The concept of energy in nonparametric statistics- Goodness-of-Fit problems and deconvolution

DISSERTATION

zur Erlangung des Grades eines Doktors

der Naturwissenschaften

vorgelegt von

Dipl.-Math. Berkan Aslan

aus Siegen

eingereicht beim Fachbereich Physik

der Universität Siegen

Siegen 2004

Gutachter der Dissertation: Prof. Dr. Günter Zech
Prof. Dr. Martin Holder

Datum der Disputation: 07. Juli 2004

Prüfer: Prof. Dr. Hans-Dieter Dahmen
Prof. Dr. Martin Holder
Prof. Dr. Günter Zech

Internetpublikation der Universitätsbibliothek Siegen: urn:nbn:de:hbz:467-727

Zusammenfassung

In dieser Arbeit wird das Energiekonzept aus der Physik in die Statistik übertragen. Die Energie von Stichproben, die aus statistischen Verteilungen gezogen werden, wird in ähnlicher Weise definiert wie für elektrostatische Punktladungen.

Ein System von zwei Punktladungsmengen mit entgegengesetztem Vorzeichen befindet sich im Zustand minimaler Energie, wenn sie der gleichen Verteilung folgen. Dieses Konzept wird zur Konstruktion von neuen nichtparametrischen, mehrdimensionalen Anpassungstests verwendet. Weiterhin wurde das Energieverfahren auf das Zwei-Stichproben Problem und die Entfaltung angewandt.

Das statistische Minimum Konzept der Energie hängt nicht von der Abstandsfunktion des elektrostatischen Potentials ab. Um die Güte der entwickelten Methoden zu erhöhen, können andere monoton fallende Abstandsfunktionen gewählt werden. Wir zeigen, dass das Verfahren für alle Abstandsfunktionen anwendbar ist, die eine positive Fouriertransformierte haben. Die vorgeschlagene Methode benötigt keine Intervallbildung. Sie hat ihre Stärken bei mehrdimensionalen Problemstellungen und ist hier herkömmlichen Verfahren in vielen konkreten Anwendungen überlegen.

Abstract

In this thesis the concept of energy is introduced from physics into statistics. The energy of samples, which are drawn from statistical distributions, is defined in a similar way as for discrete charge density distributions in electrostatics.

A system of two sets of point charges with opposite sign is in a state of minimum energy if they are equally distributed. This property is used to construct new nonparametric, multivariate Goodness-of-Fit tests, to check whether two samples belong to the same parent distribution and to deconvolute distributions distorted by measurement.

The statistical minimum energy configuration does not depend on the application of the one-over-distance power law of the electrostatic potential. To increase the power of the new approach other monotonic decreasing distance functions may be chosen. We prove that the new energy technique is applicable to all distance functions which have positive Fourier transforms. The proposed approach is binning-free. It is especially powerfull in multidimensional applications and superior to most of the common statistical methods in many concrete situations.

Contents

Chapter 1

Introduction

In this work a new method is proposed which allows to construct nonparametric, multivariate, binning-free Goodness-of-Fit (GoF) tests, two-sample tests and multivariate, binning-free unfolding. The method introduces a statistical energy, in analogy to electrostatics.

In practice it appears often that one wants to decide whether the measurements possibly came from a given distribution or not. Statistical tests that address this type of problems are GoF tests. GoF tests have been developed mostly for univariate distributions and, except for the case of multivariate normality, very few tests for multivariate GoF problem can be found in the literature. In principle, power divergence statistics, where Pearson's χ^2 statistic is a member of this family, can be applied for testing the GoF of any multivariate distribution. Power divergence statistics are very simple and need only limited computational power, but they suffer from some serious drawbacks: in how many bins must the measurements be grouped, where and how must the bin boundaries be placed? In the literature univariate GoF tests are proposed which avoid these drawbacks. Many of these tests are based on the empirical distribution function (EDF).

The problem of deciding whether or not a given sample may have been generated by a specified distribution is sometimes also known as the one-sample GoF problem. This is however not the only GoF problem. Another important member is the two-sample GoF problem or briefly two-sample problem, where the question is to test the hypothesis that two samples come from the same distribution. Most of the above mentioned GoF tests, Pearson's test and some EDF tests, are extended to this setting as well.

Another problem is the problem of unfolding, i.e. the correction of distributions which are distorted by measurement errors. Unfolding the measurement errors from a measured distribution is a frequently occurring task in high energy physics. It has been widely discussed in the literature, however, multivariate, binning-free unfolding problem seems to have received little attention in the literature.

Most of the tests considered in this thesis are nonparametric, omnibus tests. A statistical test is called nonparametric if its applicability does not depend on the particular null hypothesis distribution and an omnibus test is a sensitive test to almost all alternatives to the null hypothesis. Within the nonparametric, omnibus tests there is no uniformly most powerful test available. Hence some tests will have better powers under some alternative hypothesis and others will have better powers under other alternatives, but none has the highest power under all alternatives. This leaves the question open for nonparametric, omnibus test with good overall power properties. Therefore there is ongoing research in field of nonparametric, omnibus tests.

We have constructed a new family of nonparametric, multivariate, omnibus tests and a new multivariate, binning-free unfolding approach which are all based on the energy of two statistical distributions. The energy of statistical distributions is defined in an analogous way as the laws of electrostatics fix it for charge distributions. The energy of suitably normalized charges of two samples, one of which is positively charged and one which is negatively charged, is minimum if the positions of the point charges of the two samples agree, i.e. under null hypothesis the energy will be a minimum and all alternatives to the null hypothesis will lead to an increase of the energy.

In Chapter 2 some basic terminology and notation, as well as a formal description of the GoF and two-sample problem are given.

In Chapter 3 an overview of the literature on tests for the GoF problem is given. There is a very vast literature on GoF tests, therefore a complete survey of all GoF tests is not given. Only those tests are presented that are relevant for tests that are developed in this thesis.

The new family of nonparametric, multivariate, omnibus tests for the GoF problem is developed in Chapter 4: the Energy tests. The conjecture of the minimum property of the energy of two distributions is proven and the consistency of the new tests is shown. In a special case the relation with the Bowman-Foster test is indicated. The results of a power study, comparing different tests, are given. This is especially of interest to understand the behavior of the tests with finite sample sizes.

For the multivariate two-sample problem, new Energy tests are presented in Chapter 5. Since it is based on the same principles as the Energy tests for the GoF problem, the presentation can be kept short. The null distribution of the two-sample Energy tests is determined by a permutation method. A power study is included to compare the new tests with some competitors. The test is also applied to a physics case. A data sample taken from a particle experiment is compared to a Monte Carlo simulation.

In Chapter 6 the energy concept is applied to unfolding. Again it is based on the same idea of the energy as a measure of compatibility of two samples. To introduce the problem of unfolding, two unfolding techniques are reviewed. The new multivariate, binning-free unfolding approach is applied to two examples, where the limitations of the commonly used methods are obvious.

A summary is given in the last Chapter 7.

Chapter 2

Introduction to statistical test theory

One of the statistical problem, which appears often in physical experiments, is to test how well the n independent measurements agree with a probability model for the experiment. This problem is usually solved by a statistical test, which compares measured values from the experiment with corresponding theoretical values derived from the model. The purpose of this chapter is to present some basic concepts of statistical test theory. We do not treat this topic in detail, since it can be found in some introductory books on statistics.

2.1 Terminology

Statistics has it own specialized terminology with words whose meaning differs from the meaning in physics. Sometimes the same term has different meaning in statistics and in physics, we often choose the statistical term. An example is the word estimate. In statistics estimate is used where physicist would say determine or measure. In physics estimate is used where statisticians would say guess. We therefore make some substitutions, see Table 2.1.

Table 2.1: Relation between statistics and physics terminology.

statistics terminology	physics terminology
observation	measurement, event
sample	data (set)
sample of size n	n measurements
sample mean	experimental mean, average
class	bin

2.2 Types of statistical hypotheses

Let X be a univariate random variable (r.v.) (for simplicity we usually confine attention to the univariate case) defined over a sample space Π, which is the set of all possible values that a realization x of X can take. We use capital letters for r.v. and lower-case for its realization. Throughout this thesis r.v.s are indicated by italic capitals (X, Y, etc.) if they are univariate and by bold capitals ($\boldsymbol{X}$, $\boldsymbol{Y}$, etc.) if they are multivariate. We denote with $X_1, X_2, \ldots, X_n$ a random sample of size n for X. In what follows we generally treat $X_1, X_2, \ldots, X_n$ as continuous, independent and identically distributed (i.i.d.) with cumulative distribution function (c.d.f.) $F(x)$ whose probability density function (p.d.f.) $f(x)$ is continuous. Throughout the complete work we assume that the distributions are continuous.

