x) = 1 + \frac{1}{\sqrt{n}}\delta(x), \tag{6.7}$$

where $\delta(x)$ is an arbitrary non-zero drift function that satisfies $\int_0^1 \delta(x)dx = 0$, and for which $f_n(x)$ is a proper density function for all $n \geq 1$. The same distributions have been considered in Section 5.3, and each has a corresponding distribution function given by

$$F_n(x) = x + \frac{1}{\sqrt{n}}\Delta(x), \tag{6.8}$$

where $\Delta(x) = \int_0^x \delta(u)du$. From Section 5.3 we know that, under this family of alternatives the empirical process $\mathbb{B}_n(x) = \sqrt{n}(\hat{F}_n(x) - x)$ converges weakly to $\mathbb{B}(x) + \int_0^x \delta(u)du$, where $\mathbb{B}(x)$ is a Brownian Bridge. These results, which are in fact from Janssen (1995), are also valid for data on the circle with unit circumference. In the context of the IBPP-plot we therefore obtain that the process $\mathbb{B}_n(x) - \mathbb{B}_n(y)$ converges weakly under the family of alternatives (6.7) to $\mathbb{B}(x) - \mathbb{B}(y) + \int_y^x \delta(u)du$. Since the mean of a Brownian bridge $\mathbb{B}(x)$ is zero, the expected value of the standardised process $\mathbb{Q}_n(x, y)$ in (6.5) under the family of contiguous alternatives equals

$$Q(x, y) = \frac{\int_y^x \delta(u)du}{\sqrt{|x - y|(1 - |x - y|)}} = \frac{\Delta(x) - \Delta(y)}{\sqrt{|x - y|(1 - |x - y|)}}. \tag{6.9}$$

Analogously, the limiting expected value of the process $\mathbb{Z}_n(x, y)$ in (6.1) equals

$$Z(x, y) = \int_y^x \delta(u)du = \Delta(x) - \Delta(y). \tag{6.10}$$

Similarly as in Section 5.3, we look for which functions $\Delta(x)$ asymptotic means $Q(x, y)$ or $Z(x, y)$ are large, or in other words for which alternatives the IBPP-plot is expected to locate the LOF. Let us suppose the data come from the family of contiguous alternatives (6.7) using the Mexican hat wavelet as drift function $\delta(x)$, i.e.

$$\delta(x) = C\left(1 - \frac{(x - \mu)^2}{\sigma^2}\right)e^{-0.5\frac{(x-\mu)^2}{\sigma^2}}, \tag{6.11}$$

where μ, σ and C are the location, the scale and the amplitude of the wave function, respectively. This particular wavelet is chosen because it can generate very local deviations from uniformity at various locations. Note that the range of the constant C is limited in order to have a valid distribution function $F_n(x)$ in (6.8). Figure 6.3 shows this wavelet function for different values of the parameters μ, σ and C, while Figure 6.4 shows the corresponding distribution functions (6.8) with $n = 1$. Note that this wave function is similar to the functions δ in (5.57), although the functions δ considered here are more smooth. To obtain densities with more than one local deviation from uniformity we can take a sum of different wavelets. For example, suppose the data come from the family of contiguous alternatives to uniformity (6.7) using a sum of two wavelets as drift function $\delta(x)$, i.e.

$$\delta(x) = C_1 \left(1 - \frac{(x-\mu_1)^2}{\sigma_1^2}\right) e^{-0.5\frac{(x-\mu_1)^2}{\sigma_1^2}} + C_2 \left(1 - \frac{(x-\mu_2)^2}{\sigma_2^2}\right) e^{-0.5\frac{(x-\mu_2)^2}{\sigma_2^2}}. \tag{6.12}$$

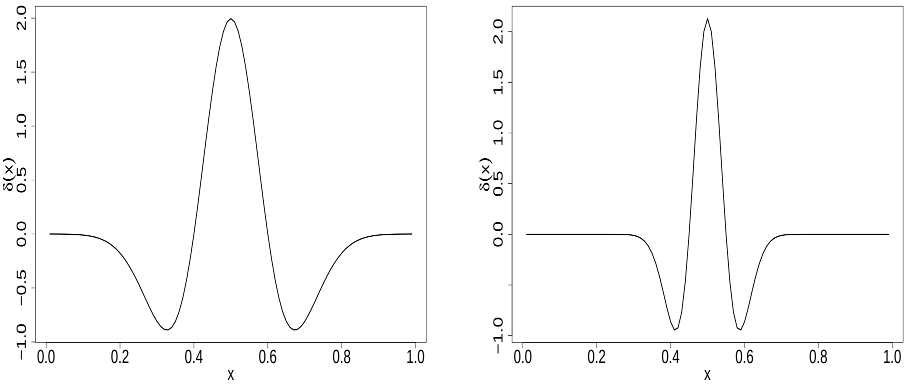
In panel (a) of Figure 6.7 we plotted such a drift function.

Figure 6.5 shows for which $\mathrm{arc}(x, y)$ the standardised process $\mathbb{Q}_n$ in (6.5) has large limiting values assuming the data come from $f_n(x)$ in (6.7) with $\delta(x)$ as in (6.11). In particular, we use heat maps with a colour legend to show $Q(x, y)$ in (6.9) as a function x and y for the function δ equal to each of the Mexican wavelet functions plotted in Figure 6.3. The same is done for the original process $\mathbb{Z}_n$ for which the plots are in Figure 6.6. Note that the limiting behaviours of both processes are very similar. We see that the location μ of the peak of the wavelet function is easily recognised by both processes $\mathbb{Q}_n$ and $\mathbb{Z}_n$. Moreover, the concentration of the expressed deviation in the plot is in accordance to the scale σ of the wavelet. Also, in case of a sum of different wavelet functions, each of the peaks is recognised and localised with the same resolution as for one wavelet function. The latter is demonstrated in Figure 6.7, which shows an example of such a δ-function (panel(a)) together with the corresponding heat maps for $Q(x, y)$ (panel (b)) and $Z(x, y)$ (panel (c)).

As mentioned above, the limiting behaviours of the original process $\mathbb{Z}_n$ and the standardised process $\mathbb{Q}_n$ are similar. These results suggest that the IBPP-plot is an effective graphical tool for localising small deviations from a distribution. To examine its ability to detect localised deviations in practice, we refer to the small-sample simulation study in Section 6.5, as well as to the real-data examples in the following section.

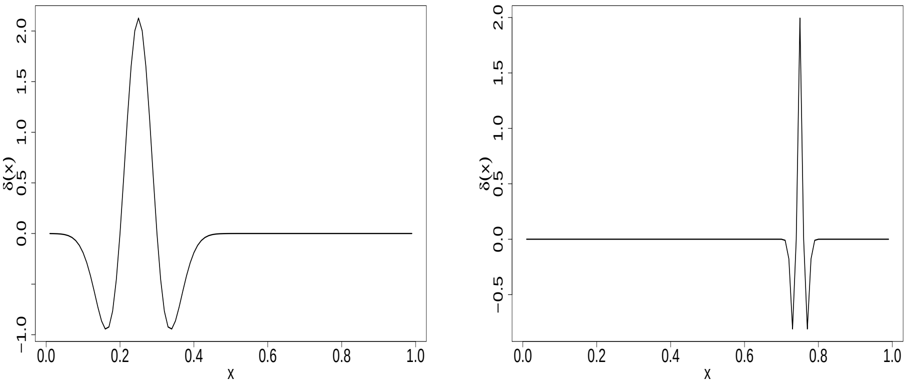
6.4 Examples

In this section we apply the new graphs to the circular data examples introduced in Section 2.2. We present both the CiPP-plot and the exploratory IBPP-plot for each example. The formal IBPP-plot is only given when the Kuiper test yields significant results. Points falling outside the dashed circle on the CiPP-plot correspond to process values larger than the critical value of the Kuiper test at the 5% significance level. We also add the standardised versions of the IBPP-plot in case we are interested in whether the data is coming from a CU distribution. For the examples with composite null hypotheses, we only give the original non-standardised IBPP-plot. The reason is that the variance becomes too complicated as mentioned above.



(a) $\mu = 0.5$, $\sigma = 0.1$ and $C = 2$

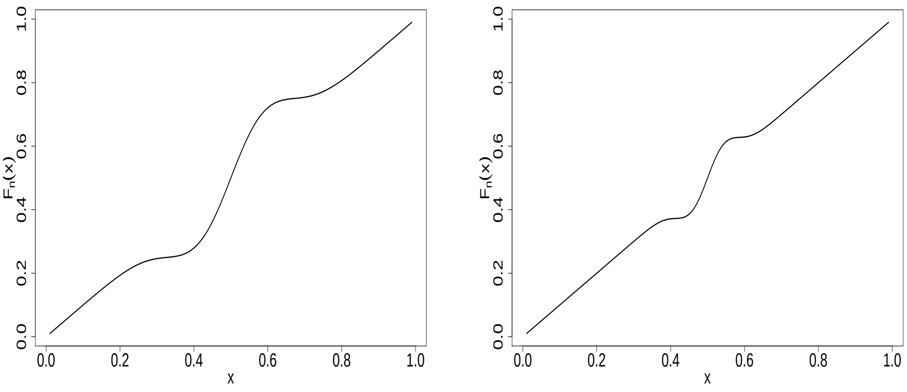
(b) $\mu = 0.5$, $\sigma = 0.05$ and $C = 2$



(c) $\mu = 0.25$, $\sigma = 0.05$ and $C = 2$

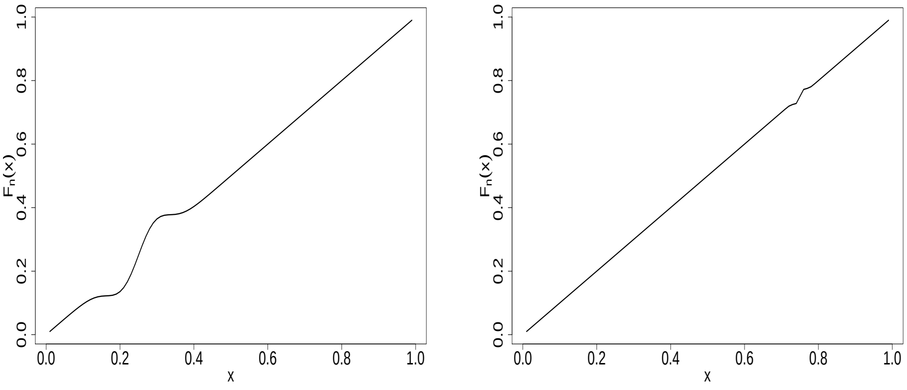
(d) $\mu = 0.75$, $\sigma = 0.01$ and $C = 2$

Figure 6.3: The Mexican hat wavelet for different parameter values of μ, σ and C.



(a) $\mu = 0.5$, $\sigma = 0.1$ and $C = 2$

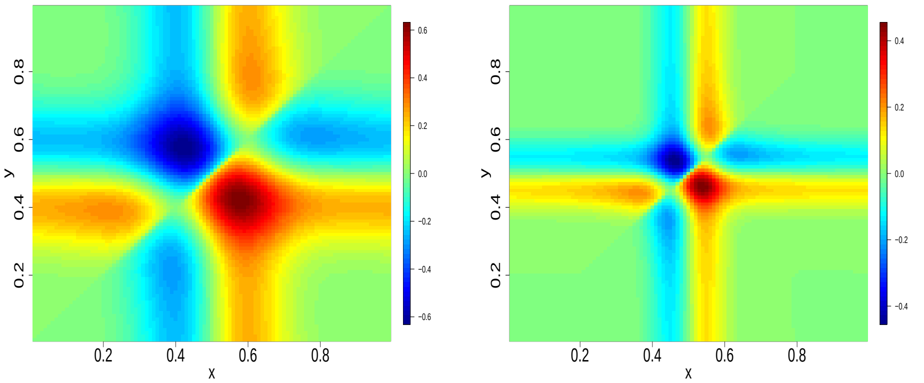
(b) $\mu = 0.5$, $\sigma = 0.05$ and $C = 2$



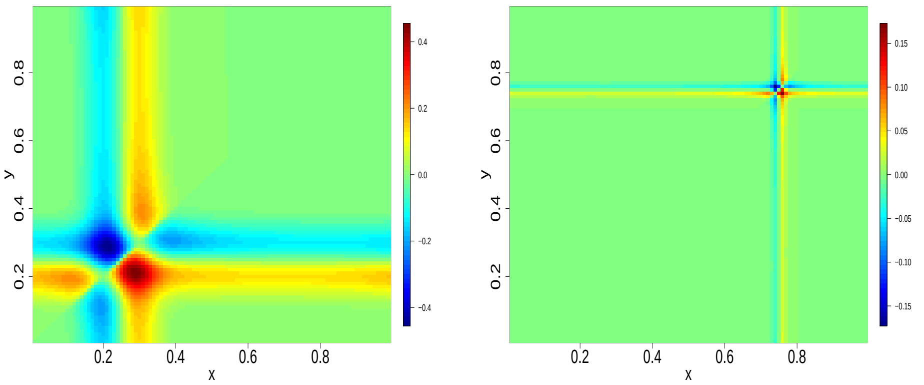
(c) $\mu = 0.25$, $\sigma = 0.05$ and $C = 2$

(d) $\mu = 0.75$, $\sigma = 0.01$ and $C = 2$

Figure 6.4: Alternative distribution functions $F_n(x)$ with $\delta(x)$ a Mexican hat wavelet for different parameter values of μ, σ and C and with $n = 1$.

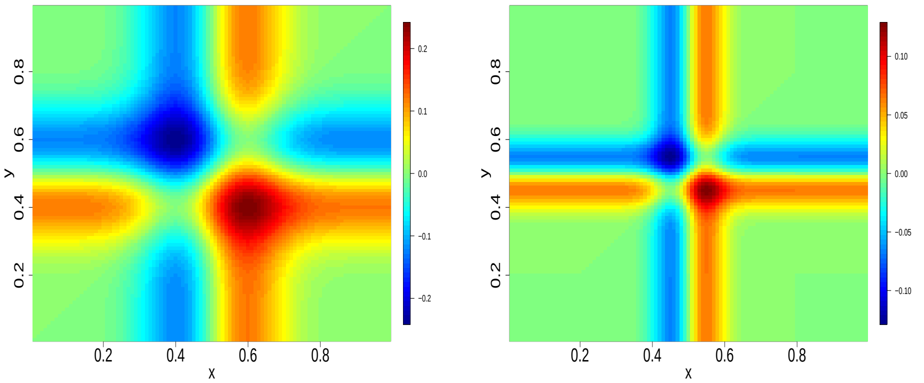


(a) $\mu = 0.5$, $\sigma = 0.1$ and $C = 0.4$ (b) $\mu = 0.5$, $\sigma = 0.05$ and $C = 0.3$

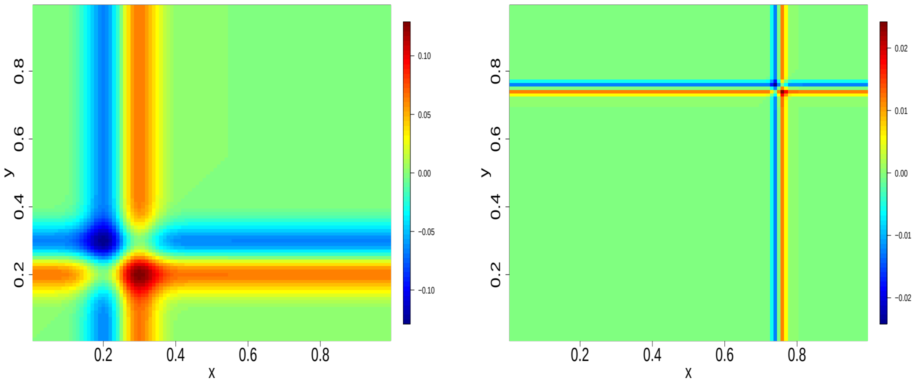


(c) $\mu = 0.25$, $\sigma = 0.05$ and $C = 0.3$ (d) $\mu = 0.75$, $\sigma = 0.01$ and $C = 1$

Figure 6.5: Limiting expected value of the standardised process $\mathbb{Q}_n$ versus its location interval $[x, y]$ assuming that the data come from the family of local alternatives (6.7) using the Mexican hat wavelet in (6.11) as noise function $\delta(x)$. The parameter values μ, σ and C for the Mexican hat wavelet take different values for the different panels.

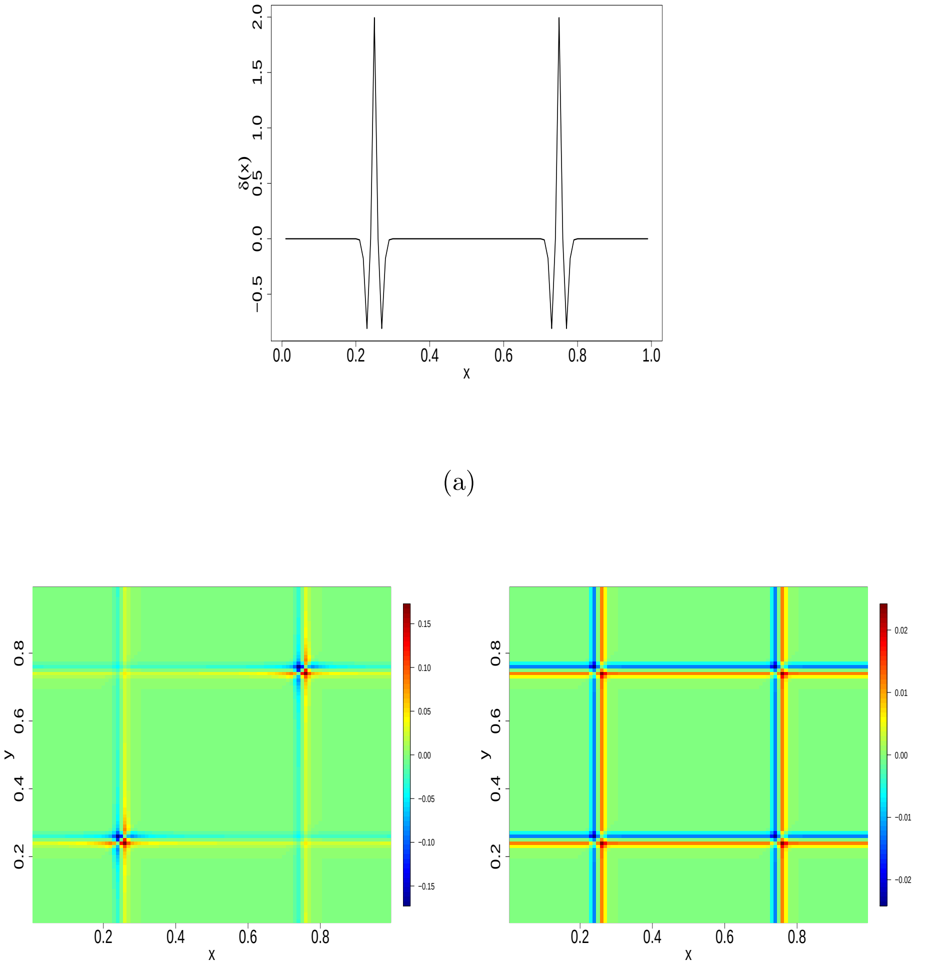


(a) $\mu = 0.5$, $\sigma = 0.1$ and $C = 0.4$ (b) $\mu = 0.5$, $\sigma = 0.05$ and $C = 0.3$



(c) $\mu = 0.25$, $\sigma = 0.05$ and $C = 0.3$ (d) $\mu = 0.75$, $\sigma = 0.01$ and $C = 1$

Figure 6.6: Limiting expected value of the process $\mathbb{Z}_n$ versus its location interval $[x, y]$ assuming that the data come from the family of local alternatives (6.7) using the Mexican hat wavelet in (6.11) as noise function $\delta(x)$. The parameter values μ, σ and C for the Mexican hat wavelet take different values for the different panels.

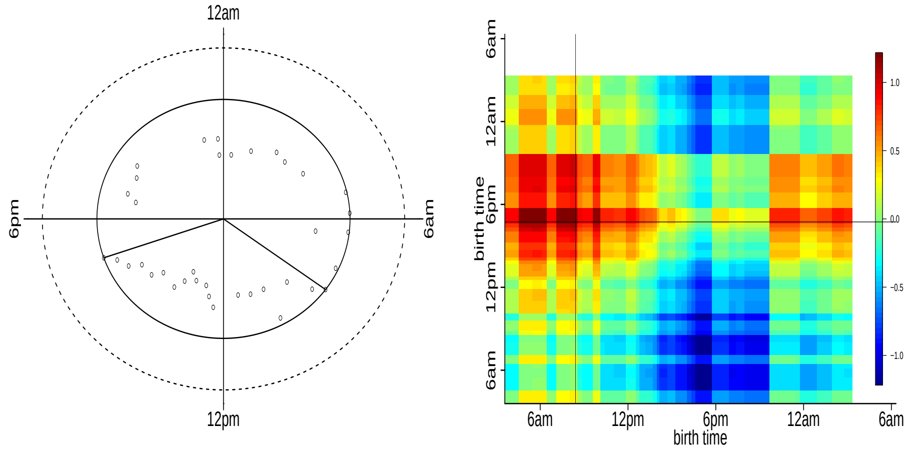


(a)

(b)

(c)

Figure 6.7: Limited expected value of $\mathbb{Q}_n$ (b) and $\mathbb{Z}_n$ (c) versus its location interval $[x, y]$ assuming that the data come from the family of local alternatives (6.7) using a sum of two Mexican hat wavelets (a) as noise function $\delta(x)$. The parameter values the two Mexican hat wavelets are $\mu = 0.25$ and 0.75, $\sigma = 0.01$ and $C = 1$.



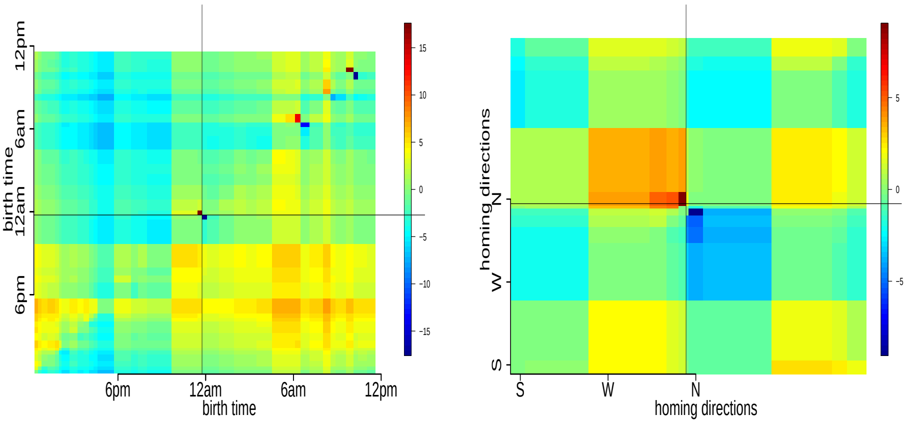
(a) (b)

Figure 6.8: Explorative Circular (a) and IBPP-plot (b) for the Birth time data.

Since the plots are also useful for linear data, we also consider some of the linear examples introduced in Section 2.1. For these examples, we confine the discussion to the original explorative and formal versions of the IBPP-plots.

6.4.1 Birth time data

Recall that for the Birth time data (Section 5.8.5), no evidence against uniformity was found based on the classical tests as well as on the circular SSPc tests. Figure 6.8 shows the explorative CiPP and the explorative IBPP-plot for the Birth time data. From both graphs we see that the maximum process value $\mathbb{Z}_n$ is obtained in about arc(8.30am,5pm), which clearly includes the largest sector on the rose diagram in panel (c) of Figure 2.7. We also see that observations between 6pm and 12am induce process values which have opposite sign to the maximum process value (see for example the smaller red or blue area which is below or above the diagonal in the IBPP-plot, respectively). This is probably because only few births appear in that period. As there is a large gap between the dashed circle on the CiPP-plot and the maximum process value, there is no reason to believe this arc is the result of a deviation from circular uniformity. However, the explorative graphs suggest a need for more observations in order to detect whether indeed there are relatively more births between 8.30am and 5pm. Panel (a) of Figure 6.9 shows the standardised explorative IBPP-plot. The latter gives a completely different impression of the location of the LOF. In particular, higher process values occur closer to the diagonal and the maximum



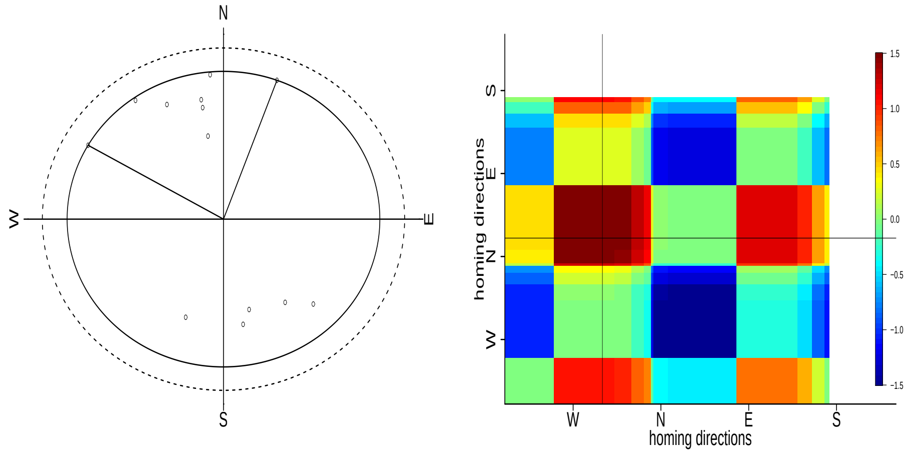
(a) (b)

Figure 6.9: Explorative standardised IBPP-plot for the Birth time data (a) and for the Homing pigeons data (b).

process value originates from an extremely small arc at 12am. Furthermore, the two large red and blue regions above the diagonal completely dissappeared. This result motivates further investigation on the ability of the standardised IBPP-plot to localise a LOF. The simulation study in Section 6.5 will indeed confirm that the standardisation causes a loss in power for global as well as local deviations.

6.4.2 Homing pigeons data

For the Homing pigeons data, the classical origin-invariant GOF tests for uniformity are non-significant. The smooth test of Bogdan et al. (2002) and the circular SPP*c* tests, however, recognise the bimodal deviation. These results were presented in Section 5.8.5. Since the Kuiper test detected no difference from uniformity, only the explorative versions of the plots are shown in Figure 6.10. The dashed circle in the CiPP-plot is much closer to the maximum process value which is obtained for the arc reaching from North-West to somewhat beyond North. The color pattern on the explorative IBPP-plot is not quite smooth. The reason may be that we only have 13 observations in this dataset. The change in colors just above the diagonal indicates that from just beyond South to just beyond West the process values are too low (too negative). The region from just beyond West to beyond North has process values that are too high, while the process values are again too low for the region from beyond



(a) (b)

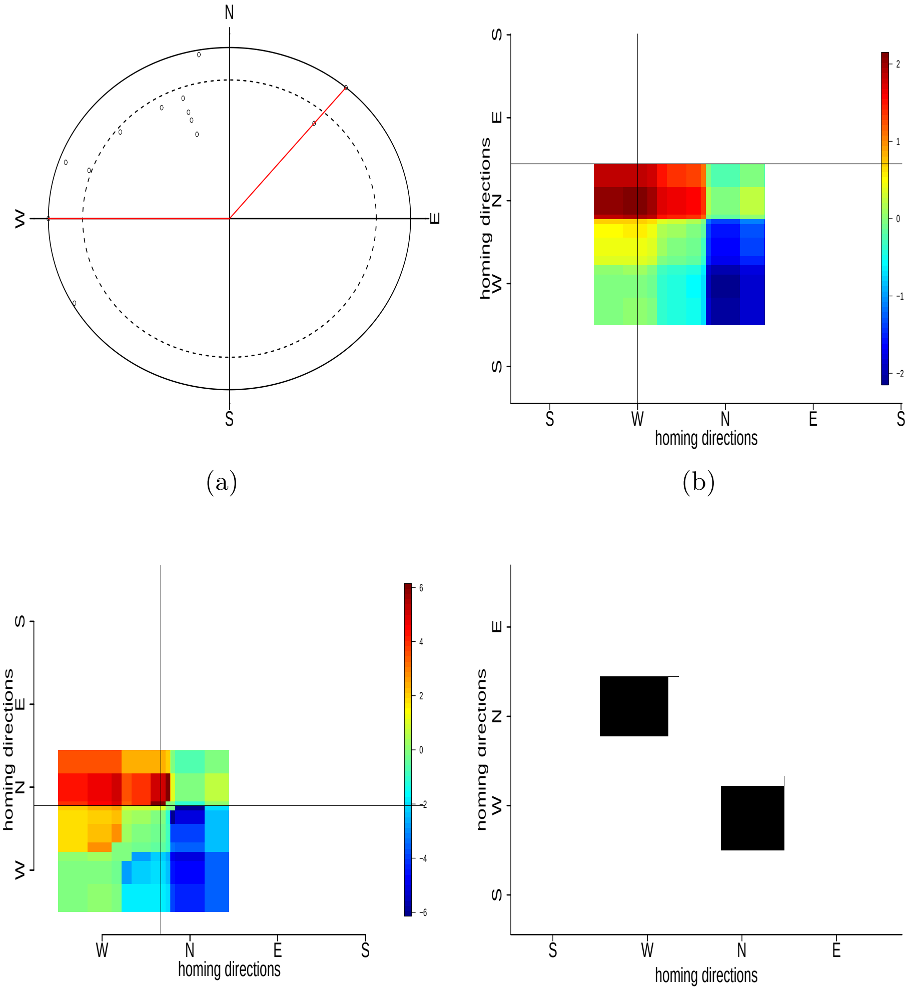
Figure 6.10: Explorative Circular (a) and IBPP-plot (b) for the Homing pigeons data.

North to beyond East. Such a pattern is typical of a bimodal dataset.

Similarly as for the Birth time data, the standardised IBPP-plot, which is given in panel (b) of Figure 6.9, shows that the typical bimodal pattern has become less expressed while the largest value of the process is again near the diagonal, for a very small arc just before North. This arc includes the three tied observations in the data.

When the data is doubled so as to see whether the birds are flying across the diagonal direction of the Valley, we obviously get a different result. The corresponding plots, including the formal IBPP plot, are presented in Figure 6.11, and we immediately see that there is now a clear deviation from uniformity. This deviation is located from West to beyond North.

Similarly as for the original Homing pigeons data, the standardised explorative IBPP-plot attains its maximum very close to the diagonal for an interval that includes the four tied observations. Also, the dark regions away from the diagonal are now lighter regions, while the light regions near the diagonal become more coloured. In Section 6.5 we will see that these shifts in colours, caused by the standardisation of the process, correspond to a loss in power of the corresponding weighted Kuiper test. This is already seen here since the weighted Kuiper test is not significant anymore and hence no formal standardised IBPP-plot is drawn. This implies that probably for small samples, the standardised process is not as useful as was suggested by its limiting behaviour in Section 6.3.



(a) (b)

(c) (d)

Figure 6.11: Explorative Circular (a), explorative original (b) and standardised (c) IBPP-plot and formal original IBPP-plot (d) for the doubled Homing pigeons data.

6.4.3 Turtles data

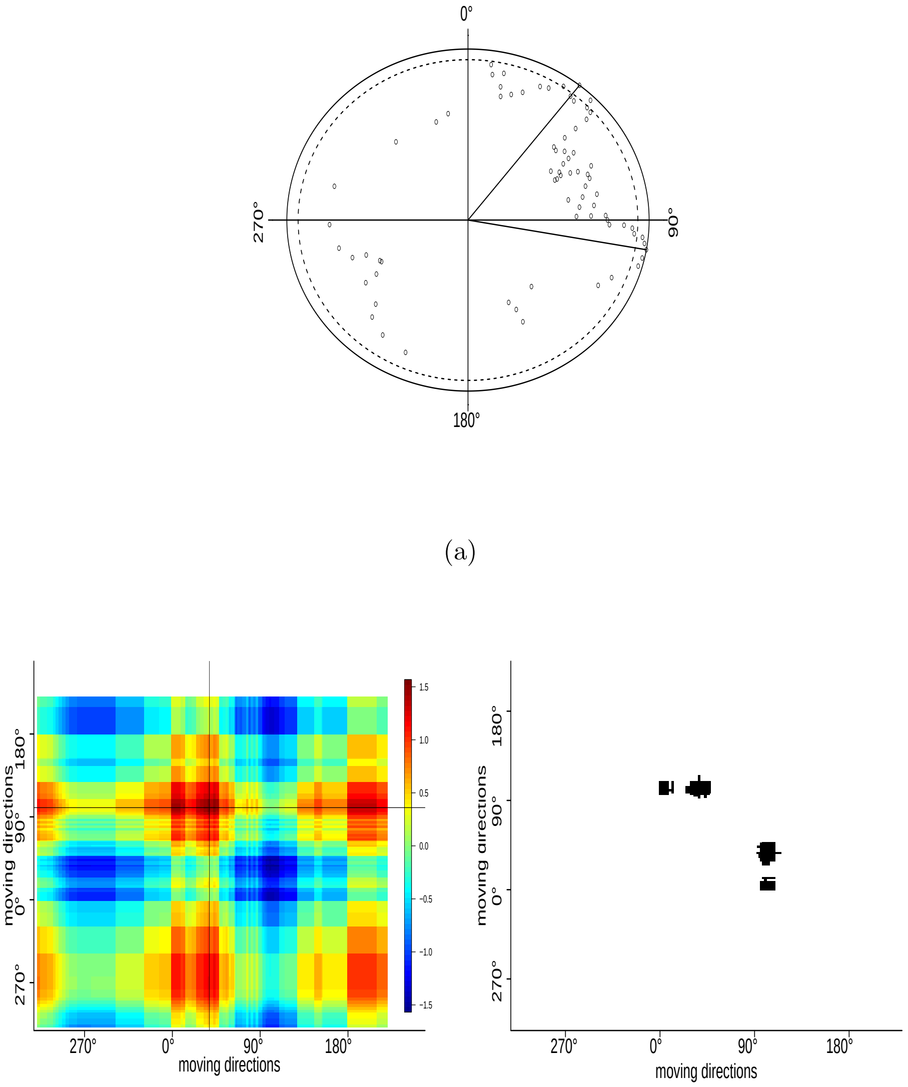
For the Turtles example, we would like to examine whether the moving directions are normally distributed. From Section 4.6.1, we concluded from the data-driven circular smooth test that the data deviate from CN primarily with respect to the second trigonometric moment. This confirmed the bimodal impression discussed in Chapter 2. Also the classical tests indicated a significant deviation from CN, but no information about the alternative can be derived from those tests. Recall that we use parametric bootstrap to find the 5% quantile of the null distribution of the Kuiper statistic. Since the von Mises distribution is a location equivariant distribution (not scale equivariant), the 100,000 bootstrap samples are generated using location μ equal to zero and concentration parameter κ equal to its MLE.

From the explorative CiPP and IBPP-plots in panels (a) and (b) of Figure 6.12, we see that the maximum significant process value of $\mathbb{Z}_n$ is obtained for the arc from about 40° to about 100°, where most observations are concentrated. In the explorative IBPP-plot, a similar pattern as for the Homing pigeons data is recognised. This seems to imply that these data deviate from circular normality in a bimodal way. In particular, the comparison distribution, which is the distribution that describes the difference between observed and hypothesised distribution, is probably bimodal. However, if we look at the formal IBPP-plot (panel (c) of Figure 6.12), we see that only two intervals of LOF are recognised, i.e. arc(0°,90°) and arc(45°,90°). Hence, the significant deviation from normality is only located in the largest cluster of the observations.

6.4.4 Ants data

For the Ants data, we concluded from the circular data-driven smooth test in Section 4.6.2 that there is much evidence against circular normality. In particular, the second order trigonometric moment, which is related to the skewness and kurtosis is responsible for this deviation. The classical GOF tests for circular normality also have highly significant results.

Figure 6.13 shows the same three plots for the Ants data. The largest significant deviation from normality is located in arc(180°,210°) (see panel (a)). Too many ants chose a direction in this interval instead of moving towards the black target placed at 180°. This causes a distribution of moving directions that is skewed to the right. Note that this confirms the impression in the explorative discussion in Section 2.2.4. More information could be obtained from the explorative and the formal IBPP-plots (see panel (b) and (c) of Figure 6.13). In particular, three adjacent regions with alternating sign are seen in the explorative version. Above the diagonal, for instance, we first notice a blue region, roughly indicating an underestimation of the density in intervals


(a)

(b)

(c)

Figure 6.12: Explorative Circular (a), explorative (b) and formal (c) IBPP-plot for the Turtles data.

that include arc(90°,180°). Then a red region is observed for intervals including arc(180°,200°). This region causes the largest overestimation of the density. Finally, the last region includes arc(220°,270°) and causes an underestimation. As is seen from the formal IBPP-plot, these three regions are significant at the 5% level.

6.4.5 Direzione data

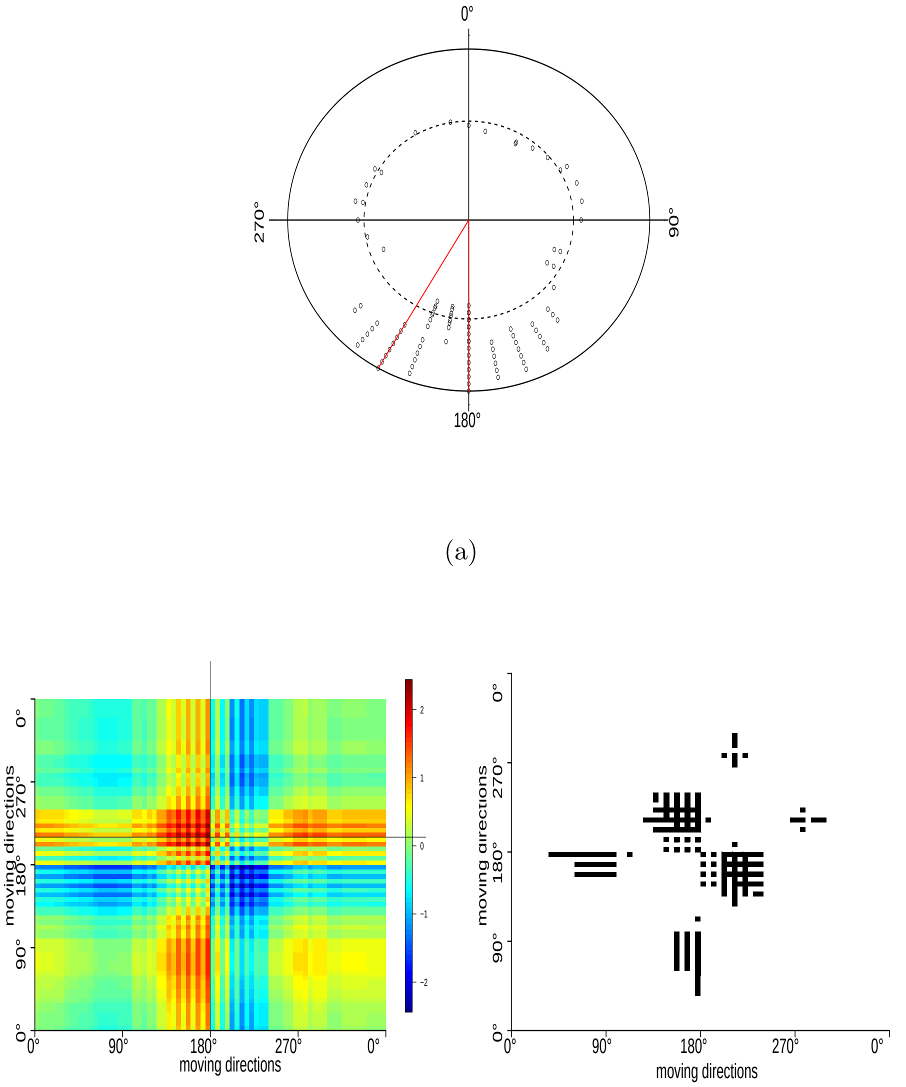
In Section 4.6.3, similar highly significant test results have been obtained for the Direzione data as for the Ants data. For the Direzione data, panels (a) and (b) of Figure 6.14 show the explorative version of, respectively, the CiPP and IBPP-plots, whereas panel (c) shows the formal analogue of the IBPP-plot. Again we are testing for circular normality. It can be seen that all $\mathbb{Y}_n(x)$ process values (panel (a)) are in the rejection region, which means that no specific LOF location can be indicated from that plot. The peak occurs around the North direction (0°), where most observations are situated. More information on the LOF location can be extracted from the explorative IBPP-plot and from the formal IBPP-plot. All significant intervals contain locations near the North direction as begin or end point. We conclude that the location of LOF is an interval containing the North direction.

6.4.6 Arrival data

As a final circular data example, we consider the Arrival data. In Section 4.6.4, the data-driven circular smooth tests indicated that there is a highly significant deviation from circular normality in at least the second and at most the fifth order trigonometric moment. On the other hand, the classical tests found no significant deviation from the CN distribution. The explorative versions of the CiPP and the IBPP-plot are in Figure 6.15. The formal IBPP-plot is not given since no significant process values were found for the Kuiper test. In accordance with the non-significance, no points are outside the dashed critical circle on the CiPP-plot in panel (a). However, the process values in the CiPP-plot as well as those in the IBPP-plot (see panel (b)) do suggest an oscillating departure from normality by the wiggly pattern. Also, a pattern with the five clusters is seen in both the exploratory IBPP and CiPP-plot. The same pattern was seen for the orthonormal series density estimate of the Arrival data presented in Section 4.6.4.

6.4.7 Contaminated Lottery data

We now consider two linear data examples. In Section 5.5.1, we have considered the Lottery data but systematically changed it in the sense that 400 is subtracted

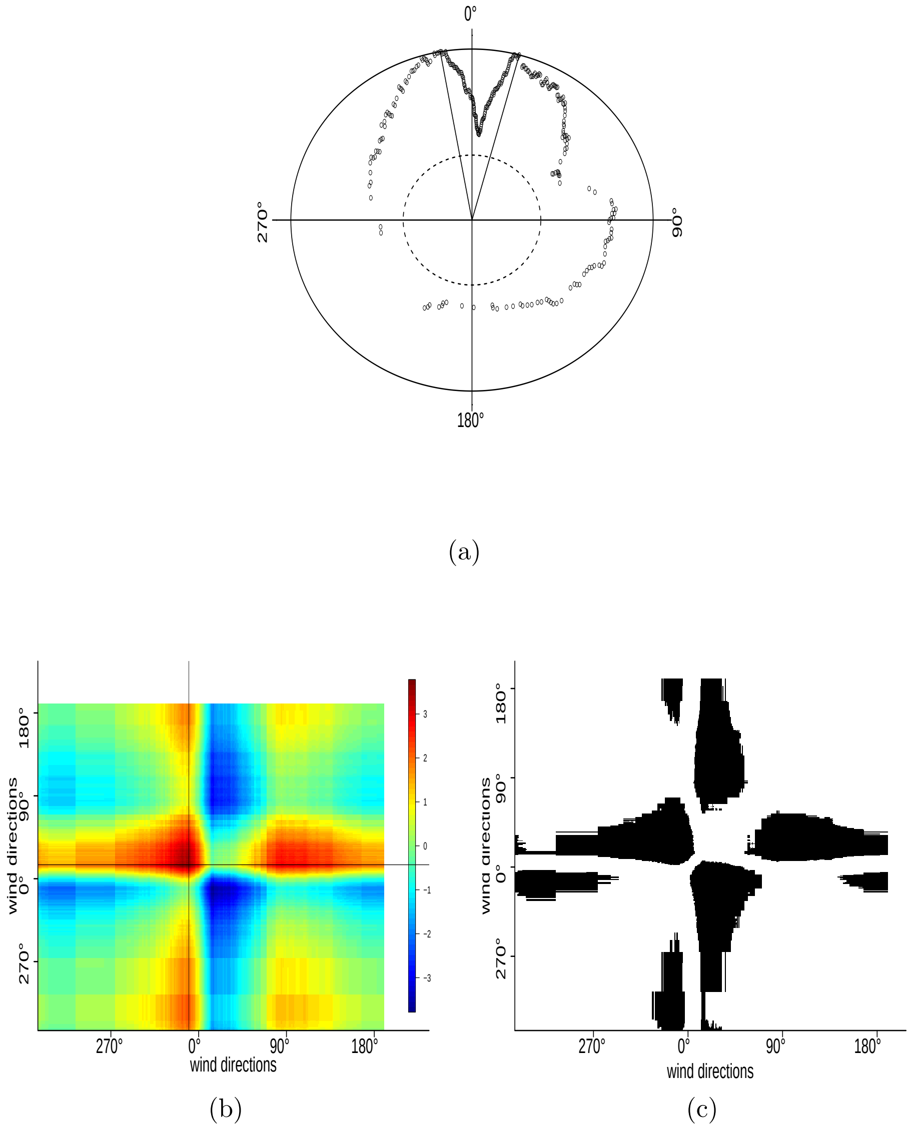


(a)

(b)

(c)

Figure 6.13: Explorative Circular (a), explorative (b) and formal (c) IBPP-plot for the Ants data.