By definition, p.d.f.s cannot be observed, they can only be induced by a set of observations. In principle a statistical hypothesis is an assertion about the p.d.f of a r.v. Let us give four examples of statistical hypotheses:

1. The r.v. X is distributed according to Gaussian with particular values of μ and σ.
2. The r.v. X is distributed according to Gaussian with particular values of μ.
3. The r.v. X is distributed according to Gaussian
4. The results of two experiments, X and Y are distributed identically.

Each of these hypotheses says something about the p.d.f. of the r.v. and is hence testable by comparison with observations. Examples 1 and 2 specify a p.d.f. and certain values for one or both of its parameters. Example 3 specifies the form of the p.d.f., but none of its parameters, and example 4 does not even specify the form of the p.d.f.

Depending on the degree of knowledge of the p.d.f. f two kinds of hypotheses are distinguished:

- parametric hypothesis
- nonparametric hypothesis

Assume that $\mathcal{P} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a parametric family of p.d.f.s, where $\boldsymbol{\Theta}$ is the parameter space and $\boldsymbol{\theta}$ is a k-vector parameter. A parametric hypothesis makes a statement about a certain set of the k parameters $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_k)$, as in examples 1 and 2. Frequently, the true functional form of the p.d.f., from which a random sample is drawn, is not known. In this case a nonparametric hypothesis makes a statement about the form of a p.d.f. rather than about a set of parameters, as in examples 3 and 4.

Further distinctions is made in the literature; Examples 1 and 2 differ in that 1 specifies all of the parameters of the p.d.f., whereas 2 specifies only a subset of the parameters. When all of the parameters $\boldsymbol{\theta}$ are specified the hypothesis is called *simple.* For an incomplete specification of $f(x;\boldsymbol{\theta})$, i.e. the form of $f(x;\boldsymbol{\theta})$ is fixed but not the values of all parameters $\boldsymbol{\theta}$=$(\theta_1, \theta_2, \ldots, \theta_k)$, the hypothesis is called *composite.* If the p.d.f. has k parameters $\theta_1, \theta_2, \ldots, \theta_k$, we can define a k-dimensional parameter space $\boldsymbol{\Theta}$. A simple hypothesis selects a unique point in this space. A composite hypothesis selects a subspace containing more than one point.

The hypothesis being tested is traditionally called the *null hypothesis* and is denoted by H_0. The validity of H_0 is checked in comparison with an *alternative* hypothesis, which we denote by H_1. Alternative hypothesis is a set of models, which do not include the model of H_0. In order to examine the measure of compatibility between the random sample $X_1, X_2, \ldots, X_n$ and a theoretical model, one constructs a *test statistic* $T(x_1, x_2, \ldots, x_n)$, which is a function of the observed values of $X_1, X_2, \ldots, X_n$ and determine in some way the conformity of the observations to the hypothesized distribution.

2.3 Tests of hypotheses

To introduce some basic definitions of a test it seems likely that it will be easier to deal with simple hypotheses than with composite ones. In this type of problem, the parameter space $\boldsymbol{\Theta}$ contains exactly two points. The following simple hypotheses are to be tested:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{against} \quad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1, \tag{2.1}$$

for some values $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ of $\boldsymbol{\theta}$. A test statistic T is a r.v., because it is a function of the observations. Hence each of the simple hypotheses H_0, H_1 will imply a given p.d.f. $g_{H_0}(T)$, $g_{H_1}(T)$ for T, see Figure 2.1.

To test a hypothesis on the basis of the random sample, we must specify a test procedure by dividing the sample space Π into two subsets. One subset, we call it R_0, contains the values of $X_1, X_2, \ldots, X_n$ for which one will accept H_0, and R_1, the complement of R_0, contains the values of $X_1, X_2, \ldots, X_n$ for which one will reject H_0 or equivalently accept H_1. The subset R_1 is referred to as the *critical region* of the test, and R_0 is called the *acceptance region.* The probability P that a random sample will fall in the critical region R_1 can be calculated, we can choose R_1 such that P is equal to some pre-chosen value α,

$$P((x_1, x_2, \ldots, x_n) \in R_1 | H_0) = \int_{T_{critical}}^{\infty} g_{H_0}(T) dT = \alpha.$$

This value α is called the *significance level* or *size* of the test. Clearly, whether we accept or reject H_0 depends on what the alternative hypothesis is. There are two

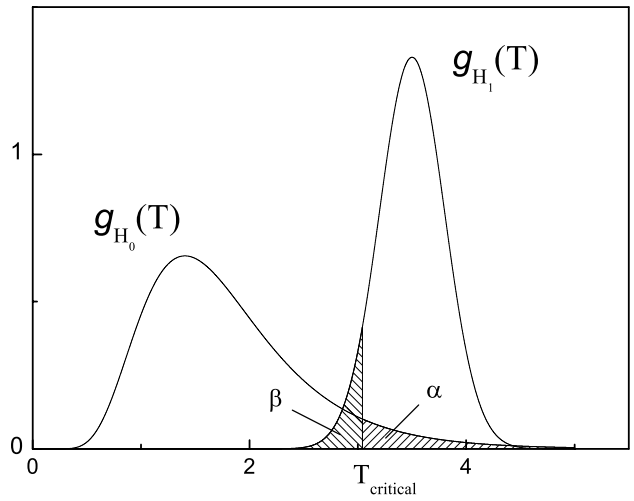

Figure 2.1: The p.d.f.s of the test statistic T under H_0 and H_1 are shown. H_0 is rejected if $T > T_{critical}$.

ways in which a mistake can made about the decision for acceptance or rejection of H_0. From Figure 2.1 we see that there is always the possibility that, even though H_0 is true, the random sample lies in R_1, in which case H_1 will be accepted. Then we have made an *error of the first type* and occurs with a probability α. On the other hand, if H_1 is true there is a possibility that the sample lies in R_0. This is referred to as an *error of the second type* and occurs with a probability β,

$$P\left((x_1, x_2, \ldots, x_n) \in R_0 | H_1\right) = \int_{-\infty}^{T_{critical}} g_{H_1}(T) dT = \beta,$$

see Figure 2.1. The complementary probability $1 - \beta$ is called the *power* of the test of H_0 against H_1, which is the probability of accepting H_1 when H_1 is true.

2.3.1 Neyman-Pearson test

Ideally, one should like a test which makes both α and β small, but it is clear from Figure 2.1 that a decreasing of α increases β and vice versa. Therefore one way in which different tests can be compared is to fix the probability of first type error to some pre-chosen value ($\alpha = 0.05$ and $\alpha = 0.01$ are common values, but throughout this thesis we use 0.05 as significance level) and then choose the test that gives the smallest probability of second type error β or equivalently the biggest power.

In case of simple hypotheses (2.1), for a given test T and a given value α, one can determine β. Repeating this procedure for various values of α, a curve $\beta(\alpha)$ can

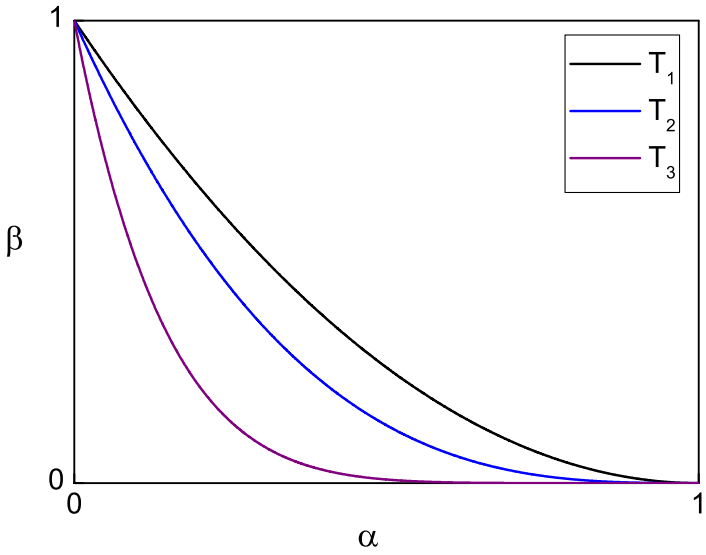

Figure 2.2: Comparison of tests in the $\alpha - \beta$ plane. The test statistic T_3 illustrates the Neyman-Pearson test.

be constructed, see Figure 2.2. A test which minimizes β for fixed α is called a *most powerful* test of size α.

The question which arises here is how to construct a most powerful test. The method of constructing a most powerful test depends on the use of a theorem which is named after two statisticians. The theorem, called the *Neyman-Pearson lemma*, states the following [1]:

> *A test for the problem* (2.1) *is a most powerful test if the critical region* R_1 *is chosen such that*
>
> $$\text{LQ}=\frac{\prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}_1)}{\prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}_0)} \begin{cases} \geq c \text{ inside } R_1, \\ < c \text{ outside } R_1, \end{cases}$$
>
> *where* c *is a constant which depends on the significance level* α.