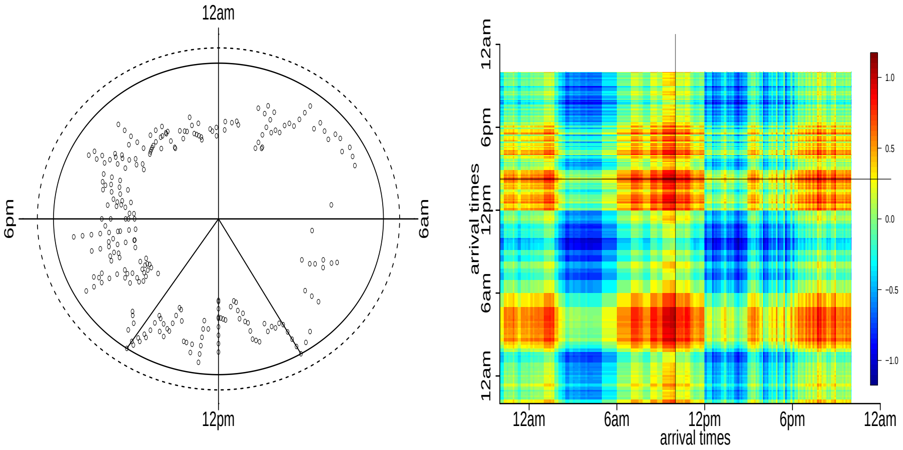


(a)

(b)

(c)

Figure 6.14: Explorative Circular (a), explorative (b) and formal (c) IBPP-plot for the Direzione data.



(a) (b)

Figure 6.15: Explorative Circular (a) and IBPP-plot (b) for the Arrival data.

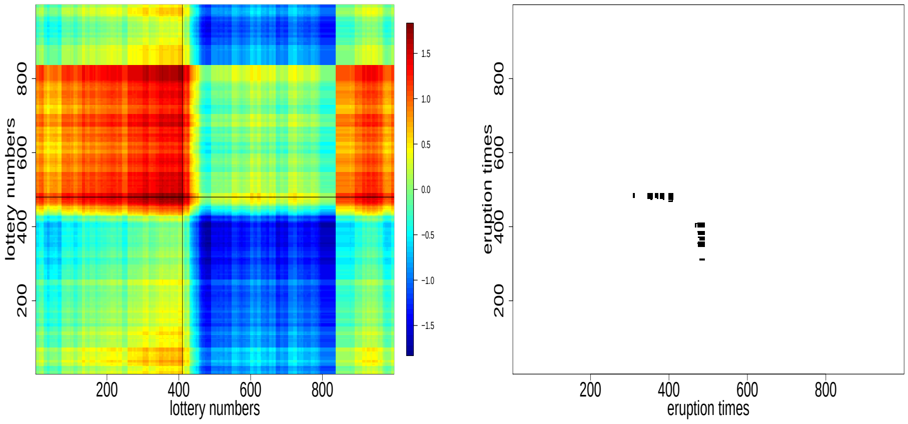
from all numbers between 800 and 875. We showed that the SSPc test was particularly useful to detect this contamination in the data.

The IBPP-plots, which are given in Figure 6.16, visualise where the contamination occurs. In particular, from panel (a), we see that intervals that include the interval [400,500] show an increased process value, while intervals that include [800,900] show a decreased process value. The formal IBPP-plot in panel (b) reveals that the Kuiper test is in fact significant at the 5% level. Moreover, the plot gives information on the location of the LOF, namely that the largest value of the process $\mathbb{Z}_n$ is located in the interval [400,450]. Note that the classical detrended PP-plot, as described troughout Chapter 2, would not indicate a deviation, since the related KS test is not significant.

6.4.8 Old Faithful geyser data

The Old Faithful geyser data, which was introduced in Section 2.1.5, showed a severe deviation from normality. We are now interested in whether the deviation from normality could still be located if we would only have a subsample of size 20 available. More specifically, we wonder if the bimodal pattern that is present in the original data, is still detectable in such a small subsample.

Before we give the results for the subsample, we first comment on the IBPP-plots of the original data, which are shown in panels (a) and (b) of Figure 6.17. Preliminary analysis for the original data revealed roughly two clusters of observations, which are also called bumps or modes. In between the two



(a) (b)

Figure 6.16: Explorative (a) and formal (b) IBPP-plot for the contaminated Lottery data.

bumps, a region with almost no concentration of observations was found, which is often referred to as a *dip*. The explorative IBPP-plot in panel (a) confirms this and shows precisely how the data deviates from normality. In particular, above the diagonal we first recognise the first bump in the distribution as the red region for intervals including [1.5,2.2], which indicates that the concentration of observations is higher than expected in case of normality. This red region is followed by a blue region for intervals including [2.5,4], which indicates that the density of the true distribution is clearly lower than for the normal distribution. The absolute value of the process also obtains its largest value for that interval where the dip is located. Finally, the second bump is recognised by the second red region in the plot, which corresponds to intervals roughly including [4,5]. The formal IBPP-plot in panel (b), which shows the rejection region for the Kuiper test, clearly indicates the intervals where the bumps and dip occur. These regions were also seen from the detrended PP-plot in panel (c) from Figure 2.5.

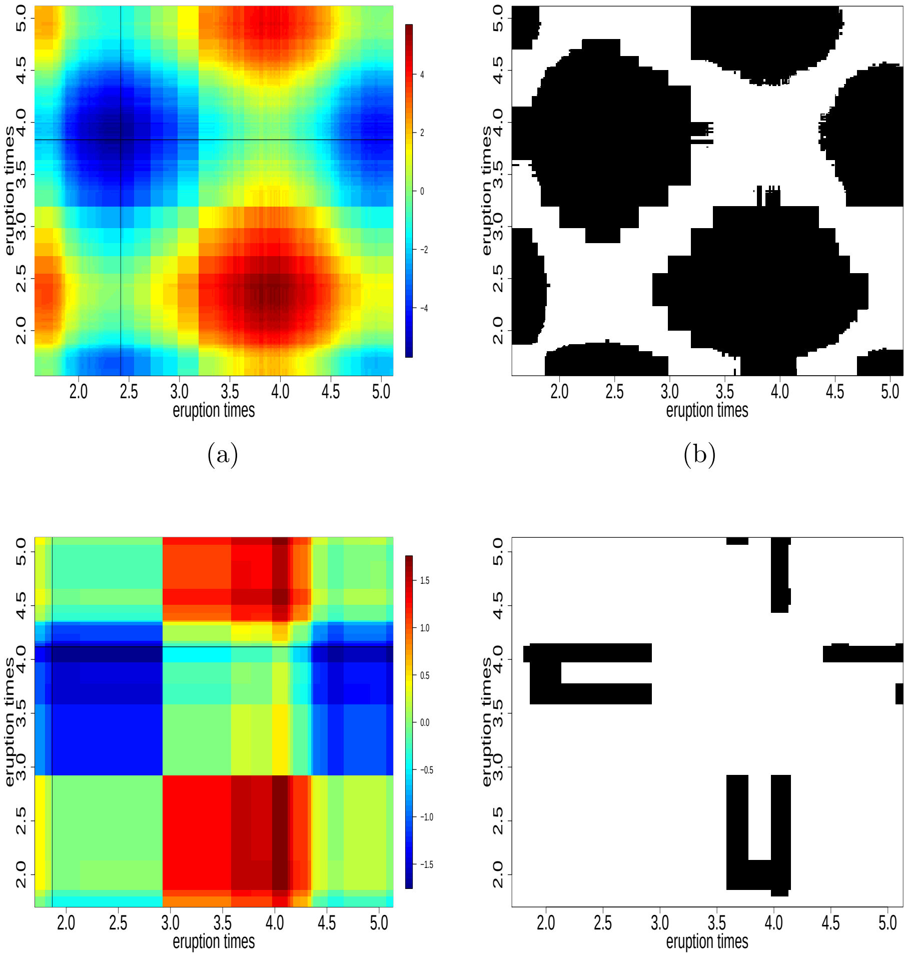
In Section 2.1.5, we have found that the KS test for the subsample was not significant anymore. This means that the related detrended PP-plot would not give useful information on the location of the bumps in the distribution. On the other hand, the new IBPP-plots here prove particularly useful to localise the bumps in the small dataset. The explorative and formal IBPP-plots are presented in panels (c) and (d) of Figure 6.17, respectively. The explorative IBPP-plot gives the same impression for the subsample as for the original sample, although, the colours are alternating less smoothly, which is due to the smaller sample size. The largest process value is observed in a larger interval, i.e. [1.5,4.2] than for the original data. The first bump is less extreme, while the second is more extreme. The formal IBPP-plot shows significant deviation at the location of the dip and the second bump. This example illustrates the usefulness of the IBPP-plot to detect deviating bumps in a distribution.

6.5 Simulation Study

In this section we present the results of an empirical simulation study that aims at demonstrating the localisation ability of the formal IBPP-plot and the formal standardised IBPP-plot. As we mentioned before and illustrated in the previous section, these plots are also applicable to linear data. Here, we take 1,000 samples from alternatives to the linear standard normal distribution of sample size 50. In particular, we use the same alternatives as in the simulation study of Section 5.6: normal distributions with either μ or σ fixed, as well as contaminated normal distributions. We expect that similar results would be obtained when simulating from the circular analogues, used in the simulation study of Section 5.8.6.

For each simulated dataset, we applied the Kuiper test as well as the standardised Kuiper test. Before presenting the results on the localisation ability of the IBPP-plots, we compare the estimated powers of the two tests for each of the alternatives. Figure 6.18 presents the power curves of both statistics for the four selected alternatives. Panel (a) shows the curves for a normal alternative with $\sigma = 1$ as a function of the mean μ, while panel (b) gives the powers for a normal alternative with $\mu = 0$ for increasing σ. Panels (c) and (d) show the powers of the contaminated normal distributions $f_{1,\gamma}$ and $f_{2,\gamma}$ versus γ, respectively. All plots show that the standardised Kuiper has lower power than the original Kuiper test. In particular, for the normal alternative with fixed $\sigma = 1$, almost zero power is found for the standardised Kuiper test. For the contaminated normal alternatives the power of the standardised Kuiper is much better, but still less than the original Kuiper test. From this we may conclude that standardising the Kuiper test statistic in the way of (6.5) is a GOF test with questionable power in practice.

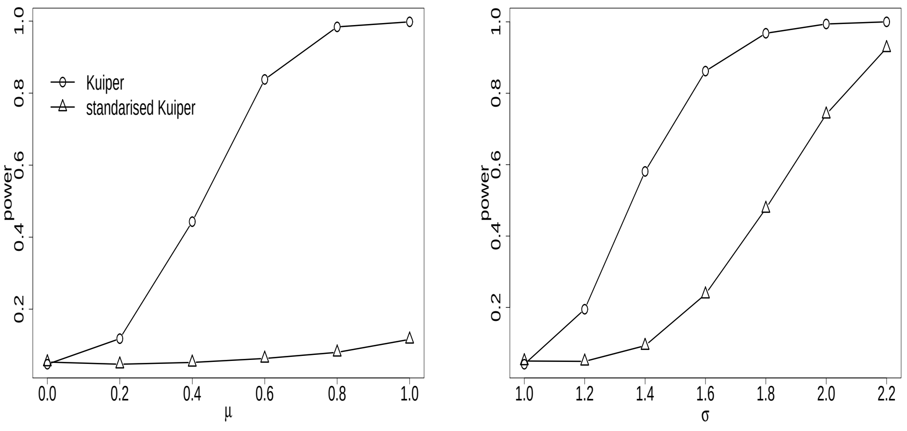
Let us now focus on the IBPP-plot. In case the Kuiper test or its standardised version rejected the null hypothesis, we recorded for which intervals the process $\mathbb{Z}_n$ and $\mathbb{Q}_n$ exceeded the 5% critical value of the corresponding test. Intervals considered were all those with observations as end points (hence the process values of $\binom{n}{2}$ intervals are computed for each sample, with $n = 50$). Let any interval (x, y) be represented by a point on the plane, as in the IBPP-



(a) (b)

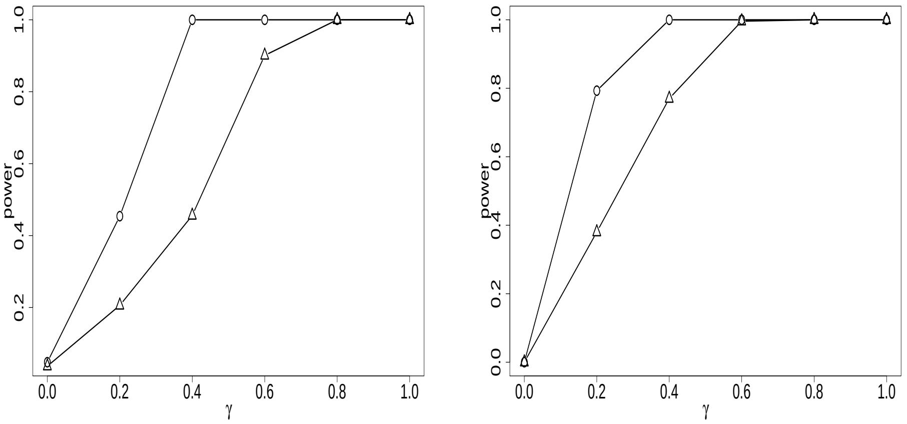
(c) (d)

Figure 6.17: Explorative (a) and formal (b) IBPP-plots for the Old Faithful geyser data. Explorative (c) and formal (d) IBPP-plots for a random subsample (n=20) of the Old Faithful geyser data.



(a) $\phi(x; \mu, 1)$

(b)$\phi(x; 0, \sigma)$



(c) $f_{1,\gamma}$

(d) $f_{2,\gamma}$

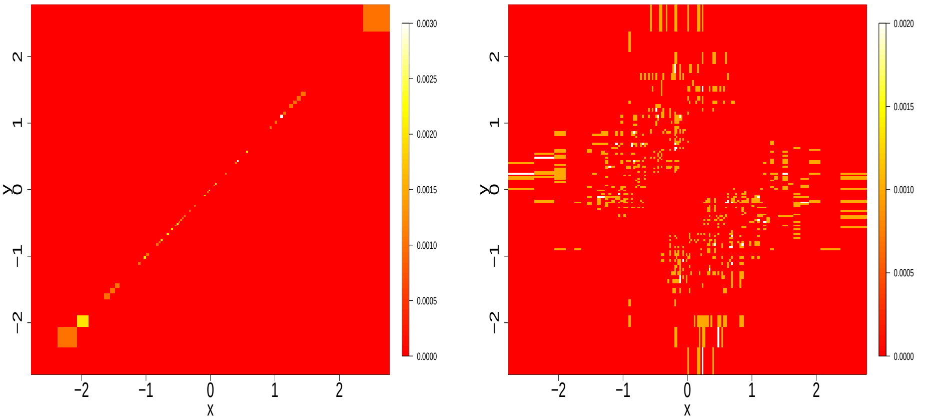
Figure 6.18: Estimated power curves of the original and the standardised Kuiper test. Panel (a) are the power curves under a normal alternative with $\sigma = 1$ versus μ. Panel (b) are the power curves under a normal alternative with $\mu = 0$ versus σ. Panels (c) and (d) are the power curves under a contaminated normal alternative with $\delta = 1$ and $\delta = 2$, respectively, versus γ.

plot. We then divided the plane into 100 × 100 cells. The size of the cells was scaled relative to the normal density, in the sense that cells in the tails of the distribution are largest. For each cell, we then counted how many of the 1,000 samples resulted in at least one rejection occurring within the cell. The relative frequencies are presented in heat maps and may be seen as estimated rejection probabilities for the respective areas on the IBPP-plots.

Figure 6.19 shows the rejection probabilities of both IBPP-plots when simulation is performed under the null hypothesis. In panel (a), we see that the nonzero rejection probabilities for the standardised IBPP-plot are only situated close to the diagonal of the plot, which correspond to very small intervals. On the other hand, the nonzero rejection probabilities for the original IBPP-plot are scattered over the whole area of the plot, except on the diagonal and in the upper left and lower right corners. The latter regions correspond to intervals over the whole range of the sample. Hence, under the null hypothesis, a type I error occurs mainly in small intervals in case of the standardised Kuiper test. On the other hand, the type I errors made by the original Kuiper test appears most frequently for intervals of intermediate size. We believe that in order to obtain an optimal power to localise any LOF, type I errors should be scattered randomly across the whole area. The original IBPP-plot succeeds in achieving this property. On the other hand, the randomly scattered impression for the standardised IBPP-plot is clearly less obvious.

Figure 6.20 shows the estimated rejection probabilities for the standardised and the original IBPP-plot for the alternative normal distribution with $\sigma = 1$ fixed and with $\mu = 0.4$ and $\mu = 1$ in the panels (a), (b) and (c), (d), respectively. We see that for the standardised IBPP-plot small intervals in the right tail has relatively high rejection probability since the shift is to the right of the standard normal distribution. On the other hand, the original IBPP-plot clearly shows large rejection probabilities for intervals which include the region where the LOF occurs. Note that these intervals are rather large, which indicates a global LOF. From these plots and by comparing the rejection probabilities we may conclude that a global LOF is better detected by the non-standardised version of the IBPP-plot.

Figure 6.21 shows the estimated rejection probabilities for the standardised and the original IBPP-plot for the normal alternative distribution with $\mu = 0$ fixed and with $\sigma = 1.4$ and $\sigma = 2.2$ in the panels (a), (b) and (c), (d), respectively. Similarly as for the normal alternatives shifted to the right, we see that the standardised IBPP-plot gives some indication that the LOF is due to a shift in variance. In particular, small intervals in both tails correspond to relatively large rejection probabilities and large intervals over the whole range also show large rejection probabilities. Therefore, we may conclude that the standardised IBPP-plot detects the variance shift correctly. For the original



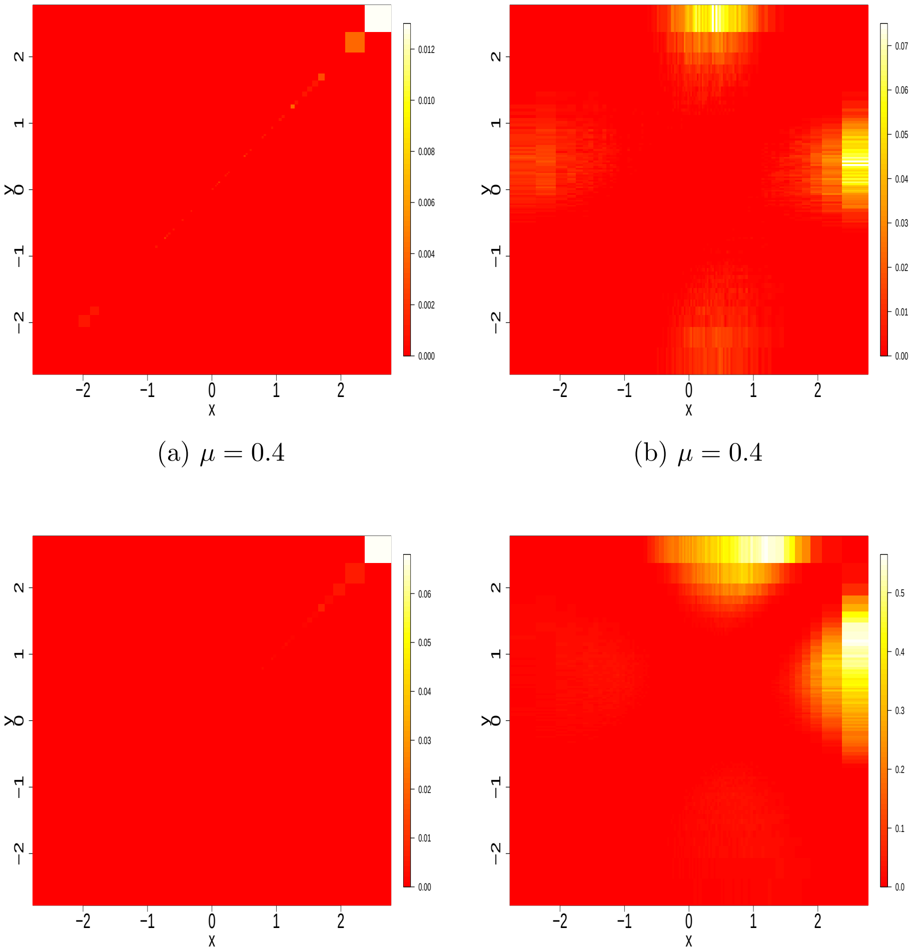
(a) Standardised IBPP-plot

(b) Original IBPP-plot

Figure 6.19: Estimated rejection probabilities to reject the null hypothesis of standard normality in each interval [x,y] for the standardised (a) and the original (b) IBPP-plot. Simulations are performed under the null hypothesis.

IBPP-plot, large rejection probabilities for symmetric intervals around zero are observed. We also see that the larger the tails of the alternative distribution, the larger the rejection intervals are. In either case, a global LOF would be concluded.

Figure 6.22 shows the estimated rejection probabilities for the IBPP-plots for the contaminated normal alternatives $f_{1,\gamma}$. Panels (a) and (c) correspond to the standardised IBPP-plots for simulations from $f_{1,0.2}$ and $f_{1,0.4}$, respectively. Panels (b) and (d) show the same cases for the original IBPP-plot. Similarly as in Section 5.6, we subtracted the true mean, which is $\delta\gamma$, from the simulated data. The location of the LOF is detected by the standardised IBPP-plot. A white dot on the diagonal at the location of the induced LOF ($\delta = 1$) shows relatively large rejection probability. The original IBPP-plot also localised the LOF, but not as precisely as the standardised IBPP-plot. In particular, the interval that corresponds to the LOF is almost always smaller for the standardised IBPP-plot than for the non-standardised IBPP-plot. For $\gamma = 0.2$, other intervals than the one purely concentrated on $\delta = 1$ show large rejection probabilities as well. Note that all those other intervals include the location of the LOF, and they all have reasonably small length. When γ is increased to 0.4, the intervals are much smaller and almost all concentrated on the location of the LOF.