This lemma allows us to design a most powerful test which is based on the ratio of the likelihood functions under the simple hypotheses H_0, H_1. The test constructed according to the Neyman-Pearson lemma is called Neyman-Pearson test. It should be noted that the most important point about this test is that it only exists for completely specified hypotheses. But in most situations such cases are usually scarce.

Example 1 *We consider the problem of testing whether a Gaussian* $N(\mu, \sigma^2)$ *with*

a known variance σ^2 has a mean $H_0 : \mu = \mu_0$ or a mean $H_1 : \mu = \mu_1 > \mu_0$. We take the logarithm of LQ and obtain

$$\begin{aligned} \ln LQ &= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_1)^2 - \sum_{i=1}^{n}(x_i-\mu_0)^2\right] \\ &= -\frac{n(\mu_1-\mu_0)}{\sigma^2}\left[-\overline{x} + \frac{\mu_1+\mu_0}{2}\right]. \end{aligned}$$

Then it yields

$$LQ \geq c \Leftrightarrow \ln LQ \geq \ln c \Leftrightarrow \overline{x} \geq \frac{\mu_0+\mu_1}{2} + \frac{\sigma^2 \ln c}{n(\mu_1-\mu_0)} = k,$$

where $\overline{x}$ is the sample mean. Since $\overline{X}$ is distributed as $N\left(\mu_0, \frac{\sigma^2}{n}\right)$ when H_0 is true, we get

$$\alpha = P\left(\overline{X} \geq k | H_0\right) = P\left(\frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \geq \frac{k-\mu_0}{\sigma}\sqrt{n} | H_0\right),$$
$$\frac{k-\mu_0}{\sigma}\sqrt{n} = u_{1-\alpha}, \quad k = \mu_0 + \frac{\sigma}{\sqrt{n}}u_{1-\alpha},$$

where $u_{1-\alpha}$ is the quantile of order $1-\alpha$ of the standard normal distribution.

2.3.2 Likelihood ratio test for composite hypotheses

If H_0 or H_1, or both, are composite hypotheses, i.e.

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Omega}_0$$
$$H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta} \backslash \boldsymbol{\Omega}_0$$

the Figure 2.2 can become a multidimensional diagram and one can only rarely find a test which is more powerful than any other test. But also for this type of problems it exists a simple procedure, the *likelihood-ratio method*, to construct 'good' tests. The likelihood-ratio test LR is defined by

$$LR = \frac{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \backslash \boldsymbol{\Omega}_0} L(x; \boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Omega}_0} L(x; \boldsymbol{\theta})},$$

where

$$L(x; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

is the likelihood function. Clearly, large values of LR lead to a rejection of H_0.

LR often produces a *uniformly most powerful* test when such exists. A test is uniformly most powerful if it is most powerful to each specific alternative.

Example 2 *X is assumed to be distributed as $N(\mu, \sigma^2)$ with unknown variance σ^2 and the hypothesis to be tested is*

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0.$$

The maximum likelihood estimate for σ^2 under H_0 is given by $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$, hence

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Omega}_0} L(x;\boldsymbol{\theta}) = \max_{\sigma^2} L\left(x;\mu_0,\sigma^2\right) = L\left(x;\mu_0,\widehat{\sigma}^2\right) = \left(\frac{1}{2\pi\widehat{\sigma}^2}\right)^{n/2} e^{-n/2}.$$

Equally the maximum likelihood estimates for μ and σ^2 under H_1 are given by $\overline{x}$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, respectively, i.e.

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(x;\boldsymbol{\theta}) = L(x;\overline{x},s^2) = \left(\frac{1}{2\pi s^2}\right)^{n/2} e^{-n/2}.$$

As a result

$$\begin{aligned} LR &= \left(\frac{\widehat{\sigma}^2}{s^2}\right)^{n/2} = \left(1 + \frac{n(\overline{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{n/2} \\ &= \left(1 + \frac{t^2}{n-1}\right)^{n/2}, \end{aligned}$$

where

$$t = \frac{\overline{x}-\mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}}\sqrt{n}.$$

t is Student's test statistic and H_0 is rejected if $|t| > t_{1-\frac{\alpha}{2}}$, where $t_{1-\frac{\alpha}{2}}$ is the quantile of order $1-\frac{\alpha}{2}$ of the Student distribution.

2.4 Goodness-of-Fit tests

A statistical test, which is designed for deciding whether or not a given random sample may have been drawn by a specified distribution F_0 is called *Goodness-of-Fit* (GoF) tests. GoF tests make inferences about the whole of a c.d.f. (or equivalently of a p.d.f.) and the difference with the tests considered in the previous section lies in the alternative hypothesis H_1, because it is just the complement of H_0, i.e. H_1 is a set of all possible alternatives to H_0.

For definiteness, let $F_0(x)$ be completely specified distribution and let $F(x)$ be the true, but unknown c.d.f. of r.v. X. Then, the GoF problem consists of testing the hypothesis

$$H_0 : F(x) = F_0(x) \text{ for all } x$$

against the general alternative

$$H_1 : F(x) \neq F_0(x) \text{ for at least one } x.$$

Note that the hypothesized distribution F_0 of the GoF problem, considered above, is completely specified, including all parameter values. A more common problem is assessing

$$H_0 : F(x) = F_0(x, \boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta},$$

where $F_0(x, \boldsymbol{\theta})$ is a specified parametric family, such as the normal. In this case the unknown parameters must be estimated from the observations before tests can be performed. In either case H_1 is the complement of H_0.

In order to perform a GoF test one has to look at the p.d.f. $g_{H_0}(T)$ of a test statistic T under H_0 and define a *p-value* for the test as, see [2]

$$P(T \text{ at least as extreme as the observed value } T_{obs}|H_0).$$

For example, if T is constructed such that large values correspond to poor agreement with H_0, then the p-value would be

$$p = \int_{T_{obs}}^{\infty} g_{H_0}(T)dT$$

and if p is small enough, then it indicates to reject H_0. The p-value is a function of the observations and is therefore itself a r.v., therefore the significance level α of a test should not be confused with the p-value, since α is a pre-chosen constant.

GoF tests that are sensitive to all types of deviation from H_0, are called *omnibus tests* and GoF tests that are especially designed to detect deviations from H_0 in the direction of specific alternatives, are called *directional tests* [3]. In most practical situations, when the null hypothesis is not true, one does not know in what way the true distribution F deviates from the specified distribution F_0. Omnibus tests have the advantage of being more generally applicable (no prior knowledge needed). Of course, the price that has to be paid for this overall sensitivity of omnibus tests is a loss in power for some specific alternatives. Directional tests against some specific alternatives would be more powerful than the omnibus tests, while against all other alternatives the omnibus test should be superior. A GoF test whose application does not depend on the hypothesized distribution F_0 is called *nonparametric* GoF test. A further property of a GoF test is that it can be *distribution-free.* It means that the distribution of the test statistic does not depend on the hypothesized distribution from which the sample was drawn. This is a nice property for a test, but nowadays the distribution of a test statistic can be easily obtained by a Monte Carlo simulation.

2.5 Two-sample GoF tests

There are times where we want to know if two samples are drawn from the same distribution and tests for this types of problems are called *two-sample GoF tests* or briefly *two-sample tests*. A natural and simple approach would be to compare the first two moments of the sample which measure location and scale. Many tests of this type can be found in the literature [4], [5], [6]. But distributions may differ in a more subtle way or the information about the distributions is not sufficient to provide any idea about what type of difference is likely to exist. Therefore, we present here a general framework for two-sample tests which compare the entire distributions of the two samples. In this framework we have two independent samples, one containing information about the distribution function F and the other containing information about the distribution function G. We are interested in testing the null hypothesis

$$H_0 : F(x) = G(x)$$

versus the alternative

$$H_1 : F(x) \neq G(x).$$

In principle the two-sample problem can be treated with similar methods as used in GoF techniques. The only difference lies in the fact that we have no informations about the underlying distributions F and G of the two samples.

Chapter 3

An overview of some relevant GoF tests

This chapter mainly deals with the most relevant GoF tests. It is however not the intention to be complete, because a listing of all published GoF tests with their important properties would fill at least a book and in addition not all GoF tests are relevant to physical experiments. A comprehensive overview of many GoF tests is given in [7]. Tests, which are considered in this chapter, are also discussed in detail in [8]

3.1 Tests based on binning

One classical approach to testing GoF is based on grouping observations into bins[1]. Tests based on binning can only be applied to continuous observations after the observations has been grouped, which already suggests that information will be lost, see Figure 3.1.

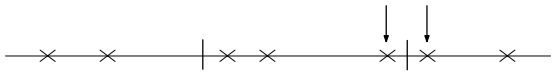

Figure 3.1: The information about the position of each observation inside the bin will be lost. That the observations indicated by the arrows are next to each other is not taken into account.

[1]The term *bin* is mostly used in physical literature, but *class* is the usual term for grouping the observations into a set of exhaustive and non-overlapping intervals.

3.1.1 Pearson's χ^2-test

Probably the best known, most frequently used and oldest GoF test is Pearson's χ^2-test. The χ^2-test is closely connected to least square fits. Under the assumption that the measurements y_i, $1 \leq i \leq n$, have a normally distributed error σ_i, the quantity

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - f_i)^2}{\sigma_i^2}, \tag{3.1}$$

where f_i is the value under H_0, is minimized in a least squares fit. From the Eq. (3.1) it is clear that if we have chosen the f_i and the σ_i correctly, then each term in the sum of Eq. (3.1) will be of order unity, and hence X^2 will be approximately equal to n. If it is, then we may conclude that the measurements are well described by the hypothesized function.

The Pearson's χ^2-test is used to test the GoF. It proceeds by binning the observations $X_1, X_2, \ldots, X_n$ from a distribution into B bins and comparing the measured frequencies with expected frequencies under H_0. More precisely, the general form of the test statistic can be written as

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i}, \tag{3.2}$$

then O_i is the number of observations that fall in the ith bin and E_i is the expected number of observations, which is given by $E_i = np_i$ where p_i is the probability content of the ith bin under H_0. The Pearson's χ^2-test (3.2) looks different from that shown in Eq. (3.1), since the variance of each observation is replaced by the mean value of O_i. Such an estimate for the variance makes sense in situations involving counting, where the counted numbers are distributed according to the Poisson distribution, for which the mean is equal to variance.

In the simplest case, Pearson's statistic is constructed for a simple null hypothesis. Then the χ^2 test statistic is asymptotically χ^2_{B-1} distributed. Often, the probabilities p_i that are specified under H_0 are still depending on an unknown parameter vector $\boldsymbol{\theta}$. These parameters need then to be estimated from the random sample. Different methods of estimating these parameters result in different tests [3]. The correct theory for a composite null hypothesis was first provided by Fisher [9], where he used maximum likelihood estimation method to estimate the parameters. Therefore, this test is often referred to as the Pearson-Fisher test and the test statistic is asymptotically χ^2_{B-k-1} distributed, where k is the number of parameters to be estimated. In both cases large values of χ^2 would indicate that the observations are not distributed according to H_0. It is sometimes asserted that H_0 should also be rejected for small values of χ^2. Arguments given for this assertion are that such small values are likely to have resulted from computational errors or overestimation of the measurement errors σ_i. But it appears on the other hand even more unlikely

to obtain a small χ^2 value using wrong hypothesis. Therefore small χ^2 should not be regarded as a reason for rejecting H_0.

The χ^2 test is very simple and needs only limited computational power. A big advantage compared to most of the other methods is that it can be applied to multidimensional histograms. There are however also serious drawbacks:

- Its power in detecting slowly varying deviations of a histogram from predictions is rather poor due to the neglect of possible correlations between adjacent bins.
- The specification of the bins is not unique.
- When the statistics is low or the number of dimensions is high, the number of observations per bin may be low. Then systematic deviations are hidden by statistical fluctuations.

3.1.2 Power divergence statistics

Pearson's χ^2-statistic is not the only statistic that is based on binning. Other well known statistics are the likelihood ratio LR and the Freeman-Tukey FT statistic. All these statistics (χ^2, LR and FT) have been embedded by [10] in the more general class of the so-called *power divergence statistics* $CR(\lambda)$, whose members are characterized through a parameter $\lambda \in \mathbb{R}$. We denote by Δ_i^λ the *power divergence* between the observed frequencies and expected frequencies in ith bin as follows

$$\Delta_i^\lambda = \frac{2}{\lambda(\lambda+1)} O_i \left[\left(\frac{O_i}{E_i} \right)^\lambda - 1 \right].$$

Δ_i^λ compares the fraction of the observed frequencies divided by the expected frequencies raised to the power λ with 1. Then the power divergence family of test statistics $CR(\lambda)$ is just the sum over all B bins of Δ_i^λ, i.e.

$$CR(\lambda) = \sum_{i=1}^{B} \Delta_i^\lambda.$$

For $\lambda = 1$ the statistic $CR(\lambda)$ reduces to Pearson's χ^2, to likelihood ratio LR when $\lambda \to 0$ and to Freeman-Tukey FT when $\lambda = -\frac{1}{2}$, respectively. Table 3.1 includes some other interesting cases.

Depending on λ, the deviation between observed frequencies O_i and expected frequencies E_i is weighted differently. The statistic of Pearson ($\lambda = 1$) is the only one which does not distinguish whether O_i with the same absolute distance lie above or below E_i. In so far, power divergences Δ_i^λ with $\lambda \neq 1$ are asymmetrical; for values $\lambda < 1$ O_i below E_i are weighted stronger, for $\lambda > 1$ O_i above E_i are weighted stronger.

Table 3.1: Certain values of λ indicate known GoF statistics

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i} = CR(1)$$

$$LR = 2\sum_{i=1}^{B} O_i \ln \frac{O_i}{E_i} = CR(0)$$

$$FT = 4\sum_{i=1}^{B} \left(\sqrt{O_i} - \sqrt{E_i}\right)^2 = CR\left(-\tfrac{1}{2}\right)$$

$$LR_m = 2\sum_{i=1}^{B} E_i \ln \frac{E_i}{O_i} = CR(-1) \text{ modified } LR$$

$$\chi^2_m = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{O_i} = CR(-2) \text{ Neyman's modified } \chi^2$$

For a simple null hypothesis $CR(\lambda)$ is asymptotically χ^2_{B-1} distributed and for a composite null hypothesis is asymptotically χ^2_{B-k-1} distributed as in case of Pearson's χ^2 statistic.

A power comparison of some interesting members of the family of power divergence statistics is given in [11].

3.2 Binning-free tests

As mentioned in previous section binning of observations loses information. Consequently we should expect tests based on binning to be inferior to tests based on each observation. Therefore this thesis is mainly focuses on construction of new nonparametric, omnibus, binning-free tests.

3.2.1 EDF-tests

A wide and diverse family of GoF-tests is based on the empirical distribution function (EDF) F_n. Before defining F_n we introduce the concept of *order statistics* denoted by $X_{(i)}$. It is just the observations X_i, which are ordered in some a way. In one dimension the order statistics obey

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

In more than one dimension it is rather arbitrary, since in more dimensional space an unique ordering scheme is missing.

F_n is defined by

$$F_n(x) = \frac{\text{number of observations } \leq x}{n}$$

or

$$\begin{array}{ll} F_n(x) = 0 & , \ x < X_{(1)} \\ F_n(x) = \frac{i}{n} & , \ X_{(i)} \leq x < X_{(i+1)} \\ F_n(x) = 1 & , \ X_{(n)} \leq x \end{array}$$

It is well known [12] that under the null hypothesis F_n is an unbiased and consistent estimator of the c.d.f. F. Even a stronger convergence holds

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right) = 1.$$

This is also known as *Glivenko-Cantelli theorem*.

EDF-tests consist of comparing F_n with F in some way. A overview of tests considered in this subsection is given in [7].

Supremum statistics

The Glivenko-Cantelli theorem suggests that the statistic

$$D = \sup_x |F_n(x) - F(x)| \tag{3.3}$$

is for any n a reasonable measure for GoF. The statistic (3.3) is called the *Kolmogorov-Smirnov statistic*. For the simple null hypothesis the limiting null distribution of D is given by

$$\lim_{n\to\infty} P\left(D \leq \frac{z}{\sqrt{n}}\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2} \quad \text{for every } z \geq 0. \tag{3.4}$$

In case of composite hypothesis there is no general asymptotic theory available, critical values must be estimated by simulation.

A closely related statistic to D is the Kuiper statistic V

$$V = D^+ + D^-,$$

where $D^+ = \sup_x [F_n(x) - F(x)]$ and $D^- = \sup_x [F(x) - F_n(x)]$. V is also useful for observations on a 'circle', for example for azimuthal distributions where the zero angle is a matter of definition.

Quadratic statistics

The Cramér-von Mises family of tests measures the integrated quadratic deviation of $F_n(x)$ from $F(x)$ suitably weighted by a weighting function ψ:

$$Q = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, \psi(x) \, dF, \tag{3.5}$$

dF can be written as $f(x)\,dx$ and hence the integration with respect to F may be regarded as averaging over the sample space.

Different choices of ψ in (3.5) result in different tests. Although many choices for ψ are allowed, in the literature mainly

$$\psi(x) \equiv 1$$
$$\psi(x) = \frac{1}{F(x)\,(1 - F(x))}$$

are chosen. The test with $\psi(x) \equiv 1$ is called the *Cramér-von Mises* (CM) *test* and $\psi(x) = \frac{1}{F(x)(1-F(x))}$ leads to the *Anderson-Darling* (AD) *test.* The weighting function in (AD) test upweights the differences between F_n and F in the tails of the distribution F. This is justified because there the experimental deviations are small.