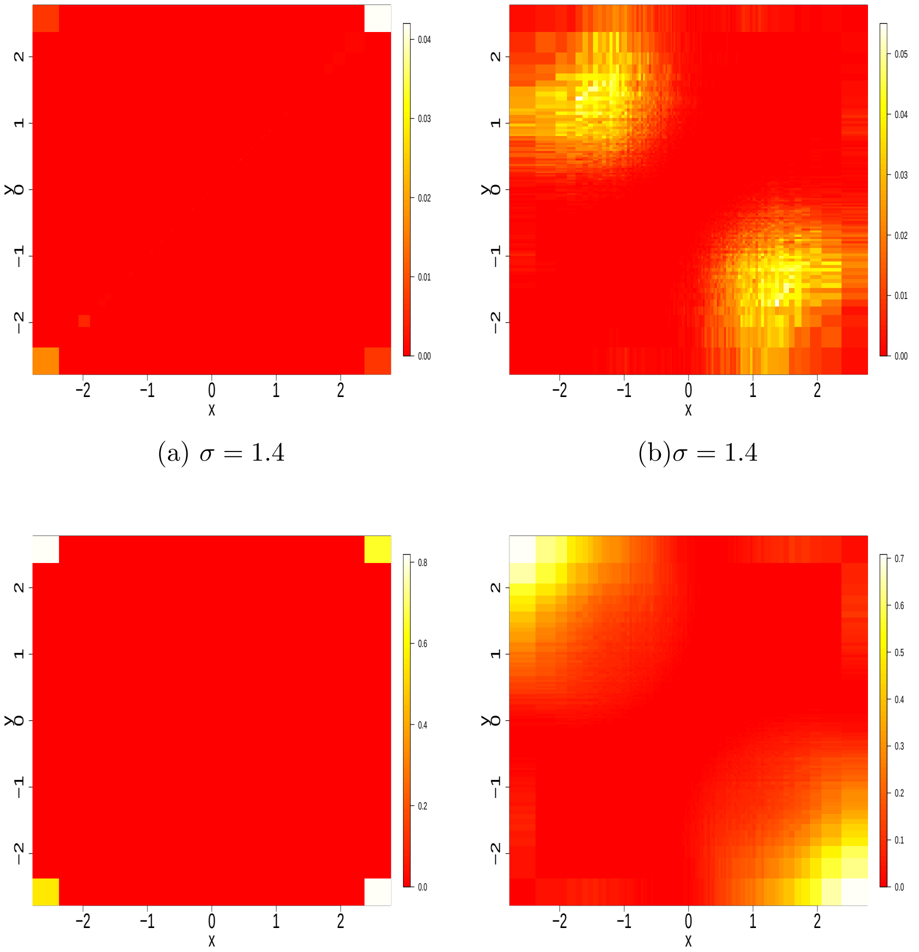
(a) $\mu = 0.4$

(b) $\mu = 0.4$

(c) $\mu = 1$

(d) $\mu = 1$

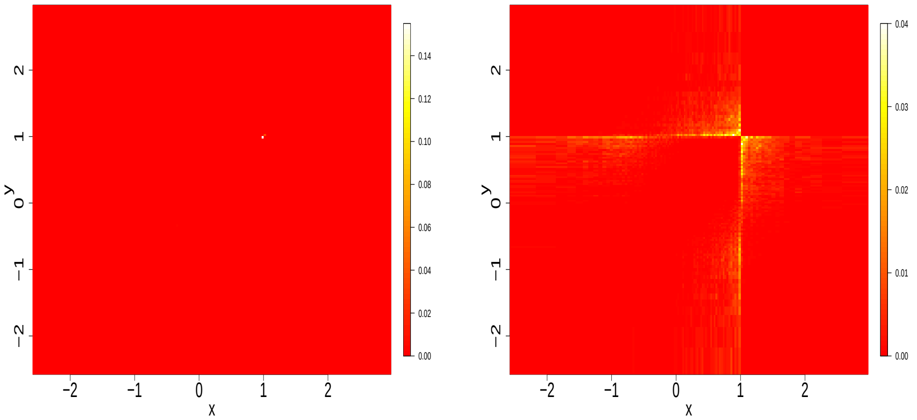
Figure 6.20: Estimated rejection probabilities to reject the null hypothesis of standard normality in each interval [x,y] for the standardised (a) and (c) and the original (b) and (d) IBPP-plot. Simulations are performed under a normal alternative with $\sigma = 1$ and μ either 0.4 ((a) and (b)) or 1 ((c) and (d)).



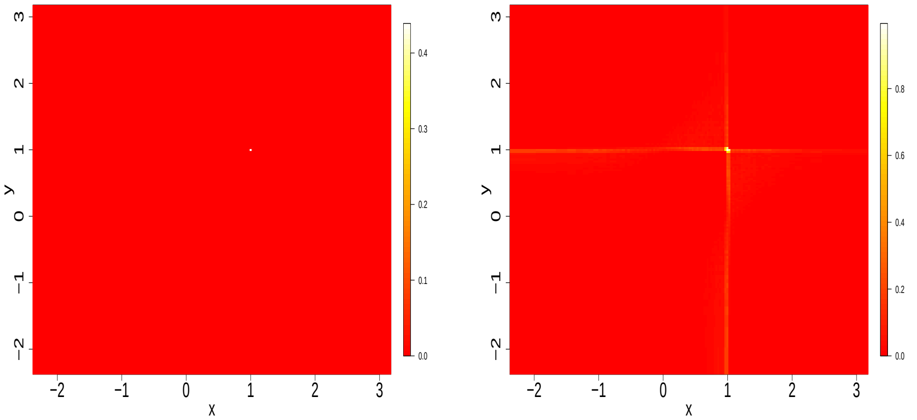
(a) $\sigma = 1.4$ (b)$\sigma = 1.4$

(c) $\sigma = 2.2$ (d) $\sigma = 2.2$

Figure 6.21: Estimated rejection probabilities to reject the null hypothesis of standard normality in each interval [x,y] for the standardised (a) and (c) and the original (b) and (d) IBPP-plot. Simulations are performed under a normal alternative with $\mu = 0$ and σ either 1.4 ((a) and (b)) or 2.2 ((c) and (d)).

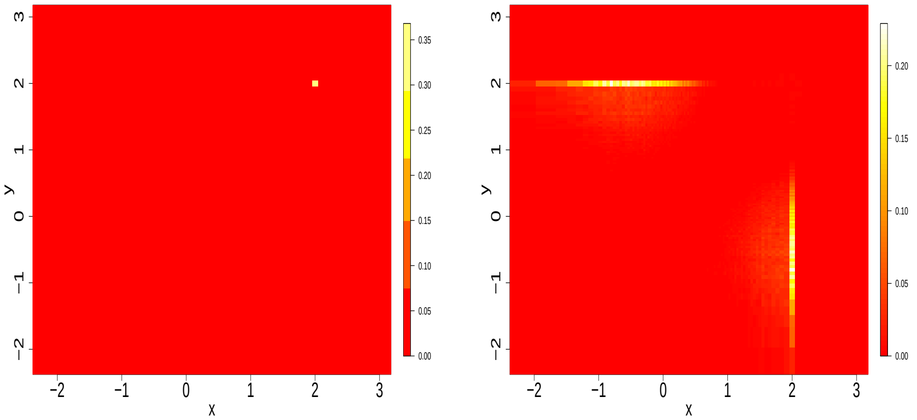


(a) $f_{1,0.2}$ (b) $f_{1,0.2}$

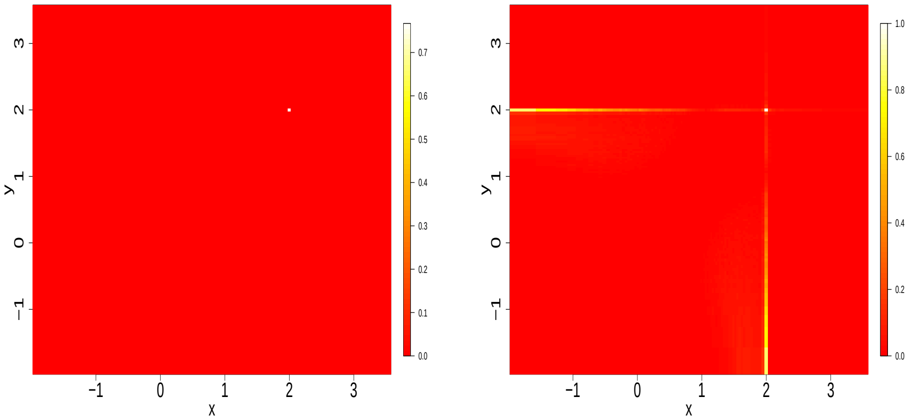


(c) $f_{1,0.4}$ (d) $f_{1,0.4}$

Figure 6.22: Estimated probabilities to reject the null hypothesis of standard normality in each interval [x,y] for the standardised (a) and (c) and the original (b) and (d) IBPP-plot. Simulations are performed under a contaminated normal alternative with $\delta = 1$ and γ either 0.2 ((a) and (b)) or 0.4 ((c) and (d)).



(a) $f_{2,0.2}$ (b) $f_{2,0.2}$



(c) $f_{2,0.4}$ (d) $f_{2,0.2}$

Figure 6.23: Estimated probabilities to reject the null hypothesis of standard normality in each interval [x,y] for the standardised (a) and (c) and the original (b) and (d) IBPP-plot. Simulations are performed under a contaminated normal alternative with $\delta = 2$ and γ either 0.2 ((a) and (b)) or 0.4 ((c) and (d)).

Figure 6.23 shows the estimated rejection probabilities for the IBPP-plots for the contaminated normal alternatives $f_{2,\gamma}$. Panels (a) and (c) again correspond to the standardised IBPP-plot for simulations from $f_{2,0.2}$ and $f_{2,0.4}$, respectively. Panels (b) and (d) show the same plots for the original IBPP-plot. As in the previous plots, the location of the LOF is clearly pin-pointed by the standardised IBPP-plot. The white dot on the diagonal at the location of the induced LOF ($\delta = 2$) is now larger. The reason is that the grid is constructed in such a way that the more the intervals are situated in the tails, the larger the cells. Nevertheless, the cell on the diagonal at location $\delta = 2$ has higher rejection probability than for the previous alternative. The original IBPP-plot also localised the LOF, although again less precise in the location than the standardised IBPP-plot. For $\gamma = 0.2$, large intervals containing the location of the LOF have large rejection probabilities. When γ is increased to 0.4, also smaller intervals that are more concentrated on the location of the LOF occur with large probability.

These simulation results indicate that the standardised and the original IBPP-plot succeed fairly well in indicating the location of the LOF. If the standardised version finds the LOF, then the plot is more precise than the original IBPP-plot. However, the associated weighted Kuiper test has lower power than the original Kuiper test and hence the original IBPP-plot may prove more valuable in many situations than the standardised plot.

6.6 The IBPP-plot as a tool for comparing density estimates

In the context of GOF, we have illustrated in Sections 3.7 and 4.5, how a density estimate naturally arises from the data-driven smooth test. In particular, if we use the Barton smooth model for the construction of the family of smooth alternatives, estimates of the parameters θ_j in the model can easily be found. Moreover, the form of such a density estimator is exactly that of the well-known orthonormal series density estimator (see e.g. Silverman, 1986). Hence, if the data-driven smooth test leads to the rejection of the null hypothesis, the order chosen by the selection criterion not only gives information about how the true model deviates from the hypothesised, it also yields an appropriate order for the associated orthonormal series density estimator. The latter can then be used to get an immediate visual impression of the deviation. The density estimate based on the linear data-driven smooth test has been illustrated in Section 3.7.2 on the Fastfood data, while an example in case of circular data has been presented in Section 4.6.4 (Arrival data). These orthonormal series estimates have been compared to a kernel density estimate. That comparison, however, was restricted to the visual inspection of the density curves.

In this section we will use the IBPP-plot as a graphical tool for assessing the quality of the fit of the proposed density estimates. We explain the general methodology in Section 6.6.1 and illustrate it on the Fastfood and the Arrival data in the Sections 6.6.2 and 6.6.3, respectively.

6.6.1 Construction

In the IBPP-plot we compare the empirical probabilities and the expected probabilities under the null hypothesis of the random variable X falling into the $\mathrm{arc}(x, y)$ or interval $[x, y]$. Suppose $x_1, \ldots, x_n$ are observations in $\mathcal{S}$. Then we computed

$$\mathbb{Z}_n(x,y) = \sqrt{n}\left(\hat{F}_n(x) - F_0(x,\boldsymbol{\beta}) - (\hat{F}_n(y) - F_0(y,\boldsymbol{\beta}))\right)$$

for each $\mathrm{arc}(x, y)$ or interval $[x, y]$ of $\mathcal{S}$, where $\boldsymbol{\beta}$ is replaced by its MLE if it is unknown.

When the null hypothesis is rejected according to the data-driven smooth test and an alternative density estimate $\hat{g}_k$ in (3.90) is obtained, we may compare the empirical probabilities and the expected probabilities under the alternative density estimate. In particular, instead of the process $\mathbb{Z}_n(x, y)$ we compute the process

$$\mathbb{S}_n(x,y) = \sqrt{n}\left(\hat{F}_n(x) - \hat{G}_k(x) - (\hat{F}_n(y) - \hat{G}_k(y))\right), \tag{6.13}$$

where $\hat{G}_k$ is the CDF of $\hat{g}_k$. Moreover,

$$K_n^i = \sup_{x,y\in\mathcal{S}} |\mathbb{S}_n(x,y)| \tag{6.14}$$

can be interpreted as a distance measure between the EDF and the estimated CDF. The smaller this distance measure, the better the density estimate $\hat{g}_k$ fits the data. Note that we could also replace $\hat{G}_k$ in (6.13) by any other estimate of the true CDF which is obtained from an appropriate density estimate. We can then plot the process values in (6.13) versus x and y in an explorative IBPP-plot for comparing the true distribution with the density estimate. We will refer to this plot as the *improved* IBPP-plot for the estimate $\hat{g}_k$. This term refers to the density estimate, which is in a sense an *improvement* to the hypothesised density $f_0(x, \boldsymbol{\beta})$. As the original IBPP-plot is sensitive to local alternatives to the null distribution, we expect the improved IBPP-plot to be sensitive to local deviations from the estimated distribution $\hat{G}_k$. We now illustrate the improved IBPP-plot on a linear as well as on a circular data example in the next two sections. The discussion is confined to the explorative version of the IBPP-plot since for the application of the formal version we need additional theoretical results on the limiting distribution of the process $\mathbb{S}_n(x, y)$, and thus is beyond the scope of this thesis.

6.6.2 Fastfood data

From Example 3.7.2 we know that the data-driven smooth test for composite normality based on the BIC yields a significant result for the Fastfood data. The test selected components up to the third order. To obtain the corresponding Hermite series density estimate we therefore need estimates for three parameters θ_j next to the MLEs for the parameters μ and σ. The MISE criterion also selected components up to the third order and therefore resulted in the same density estimate. In the same section, we have also given the kernel density estimate and the Legendre series estimates based on the BIC and the MISE criterion. For the Legendre series the BIC criterion selected components up to the third order while the MISE criterion selected components up to the fourth order.

The improved IBPP-plots of all these estimates are in Figure 6.25, while the original IBPP-plot is shown in Figure 6.24. All plots are drawn on the same colour scale to enable meaningful comparison. The minimum and maximum process values are indicated in each of the legends. The maximum value corresponds to the value of the statistics K_n^i and K_n in the improved and the original IBPP-plots, respectively. The largest among these values is obviously $K_n = 1.607$ corresponding to the original IBPP-plot. This plot shows a red region above the diagonal which indicates that the empirical probabilities are larger than the probabilities expected under normality. This happens for intervals including [90,150]. On the other hand, intervals including [150,350] induce negative values of the process, which indicates that the empirical probabilities are smaller than expected. This blue region is larger but less extreme. Such an impression is typical for a skewed deviation from the null distribution.

The improved IBPP-plots in Figure 6.25 now reveal to which extent the various density estimates modeled this skewness correctly. The maximum process values K_n^i for each of the density estimates are listed in increasing order as

Kernel UCV	1.12
Legendre series MISE	1.16
Hermite series BIC & MISE	1.18
Legendre series BIC	1.42.

From the corresponding improved IBPP-plot of the Legendre series BIC estimate, we see that the skewed impression is still present. Comparing both Legendre estimates, we may conclude that the inclusion of the order four term, as suggested by MISE criterion but not by BIC, makes a considerable difference.

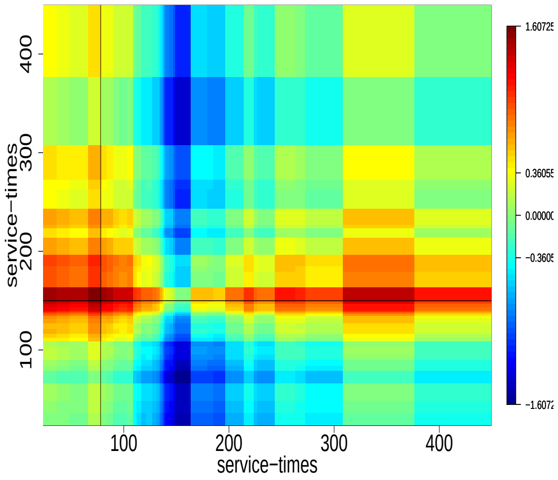


Figure 6.24: Explorative IBPP-plot for the Fastfood data.

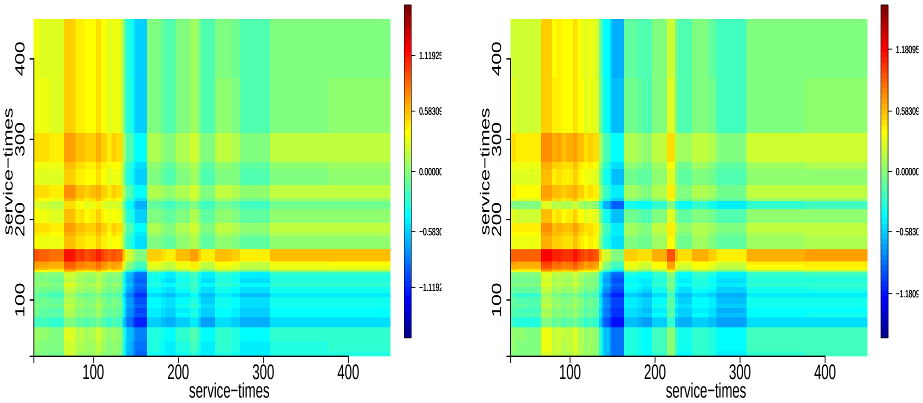
The kernel density estimate comes out as the estimate which best follows the data, but the differences with the Legendre series estimate based on the MISE criterion and the Hermite series estimate are minor. Note that this conclusion is only an explorative one, and no formal decision can be made from these plots.

6.6.3 Arrival Data

The improved IBPP-plots for the density curves fitted to the Arrival data in Section 4.6.4 are presented in Figures 6.26 and 6.27. For comparison purposes we show the original IBPP-plot in panel (a) of 6.26 and all plots are based on the same colour scale. The K_n^i value for the kernel density estimate is equal to 0.94, while those of the orthonormal series density estimates are

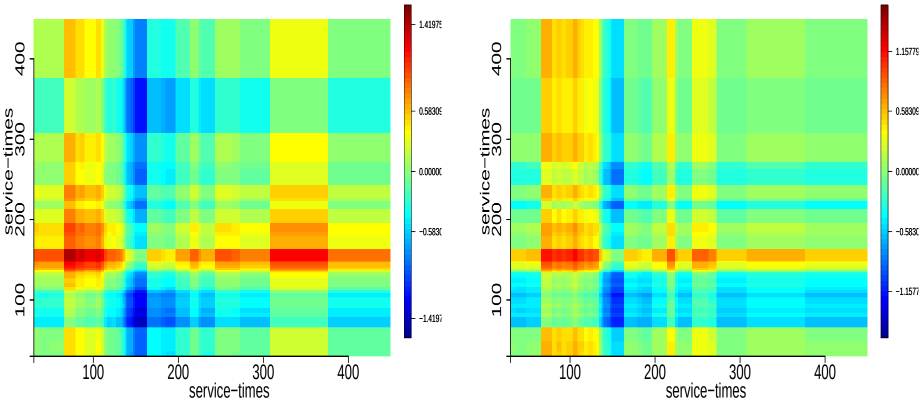
	CU	CN
BIC	0.96	1.05
AIC	0.71	0.72
MISE	0.96	0.95,

while the corresponding selected orders are



(a) Kernel UCV

(b) Hermite Series BIC and MISE



(c) Legendre BIC

(d) Legendre MISE

Figure 6.25: The explorative IBPP-plots for the nonparametric density estimates of the Fastfood data. (a) the kernel density estimator with bandwidth chosen by unbiased cross validation. (b) the Hermite series density estimator with BIC and MISE order selection criterion, and Legendre series estimator with BIC (c) and MISE (d) order selection criterion.

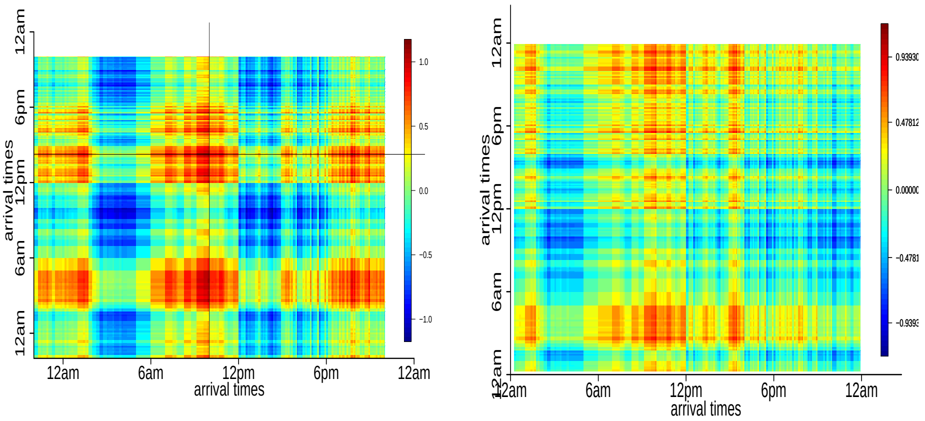
	CU	CN
BIC	1	2
AIC	5	5
MISE	1	3.

It is expected that, using either Legendre or Hermite series estimates, the estimate better approximates the data as the order of the series increases. Hence, as the table above shows, the K_n^i values decrease as the order increases for each of the two types of series estimators. The question is now two-fold. On the one hand, we are interested in whether there is a difference in performance between the two series estimates. In particular, is there a considerable difference in using either the CU or the CN distribution as a starting distribution in the series estimators? Moreover, what is their performance compared to the classical kernel density estimate? On the other hand, we want to know whether the inclusion of more parameters makes a considerable difference in the sense that less systematic deviation is present in the improved IBPP-plots.

Since the AIC criterion selected the largest order (5) for both series estimates it has also the smallest K_n^i values. On the corresponding improved IBPP-plots we see that there is a considerable difference between those estimates and e.g. the CU series estimates of order one. Even the selection of order two for the CN series estimate still shows some systematic deviations. The pattern of its corresponding improved IBPP-plot is similar to that of the kernel density estimate. Furthermore, we see that the pattern of the improved IBPP-plot of the the CN series with order 3 (MISE) is only slightly different from the series with AIC selection criterion.

To compare the CU series with the CN series, we first mention that for this example, the selected orders are different for the BIC and the MISE. The CN series choose higher orders, and therefore their corresponding improved IBPP-plots result in patterns that show less systematic error. This seems to confirm that choosing a starting distribution that approaches better the true distribution probably results in a better density estimate (see also Sections 3.7 and 4.5).

We may conclude from the IBPP-plots that the CN series estimate with the MISE criterion might be a good choice of density estimate, since the systematic deviation has considerably diminished and overfitting seems to have been avoided. The series estimates based on the AIC criterion, on the other hand, probably do suffer from overfitting, while the other improved IBPP-plots still display some systematic deviation between estimated distribution and true distribution.