A modification of CM is the Watson statistic U^2 defined by

$$U^2 = n \int_{-\infty}^{\infty} \left\{ F_n(x) - F(x) - \int_{-\infty}^{\infty} [F_n(x) - F(x)] \; dF(x) \right\}^2 dF(x),$$

which can also be applied for observations on a circle.

3.2.2 The Neyman Smooth test

For a simple H_0 where the form of $F_0(x)$ is completely specified, such as uniformity or normality, many existing tests, for example entropy tests [13], [14], could be employed in a straightforward manner. As an application of the likelihood ratio method to construct a test statistic we present a powerful test, *the Neyman Smooth* (NS) *test*, for the uniform distribution. Before defining the test statistic NS we introduce here a transformation method. With a *probability integral transformation*

$$z = F(x) \tag{3.6}$$

a given sample from a full specified distribution can be transformed to a sample from a uniform distribution. Therefore tests based on the transformed sample

$F_0(X_1), F_0(X_2), \ldots F_0(X_n)$ leading to tests for uniformity on the interval $[0, 1]$. From this point of view a GoF problem with a simple H_0 is equivalent to the problem

$$H_0 : F(x) = \text{uniform}$$
$$H_1 : F(x) \neq \text{uniform}$$

by using the transformation (3.6).

Note, however, that the transformation (3.6) does not necessarily conserve all interesting features of the GoF problem. For example, in a lifetime distribution an excess of observations at small and large lifetimes may be judged differently but are treated similarly after a probability integral transformation.

The test NS is different from all previously discussed tests, since the alternative hypothesis is parametrized. The alternative hypothesis density has the functional form

$$g_k(z) = C\left(\theta_1, \theta_2, \ldots, \theta_k\right) \exp\left[\sum_{i=1}^{k} \theta_i \pi_i(z)\right],$$

where π_i are the Legendre polynomials of order i, $C\left(\theta_1, \theta_2, \ldots, \theta_k\right)$ is a normalization constant and θ_i are free parameters. In the literature g_k is known as *exponential family*. The term 'smooth' refers to the characteristic that the specified distribution in H_0 is imbedded in a family of alternatives g_k which varies smoothly with the parameters $\theta_1, \theta_2, \ldots, \theta_k$. From the form of g_k it is clear that the test for uniformity reduces to

$$H_0 : \theta_i = 0 \quad \text{for all } i,$$
$$H_1 : \text{at least one of } \theta_i \neq 0.$$

Since the distribution function of the family of the alternatives is explicitly given, likelihood ratio test may be applied directly and it leads to the test statistic

$$NS = \frac{1}{n} \sum_{i=1}^{k} \left(\sum_{j=1}^{n} \pi_i(z_j) \right)^2 .$$

The number k of parameters is selected by the user. It seems that the most difficult problem connected with the application of NS is the choice of k. This problem is very similar to the problem of choosing the number of bins in the χ^2 GoF test. In the literature $k = 2$ and $k = 4$ are recommended, see [7]. NS is asymptotically distributed as χ^2 with k degrees of freedom and large values of NS lead to a rejection of H_0.

3.2.3 Tests based on density estimation

The most well known density estimator $\widehat{f}_n$ is the histogram. By allowing the bin width to vary another class of density estimator, *the kernel density estimator* (KDE), can be obtained. A KDE is constructed by centering a smooth *kernel* function about the observations and summing the heights of the kernels at each observation. The KDE $\widehat{f}_n$ can then be written as

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h\left(x, X_i; h\right), \end{aligned} \tag{3.7}$$

where K is itself a probability density, called kernel function, whose variance is controlled by the parameter h which is called *smoothing parameter*, see [15]. From (3.7) it is clear that a KDE inherits all the continuity and differentiability properties of the used kernel K. Therefore it is often convenient to use for K a normal density function. A KDE is illustrated in the Fig.3.2

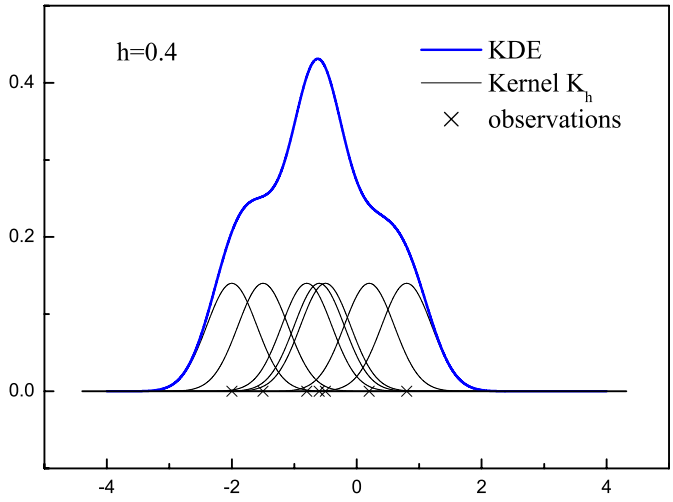

Figure 3.2: Kernel estimate using Gaussian kernel with smoothing parameter 0.4.

It should be noted that the behavior of a KDE is affected by the choice of h. When h is small the estimate displays spurious fine structure. When h is too large all detail is obscured. The effect of varying of h is illustrated in Fig.3.3

EDF test statistics are in some sense measures of the difference between the EDF and hypothesized distribution. A similar approach with density estimates can

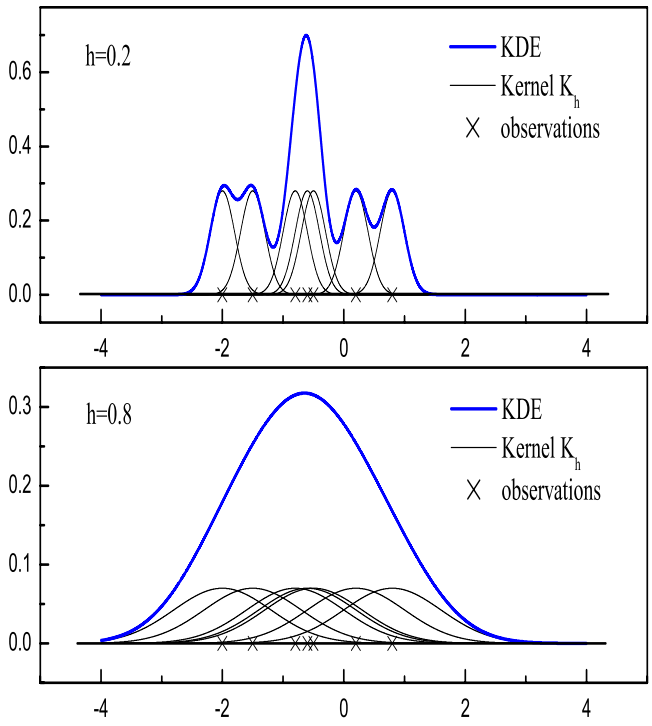

Figure 3.3: Kernel estimate using Gaussian kernel with two different smoothing parameter.

also be applied as test procedure. Instead of the EDF the density estimate $\widehat{f_n}$ is compared to the hypothesized p.d.f. For example, the idea of the Anderson-Darling test statistic with kernel density estimate $\widehat{f_n}$ leads to the test statistic

$$\int \frac{\left[f(x) - \widehat{f_n}(x)\right]^2}{Var\left[\widehat{f_n}(x)\right]} f(x)\, dx.$$

As shown in [16], $Var\left[\widehat{f_n}(x)\right]$ is asymptotically proportional to $f(x)$. Therefore we get the L_2 error of the kernel density estimate as a test statistic:

$$\int \left[f(x) - \widehat{f_n}(x)\right]^2 dx. \tag{3.8}$$

The idea of using density estimators for GoF tests goes back to [17] and [18].

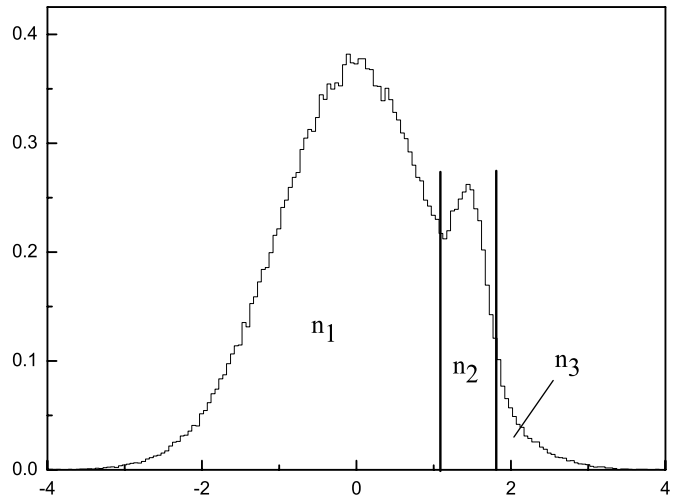

Figure 3.4: Searching for the maximum of the test statistic R_3 with three variable intervals, where the interval boundaries are placed at the observations.

3.3 Three region test

Often experimental distributions have local clusters as a result of statistical fluctuations. Sometimes though, the clustering is the result of a new physical effect. For this process of *bump hunting* we have designed a test which subdivides the variable space into three pieces, containing n_1, n_2 and $n_3 = n - n_1 - n_2$ observations, such that the deviation between observed sample and prediction from H_0 is maximum, see Figure 3.4.

The test statistic which we have developed is given by

$$R_3 = \sup_{n_1,n_2} \left[w_1 \left(n_1 - np_1\right)^2 + w_2 \left(n_2 - np_2\right)^2 + w_3 \left(n_3 - np_3\right)^2 \right],$$

where np_i are the expected values and w_i weights depending on np_i. In the power simulation study R_3 is carried out with weights equal to one. We did not investigate the consistency and biasness of this special test for bump hunting.

Clearly, the three region test can be extended to test for densities containing high frequency components, but in physical applications one is mainly confronted with slowly varying deviations between hypothesized distribution and observed distribution whereas in other fields where for example time series are investigated, high frequency distortions are more likely.

3.4 Multivariate normality tests

In sociology or medicine it is rather common to deal with samples of observations which are drawn from a multivariate normal distribution. Therefore much of multivariate GoF tests have been developed mostly for multinormality and a current review of the literature has revealed at least 50 tests for multivariate normality. We select 4 promising tests of multivariate normality.

Mardia's tests

The third standardized moment characterizes the skewness of a distribution. The skewness of a univariate normal distribution is 0. The fourth standardized moment characterizes the kurtosis of a distribution. The kurtosis of a univariate normal distribution is 3. Therefore univariate sample measures of skewness and kurtosis may be used for testing univariate normality. In [19] a generalization of these statistics to test a hypothesis of multivariate normality is given. The test statistic for multivariate skewness is

$$b_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(\mathbf{X}_i - \overline{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_j - \overline{\mathbf{X}}) \right]^3 .$$

and the corresponding test statistic for kurtosis is

$$b_2 = \frac{1}{n} \sum_{i=1}^{n} \left[(\mathbf{X}_i - \overline{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}) \right]^2 ,$$

where $\mathbf{S}$ is the sample covariance matrix.

Neyman smooth test

In [3] a multivariate version of Neyman smooth test for multivariate normality is given. The formulation of the test statistic would fill at least two pages, we refer to [3].

BHEP test

[20] proposed a test, which is called BHEP test, of multivariate normality based on the following statistic:

$$T_\beta = \int |P(t) - Q(t)|^2 \, \varphi_\beta(t) \, dt,$$

where $P(t)$ is the characteristic function of the multivariate normal distribution, $Q(t)$ is the empirical characteristic function, $\varphi_\beta(t)$ is a kernel function which was chosen to be $N\left(0, \beta^2 \mathbf{I}_d\right)$ ($\mathbf{I}_d = d \times d$ unit matrix) in order to obtain a simple closed

form expression for T_β and β is a smoothing parameter. Carrying out the integration leads to

$$T_\beta = \frac{1}{n}\sum_{i,j=1}^{n} \exp\left(-\frac{\beta^2}{2}\left[D_{ii} - 2D_{ij} + D_{jj}\right]\right) + \frac{n}{\left(1+2\beta^2\right)^{d/2}} + \\ - \frac{2}{\left(1+\beta^2\right)^{d/2}} \sum_{i=1}^{n} \exp\left(-\frac{\beta^2 D_{ii}}{2\left(1+\beta^2\right)}\right),$$

where

$$D_{ij} = (\mathbf{X}_i - \overline{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{X}_j - \overline{\mathbf{X}}), \quad 1 \leq i, j \leq n,$$

and d is the dimension of the variate space.

The statistic T_β was proposed by [21] for the univariate case $d = 1$. An extension for the multivariate case $d > 1$ is given in [22] for $\beta = 1$. By [23] the term BHEP tests is introduced with reference to the four authors of the papers [21] and [22].

Chapter 4

The quantity *energy* as a GoF test

Tests which are based on the order statistics (some of them are discussed in previous chapter) play an important role in univariate GoF problems. If one tries to generalize order statistics to the multivariate case, one is faced with the problem of defining an ordering scheme. Because of the absence of a natural linear order in $\mathbb{R}^d$ it is not clear how to define a multivariate order statistic in a meaningful way. Therefore the extension of the EDF tests to the multivariate case is difficult. However tests based on binning suffer from the arbitrariness of binning and from lack of power for small samples, since a high dimensional space is essentially empty which is known in the literature under the term *curse-of-dimensionality* [24]. In this chapter we propose a new class of nonparametric, multivariate, omnibus GoF tests which avoid ordering and binning of the observations. The new class of tests is called Energy tests, because the definition of the test statistic is closely related to the energy of electric charge distributions. The Energy tests are essentially powerful in multivariate testing problems.

4.1 The interaction energy of a system of charges

The energy of a system of interacting charges can be calculated simply in the following way. Let a charge e_1 be fixed at $\mathbf{r}_1$. A second charge e_2, which was at infinity, is displaced to a point $\mathbf{r}_2$ located at a distance $|\mathbf{r}_2 - \mathbf{r}_1|$ from the first charge. In this case we must do work W_{12} against the forces of the field of the first charge:

$$W_{12} = -e_2 \int_{\infty}^{\mathbf{r}_2} \mathbf{E}_1 \, d\mathbf{r} = e_2 \left(\varphi_1(\mathbf{r}_2) - \varphi_1(\infty)\right),$$

where $\mathbf{E}_1$ is the field of the first charge. $\varphi_1(\mathbf{r}_2)$ represents the potential of $\mathbf{E}_1$ at the point $\mathbf{r}_2$, which is given by

$$\varphi_1(\mathbf{r}_2) = \frac{1}{4\pi\epsilon_0} \frac{e_1}{|\mathbf{r}_2 - \mathbf{r}_1|}.$$

Hence the work W_{12} of displacement of the second charge is equal to

$$W_{12} = e_2\varphi_1(\mathbf{r}_2) = \frac{1}{4\pi\epsilon_0}\frac{e_1e_2}{|\mathbf{r}_2 - \mathbf{r}_1|},$$

since the potential of the field of the first charge at infinity is equal to 0.

If a third charge e_3 is added to a system of two charges, one has to expend the work

$$W_{123} = \frac{1}{4\pi\epsilon_0}\frac{e_1e_3}{|\mathbf{r}_3 - \mathbf{r}_1|} + \frac{1}{4\pi\epsilon_0}\frac{e_2e_3}{|\mathbf{r}_3 - \mathbf{r}_2|}.$$

Continuing with such a procedure for a system of n charges, it is necessary to produce the work

$$\begin{aligned} W &= \frac{1}{4\pi\epsilon_0}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{e_ie_j}{|\mathbf{r}_i - \mathbf{r}_j|} \\ &= \frac{1}{8\pi\epsilon_0}\sum_{i\neq j}^{n}\frac{e_ie_j}{|\mathbf{r}_i - \mathbf{r}_j|}. \end{aligned} \tag{4.1}$$

Passing over from point charges to a continuous charge density distribution ρ, we can write (4.1) in the form

$$W = \frac{1}{8\pi\epsilon_0}\int\int\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\,d\mathbf{r}d\mathbf{r}'. \tag{4.2}$$

The total energy W of an external continuous charge density distribution ρ_{ex} with a continuous charge density distribution ρ can be easily obtained by the Eq. (4.2):

$$W = \frac{1}{8\pi\epsilon_0}\int\int\frac{[\rho(\mathbf{r}) + \rho_{ex}(\mathbf{r})]\,[\rho(\mathbf{r}') + \rho_{ex}(\mathbf{r}')]}{|\mathbf{r} - \mathbf{r}'|}\,d\mathbf{r}d\mathbf{r}'. \tag{4.3}$$

In Figure 4.1 both continuous charge density distributions ρ_{ex} and ρ are illustrated.

We now consider a system of a positively continuous charge density distribution ρ and a negatively external continuous charge density distribution $-\rho_{ex}$ with

$$\int[\rho(\mathbf{r}) - \rho_{ex}(\mathbf{r})]\,d\mathbf{r} = 0, \tag{4.4}$$

i.e. we fix the total charge to zero. Physics tell us that for that case the energy of this system is zero, i.e. the state of minimum energy, the vacuum, is free of charges. In this thesis we will use this property to compare statistical distributions.

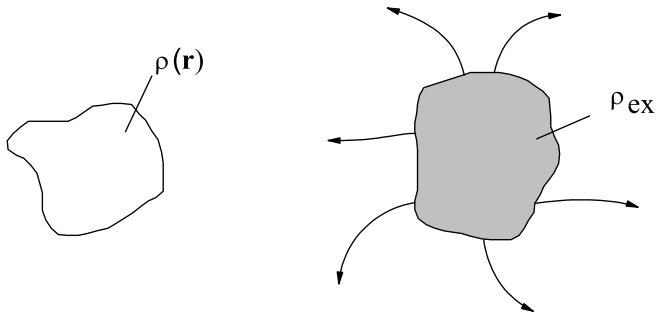

Figure 4.1: Illustration of the interaction of two continuous charge density distributions.