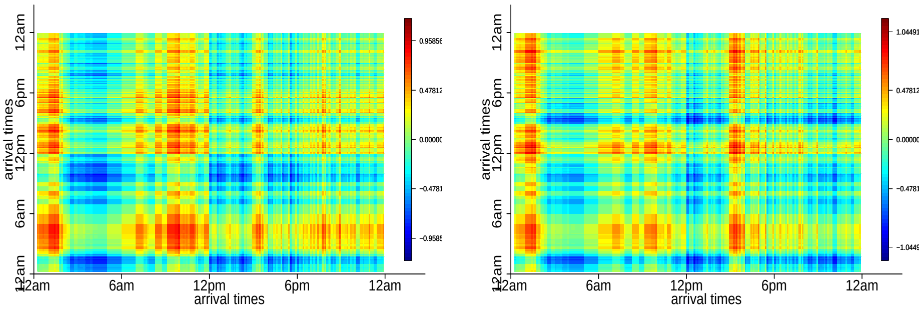
(a) (b)

Figure 6.26: The original IBPP-plot (a) and the IBPP-plot for checking the compatibility of the kernel density estimate (b) for the Arrival data.

6.7 Discussion

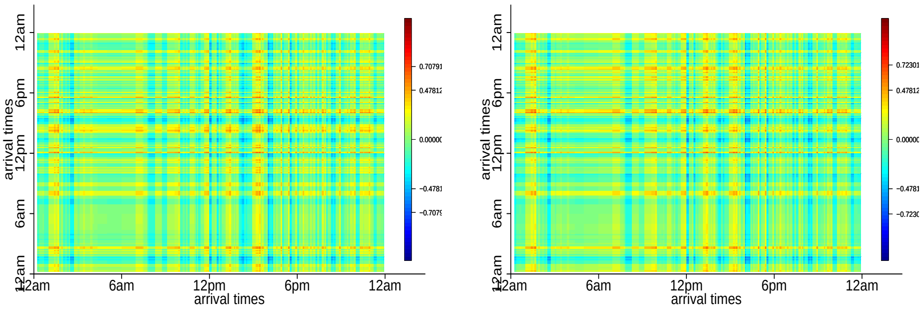
A graphical tool for LOF on a circle is developed, called the IBPP-plot. Two types of the plot, which is useful to detect and locate LOF are described. Both are based on the Kuiper test. Therefore they are origin-invariant and may be used as a formal diagnostic tool. A standardised version of the IBPP-plot is described as well. This plot has also good properties to localise small deviations with a high precision. However, the related standardised version of the Kuiper test has a considerable power loss to detect both local and global LOF.

The methods described in this chapter are applicable to linear data as well. Furthermore, it was illustrated how an adapted version of the IBPP-plot can be constructed to enable an explorative comparison between different density estimates.



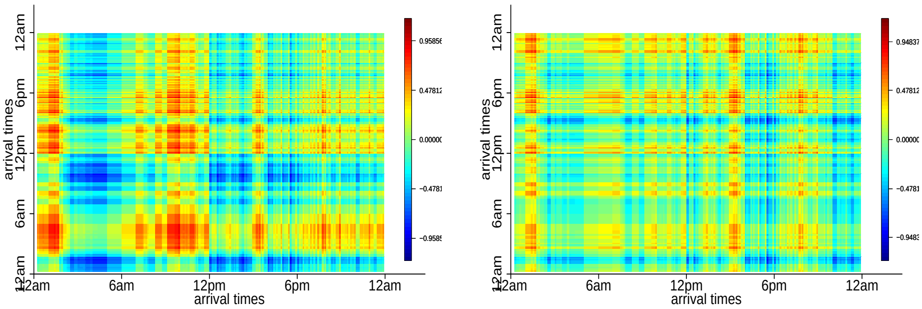
(a) CU series BIC

(b) CN series BIC



(c) CU series AIC

(d) CN series AIC



(e) CU series MISE

(f) CN series MISE

Figure 6.27: The IBPP-plots for checking the compatibility of the non-parametric density estimates for the Arrival data.

CHAPTER 7

Conclusions and further research

An important statistical question is whether a sample of observations agrees with a certain prespecified distribution or family of distributions. To deal with this kind of statistical problems, we can either use formal goodness-of-fit (GOF) tests or explorative graphical tools. It is however recommended to apply both simultaneously. The sample space from which the observations are drawn is usually the real line, but data on a circle also arise in many fields. GOF methods for this kind of data need to be origin-invariant, since their conclusions should not depend on the chosen origin.

In this thesis, three contributions to the statistical analysis of linear and circular data are presented. In this chapter we discuss the results together with possible further research topics.

Applying smooth tests to solve the GOF problem for linear distributions has the advantage that the components in the orthogonal decomposition of the corresponding score statistic often lead to easy interpretation and sum up to a test statistic with limiting omnibus features. The difficulty with the construction of smooth tests for circular distributions is to find appropriate orthonormal polynomials, because these are usually described in the complex field. Therefore, we have first defined the circular observations and the family of order k smooth alternatives on the field of complex numbers. Nevertheless, the latter family is constructed so that its functions are real-valued and proper densities. The score test for the complex parameters in the model is also based on a real-valued statistic which is asymptotically χ^2 distributed, and has an interpretation in terms of trigonometric moment deviations. This construction and the general theory of orthonormal polynomials on the unit circle (e.g. Simon, 2005) leads to a new class of smooth GOF tests for circular distributions. The tests are called the *complex smooth tests* since they are constructed using the “complex” framework described above. We have shown that the complex smooth model can be rewritten as a real smooth model for which the score statistic is equal to the former complex score statistic. Hence, in some sense this class of tests generalises the methodology of Rayner and Best (1989) for smooth tests on the real line.

Since we apply the test to circular data, the origin-invariance property needs to be checked. When the smooth test is not origin-invariant, we propose to subtract the circular mean direction from each observation before computing the test statistic.

For circular uniformity and circular normality we have given the explicit form of the smooth tests and its asymptotic distribution. In case of testing for circular uniformity we have explained how this construction leads to the smooth test of Bogdan et al. (2002). We have also shown that in case of testing for circular normality, our test generalizes the test proposed by Barndoff-Nielsen and Cox (1979).

Similarly as for a linear smooth test, the choice of the order of the family of alternatives in the smooth model is crucial to obtain optimal power. To overcome the problem of choosing the order, a data-driven version of the complex smooth test is discussed. Both the AIC and BIC selection rules have been considered to make an appropriate choice on the order of the complex smooth model, resulting in two versions of the data-driven smooth test for circular distributions against a general class of order k smooth alternatives. The parametric bootstrap was used to approximate the null distributions for the data-driven statistics. The complex data-driven smooth test for the CN distribution has been applied on real data examples. It has been demonstrated that, if the null hypothesis is rejected, the components of the smooth test may contain interesting information about how the true distribution deviates from the hypothesised. Some characteristics of the data-driven smooth test for circular normality have been investigated in a simulation study, which showed that it has good power against many different alternatives. In particular, the data-driven smooth test based on the AIC criterion has good power against higher order alternatives, while the test based on the BIC criterion has good power against low order alternatives.

In this thesis it has also been illustrated by means of an example how the application of the data-driven smooth test naturally leads to a nonparametric estimate of the true circular density. The result is essentially an orthonormal series density estimator, i.e. graph which can reveal how the true distribution deviates from the hypothesised. In that sense, the interpretation of the results from the test can be visualised.

Regarding future research perspectives, it would be interesting to study complex smooth tests for univariate circular distributions other than the CU or CN distribution, or for multivariate circular distributions. Developing a smooth test for multivariate circular distributions such as the von Mises-Fisher distribution on a multi-dimensional sphere, will also require a complex framework. Rayner and Best (1989) considered the smooth test for a multivariate normal distribution. Their approach can be generalised, resulting in a complex smooth test for multivariate circular normality. Furthermore, smooth tests for discrete circular distributions would be interesting as well since many circular data examples involve categorised data. Rayner and Best's (1989) test for categorised data can similarly be used as a basis to construct a complex smooth test for circular discrete distributions. The main issue for each of the previous generalisations is finding appropriate polynomials orthonormal to the hypothesised distributions. Also, origin-invariance will need to be checked and useful adaptations should be proposed if necessary. We presume that the adaptation will reduce to changing the origin to a sensible mean direction.

For the von Mises distribution with concentration parameter larger than 1, we have encountered problems with the smooth test statistic, in the sense that it could not easily be decomposed into orthogonal components. This problem was also reflected in the estimation of the parameters in the orthonormal series density estimator. We therefore aim to look for methods to find the parameter estimates such that the density estimate easily follows. If we find a solution to this problem, it will immediately enable us to interpret how the true distribution deviates from the hypothesised from the visual inspection of the density estimate.

We have also presented some new results on the integral version of the class of GOF tests proposed by Thas (2001). The tests are constructed by integrating out the Pearson χ^2 statistic over all possible partitions of the sample space in c cells. The degrees of freedom of Pearson's statistic are directly related to the indexing parameter of our new class, the SSP size c. The resulting tests are therefore called the linear SSPc tests. The tests are generalisations of the Anderson-Darling test, which is included in the class by taking $c = 2$. The methodology may be used to obtain similar extensions to the tests proposed by Zhang (2002) and Einmahl and McKeague (2003).

The construction of the linear SSPc test statistics and their asymptotic null distributions were given, and omnibus consistency was proved. We have written the linear SSP3 statistic as a V-statistic so as to find an appropriate decomposition in terms of Legendre polynomials. This decomposition led to the limiting distribution of the statistic under contiguous alternatives. The limiting values of the linear SSP3 and SSP4 statistics under a particular family of local alternatives have been studied, from which we found that the statistics have relatively high values for "local" alternatives, i.e. alternatives which deviate from the null distribution in small intervals of the sample space only.

To avoid the problem of choosing the right value for the indexing parameter c, we have proposed a data-driven version of the test. Simulations confirmed that the selection rule succeeds quite well in selecting a good choice for c. The usefulness of those data-driven tests has also been demonstrated on real data examples. The power study for the linear SSPc tests indicated that a substantial power gain may result from choosing some $c > 2$. On the other hand, the study also showed that for some alternatives, the highest power is obtained with $c = 2$. The weight functions that are involved in the test statistic, as well as the limiting behaviour and the simulation results, suggest that the new tests are very sensitive to deviations from the hypothesised distribution F_0 in small intervals of the support of F_0. Furthermore, this sensitivity increases with increasing SSP size c.

Extensions to composite null hypothesis have been described as well. In particular, the theory of the new class of GOF tests to composite null hypotheses is based on the estimated empirical process. If the nuisance parameters are estimated by asymptotically linear estimators, this estimated empirical process converges weakly to a Gaussian process with known covariance function. However, the limiting Gaussian process is quite complicated. Therefore, we suggest using parametric bootstrap to obtain the approximate null distribution.

The new class of GOF tests for linear data has been extended to a similar class of GOF tests for circular data. This was done by making the class of statistics origin-invariant. We have simply integrated out all possible origins resultingin the origin-invariant class of statistics. The resulting type of tests is called the circular SSPc tests and reduces to Rothman's test (1972) if $c = 2$.

The limiting null distribution of the linear SSPc statistic has been derived and computational formulae for SSP size $c = 2, 3$ and 4 have been found. The formulae for the circular analogues were then found by rewriting the statistic as a linear combination of V-statistics. This methodology can be extended to any SSP size.

The data-driven version and its asymptotic theory are similar as in the linear case. A simulation study indicated that the circular SSPc test has the same power characteristics as the linear SSPc test, though the differences between their powers are less pronounced than in the linear case. The power study showed that the SSPc tests perform at least as good as their competitors for all alternatives considered.

Additional simulations for both the linear and the circular SSPc tests, as well as the data-driven versions would be desirable for the context of composite null hypotheses. Indeed, it would be nice to know whether the localising small sample properties remain valid in the composite case.

In the final part of the thesis we have developed and discussed the IBPP-plot, which is a useful graphical tool for detecting and localising LOF on the circle. Two types of plots have been described, both of which are based on the Kuiper statistic. Similarly as for the PP-plot, the IBPP-plot is thus related to a formal statistical test, which is particularly intersting since the results of that test can be derived from the graph. Hence, the conclusions obtained from that graph are objective. This is in contrast to most other graphial tools, which are merely explorative and hence subjective. A standardised version of the IBPP-plot has been described as well. However, the related standardised version of the Kuiper test shows a considerable power loss to detect both local and global LOF as compared to the unstandardised version. Further research is needed to study the standardised version and to examine whether alternative standardisation techniques may improve the procedure.

All methods for circular data are applicable to linear data as well. Furthermore, we have extended the use of the IBPP-plot and demonstrated that an adapted version enables explorative comparison between different appropriate density estimates.

A formal version of this type of IBPP-plot can be developed if more research is performed on the asymptotic properties of the process used for the construction of that plot.

APPENDIX A

Orthonormal polynomials

A.1 Hermite polynomials

The Hermite polynomials for the standard normal distribution are defined by

$$H_k(x) = (-1)^n e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \tag{A.1}$$

The first Hermite polynomials are,

$$\begin{aligned}
H_0(x) &= 1 \\
H_1(x) &= x \\
H_2(x) &= x^2 - 1 \\
H_3(x) &= x^3 - 3x \\
H_4(x) &= x^4 - 6x^2 + 3 \\
H_5(x) &= x^5 - 10x^3 + 15x \\
H_6(x) &= x^6 - 15x^4 + 45x^2 - 15
\end{aligned}$$

We have that

$$\int_{-\infty}^{\infty} H_k(x) H_l(x)\, e^{-x^2/2}\, dx = n!\sqrt{2\pi}\delta_{kl} \tag{A.2}$$

A.2 Legendre polynomials

The orthogonal Legendre polynomials on $[0, 1]$ are defined by

$$P_k(x) = \frac{1}{k!}\frac{d^k}{dx^k}(x(x-1))^k \tag{A.3}$$

or

$$P_k(x) = \sum_{l=0}^{k}(-1)^{k-l}\binom{k}{k-l}\binom{k+l}{k}x^l \tag{A.4}$$

The first Legendre polynomials are,

$$\begin{aligned}
P_0(x) &= 1\\
P_1(x) &= 2x-1\\
P_2(x) &= 16x^2-6x+1\\
P_3(x) &= 20x^3-30x^2+12x-1
\end{aligned}$$

Some properties of these polynomials are

$$\int_0^1 P_k(x)P_l(x)dx = 0, \ \ k \neq l \tag{A.5}$$

$$\int_0^1 P_k(x)dx = 0, \ \ l < k \tag{A.6}$$

$$\int_0^1 x^l P_k(x)dx = \frac{l!}{(l-k)!}\frac{l!}{(l+k+1)!}, \ \ l \geq k \tag{A.7}$$

$$\int_0^1 P_k^2(x)dx = \frac{1}{2k+1}, \ \ k \geq 0 \tag{A.8}$$

$$P_k(x) = (-1)^k P_k(1-x), k \geq 0 \tag{A.9}$$

and for all $0 < x < 1$ and $k \geq 1$

$$|P_k(x)| \geq \frac{1}{\sqrt{\pi k}}(x(1-x))^{-1/4}. \tag{A.10}$$

From (A.8) we obtain the orthonormal Legendre polynomials, denoted by $L_k, k = 0, 1, \ldots$

$$L_k(x) = \sqrt{2k+1}P_k(x) \tag{A.11}$$

APPENDIX B

Proofs

B.1 Proof of Theorem 5.1

Let

$$A_c = \int_0^1 ((1-x)^{c-1} + x^{c-1}) \frac{\mathbb{B}^2(x)}{x(1-x)} dx \tag{B.1}$$

and

$$U_c = \int_0^1 \int_0^1 \frac{(1-(x \vee y))^{c-2} - (x \wedge y)^{c-2}}{(1-(x \vee y)) - (x \wedge y)} \frac{(\mathbb{B}(x) - \mathbb{B}(y))^2}{|x-y|} dx dy \tag{B.2}$$

and $T_c = (c-1)A_c + \binom{c-1}{2} U_c$. The convergence of $T_{c,n}$ is obtained if we can prove that for the Skorokhod construction of $\mathbb{B}(.)$,

$$|T_{c,n} - T_c| \xrightarrow{p} 0 \text{ as } n \to \infty.$$

First, note that

$$|T_{c,n} - T_c| \leq (c-1)|A_{c,n} - A_c| + \binom{c-1}{2} |U_{c,n} - U_c|. \tag{B.3}$$

We will proceed by showing that both $|A_{c,n} - A_c|$ and $|U_{c,n} - U_c|$ are asymptotically negligible.

Since $(1-x)^{c-1} + x^{c-1} \leq 1$ for all $c \geq 2$ and $0 \leq x \leq 1$,

$$|A_{c,n} - A_c| \leq \left| \int_0^1 \frac{\mathbb{B}_n^2(x) - \mathbb{B}^2(x)}{x(1-x)} dx \right|.$$

Shorack and Wellner (1986) proved that $\left|\int_0^1 \frac{\mathbb{B}_n^2(x)-\mathbb{B}^2(x)}{x(1-x)}dx\right| \xrightarrow{p} 0$ as $n \to \infty$. Hence $|A_{c,n} - A_c| \xrightarrow{p} 0$.

Suppose $x \leq y$, then

$$\begin{aligned}\frac{(1-(x \vee y))^{c-2} - (x \wedge y)^{c-2}}{(1-(x \vee y)) - (x \wedge y)} &= x^{c-3} + x^{c-4}(1-y) + \ldots + (1-y)^{c-3} \\ &= \sum_{d=3}^{c} x^{c-d}(1-y)^{d-3}\end{aligned}$$

Hence,

$$U_{c,n} = \sum_{d=3}^{c} \int_0^1 \int_0^1 (x \wedge y)^{c-d}(1-(x \vee y))^{d-3}(1-|x-y|)\frac{(\mathbb{B}_n(x)-\mathbb{B}_n(y))^2}{|x-y|(1-|x-y|)}dxdy.$$

Since for all $c \geq 3$, all $d = 3, \ldots, c$ and all x, y, we have

$$(x \wedge y)^{c-d}(1-(x \vee y))^{d-3}(1-|x-y|) \leq 1,$$

it follows that

$$|U_{c,n} - U_c| \leq \left|\int_0^1 \int_0^1 \frac{(\mathbb{B}_n(x)-\mathbb{B}_n(y))^2 - (\mathbb{B}(x)-\mathbb{B}(y))^2}{|x-y|(1-|x-y|)}dxdy.\right| \tag{B.4}$$

Since Shorack and Wellner (1982) have shown that the right hand side of (B.4) converges to zero in probability, we find $|U_{c,n} - U_c| \xrightarrow{p} 0$. Finally, by (B.3), we conclude for the Skorokhod construction

$$|T_{c,n} - T_c| \xrightarrow{p} 0 \text{ as } n \to \infty$$

.

B.2 Proof of Theorem 5.2

Omnibus consistency of the test based on $T_{c,n}$ is easily established by recognising that for all finite $c \geq 2$ and all nested subsets $D_c \subset D_{c+1}$,

$$P_{c+1,n}(D_c) \geq P_{c,n}(D_{c+1})$$

with probability 1 (see e.g. Section 5.1 in Cressie & Read, 1984). It follows that also $T_{c+1,n} \geq T_{c,n}$ with probability 1. Since the AD test is omnibus consistent (in particular, $T_{2,n}$ becomes unbounded as $n \to \infty$), omnibus consistency of the SSPc test follows immediately.

B.3 Proof of Lemma 5.1

First, we have in (5.27)

$$\frac{u \vee v}{u \wedge v}\frac{1}{|u-v|} = \frac{1}{|u-v|}\Big(1+\frac{u\vee v}{u\wedge v}-1\Big) = \frac{1}{|u-v|}+\frac{1}{u\wedge v},$$

$$\frac{1-u\wedge v}{1-u\vee v}\frac{1}{|u-v|} = \frac{1}{|u-v|}\left(1+\frac{1-u\wedge v}{1-u\vee v}-1\right) = \frac{1}{|u-v|}+\frac{1}{1-u\vee v}$$

and

$$\mathbb{B}_x(u\wedge v)\mathbb{B}_y(u\wedge v)+\mathbb{B}_x(u\vee v)\mathbb{B}_y(u\vee v)-\mathbb{B}_x(u\wedge v)\mathbb{B}_y(u\vee v)-\mathbb{B}_x(u\vee v)\mathbb{B}_y(u\wedge v)$$

$$= (\mathbb{B}_x(u\wedge v)-\mathbb{B}_x(u\vee v))(\mathbb{B}_y(u\wedge v)-\mathbb{B}_y(u\vee v))$$

Hence, in (5.27)

$$\Psi_{\mathrm{SSP3}}(x,y) = \Phi(x,y)+\Omega(x,y), \tag{B.5}$$

where

$$\Phi(x,y) = \frac{1}{n}\int_0^1\int_0^1\left(\frac{\mathbb{B}_x(u\wedge v)\mathbb{B}_y(u\wedge v)}{u\wedge v}+\frac{\mathbb{B}_x(u\vee v)\mathbb{B}_y(u\vee v)}{1-u\vee v}\right)dudv$$

$$\Omega(x,y) = \frac{1}{n}\int_0^1\int_0^1(\mathbb{B}_x(u\wedge v)-\mathbb{B}_x(u\vee v))(\mathbb{B}_y(u\wedge v)-\mathbb{B}_y(u\vee v))\frac{1}{|u-v|}dudv$$

Let Φ_1 and Φ_2 be the first and the second term in $\Phi(x,y)$, respectively. Then, we write

$$\begin{aligned}
\Phi_1 &= \frac{1}{n}\int_0^1\int_0^1\frac{\mathbb{B}_x(u\wedge v)\mathbb{B}_y(u\wedge v)}{u\wedge v}dudv \\
&= 2\frac{1}{n}\int_0^1\int_0^v\frac{\mathbb{B}_x(u)\mathbb{B}_y(u)}{u}dudv \\
&= 2\int_0^1\int_0^v\frac{1}{u}(I(x\le u)-u)(I(y\le u)-u)dudv \\
&= 2\int_0^1\int_0^v\frac{1}{u}I(x\le u)I(y\le u)dudv+2\int_0^1\int_0^v ududv \\
&\quad -2\int_0^1\int_0^v I(x\le u)dudv-2\int_0^1\int_0^v I(y\le u)dudv.
\end{aligned}$$

Further

$$\begin{aligned}
\int_0^1 \int_0^v \frac{1}{u} I(x \leq u) I(y \leq u) du dv &= \int_{x\vee y}^1 \int_{x \vee y}^v \frac{1}{u} du dv \\
&= \int_{x\vee y} (\ln v - \ln(x \vee y) dv \\
&= -1 + x \vee y - \ln(x \vee y), \\
\int_0^1 \int_0^v u du dv &= \frac{1}{6}, \\
\int_0^1 \int_0^v I(x \leq u) &= \int_x^1 \int_x^v du dv = \frac{1}{2}(1-x)^2
\end{aligned}$$

and

$$\int_0^1 \int_0^v I(x \leq u) = \frac{1}{2}(1-y)^2.$$

Therefore,

$$\Phi_1 = 2\left(-1 + x \vee y - \ln(x \vee y) - \frac{1}{2}(1-x)^2 - \frac{1}{2}(1-y)^2 + \frac{1}{6}\right).$$

Similarly we find

$$\Phi_2 = 2\left(-1 + (1-x) \vee (1-y) - \ln((1-x) \vee (1-y)) - \frac{1}{2}x^2 - \frac{1}{2}y^2 + \frac{1}{6}\right).$$

As a result, we obtain the formula

$$\begin{aligned}
\Phi(x,y) &= 2\Big(-\ln(x \vee y) - \ln((1-x) \vee (1-y)) \\
&\quad + x \vee y + (1-x) \vee (1-y) + x(1-x) + y(1-y) - \frac{8}{3}\Big). \qquad \text{(B.6)}
\end{aligned}$$

Since for all $0 \leq x \leq 1$

$$\int_0^1 \ln(x \vee y) dy = x - 1, \int_0^1 x \vee y dy = \frac{1}{2}(1 + x^2) \text{ and } \int_0^1 y(1-y) dy = \frac{1}{6}$$

the function Φ possesses the property of degeneracy, i.e

$$\int_0^1 \Phi(x,y) dy = 0. \qquad \text{(B.7)}$$

Moreover, in (B.6) we see that

$$\Phi(x,y) = \Phi(y,x) \text{ and } \Phi(x,y) = \Phi(1-x, 1-y). \qquad \text{(B.8)}$$

Further, we rewrite $\Omega(x,y)$ in (B.5) as

$$
\begin{aligned}
\Omega(x,y) &= \frac{1}{n}\int_0^1\int_0^1 (\mathbb{B}_x(u\wedge v)-\mathbb{B}_x(u\vee v))(\mathbb{B}_y(u\wedge v)-\mathbb{B}_y(u\vee v))\frac{1}{|u-v|}dudv\\
&= \frac{2}{n}\int_0^1\int_0^v (\mathbb{B}_x(u)-\mathbb{B}_x(v))(\mathbb{B}_y(u)-\mathbb{B}_y(v))\frac{1}{v-u}dudv\\
&= 2\int_0^1\int_0^v \left(v-u-I(u<x\le v)\right)\left(v-u-I(u<y\le v)\right)\frac{1}{v-u}dudv\\
&= 2\int_0^1\int_0^v I(u<x\le v)I(u<y\le v)\frac{1}{v-u}dudv + 2\int_0^1\int_0^v (v-u)dudv\\
&- 2\int_0^1\int_0^v I(u<x\le v)dudv - 2\int_0^1\int_0^v I(u<y\le v)dudv
\end{aligned}
$$

Here

$$
\begin{aligned}
\int_0^1\int_0^v I(u<x\le v)I(u<y\le v)\frac{1}{v-u}dudv &= \int_{x\vee y}^1\int_0^{x\wedge y}\frac{1}{v-u}dudv\\
&= \int_{x\vee y}^1 (\ln v - \ln(v-x\wedge y))\\
&= |x-y|\ln|x-y| - x\vee y\ln(x\vee y)\\
&- ((1-x)\wedge(1-y))\ln((1-x)\wedge(1-y)),\\
\int_0^1\int_0^v (v-u)dudv &= \frac{1}{6},\\
\int_0^1\int_0^v I(u<x\le v)dudv &= \int_x^1\int_0^x du = x(1-x),\\
\int_0^1\int_0^v I(u<y\le v)dudv &= y(1-y).
\end{aligned}
$$

As a result

$$
\begin{aligned}
\Omega(x,y) \quad = \quad & 2\Bigg(|x-y|\ln|x-y|\\
& -x\vee y\ln(x\vee y) - ((1-x)\wedge(1-y))\ln((1-x)\wedge(1-y))\\
& -x(1-x)-y(1-y)+\frac{1}{6}\Bigg). \qquad \text{(B.9)}
\end{aligned}
$$

Again, for all $0\le x\le 1$

$$
\begin{aligned}
\int_0^1 |x-y|\ln|x-y|dy &= \frac{1}{2}\left(x^2\ln x + (1-x)^2\ln(1-x)\right) - \frac{1}{4}\left(x^2+(1-x)^2\right),\\
\int_0^1 x\vee y\ln(x\vee y)dy &= \frac{1}{2}x^2\ln x + \frac{1}{4}(x^2-1)
\end{aligned}
$$

Hence, the function $\Omega(x,y)$ is degenerate as well, i.e.

$$\int_0^1 \Omega(x,y)dy = 0. \tag{B.10}$$

Evidently, we have

$$\Omega(x,y) = \Omega(y,x) \text{ and } \Omega(x,y) = \Omega(1-x, 1-y). \tag{B.11}$$

From (B.5), (B.6) and (B.9), we have (5.28). Similarly, from (B.5), (B.7) and (B.10), it follows that $\Phi(x,y)$ is degenerate. Together with (B.8) and (B.11), we proved (5.29).

B.4 Proof of Lemma 5.2

In this proof the index SSP3 is omitted for notational comfort. As mentioned before we only need to compute the coefficients Ψ_{kl} in (5.41) for the kernel Ψ in (5.27). Note that these coefficients are invariant w.r.t. permutations of indices, i.e. $\Psi_{kl} = \Psi_{lk}$.

We will first prove that these coefficients are zero for all odd $k-l \geq 1$. This is done by rewriting the kernel Ψ in (5.40) as

$$\begin{aligned}\Psi(x,y) &= \sum_{k=1}^{\infty}(2k+1)^2\Psi_{kk}P_k(x)P_k(y)\\ &\quad + \sum_{1\leq k\neq l<\infty}(2k+1)(2l+1)\Psi_{kl}P_k(x)P_l(y)\end{aligned}$$

for which the second term can be rewritten as

$$\begin{aligned}&\frac{1}{2}\sum_{1\leq k\neq l<\infty}(2k+1)(2l+1)\Psi_{kl}[P_k(x)P_l(y)+P_k(y)P_l(x)]\\ =&\sum_{1\leq k<l<\infty}(2k+1)(2l+1)\Psi_{kl}[P_k(x)P_l(y)+P_k(y)P_l(x)]\\ =&\sum_{k=2}^{\infty}\sum_{l=1}^{k-1}(2k+1)(2l+1)\Psi_{kl}[P_k(x)P_l(y)+P_k(y)P_l(x)]\end{aligned}$$

Hence,

$$\begin{aligned}\Psi(x,y) &= \sum_{k=1}^{\infty}(2k+1)^2\Psi_{kk}P_k(x)P_k(y) \qquad (B.12)\\ &\quad +\sum_{k=2}^{\infty}\sum_{l=1}^{k-1}(2k+1)(2l+1)\Psi_{kl}[P_k(x)P_l(y)+P_k(y)P_l(x)]\end{aligned}$$

By the virtue of properties (5.29) and (A.9) we have

$$\begin{aligned}
\Psi_{kl} &= \int_0^1\int_0^1 \Psi(x,y)P_k(x)P_l(y)dxdy \\
&= (-1)^{k+l}\int_0^1\int_0^1 \Psi(1-x,1-y)P_k(1-x)P_l(1-y)dxdy \\
&= (-1)^{k+l}\int_0^1\int_0^1 \Psi(x,y)P_k(x)P_l(y)dxdy \\
&= (-1)^{k-l}\Psi_{kl}.
\end{aligned}$$

We see that

$$\text{for all odd } k-l \geq 1, \quad \Psi_{kl} = 0 \tag{B.13}$$

In the following, we will consider Ψ_{kl} for all even $k-l \geq 0$. By the definition of Ψ_{kl} (see (5.41)) we write

$$\begin{aligned}
\Psi_{kl} &= \int_0^1 P_k(x)\int_0^1 \Psi(x,y)P_l(y)dydx \\
&= 2\int_0^1 P_k(x)\int_0^1 |x-y|\ln|x-y|P_l(y)dydx \\
&\quad -2\int_0^1 P_k(x)\int_0^1 (x\vee y)\ln(x\vee y)P_l(y)dydx \\
&\quad -2\int_0^1 P_k(x)\int_0^1 ((1-x)\vee(1-y))\ln((1-x)\vee(1-y))P_l(y)dydx \\
&\quad +2\int_0^1 P_k(x)\int_0^1 (x\vee y+(1-x)\vee(1-y))P_l(y)dydx \\
&\quad -2\int_0^1 P_k(x)\int_0^1 (\ln(x\vee y)+\ln((1-x)\vee(1-y)))P_l(y)dydx. \qquad (\text{B.14})
\end{aligned}$$

Let $\Psi_1,\ldots,\Psi_5$ be each of five integrals in (B.14), respectively, i.e.

$$\Psi_{kl} = 2\Psi_1 - 2\Psi_2 - 2\Psi_3 + 2\Psi_4 - 2\Psi_5. \tag{B.15}$$

Here each integral is computed explicitly,

$$\begin{aligned}
\Psi_1 = &\int_0^1 P_k(x)\int_0^x (x-y)\ln(x-y)P_l(y)dydx \\
+ &\int_0^1 P_k(x)\int_x^1 (y-x)\ln(y-x)P_l(y)dydx
\end{aligned}$$

After the change of the integrand variables $x\to 1-x, y\to 1-y$ in the second line and by (A.9)

$$= 2\int_0^1 P_k(x)\int_0^x (x-y)\ln(x-y)P_l(y)dydx$$

After the change $y \to xy$

$$
\begin{aligned}
&= 2\int_0^1 P_k(x)x^2 \ln x \int_0^1 (1-y)P_l(xy)dydx \\
&+ 2\int_0^1 P_k(x)x^2 \int_0^1 (1-y)\ln(1-y)P_l(xy)dydx,
\end{aligned}
\tag{B.16}
$$

$$
\begin{aligned}
\Psi_2 + \Psi_3 = &\int_0^1 P_k(x)\int_0^x (x\vee y)\ln(x\vee y)P_l(y)dydx \\
+ &\int_0^1 P_k(x)\int_0^1 ((1-x)\vee(1-y))\ln((1-x)\vee(1-y))P_l(y)dydx
\end{aligned}
$$

After the change $x \to 1-x, y \to 1-y$ and by (A.9)

$$
\begin{aligned}
&= 2\int_0^1 P_k(x)\int_0^x (x\vee y)\ln(x\vee y)P_l(y)dydx \\
&= 2\int_0^1 P_k(x)x\ln x\int_0^x P_l(y)dydx \\
&+ 2\int_0^1 P_k(x)\int_x^1 y\ln yP_l(y)dydx \\
&= 2\int_0^1 P_k(x)x\ln x\int_0^x P_l(y)dydx \\
&- 2\int_0^1 P_k(x)\int_0^x y\ln yP_l(y)dydx
\end{aligned}
$$

After the change $y \to xy$

$$
\begin{aligned}
&= 2\int_0^1 P_k(x)x^2 \ln x \int_0^1 (1-y)P_l(xy)dydx \\
&- 2\int_0^1 P_k(x)x^2 \int_0^1 y\ln yP_l(xy)dydx,
\end{aligned}
\tag{B.17}
$$

$$\begin{aligned}\Psi_4 &= \int_0^1 P_k(x) \int_0^1 (x \vee y + (1-x) \vee (1-y)) P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) \int_0^1 x \vee y P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) x \int_0^x P_l(y) dy dx \\ &+ 2 \int_0^1 P_k(x) \int_x^1 y P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) x \int_0^x P_l(y) dy dx \\ &- 2 \int_0^1 P_k(x) \int_0^x y P_l(y) dy dx\end{aligned}$$

After the change $y \to xy$

$$= 2 \int_0^1 P_k(x) x^2 \int_0^1 (1-y) P_l(xy) dy dx, \tag{B.18}$$

$$\begin{aligned}\Psi_5 &= \int_0^1 P_k(x) \int_0^1 (\ln(x \vee y) + \ln((1-x) \vee (1-y))) P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) \int_0^1 \ln x \vee y P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) \ln x \int_0^x P_l(y) dy dx \\ &+ 2 \int_0^1 P_k(x) \int_x^1 \ln y P_l(y) dy dx \\ &= 2 \int_0^1 P_k(x) \ln x \int_0^x P_l(y) dy dx \\ &- 2 \int_0^1 P_k(x) \int_0^x \ln y P_l(y) dy dx\end{aligned}$$

After the change $y \to xy$

$$= -2 \int_0^1 P_k(x) x \int_0^1 \ln y P_l(xy) dy dx, . \tag{B.19}$$

From (B.14)-(B.16), it follows that for all even $k-l = 0, 2, 4, \dots$

$$\begin{aligned}\Psi_{kl} &= 4\int_0^1 P_k(x)x^2 \int_0^1 (1-y)\ln(1-y)P_l(xy)dydx \\ & \quad 4\int_0^1 P_k(x)x^2 \int_0^1 (y\ln y + 1 - y)P_l(xy)dydx \\ & \quad 4\int_0^1 P_k(x)x \int_0^1 \ln y P_l(xy)dydx. \end{aligned} \tag{B.20}$$

By the definition of the Legendre polynomials (see (A.4)), we have

$$P_l(xy) = \binom{2l}{l}x^l y^l - \binom{l}{1}\binom{2l-1}{l}x^{l-1}y^{l-1} + \binom{l}{2}\binom{2l-2}{l}x^{l-2}y^{l-2} + \dots \tag{B.21}$$

By substituting (B.21) in (B.20) and taking account of (A.6) we see that

$$\text{for all } 1 \le l \le k-4, \;\; \Psi_{kl} = 0, \tag{B.22}$$

for $l = k-2 \ge 1$

$$\begin{aligned}\Psi_{k,k-2} &= 4\binom{2k-4}{k-2}\int_0^1 x^k P_k(x) \int_0^1 (1-y)\ln(1-y)y^{k-2} \\ & \quad +4\binom{2k-4}{k-2}\int_0^1 x^k P_k(x) \int_0^1 (y\ln y + 1 - y)y^{k-2} \end{aligned} \tag{B.23}$$

and for $l = k \ge 1$

$$\begin{aligned}\Psi_{k,k} &= 4\binom{2k}{k}\int_0^1 P_k(x)x^{k+2}\int_0^1 (1-y)\ln(1-y)y^k dydx \\ & \quad -4\binom{k}{1}\binom{2k-1}{k}\int_0^1 P_k(x)x^{k+1}\int_0^1 (1-y)\ln(1-y)y^{k-1}dydx \\ & \quad +4\binom{k}{2}\binom{2k-2}{k}\int_0^1 P_k(x)x^{k}\int_0^1 (1-y)\ln(1-y)y^{k-2}dydx \\ & \quad +4\binom{2k}{k}\int_0^1 P_k(x)x^{k+2}\int_0^1 (y\ln y + 1 - y)y^k dydx \\ & \quad -4\binom{k}{1}\binom{2k-1}{k}\int_0^1 P_k(x)x^{k+1}\int_0^1 (y\ln y + 1 - y)y^{k-1}dydx \\ & \quad +4\binom{k}{2}\binom{2k-2}{k}\int_0^1 P_k(x)x^{k}\int_0^1 (y\ln y + 1 - y)y^{k-2}dydx \\ & \quad +4\binom{2k}{k}\int_0^1 P_k(x)x^{k+1}\int_0^1 \ln(y)y^k dydx \\ & \quad -4\binom{k}{1}\binom{2k-1}{k}\int_0^1 P_k(x)x^{k}\int_0^1 \ln(y)y^{k-1}dydx. \end{aligned} \tag{B.24}$$

In (B.23) and (B.24) all integrals are one-type. For example,

$$\int_0^1 \ln(1-y) y^k dy = -\frac{1}{k+1}\sum_{p=1}^{k+1}\frac{1}{p}, \quad k \geq 0$$

$$\int_0^1 \ln(y) y^k dy = -\frac{1}{(k+1)^2}, \quad k \geq 0$$

$$\int_0^1 (y\ln y + 1 - y) y^k dy = -\frac{1}{(k+1)^2(k+1)}, \quad k \geq 0$$

$$\int_0^1 ((1-y)\ln(1-y) + y\ln y + 1 - y) y^{k-2} dy = \begin{cases} -\frac{1}{k(k-1)}\sum_{p=2}^{k-1}\frac{1}{p}, & k \geq 3 \\ 0 & k = 2 \end{cases}$$

We use the formula (A.7) for $l = k, k+1$ and $k+2$. For example for $l = k$, it becomes

$$\int_0^1 P_k(x) x^k = \frac{k!}{(2k)!}\frac{k!}{2k+1}.$$

From (B.23) we have for $k \geq 3$

$$\begin{aligned} \Psi_{kk} &= 4\binom{2k-4}{k-2}\frac{k!}{2k!}\frac{k!}{2k+1}\left(-\frac{1}{(k-1)k}\sum_{p=2}^{k-1}\frac{1}{p}\right) \\ &= -\frac{1}{(2k-3)(2k-1)(2k+1)}\left(\sum_{p=2}^{k-1}\frac{1}{p}\right) \\ &= -\frac{\sigma_k}{(2k-3)(2k-1)(2k+1)}. \end{aligned} \tag{B.25}$$

From (B.24) we obtain for $k \geq 2$

$$\begin{aligned}
\Psi_{kk} &= 4\binom{2k}{k}\frac{(k+2)!(k+2)!}{2(2k+3)!}\left(-\frac{1}{(k+1)(k+2)}\sum_{p=2}^{k+1}\frac{1}{p}\right) \\
&\quad -4\binom{k}{1}\binom{2k-1}{k}\frac{(k+1)!(k+1)!}{(2k+2)!}\left(-\frac{1}{k(k+1)}\sum_{p=2}^{k}\frac{1}{p}\right) \\
&\quad +4\binom{k}{2}\binom{2k-2}{k}\frac{k!k!}{(2k)!(2k+1)}\left(-\frac{1}{(k-1)k}\sum_{p=2}^{k-1}\frac{1}{p}\right) \\
&\quad -4\frac{(2k)!}{k!k!}\frac{(k+1)!(k+1)!}{(2k+2)!}\frac{1}{(k+1)^2} \\
&\quad +4\frac{k!}{(k-1)!}\frac{(2k-1)!}{(k-1)!k!}\frac{k!k!}{(2k)!(2k+1)}\frac{1}{k^2} \\
&= -\frac{k+2}{(2k+1)(2k+3)}\left(\sum_{p=2}^{k+1}\frac{1}{p}\right)+\frac{1}{2k+1}\left(\sum_{p=2}^{k}\frac{1}{p}\right) \\
&\quad -\frac{k-1}{(2k-1)(2k+1)}\left(\sum_{p=2}^{k-1}\frac{1}{p}\right)+\frac{2}{k(k+1)(2k+1)} \\
&= -\frac{k+2}{(2k+1)(2k+3)}\left(\sigma_k+\frac{1}{k}+\frac{1}{k+1}\right) \\
&\quad +\frac{1}{2k+1}\left(\sigma_k+\frac{1}{k}\right)-\frac{k-1}{(2k-1)(2k+1)}\sigma_k+\frac{2}{k(k+1)(2k+1)} \\
&= \left(\frac{1}{2k+1}-\frac{k+2}{(2k+1)(2k+3)}-\frac{k-1}{(2k-1)(2k+1)}\right)\sigma_k \\
&\quad +\frac{1}{k(2k+1)}-\frac{k+2}{k(2k+1)(2k+3)}-\frac{k+2}{(k+1)(2k+1)(2k+3)} \\
&\quad +\frac{2}{k(k+1)(2k+1)} \\
&= \frac{2\sigma_k}{(2k-1)(2k+1)(2k+3)}+\frac{4k+7}{k(k+1)(2k+1)(2k+3)} \qquad \text{(B.26)}
\end{aligned}$$

From (B.24) we obtain for $k = 1$ that

$$\Psi_{11} = \frac{1}{3} - \frac{1}{10} \qquad \text{(B.27)}$$

From (B.12)-(B.13) and (B.22) it follows

$$\begin{aligned}\Psi(x,y) &= \sum_{k=1}^{\infty}(2k+1)^2\Psi_{kk}P_k(x)P_k(y)\\ &= +\sum_{k=3}^{\infty}(2k+1)(2k-3)\Psi_{k,k-2}(P_k(x)P_{k-2}(y)+P_{k-2}(x)P_k(y))\end{aligned}$$

after the change $k \to k+2$ in the second sum

$$\begin{aligned}&= \sum_{k=1}^{\infty}(2k+1)^2\Psi_{kk}P_k(x)P_k(y)\\ &= \sum_{k=1}^{\infty}(2k+5)(2k+1)\Psi_{k+2,k}(P_k(x)P_{k+2}(y)+P_{k+2}(x)P_k(y))\end{aligned}$$

by virtue of (B.25) and (B.26)

$$\begin{aligned}&= 9\frac{7}{30}P_1(x)P_1(y) + \sum_{k=2}^{\infty}\left(\frac{(2k+1)(4k+7)}{k(k+1)(2k+3)}\right)P_k(x)P_k(y)\\ &-\sum_{k=1}^{\infty}\frac{\sigma_{k+2}}{2k+3}(P_k(x)P_{k+2}(y)+P_{k+2}(x)P_k(y)).\end{aligned}$$

This leads to (5.48) and hence Lemma 5.2 is proved.

B.5 Proof of Theorem 5.7

$$\mathrm{P}\left[C_n \neq c_m\right] \quad = \quad \sum_{c\in\Gamma\backslash\{c_m\}} \mathrm{P}\left[C_n = c\right].$$

Next we make use of the characteristic that a selected order equal to c implies that the order c *beats* the minimal order c_m, and hence $T_{c,n} - 2(c-1)\ln a_n > T_{c_m,n} - 2(c_m-1)\ln a_n$. Let $d = (c-1)-(c_m-1)$. Then,

$$\begin{aligned}\mathrm{P}\left[C_n \neq c_m\right] \quad &\leq \quad \sum_{c\in\Gamma\backslash\{c_m\}} P\left[T_{c,n} - 2(c-1)\ln a_n > \right.\\ &\qquad\qquad \left. T_{c_m,n} - 2(c_m-1)\ln a_n\right]\\ &\leq \quad \sum_{c\in\Gamma\backslash\{c_m\}} \mathrm{P}\left[T_{c,n} - T_{c_m,n} > 2d\ln a_n\right]\\ &\leq \quad \sum_{c\in\Gamma\backslash\{c_m\}} \mathrm{P}\left[T_{c,n} > 2d\ln a_n\right],\end{aligned}$$

where in the last step we made use of the fact that $T_{c_m,n} \geq 0$. Let $\mu_{c,n} = \mathrm{E}_0\left[T_{c,n}\right]$ and $\nu_{c,n} = \mathrm{Var}_0\left[T_{c,n}\right]$ (the index 0 refers to the hypothesised distribution F_0). Then, we continue by subtracting $\mu_{c,n}$, taking the absolute value and applying Chebychev's inequality.

$$\begin{aligned} \mathrm{P}\left[C_n \neq c_m\right] &\leq \sum_{c\in\Gamma\setminus\{c_m\}} \mathrm{P}\left[T_{c,n} - \mu_{c,n} > 2d\ln a_n - \mu_{c,n}\right] \\ &\leq \sum_{c\in\Gamma\setminus\{c_m\}} \mathrm{P}\left[\left|T_{c,n} - \mu_{c,n}\right| > 2d\ln a_n - \mu_{c,n}\right] \\ &\leq \sum_{c\in\Gamma\setminus\{c_m\}} \frac{\nu_{c,n}}{\left(2d\ln a_n - \mu_{c,n}\right)^2}. \end{aligned} \tag{B.28}$$

From the definition of $T_{c,n}$ (Equation 5.8) it is seen that $\mu_{c,n} \to (c-1)$ and that $\nu_{c,n} \to (c-1)\sigma$ ($\mathrm{Var}_g\left[T_{2,n}\right] \to \sigma$ as $n \to \infty$) as $n \to \infty$. σ is finite and c is assumed to be finite. Furthermore is was assumed that $a_n \to \infty$ as $n \to \infty$. Hence each term in Equation B.28 converges to zero. Since $\#\Gamma$ is finite, only a finite number of terms appear in Equation B.28. Thus, we have

$$\mathrm{P}\left[C_n \neq c_m\right] \to 0 \tag{B.29}$$

as $n \to \infty$, which completes the proof.