4.2 The Energy tests

4.2.1 The idea

The energy W, which is given by (4.3), can be used simply as a test statistic in the following way. We consider a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a system of positive charges and a second sample $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ as a system of negative charges. The charges are normalized such that each sample contains a total charge of one unit. From electrostatics we know that in the limit of where n, m tend to infinity, the total potential energy of the pooled sample computed for a potential following an one-over-distance law will be minimum if both charge samples have the same distribution. In this limit any displacement of charges would increase the energy. We use this property to construct a binning-free multivariate test procedure.

4.2.2 The new test statistics

Corresponding to (4.3) with (4.4) we define a quantity ϕ, the energy, which measures the difference between two p.d.f.s $f_0(\mathbf{x})$ and $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, by

$$\phi = \frac{1}{2} \int \int [f(\mathbf{x}) - f_0(\mathbf{x})] \, [f(\mathbf{x}') - f_0(\mathbf{x}')] \, R(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'. \tag{4.5}$$

Here and in what follows, an unspecified integral denotes integration over $\mathbb{R}^d$. The distance function $R(\mathbf{x}, \mathbf{x}')$ is a monotonically decreasing function of the Euclidian distance $|\mathbf{x} - \mathbf{x}'|$. Relation (4.5) with $R(\mathbf{x}, \mathbf{x}') = 1/|\mathbf{x} - \mathbf{x}'|$ is proportional to the

electrostatic energy of two charge distributions f and f_0 of opposite sign which is minimum if the charges neutralize each other. In this thesis we have considered three different distance functions R which are discussed in section 4.4. The coefficient $\frac{1}{2}$ in (4.5) is introduced because the same terms, corresponding to $\mathbf{x}$ and $\mathbf{x}'$, are considered twice.

We want to compare a p.d.f. $f(\mathbf{x})$ to a reference p.d.f. $f_0(\mathbf{x})$ which we consider as being fixed, i.e. our GoF problem can be formulated as

$$H_0 : f(\mathbf{x}) = f_0(\mathbf{x})$$
$$H_1 : f(\mathbf{x}) \neq f_0(\mathbf{x}).$$

The vacuum is free of charges and therefore it is characterized by the minimum of energy. Hence, ϕ must equal zero under H_0 and otherwise positive. Mathematically, this assertion is not immediately obvious. We will see in the next section that the test statistic (4.5) is a non-negative functional under some constraints of the distance function $R(\mathbf{x}, \mathbf{x}')$.

The formula in Eq. (4.5) is however not convenient for computation. Expanding (4.5)

$$\begin{aligned} \phi &= \frac{1}{2}\int\int \left[f(\mathbf{x})f(\mathbf{x}') + f_0(\mathbf{x})f_0(\mathbf{x}') - 2f(\mathbf{x})f_0(\mathbf{x}')\right] R(|\mathbf{x}-\mathbf{x}'|)d\mathbf{x}d\mathbf{x}' \\ &= \frac{1}{2}E_1 + \frac{1}{2}E_2 - E_3 \end{aligned} \tag{4.6}$$

we obtain three terms which are the expectation values of R.

Statistics is concerned with finite samples. To obtain the Energy statistic ϕ_n of a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ relative to f_0 we consider the first term E_1 in (4.6). This is the expectation value

$$E_1 = E(R) = \int g(\mathbf{y})R(\mathbf{y})d\mathbf{y}$$

of the distance function $R(\mathbf{y})$ relative to the p.d.f. $g(\mathbf{y})$, where $\mathbf{y} = (\mathbf{x}, \mathbf{x}')$ and $g(\mathbf{y}) = f(\mathbf{x})f(\mathbf{x}')$. We accept only distance functions R for which its expectation value $E(R)$ and variance $V(R)$

$$V(R) = \int g(\mathbf{y})\left(R(\mathbf{y}) - E(R)\right)^2 d\mathbf{y}$$

exist. From the law of large numbers follows that the sample mean of n observations from a distribution with mean μ and finite variance will converge towards μ as n becomes large, see [25], [26]. Hence E_1 can consistently be estimated from a corresponding sample mean.

Note that to obtain one observation $\mathbf{y}$ two independent observations $\mathbf{x}_i$, $\mathbf{x}_j$ both drawn from $f(\mathbf{x})$ are required. Thus for a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, we can construct the sampling version E_{1n} of E_1 by splitting the sample into two parts in the following way:

$$E_{1n} = \frac{2}{n} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} R(|\mathbf{x}_i - \mathbf{x}_j|). \tag{4.7}$$

The splitting of the sample used in (4.7) is arbitrary. We can improve the precision of the sample mean by averaging over all possible splittings of the sample into two parts

$$E_{1n} = \frac{1}{n(n-1)} \sum_{i<j}^{n} R(|\mathbf{x}_i - \mathbf{x}_j|).$$

E_{1n} converges to E_1 in the limit of large n, since it is a consistent estimator of E_1. This leads to the energy statistic ϕ_n of a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$

$$\begin{aligned}\phi_n = &\frac{1}{n(n-1)} \sum_{i<j}^{n} R(|\mathbf{x}_i - \mathbf{x}_j|) + \frac{1}{2} \int \int f_0(\mathbf{x}) f_0(\mathbf{x}') R(|\mathbf{x} - \mathbf{x}'|) d\mathbf{x} d\mathbf{x}' + \\ &- \frac{1}{n} \sum_{i=1}^{n} \int f_0(\mathbf{x}') R(|\mathbf{x}_i - \mathbf{x}'|) d\mathbf{x}'. \end{aligned} \tag{4.8}$$

Since the evaluation of ϕ_n usually requires a sum over difficult integrals, we prefer to represent f_0 by a sample $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$, usually generated through a Monte Carlo simulation. In many experimental situations f_0 is anyhow available only in form of Monte Carlo simulations. Statistical fluctuations of the simulation are negligible if m is large compared to the observed sample size n, typically $m \geq 10n$. The sampling version of (4.6) can easily be obtained from two samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ drawn from f and f_0, respectively:

$$\begin{aligned}\phi_{nm} = &\frac{1}{n(n-1)} \sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \frac{1}{m(m-1)} \sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) + \\ &- \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right). \end{aligned} \tag{4.9}$$

4.2.3 The distance function

A discrepancy between a sample and the theoretical description can be of different origin. The problem may be in the theory which is wrong or the sample may be biased by measurement errors or by background contamination. In physical experiments we mainly have the latter situation. Even though the statistical description is

the same in both cases the choice of the specific test may be different. In our application in physics we are mainly confronted with slowly varying deviations between f and f_0 whereas in other fields where for example time series are investigated, high frequency distortions are more likely. Therefore R should be adjusted to a specific statistical problem. For slowly varying deviations between f and f_0 a long range distance function would be preferred.

4.2.4 Normalization of the distances

The Euclidean distances between two points $\mathbf{z}_i$ and $\mathbf{z}_j$ in $\mathbb{R}^d$ is

$$|\mathbf{z}_i - \mathbf{z}_j| = \sqrt{\sum_{k=1}^{d}(z_{i,k} - z_{j,k})^2} \tag{4.10}$$

with projections $z_{i,k}$ and $z_{j,k}$, $k = 1, \ldots, d$, of the vectors $\mathbf{z}_i$ and $\mathbf{z}_j$.

Since the relative scale of the different variates usually is arbitrary, we propose to normalize the projections by the following transformation

$$z_{ik}^{*} = \frac{z_{i,k} - m_k}{s_k} \quad , \quad \begin{matrix} i = 1, \ldots, n+m \\ k = 1, \ldots, d \end{matrix} \quad ,$$

where m_k, s_k are the empirical estimate of the mean and variance of the projection $z_{1,k}, \ldots, z_{(n+m),k}$ of the coordinates of the observations of the pooled sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_{n+m}) = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$. In this way we avoid that a single projection dominates the value of the energy and that other projections contribute only little to it.

To allow a direct comparison of the Energy tests with some selected tests from the literature, we did not apply this transformation in our power studies.