B.6 Proof of Theorem 5.8

We proceed as Inglot et al. (1997). Under H_0, for all $x \in \mathcal{S}$,

$$\begin{aligned} \mathrm{P}\left[T_{C_n,n} \leq x\right] &= \mathrm{P}\left[T_{C_n,n} \leq x, C_n = c_m\right] + \mathrm{P}\left[T_{C_n,n} \leq x, C_n \neq c_m\right] \\ &= \mathrm{P}\left[T_{C_n,n} \leq x | C_n = c_m\right]\mathrm{P}\left[C_n = c_m\right] + \mathrm{P}\left[T_{C_n,n} \leq x, C_n \neq c_m\right] \end{aligned}$$

Thus, with $\mathrm{P}\left[T_{C_n,n} \leq x | C_n = c_m\right] = \mathrm{P}\left[T_{c_m,n} \leq x\right]$,

$$\mathrm{P}\left[T_{C_n,n} \leq x\right] = \mathrm{P}\left[T_{c_m,n} \leq x\right]\mathrm{P}\left[C_n = c_m\right] + \mathrm{P}\left[T_{C_n,n} \leq x, C_n \neq c_m\right],$$

where, by Equation B.29, the last term tends to zero, and $\mathrm{P}\left[C_n = c_m\right] \to 1$ as $n \to \infty$. Thus, the result follows immediately from the asymptotic null distribution of $T_{c,n}$ with $c = c_m$.

APPENDIX C

Datasets

C.1 Lottery data

162	671	933	414	788	730	817	33	536
875	670	236	473	167	877	980	316	950
456	92	517	557	956	954	104	178	794
278	147	773	437	435	502	610	582	780
689	562	964	791	28	97	848	281	858
538	660	972	671	613	867	448	738	966
139	636	847	659	754	243	122	455	195
968	793	59	730	361	574	522	97	762
431	158	429	414	22	629	788	999	187
215	810	782	47	34	108	986	25	644
829	630	315	567	919	331	207	412	242
607	668	944	749	168	864	442	533	805
372	63	458	777	416	340	436	140	919
350	510	572	905	900	85	389	473	758
444	169	625	692	140	897	672	288	312
860	724	226	884	508	976	741	476	417
831	15	318	432	241	114	799	955	833
358	935	146	630	830	440	642	356	373
271	715	367	393	190	669	8	861	108
795	269	590	326	866	64	523	862	840
219	382	998	4	628	305	747	247	34
747	729	645	856	974	24	568	24	694
608	480	410	729	947	293	53	930	223
203	677	227	62	455	387	318	562	242
428	968							

C.2 Lew data

-213	-564	-35	-15	141	115	-420	-360
203	-338	-431	194	-220	-513	154	-125
-559	92	-21	-579	-52	99	-543	-175
162	-457	-346	204	-300	-474	164	-107
-572	-8	83	-541	-224	180	-420	-374
201	-236	-531	83	27	-564	-112	131
-507	-254	199	-311	-495	143	-46	-579
-90	136	-472	-338	202	-287	-477	169
-124	-568	17	48	-568	-135	162	-430
-422	172	-74	-577	-13	92	-534	-243
194	-355	-465	156	-81	-578	-64	139
-449	-384	193	-198	-538	110	-44	-577
-6	66	-552	-164	161	-460	-344	205
-281	-504	134	-28	-576	-118	156	-437
-381	200	-220	-540	83	11	-568	-160
172	-414	-408	188	-125	-572	-32	139
-492	-321	205	-262	-504	142	-83	-574
0	48	-571	-106	137	-501	-266	190
-391	-406	194	-186	-553	83	-13	-577
-49	103	-515	-280	201	300	-506	131
-45	-578	-80	138	-462	-361	201	-211
-554	32	74	-533	-235	187	-372	-442
182	-147	-566	25	68	-535	-244	194
-351	-463	174	-125	-570	15	72	-550
-190	172	-424	-385	198	-218	-536	96

C.3 Chemical concentration data

452	184	115	315	139	177	214	356
166	246	177	289	175	296	205	324
260	188	208	109	204	89	320	256
138	198	191	193	316	122	305	203
396	250	230	214	46	256	204	150
218	261	143	229	173	132	175	236
220	212	119	144	147	171	216	232
216	164	185	216	199	236	237	206
87							

C.4 Fastfood data

54	108	115	129	92	138	43	141	110	118
78	88	340	230	138	177	150	125	80	148
205	413	276	146	188	99	134	30	182	223
135	269	224	257						

C.5 Old Faithful geyser data

3.600	1.800	3.333	2.283	4.533	2.883	4.700	3.600	1.950	4.350	1.833	3.917
4.200	1.750	4.700	2.167	1.750	4.800	1.600	4.250	1.800	1.750	3.450	3.067
4.533	3.600	1.967	4.083	3.850	4.433	4.300	4.467	3.367	4.033	3.833	2.017
1.867	4.833	1.833	4.783	4.350	1.883	4.567	1.750	4.533	3.317	3.833	2.100
4.633	2.000	4.800	4.716	1.833	4.833	1.733	4.883	3.717	1.667	4.567	4.317
2.233	4.500	1.750	4.800	1.817	4.400	4.167	4.700	2.067	4.700	4.033	1.967
4.500	4.000	1.983	5.067	2.017	4.567	3.883	3.600	4.133	4.333	4.100	2.633
4.067	4.933	3.950	4.517	2.167	4.000	2.200	4.333	1.867	4.817	1.833	4.300
4.667	3.750	1.867	4.900	2.483	4.367	2.100	4.500	4.050	1.867	4.700	1.783
4.850	3.683	4.733	2.300	4.900	4.417	1.700	4.633	2.317	4.600	1.817	4.417
2.617	4.067	4.250	1.967	4.600	3.767	1.917	4.500	2.267	4.650	1.867	4.167
2.800	4.333	1.833	4.383	1.883	4.933	2.033	3.733	4.233	2.233	4.533	4.817
4.333	1.983	4.633	2.017	5.100	1.800	5.033	4.000	2.400	4.600	3.567	4.000
4.500	4.083	1.800	3.967	2.200	4.150	2.000	3.833	3.500	4.583	2.367	5.000
1.933	4.617	1.917	2.083	4.583	3.333	4.167	4.333	4.500	2.417	4.000	4.167
1.883	4.583	4.250	3.767	2.033	4.433	4.083	1.833	4.417	2.183	4.800	1.833
4.800	4.100	3.966	4.233	3.500	4.366	2.250	4.667	2.100	4.350	4.133	1.867
4.600	1.783	4.367	3.850	1.933	4.500	2.383	4.700	1.867	3.833	3.417	4.233
2.400	4.800	2.000	4.150	1.867	4.267	1.750	4.483	4.000	4.117	4.083	4.267
3.917	4.550	4.083	2.417	4.183	2.217	4.450	1.883	1.850	4.283	3.950	2.333
4.150	2.350	4.933	2.900	4.583	3.833	2.083	4.367	2.133	4.350	2.200	4.450
3.567	4.500	4.150	3.817	3.917	4.450	2.000	4.283	4.767	4.533	1.850	4.250
1.983	2.250	4.750	4.117	2.150	4.417	1.817	4.467				

C.6 Birth time data

7.02pm,	11.08pm,	3.56am,	8.12pm,	8.40am,	12.25pm,
1.24am,	8.25am,	2.02pm,	11.46pm,	10.07am,	1.53pm,
6.45pm,	9.06am,	3.57pm,	7.40am,	3.02am,	10.45am,
3.06pm,	6.26am,	4.44pm,	12.26am,	2.17pm,	11.45pm,
5.08am,	5.49am,	6.32am,	12.40pm,	1.30pm,	12.55pm,
3.22pm,	4.09pm,	7.46pm,	2.28am,	10.06am,	11.19am,
4.31pm					

C.7 Homing pigeons data

$$20^\circ, 135^\circ, 145^\circ, 165^\circ, 170^\circ, 200^\circ, 300^\circ, 325^\circ, 335^\circ, 350^\circ, 350^\circ, 350^\circ \text{ and } 355^\circ$$

C.8 Turtles data

8	9	13	13	14	18	22	27	30	34
38	38	40	44	45	47	48	48	48	48
50	53	56	57	58	58	61	63	64	64
64	65	65	68	70	73	78	78	78	83
83	88	88	88	90	92	92	93	95	96
98	100	103	106	113	118	138	153	153	155
204	215	223	226	237	238	243	244	250	251
257	268	285	319	343	350				

C.9 Ants data

330	290	60	200	200	180	280	220	190	180
180	160	280	180	170	190	180	140	150	150
160	200	190	250	180	30	200	180	200	350
200	180	120	200	210	130	30	210	200	230
180	160	210	190	180	230	50	150	210	180
190	210	220	200	60	260	110	180	220	170
10	220	180	210	170	90	160	180	170	200
160	180	120	150	300	190	220	160	70	190
110	270	180	200	180	140	360	150	160	170
140	40	300	80	210	200	170	200	210	190

C.10 Direzione data

356.800	59.180	8.600	41.250	126.100	26.380	36.370	82.900	21.460
111.600	4.478	8.690	18.720	5.213	5.299	356.700	2.718	351.600
359.800	353.300	357.300	345.000	351.700	358.300	0.452	308.100	303.700
322.300	43.980	76.500	351.900	12.700	357.100	133.600	207.000	28.190
350.900	0.833	0.125	26.320	12.700	30.490	25.990	165.000	49.580
8.130	6.603	3.280	5.688	51.230	351.600	5.484	351.900	8.510
339.000	7.500	61.220	45.910	155.800	123.500	38.050	46.320	3.608
4.490	6.423	11.390	2.114	6.314	348.000	19.430	5.219	3.133
11.700	341.200	3.645	30.120	357.100	14.920	0.574	1.482	6.545
47.930	85.600	25.340	91.000	22.070	1.016	11.540	38.880	357.700
44.730	35.070	13.340	1.061	7.380	34.240	9.330	26.500	5.166
55.030	48.050	21.930	4.281	347.900	89.100	8.100	1.805	323.700
123.100	59.040	28.840	42.140	0.361	350.300	348.800	42.790	46.600
113.200	137.900	97.700	307.200	314.100	312.300	320.000	304.600	303.400
324.100	318.000	333.100	356.400	9.910	2.213	12.970	23.490	0.811
357.400	352.500	31.290	11.760	353.600	89.900	138.400	122.400	10.460
1.063	13.420	333.200	329.700	7.940	344.700	59.640	136.200	99.100
345.000	11.550	4.690	1.147	2.585	87.200	14.900	12.210	11.510
22.840	354.700	20.370	358.800	300.800	323.600	356.700	350.400	359.200
7.390	33.130	143.100	199.700	115.000	117.100	146.300	23.970	9.290
356.200	10.710	352.900	5.984	10.870	109.000	51.060	359.300	6.938
12.510	345.200	186.300	178.900	283.100	12.680	198.200	198.200	151.500
87.100	205.600	12.670	23.350	110.600	85.400	81.700	354.900	335.900
355.200	346.100	36.950	349.900	356.300	353.700	327.600	340.100	323.400
261.500	356.800	339.500	148.400	101.900	17.700	18.430	12.520	39.960
94.000	31.000	110.600	19.750	355.300	344.400	332.600	343.700	343.600
337.500	205.200	355.800	32.840	3.879	318.700	29.130	12.640	359.000
349.300	14.230	29.630	265.500	138.000	61.690	74.100	159.500	196.300
167.300	341.900	2.127	89.300	98.200	112.700	99.600	354.500	352.300
307.700	17.330	60.360	150.900	13.750	13.260	167.300	137.500	58.970
8.450	2.462	27.370	5.342	15.150	289.500	141.300	26.380	17.170
29.370	296.300	95.500	23.000	52.730	45.890	332.800	10.270	34.400
21.840	6.574	313.300	21.190	9.630	352.400	354.100	335.700	58.580
34.920	291.500	105.100	6.438	327.400	36.820	65.980	2.350	10.640
29.230	9.320	47.910	34.740					

C.11 Arrival data

11.00	17.00	23.15	10.00	12.00	08.45	16.00	10.00	15.30	20.20
04.00	12.00	02.20	12.00	05.30	07.30	12.00	16.00	16.00	01.30
11.05	16.00	19.00	17.45	20.20	21.00	12.00	12.00	18.00	22.00
22.00	22.05	12.45	19.30	18.45	16.15	16.00	20.30	23.40	20.20
18.45	16.30	22.00	08.45	19.15	15.30	12.00	18.15	14.00	13.00
23.00	19.15	22.00	10.15	12.30	18.15	21.05	21.00	00.30	01.45
12.20	14.45	22.30	12.30	13.15	17.30	11.20	17.30	23.00	10.55
13.30	11.00	18.30	11.05	04.00	07.30	20.00	21.30	06.30	17.30
20.45	22.00	20.15	21.00	17.30	19.50	02.00	01.45	03.40	04.15
23.55	03.15	19.00	21.45	21.30	00.45	02.30	15.30	21.00	08.45
14.30	17.00	03.30	15.45	17.30	14.00	02.00	11.30	17.30	17.10
21.20	03.00	13.30	23.00	20.10	23.15	20.00	16.00	18.30	21.00
21.10	17.00	13.25	15.05	14.10	19.15	14.05	22.40	09.30	17.30
12.30	17.30	14.30	16.00	14.10	14.00	15.30	04.30	11.50	11.55
15.20	15.40	11.15	02.15	11.15	21.30	03.00	00.40	10.00	09.45
23.45	10.00	07.50	13.30	12.30	13.45	19.30	00.15	07.45	15.20
18.40	19.50	23.55	01.45	10.50	07.50	15.30	18.00	23.05	19.30
19.00	16.10	10.00	02.30	22.00	21.50	19.10	11.45	15.45	16.30
18.30	10.05	20.00	13.35	16.45	02.15	20.30	14.00	21.15	18.45
14.05	14.15	01.15	01.45	18.00	14.15	15.15	16.15	10.20	13.35
17.15	19.50	22.45	07.25	17.00	12.30	23.15	10.30	13.45	02.30
12.00	15.45	17.00	17.00	01.30	20.15	12.30	15.40	03.30	18.35
13.30	16.40	18.00	20.00	11.15	16.40	13.55	21.00	07.45	22.30
16.40	23.10	19.15	11.00	00.15	14.40	15.45	12.45	17.00	18.00
21.45	16.00	12.00	02.30	12.55	20.20	10.30	15.50	17.30	20.00
02.00	01.45	01.45	02.05						

Bibliography

Abrahamson, I. G. (1967). Exact bahadur efficiencies for the kolmogoros-smirnov and kuiper one- and two-sample statistics. *The annals of mathematical statistics*, *38*(5), 1475-1490.

Agostinelli, C. (2006). Robust estimation for circular data. *Computational statistics and data analysis,, doi:10.1016/j.csda.2006.11.02*.

Ahmad, I. A. (1993). Modification of some goodness-of-fit statistics to yield asymptotically normal null distributions. *Biometrika*, *80*(2), 466-472.

Anderson, G. L., & de Figueiredo, R. J. P. (1980). An adaptive orthogonal-series estimator for probability density functions. *The annals of statistics*, *8*(2), 347-376.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *The annals of mathematical statistics*, *23*(2), 193-212.

Anje, B. (1968). A simple test for uniformity of a circular distribution. *Biometrika*, *55*(2), 343-354.

Azzalini, A., & Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied statistics*, *39*, 357-365.

Babu, G. J., & Rao, C. R. (2004). Goodness-of-fit when parameters are estimated. *Skankhyā*, *66*(1), 63-74.

Barndoff-Nielsen, O., & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the royal statistical society Series B (Methodological)*, *41*(3), 279-312.

Barr, D. R., & Shudde, R. H. (1973). A note on the kuiper's v_n statistic. *Biometrika*, *60*(3), 663-664.

Barton, D. (1955). A form of neyman's ψ_k^2 test of goodness-of-fit applicable to grouped and discrete data. *Skandinavisk aktuarietidskrift*, *38*, 1-16.

Barton, D. E. (1953). On neyman's smooth test of goodness-of-fit and its power with respect to a particular system of alternatives. *Skandinavisk aktuarietidskrift*, *36*, 24-63.

Batschelet, E. (1981). *Circular statistics in biology.* Academic press, London.

Best, D. J., & Fisher, N. I. (1981). The bias of the maximum likelihood estimators of the von mises-fisher concentration parameters. *Communications in statistics, simulation and computation*, *10*, 493-502.

Bogdan, M., Bogdan, K., & Futschik, A. (2002). A data-driven smooth test for circular data. *Annals of the institute of statistical mathematics*, *54*(1), 29-44.

Buckland, S. T. (1992). Fitting density functions with polynomials. *Applied statistics*, *41*(1), 63-76.

Cencov, N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.*, *3*, 1559-1562.

Chihara, T. (1978). *An introduction to orthogonal polynomials.* New York: Gordon and Breach.

Claeskens, G., & Hjort, N. L. (2004). Goodness of fit via non-parametric likelihood ratios. *Scandinavian journal of statistics*, *31*, 487-513.

Clutton-Brock, M. (1990). Density estimation using exponentials of orthogonal series. *Journal of the american statistical association*, *85*(411), 760-764.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk aktuarietidskrift*, *11*(13-74), 141-180.

Cressie, N., & Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the royal statistical society Series B*, *46*, 440-464.

D' Agostino, R., & Stephens, M. (1986). *Goodness-of-fit techniques.* New York: USA: Marcel Dekker.

Dallal, G. E. (1986, November). An analytical approximation to the distribution of lilliefors test statistic for normality. *The american statistician*, *40*(4), 294-296.

Diggle, P. J., & Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the american statistical association*, *81*(393), 230-233.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley and Sons, Inc.

Drew, J. H., Glen, A. G., & Leemis, L. M. (1998). Computing the cumulative distribution function of the kolmogorov-smirnov statistic. *Computational statistics and data analysis*, *34*, 1-15.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The annals of statistics*, *1*(2), 279-290.

Durbin, J., & Knott, M. (1972). Components of cramér-von mises statistics i. *Journal of the royal statistical society Series B*, *34*, 290-307.

Einmahl, J., & McKeague, I. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, *9*, 267-290.

Fernández-Durán, J. J. (2004). Circular distribution based on nonnegative trigoniometric sums. *Biometrics*, *60*, 499-503.

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

Fisher, R. (1924). The conditions under which χ^2 measures the discripancy between observations and hypothesis. *Journal of the royal statistical society*, *87*, 442-450.

Gajek, L. (1986). On improving density estimators which are not bona fide functions. *The annals of statistics*, *14*(4), 1612-1318.

Glad, I. K., & Hjort, N. L. (2003). Correction of density estimators that are not densities. *Scandinavian journal of statistics*, *30*, 415-427.

Greenwood, P. E., & Nikulin, M. S. (1996). *A guide to chi-squared testing.* New York: John Wiley and Sons, Inc.

Gregory, G. G. (1977). Large sample theory for u-statistics and tests of fit. *The annals of statistics*, *5*(1), 110-123.

Hall, P., Watson, G., & Cabrera, J. (1990). Kernel density estimation with spherical data. *Biometrika*, *74*(4), 751-762.

Hamdam, M. (1962). The powers of certain smooth tests of goodness-of-fit. *Australian journal of statistics*, *4*, 25-40.

Härdle, W. (1991). *Smoothing thechniques with implementation in s.* New York: Springer.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests.* Springer.

Henze, N. (1997). Do components of smooth tests of fit have diagnostic properties? *Metrika*, *45*, 121-130.

Henze, N., & Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Australian journal of statistics*, *38*(1), 61-74.

Hjort, N. L., & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The annals of statistics*, *23*(3), 882-904.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods.* Wiley series in probability and statistics.

Inglot, T., Kallenberg, C. M., Wilbert, & Ledwina, T. (1997, November). Data driven smooth tests for composite hypothesis. *The annals of statistics*, *25*(3), 1222-1250.

Inglot, T., & Ledwina, T. (1996). Asymptotic optimality of data driven neyman's test. *Annals of statistics*, *24*, 1982-2019.

Jammalamadaka, S. R., & Kozubowski, T. (2003). A new family of circular models: The wrapped laplace distributions. *Advances and applications in statistics*, *3*(1), 77-103.

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

Jander, R. (1957). Die optische ichtungsorientierung der roten waldameise (*formica rufa*). *Zeitschrift fur vergleichende physiologie*, *40*, 162-238.

Janic-Wróblewska. (2004). Data-driven smooth test for a location-scale family. *Statistics*, *38*, 337-355.

Janssen, A. (1995). Principal component decomposition of non-parametric tests. *Probability theory and related fields*, *101*, 193-209.

Janssen, P., Swanepoel, J., & Veraverbeke, N. (2005). Bootstrapping modified goodness-of-fit statistics with estimated parameters. *Statistics and probability letters*, *71*, 111-121.

Kac, J., & Siegert, A. (1947). An explicit representation of a stationary gaussian process. *Annals of mathematical statistics*, *18*, 438-442.

Kallenberg, W. C. M., & Ledwina, T. (1995). Consistency and monte carlo simulation of a data driven version of smooth goodness-of-fit tests. *Annals of statistics*, *23*, 1594-1608.

Kallenberg, W. C. M., & Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the american statistical association*, *92*(439), 1094-1104.

Klar, B. (2000). Diagnostic smooth test of fit. *Metrika*, *52*, 237-252.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Gior. Ist. Ital. Attuari*, *4*, 83-91.

Kopecky, K. J., & Pierce, A., Donald. (1979). Efficiency of smooth goodness-of-fit tests. *Journal of the american statistical association*, *74*(366), 393-397.

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Proceedings of the koninklijke nederlandse akademie van wetenschappen*, *63*, 38-47.

Lancaster, H. O. (1969). *The chi-squared distribution.* New York: John Wiley.

Ledwina, T. (1994, September). Data-driven version of neyman's smooth test of fit. *Journal of the american statistical association*, *89*(427), 1000-1005.

Lee, A. (1990). *U-statistics.* New York: Marcel Dekker.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypothesis* (Third ed.). Springer.

Lilliefors. (1967). On the kolmgorov-smirnov test for normality with mean and variance unknown. *Journal of the american statisical association*, *62*, 399-402.

Lockhart, R. A., & Stephens, M. A. (1985). Tests of fit for the von mises distribution. *Biometrika*, *72*(3), 647-652.

Lund, U., & Agostinelli, C. (2005, may). *The circular package.*

Lund, U., & Jammalamadaka, S. R. (2000). An entropy-based test for goodness-of-fit of the von mises distribution. *Journal of statistical computation and simulation*, *67*, 319-332.

Magnello, M. E. (1998). Karl pearson's mathematization of inheritance: from ancestral heredity to mendelian genetics (1895-1909). *Annals of science*, *55*, 35-94.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics.* John Wiley and Sons, Inc.

Massey. (1951). The distribution ofthe maximum deviation between two sample cumulative step functions. *Annals of mathematical statistics*, *22*, 125-128.

Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the theory of statistics* (3rd ed.). Tokyo: McGraw-Hill Kogakusha.

Mudholkar, G. S., Kollia, G. D., Lin, C. T., & Patel, K. R. (1991). A graphical procedure for comaring goodness-of-fit tests. *Journal of the royal statistical society Series B (Methodological)*, *53*(1), 221-232.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statisical models* (4th ed.). Mc Graw-Hill.

Neyman, J. (1937). Smooth goodness-of-fit. *Skandinavisk aktuarietidskrift*, *20*, 149-199.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case o a correlated system of variables is such that it can be reasonable supposedto have arisen from random sampling. *Philosophical magazine*, *50*, 157-175.

Periwal, V., & Shewitz, D. (1990). Unitary-matrix models as exaclty solvable string theories. *Physical review letters*, *64*, 1326-1329.

van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge, UK: Cambridge university press.

Pyke, R. (1965). Spacings. *Journal of the royal statistical society Series B (Methodological)*, *27*(3), 395-449.

Rao, J. S. (1969). *Some contributions to the analysis of circular data.* Unpublished doctoral dissertation, Indian statistical institution, Calcutta.

Rao, J. S. (1972). Some variants of chi-square for testing uniformity on the circle. *Zeitschrift fur wahrscheinlichkeitstheorie and vervandt gebiete*, *22*, 33-44.

Rayleigh, L. (1919). On the problem of random vibrations, and of random flights in one, two or three dimensions. *Philosophical magazine and journal of science*, *37*(220), 321-347.

Rayner, J., & Best, D. (1989). *Smooth tests of goodness-of-fit.* Oxford university press.

Risebrough, R. (1972). Effects of environmental pollutants upon animal ither that man. In *Proceedings of the berkeley syposium on mathematics and* *statistics* (Vol. VI, p. 443-463). Berkeley, CA, USA: University of california university press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of mathematical statistics*, *27*, 832-837.

Rothman, E. D. (1972). Tests for uniformity of a circular distribution. *Sankhy, A*, *34*, 23-32.

Schreier, P. J., & Scharf, L. L. (2003). Second-order analysis of improper complex random vectors and processes. *IEEE transactions on signal processing*, *51*(3), 714-725.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*, 461-464.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice and visualization.* New York: Wiley.

Shorack, G. R., & Wellner, J. A. (1982). Limit theorms and inequalities for the uniform empirical process indexed by intervals. *The annals of probability*, *10*(3), 639-652.

Shorack, G. R., & Wellner, J. A. (1986). *Empirical processes with applications to statistics.* John Wiley and Sons, Inc.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall.

Simon, B. (2005). *Orthogonal polynomials on he unit circle* (Vol. 54). Amer. Math. Soc. Colloq. Publ.

Smirnov, N. (1939). Sur les ecarts de la courbe de distribution empirique (in russian). *Rec. Math.*, *6*, 3-26.

Stephens, M. A. (1969). *Techniques for directional data* (Technical report No. 150). Department of statistics, Stanford university.

Stephens, M. A. (1970). Use of the kolmogorov smirnov, cramer von mises and related staistics without extensive tables. *Journal of the royal statistical society Series B (Methodological)*, *32*(1), 115-122.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The annals of statistics*, *4*(2), 357-369.

Stute, W., Gonzáles-Manteiga, W., & Presedo-Quindimil, M. (1993). Bootstrap based goodness-of-fit tests. *Metrika*, *40*, 243-256.

Szegö, G. (1975). *Orthogonal polynomials* (Fourth ed., Vol. 23). Providence RI: American mathematical society. Colloquium publications.

Tarter, M., & Kronmal, R. (1976). An introduction to the implementation and theory of nonparametric density estimation. *The american statistician*, *30*(3), 105-112.

Thas, O. (2001). *Nonparametrical tests based on sample space partitions.* Unpublished doctoral dissertation, Ghent university.

Thas, O., & Ottoy, J.-P. (2003a). An extension of the anderson-darling k-sample test to arbitrary sample space partition sizes. *Journal of statistical computation and simulation*, *00*, 1-15.

Thas, O., & Ottoy, J.-P. (2003b). Some generalizations of the anderson-darling statistic. *Statistics and probability letters*, *64*, 255-261.

Thomas, D. R., & Pierce, D. A. (1979). Neyman's smooth goodness-of-fit test when the hypothesis is composite. *Journal of the american statisical association*, *74*(366), 441-445.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the royal statistical society Series B (Methodological)*, *38*(1), 54-59.

Venables, W., & Ripley, B. (1997). *Modern applied statistics with s-plus.* Springer.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung.* Vienna, Austria: Deuticke.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of mathematical statistics*, *18*, 309-348.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, *48*(1 and 2), 109-114.

Watson, G. S. (1967). Another test for uniformity of a cicular distribution. 675-677.

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the royal statistical society Series B*, *64*(2), 281-294.

Summary

An important statistical question is whether a sample of observations agrees with a certain prespecified distribution or family of distributions. To deal with this kind of statistical problems, it is recommended to apply both formal goodness-of-fit (GOF) tests or explorative graphical tools simultaneously. The sample space from which the observations are drawn is usually the real line, but data on a circle also arise in many fields. GOF methods for this kind of data need to be origin-invariant, since their conclusions should not depend on the chosen origin.

In this thesis, three contributions to the statistical analysis of linear and circular data are presented.

Applying smooth tests to solve the GOF problem for linear distributions has the advantage that the components in the orthogonal decomposition of the corresponding score statistic ofthen lead to easy interpretation and sum up to a test statistic with limiting omnibus features. The difficulty with the construction of smooth tests for circular distributions is to find appropriate orthonormal polynomials, because these are usually described in the complex field. We used the "complex" framework and the general theory of orthonormal polynomials on the unit circle (e.g. Simon, 2005) to construct a new class of smooth GOF tests for circular distributions, which is called the class of *complex smooth tests*. This class of tests generalises the framework of Rayner and Best (1989) for smooth tests on the real line.

Since we apply the test to circular data, the origin-invariance property needs to be checked. In case the smooth test is not origin-invariant, we proposed to subtract the circular mean direction from each observation before computing the test statistic.

For circular uniformity and circular normality we gave the explicit form of the smooth tests and its asymptotic distribution. We explained how, in case

of testing for circular uniformity, this construction leads to the smooth test of Bogdan et al. (2002). We have also shown that in case of testing for circular normality, our test generalizes the test proposed by Barndoff-Nielsen and Cox (1979).

Similarly as for a linear smooth test, the choice of the order of the family of alternatives in the smooth model is crucial to obtain good power. To overcome the problem of choosing the order, two data-driven version of the complex smooth test are discussed. The parametric bootstrap was used to approximate the null distributions for the data-driven statistics. The complex data-driven smooth test for the CN distribution has been applied on real data examples. It has been demonstrated that, if the null hypothesis is rejected, the components of the smooth test may contain interesting information about how the true distribution deviates from the hypothesised. Some characteristics of the data-driven smooth test for circular normality have been investigated in a simulation study, which showed that it has good power against many different alternatives.

In this thesis it has also been illustrated by means of an example how the application of the data-driven smooth test naturally leads to a nonparametric estimate of the true circular density. The result is essentially an orthonormal series density estimator, i.e. graph which can reveal how the true distribution deviates from the hypothesised. In that sense, the interpretation of the results from the test can be visualised.

We have also presented some new results on the integral version of the class of GOF tests for simple linear null hypothesis proposed by Thas (2001). The new versions of the tests are constructed by integrating out the Pearson χ^2 statistic over all possible partitions of the sample space in c cells. The degrees of freedom of Pearson's statistic are directly related to the indexing parameter of our new class, the SSP size c. The resulting tests are therefore called the *linear* SSPc tests. The tests are generalisations of the Anderson-Darling test, which is included in the class by taking $c = 2$.

The construction of the linear SSPc test statistics and their asymptotic null distributions were given, and omnibus consistency was proved.

To avoid the problem of choosing the right value for the indexing parameter c, we have proposed a data-driven version of the test. Simulations confirmed that the selection rule succeeds quite well in selecting a good choice for c. The weight functions that are involved in the test statistic, as well as the limiting behaviour and the simulation results, suggest that the new tests are very sensitive to deviations from the hypothesised distribution F_0 in small intervals of the support of F_0. Furthermore, this sensitivity increases with increasing SSP size c.

Extensions to composite null hypothesis were described as well. In particular, the use of the new class of GOF tests to composite null hypotheses is based on the estimated empirical process. The limiting Gaussian process is quite complicated, however, and we used parametric bootstrap to obtain the null distribution.

The new class of GOF tests for linear data was adapted to a similar class of GOF tests for circular data. This was done by making the class of statistics origin-invariant. We simply integrated out all possible origins to obtain the origin-invariant class of statistics. The resulting type of tests is called the *circular* SSPc tests and reduces to Rothman's test (1972) if $c = 2$.

The limiting null distribution of the circular SSPc test was derived and computational formulae for SSP size $c = 2, 3$ and 4 were found.

The data-driven version and its asymptotic theory are similar as in the linear case. A simulation study indicated that the circular SSPc test has the same power characteristics as the linear SSPc test, although the differences between their powers are less pronounced than in the linear case. The power study showed that the SSPc tests perform at least as good as their competitors for all alternatives considered.

In the final part of the thesis we have developed and discussed the IBPP-plot, which is a useful graphical tool for detecting and localising LOF on the circle. Two types of plots have been described, both of which are based on the Kuiper test. Similarly as for the PP-plot, the IBPP-plot is thus related to a formal statistical test, which is particularly intersting since the results of that test can be derived from the graph. Hence, the conclusions obtained from that graph are objective in contrast to most other graphial tools which are merely explorative and hence subjective.

All methods for circular data are applicable to linear data as well and thus so is the IBPP-plot. Furthermore, we have extended the use of the IBPP-plot and demonstrated that an adapted version enables explorative comparison between different density estimates.

Samenvatting

Een belangrijke en vaak voorkomende statistische vraag is of een steekproef van observaties al dan niet in overeenstemming is met een welbepaalde distributie, of familie van distributies. Zulke statistische problemen worden doorgaans behandeld tegelijkertijd op basis van formele toetsen voor aanpassing (*goodness-of-fit, GOF*) als op basis van explorerende grafische technieken. De steekproefruimte waaruit de observaties worden getrokken is gewoonlijk de reële rechte, maar ook data op een cirkel komt in vele toepassingsgebieden voor. Toetsen voor aanpassing voor zulke circulaire data dienen oorsprongsinvariant te zijn, opdat de conclusies niet zouden afhangen van de gekozen oorsprong.

In deze thesis worden drie bijdragen geleverd op het gebied van de statistische analyse van lineaire en circulaire data.

Het toepassen van zogenaamde gladde toetsen (*smooth tests*) ter behandeling van het GOF probleem voor lineaire distributies heeft als voordeel dat de componenten in de orthogonale decompositie van de overeenkomstige scorestatistiek leiden tot duidelijke interpretatie. Samengeteld vormen de componenten bovendien in de limiet een toets met omnibus kenmerken. De moeilijkheid die gepaard gaat met het opstellen van gladde toetsen voor circulaire data is het vinden van geschikte orthonormale veeltermen, aangezien deze doorgaans moeten gezocht worden in het complexe veld. Het "complexe" kader en de algemene theorie van orthonormale veeltermen op de eenheidscirkel (bv. Simon, 2005) werden gebruikt om een nieuwe klasse van gladde toetsen van aanpassing op te stellen voor circulaire distributies, namelijk de klasse van *complexe gladde toetsen*. De klasse van toetsen veralgemeent het werk van Rayner en Best (1989) in verband met gladde toetsen op de reële rechte.

Aangezien het toetsen voor circulaire data betreft, moet de oorsprongsinvariantie nagegaan worden. Indien de toets in eerste instantie niet oorsprongsinvariant is, stellen we voor om dit op te lossen door de circulaire gemiddelde richting af te trekken van de observaties alvorens de toetsingsstatistiek te berekenen.

Voor circulaire uniformiteit en circulaire normaliteit hebben we de expliciete vorm gegeven van de gladde toets, alsook zijn asymptotische verdeling. We hebben aangetoond dat ingeval van circulaire uniformiteit de toets zich herleidt tot de toets van Bogdan et al. (2002). Ingeval van normaliteit zijn onze toetsen een veralgemening van de toets voorgesteld door Barndorff-Nielsen en Cox (1979).

Net zoals bij de lineaire gladde toetsen is de keuze van de orde van de familie van alternatieven van groot belang opdat de toets optimaal krachtig zou zijn. Om die keuze niet op een subjective manier te moeten maken, hebben we twee data-gedreven versies van de complexe gladde toetsen besproken. De nuldistributies van de data-gedreven statistieken kunnen worden bekomen door middel van parametrische bootstrap. We hebben de data-gedreven complexe gladde toetsen voor normaliteit uitgebreid toegepast op bestaande data voorbeelden. We hebben zo bijvoorbeeld aangetoond dat, bij verwerping van de nulhypothese, de componenten van de gladde toetsingsstatistiek interessante informatie kunnen bevatten omtrent de aard van het verschil tussen de werkelijke distributie en de hypothetische distributie. Verder werden verschillende karakteristieken van de data-gedreven gladde toetsen voor circulaire normaliteit onderzocht in een simulatiestudie, dewelke heeft aangetoond dat de toetsen relatief krachtig zijn voor vele verschillende alternatieven.

Tenslotte werd ook geillustreerd, door middel van een voorbeeld, hoe de toepassing van de data-gedreven gladde toets op een natuurlijke wijze leidt tot een niet-parametrische schatting van de werkelijke circulaire dichtheid. Het betreft dan in feite een orthonormale reeksschatter, en een grafische voorstelling hiervan kan helpen uitwijzen hoe de werkelijke distributie verschild van de hypothetische. In deze zin kunnen de interpretaties van de resultaten dus gevisualiseerd worden.

In het volgende deel van de thesis werden enkele nieuwe resultaten voorgesteld betreffende de integraalversie van de klasse van toetsen van aanpassing voorgesteld door Thas (2001). De nieuwe versies van deze toetsen werden geconstrueerd door het integreren van de Pearson χ^2-statistiek over alle mogelijke partities van de steekproefruimte (*sample space partitions, SSP*) in c cellen. Het aantal vrijheidsgraden van de Pearson statistiek is rechtstreeks gerelateerd aan de indexparameter van onze nieuwe klasse van toesten, namelijk de SSP grootte c. De resulterende toetsen worden daarom *lineaire* SSPc toetsen genoemd. De toetsen zijn veralgemeningen van de Anderson-Darling toets, die overeenkomt met de SSPc toets ingeval $c = 2$.

Naast de constructie van de SSPc toetsen werd ook hun asymptotische verdeling afgeleid en werd omnibus consistentie bewezen. Om het probleem te omzeilen van de keuze van de waarde van de indexparameter c, hebben we opnieuw een data-gedreven versie van de toets voorgesteld. Simulaties hebben bevestigd dat de data-gedreven selectie succesvol is in het vinden van de geschikte keuze voor c. De gewichtsfuncties die in de toetsingsstatistiek voorkomen, het onderzochte limietgedrag en de simulatieresultaten duiden er alle op dat de nieuwe toetsen specifiek erg gevoelig zijn voor lokale afwijkingen van de veronderstelde distributie. Deze gevoeligheid wordt nog groter met toenemende SSP grootte c.

De SSPc toetsen werden in eerste instantie ingevoerd voor enkelvoudige hypothesen, maar we hebben vervolgens de uitbreiding naar samengestelde hypothesen beschreven. Het gebruik van de de nieuwe klasse van GOF toetsen voor samengestelde hypothesen is gebaseerd op het geschatte empirische proces. Het Gaussische proces dat de limiet vormt is erg gecompliceerd en we hebben daarom de bootstrap gebruikt om de nuldistributie van de toetsingsstatistiek te bekomen.

Daarnaast werd de nieuwe klasse van GOF toetsen voor lineaire data aangepast om tot een gelijkaardige klasse te komen van toetsen voor circulaire data. De nodige aanpassing betrof het oorsprongsinvariant maken van de statistieken, hetgeen we hebben bewerkstelligd door te integreren over alle mogelijke oorsprongen. Naar het resulterende type van toetsen werd dan verwezen als de klasse van *circulaire* SSPc toetsen. Ingeval $c = 2$ wordt de bestaande toets van Rothman (1972) bekomen.

We hebben de asymptotische nuldistributie van de circulaire SSPc toets afgeleid en computationele formules bekomen voor de gevallen $c = 2, 3$ en 4. De data-gedreven versie en zijn asymptotische distributie zijn gelijkaardig aan het lineaire geval. Een simulatiestudie wees aan dat betreffende de gevoeligheid en kracht van de toetsen, de circulaire versies ongeveer dezelfde karakteristieken vertonen als hun lineaire tegenhangers, al is de gevoeligheid nu minder afhankelijk van de indexparameter c. Een belangrijke conclusie van de simulatiestudie was ook dat de SSPc toetsen minstens even krachtig (en vaak krachtiger) zijn als alle andere onderzochte toetsen, tenminste voor de families van alternatieven die beschouwd werden in de studie. Alle methoden uit dit deel van de thesis werden eveneens toegepast op bestaande data voorbeelden.

In het laatste deel van de thesis hebben we de zogenaamde *Interval-Based Probability-Probability-plot* of IBPP-plot, ontwikkeld en besproken. Deze grafische voorstelling is een nuttige techniek om gebrek aan aanpassing (*lack-of-fit, LOF*) van een distributie op de cirkel te detecteren en te lokaliseren. Twee versies van de techniek werden besproken, dewelke beide gebaseerd zijn de Kuiper toets. Net als de klassieke PP-plot is de IBPP-plot dus gerelateerd aan een formele statistische toets, hetgeen interessant omdat zodoende de resultaten van de toets kunnen afgeleid worden van de grafische voorstelling. De conclusies van de IBPP-plot zijn daarom objectief, terwijl de meeste andere grafische voorstellingen enkel explorerend van aard zijn en dus subjectief.

Alle methoden voor circulaire data zijn evenzeer toepasbaar op lineaire data en dat geldt dus ook voor de IBPP-plot, zoals we hebben geïllustreerd met enkele voorbeelden. Tenslotte hebben we de toepassing van de IBPP-plot uitgebreid en hebben we aangetoond dat een aangepaste versie de mogelijkheid biedt tot een explorerende vergelijking tussen verschillende dichtheidsschattingen.

Curriculum Vitae

• Personal Information

Born: June 29 1977, Ghent (Belgium)

Address: Langebilkstraat 20, B-9032 Gent-Wondelgem, Belgium

Nationality: Belgian

e-mail: heidi.wouters@ugent.be

• Studies

1995-1999: Mathematics at Ghent University.

Obtained degree in Mathematics, option Pure Mathematics.

1997-1999: Qualified Teacher's Degree for secondary education section 2 in Mathematics.

2005: Master in Biostatistics at Hasselt University.

• Career

1999-2000: Teaching and Research Assistant at the University of Limburg.

2000-2001: Research Assistant at Ghent university, faculty of Bio-science engineering, Dept. of Applied Mathemathics, Biometrics and Process Control.

2001-2007: Teaching and Research Assistant at Ghent university, faculty of Bio-science engineering, Dept. of Applied Mathemathics, Biometrics and Process Control.

• Teaching Activities:

2000-2001 and 2005-2006: Algebra and Analytical geometry
2000-2001: Calculus
2000-2006: Probabilistic Models
2000-2001: Data analysis
2002-2006: Statistical Data analysis
2002-2006: Experimental Design
2002-2003: Applied Statistics for the Food Sciences
2004-2006: Statistics

- **Teaching Activities for IVPV (Instituut Voor Permanente Vorming):**

2001, 2003, 2005: Module 2: Regression Analysis

2002: Module 1: Basic Course in statistics and Module 0: Introduction to Splus

- **Other Activities:**

2002-2004: adviser for several graduate student theses.

2001-2006: adviser for statistical consulting within the faculty of Bio-science engineering.

2006: three month Project at statistics department of University of British Columbia in Vancouver (local supervisor Prof. Raphael Gottardo).

- **Conferences and workshops:**

8th Annual Meeting of the Belgian Statistical Society, Herbeumont, Belgium, 4-6 oktober 2000: participation

9th Annual Meeting of the Belgian Statistical Society, Oostende, Belgium, 12-13 october 2001: participation

Workshop on Resampling Methods, Ghent, 5 April 2002: participation

10th Annual Meeting of the Belgian Statistical Society, Kerkrade, Nederland, 18-19 oktober 2002: oral presentation: “On The Link Between EDF and Smooth Goodness-of-fit Tests”.

Joint Statistical Meeting 2003, San Francisco, California, USA, August 3-7, 2003: oral presentation: “Interval-based PP-plot: a graphical Tool to localize regions of lack of fit”.

11th Annual Meeting of the Belgian Statistical Society, La Roche-en-Ardennes, Belgium, 10-11 oktober 2003: participation

16th Symposium of IASC, COMPSTAT 2004, Prague, Czech Republic, 23-28 august, poster presentation: “The SignUm Plot: A Simple Graphical Tool for Bump Hunting”.

- **List of Publications:**

  - Wouters, H., Thas, O. and Ottoy, J.P. (2002). On the Link between EDF and Smooth GOF tests. In: proceedings of the 10th Annual Meeting of the Belgian Statistical Society, Kerkrade, Netherlands, 18-19 October 2002.
  - Thas, O., Wouters, H., Ottoy, J.P. Localized Pearson Chi-squared Goodness-

of-fit Tests. (submitted).

- Dedecker, A.P., Janssen, K., Wouters, H., Thas O., Goethals, P.L.M. and De Pauw, N. Assessment of sampling variability of macroinvertebrate communities in rivers. (submitted)
- Smagghe, G., Van Loocke K., De Regge N., Wouters H. (2004) Compartment Model for Kinetics and Toxicity of Insecticides utilizing caterpillar Midguts ex vivo proceedings QSAR 2004 Istanbul ADME exvivo.
- Wouters, H., Thas, O., Ottoy, J.P. Data Driven Smooth Tests and a Diagnostic Tool for Lack-of-Fit for Circular Data. (submitted).