4.3 Proof of the minimum property of ϕ

The Energy test statistic is based on the assertion that the energy ϕ is zero only for H_0 and positive for H_1. To show that $\phi > 0$ implies $f(\mathbf{x}) \neq f_0(\mathbf{x})$ we substitute $h(\mathbf{x}) = f(\mathbf{x}) - f_0(\mathbf{x})$ and obtain from Eq. (4.5)

$$\phi = \frac{1}{2} \int \int h(\mathbf{x}) h(\mathbf{x}') R(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'. \tag{4.11}$$

We replace in the Eq. (4.11) the distance function $R(\mathbf{x}, \mathbf{x}') = R(|\mathbf{x} - \mathbf{x}'|)$, which also called kernel, by its Fourier integral $\tilde{R}(\mathbf{k})$

$$R(|\mathbf{x} - \mathbf{x}'|) = \int \tilde{R}(\mathbf{k}) e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} d\mathbf{k}$$

and obtain

$$\begin{aligned}\phi &= \frac{1}{2}\int\int\int h(\mathbf{x})h(\mathbf{x}')\tilde{R}(\mathbf{k})e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}d\mathbf{x}d\mathbf{x}'d\mathbf{k} \\ &= \frac{1}{2}\int \left|\tilde{h}(\mathbf{k})\right|^2 \tilde{R}(\mathbf{k})d\mathbf{k} \end{aligned} \tag{4.12}$$

where $\tilde{h}(\mathbf{k})$ is the Fourier transform of $h(\mathbf{x})$.

If

$$\tilde{R}(\mathbf{k})>0$$

we get from Eq. (4.12) that ϕ is zero only for $h(\mathbf{x}) = f(\mathbf{x}) - f_0(\mathbf{x}) \equiv 0$ and positive for $h(\mathbf{x}) \neq 0$. Therefore the new test statistics are only applicable with decreasing distance functions which have positive Fourier transforms or equivalently the kernel $R(\mathbf{x}, \mathbf{x}')$ must be positive definite. Clearly, we may also use monotonically increasing distance functions which should have negative Fourier transformation. For these distance functions $f(\mathbf{x}) \neq f_0(\mathbf{x})$ would imply $\phi < 0$.

Note that the minimum energy requirement for the equality of f and f_0 is strictly correct only when the size of Monte Carlo sample, m, is equal to n, where n is the size of the observed sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ which is drawn from f. For the general case with a continuous distribution f_0 or Monte Carlo sample and observed sample of different size, the optimum agreement of the two distributions is not well defined and there is a slight dependence of the minimum energy configuration on the distance function. The assertion that the Energy statistic ϕ_{nm} is zero only for H_0 is valid when n, m tend to infinity. For small n, m we did not success to proof this assertion. In special case of $R(\mathbf{x}, \mathbf{x}') = |\mathbf{x} - \mathbf{x}'|$, where the Fourier transform of $|.|$ is negative, the assertion of the maximum condition of the energy for $n = m$ is proven in [27] where one can see the difficulty of this problem even for a simple distance function. This assertion was originally formulated by [28] in the form of a question as

> *For equal numbers of black and white points in euclidean space the sum of the pairwise distances between points of equal color is less than or equal to the sum of the pairwise distances between points of different color, and equality holds only in the case when black and white points coincide.*

We assert that for $n = m$ the minimum condition still applies if we have to replace the factors $1/(n-1)$ and $1/(m-1)$ by $1/n$ in (4.9):

$$\begin{aligned}\phi_{nn} = &\frac{1}{n^2}\sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \frac{1}{n^2}\sum_{i<j}^{n} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) + \\ &- \frac{1}{n^2}\sum_{i,j=1}^{n} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right).\end{aligned}$$

To demonstrate the minimum condition which leads to $\phi_{nn} = 0$ for two samples of equal sizes which coincide, we apply an infinitesimal shift to one observation $\mathbf{x}_i - \mathbf{y}_i = \delta\mathbf{x}_i$. Note that only the pair $\mathbf{x}_i, \mathbf{y}_i$ contributes to the energy, all other terms cancel, i.e. the change of the total energy is given by

$$\Delta\phi_{nn} = \frac{1}{n^2}\left[R\left(|\delta\mathbf{x}_i|\right) - R(0)\right].$$

Since R decreases with its argument, $R\left(|\delta\mathbf{x}_i|\right) - R(0) < 0$, we have found a local minimum of the energy. For $n \neq m$ this conclusion is not obvious, however experimentally the test statistic ϕ_{nm} has shown to be powerfull also in this case.

4.4 Some selected distance functions

We note that the minimum energy configuration does not depend on the application of the one-over-distance power law of electrostatics. We may apply a class of distance functions with the requirement that R has to decrease monotonically and the Fourier transform of R must be positive. Therefore different choice of R leads to different tests.

We know that the Fourier transform of a Dirac Delta function is positive, hence we can set $R\left(\mathbf{x}, \mathbf{x}'\right) = \delta(\mathbf{x} - \mathbf{x}')$ in (4.5) and we obtain

$$\phi = \frac{1}{2}\int \left[f(\mathbf{x}) - f_0(\mathbf{x})\right]^2 d\mathbf{x}.$$

For $R\left(\mathbf{x}, \mathbf{x}'\right) = \delta(\mathbf{x} - \mathbf{x}')$ the ϕ reduces to the integrated quadratic difference of the two p.d.f.s. The Dirac Delta function is not well suited for testing GoF, since we are not only interested in local differences. For that reason we have considered three different distance functions which correlate different locations:

$$R\left(|\mathbf{x} - \mathbf{x}'|\right) = \frac{1}{|\mathbf{x} - \mathbf{x}'|^{\kappa}}\ ,\ 0 < \kappa < d/2 \tag{4.13}$$

$$R\left(|\mathbf{x} - \mathbf{x}'|\right) = -\ln\left(|\mathbf{x} - \mathbf{x}'|\right) \tag{4.14}$$

$$R\left(|\mathbf{x} - \mathbf{x}'|\right) = e^{-|\mathbf{x}-\mathbf{x}'|^2/\left(2s^2\right)} \tag{4.15}$$

The first type of the distance function is motivated by the analogy to electrostatics, the second is long range and the third emphasizes a limited range for the correlation between different observations. The power κ of the denominator in Eq. (4.13) and the parameter s in Eq. (4.15) may be chosen differently for different dimensions of the sample space and different applications. For example, for long range distortions a small value of κ and for short range deviations a large value of κ would be used.

The inverse power law and the logarithm have singularity at $\mathbf{x} = \mathbf{x}'$. In principle these singularities are not a problem since the expectation values exist, however due

to rounding, limited resolution in measurement scales, and so forth $\mathbf{x} = \mathbf{x}'$ might occur. They can be easily handled by introducing a cut-off parameter ε. Very small distances, however, should not be weighted too strongly, since distortions with sharp peak are not expected and usually inhibited by finite experimental resolution. To prevent this, a simple remedy is to reset any such distances of length zero to some small positive value ε. The value of the cut-off parameter ε is not critical, it should be of the order of the average distance in the regions where the f_0 is maximum and not less than the experimental resolution. A cut-off parameter is also desirable for another reason. Comparing two finite samples modifications of the distances by values which are very small compared to the average distance of the most dense regions should not matter. The choice of the cut-off parameter could be left to the user of the test. We have not observed a statistically significant change of the power of the Energy tests when we varried ε by an order of magnitude.

It is well known that the Fourier transform of the Gaussian function is also Gaussian and hence is positive. We have to show that the Fourier transforms of the other two distance functions are positive. The distance functions considered in this thesis are symmetrical functions and it is known if $R(|\mathbf{r}|)$ is a function only of the modulus of $\mathbf{r}$, then its Fourier transform $F(\mathbf{k})$ is also a function only of the modulus of $\mathbf{k}$, see [29].

For the function $R(|\mathbf{r}|) = R(r) = \frac{1}{r^\kappa}$ with $d > \kappa$, where d is the dimension of $\mathbf{r}$, the Fourier transformation $F(\mathbf{k})$ is [30]:

$$F(k) = 2^{d-\kappa}\pi^{d/2}\frac{\Gamma\left(\frac{d-\kappa}{2}\right)}{\Gamma\left(\frac{\kappa}{2}\right)}k^{\kappa-d} > 0$$

with $k = |\mathbf{k}|$.

Note that the second moments

$$\int\int f(x)f(x')R^2(|x-x'|)dxdx', \int\int f(x)f_0(x')R^2(|x-x'|)dxdx'$$

for the specific kernel $R(|\mathbf{r}|) = R(r) = \frac{1}{r^\kappa}$ with $d > 2\kappa$ are finite because $R^2(|\mathbf{r}|) = R^2(r) = \frac{1}{r^{2\kappa}}$ is again a positive definite kernel of the same type. Hence ϕ_{nm} is a consistent estimator of ϕ with the distance function $R(|\mathbf{r}|) = R(r) = \frac{1}{r^\kappa}$, $d > 2\kappa$.

The logarithmic distance function $R(r) = -\ln(r)$ can be considered as the $\kappa \to 0$ limit of the power law distance function (4.13), i.e. it yields

$$-\ln r = \lim_{n\to\infty} n\left(\left(\frac{1}{r}\right)^{1/n} - 1\right).$$

The corresponding test quantiles of ϕ_{nm} are invariant under a linear transformation $r \to ar$. In the remainder of this thesis, we will restrict the attention mainly to the logarithmic distance function.

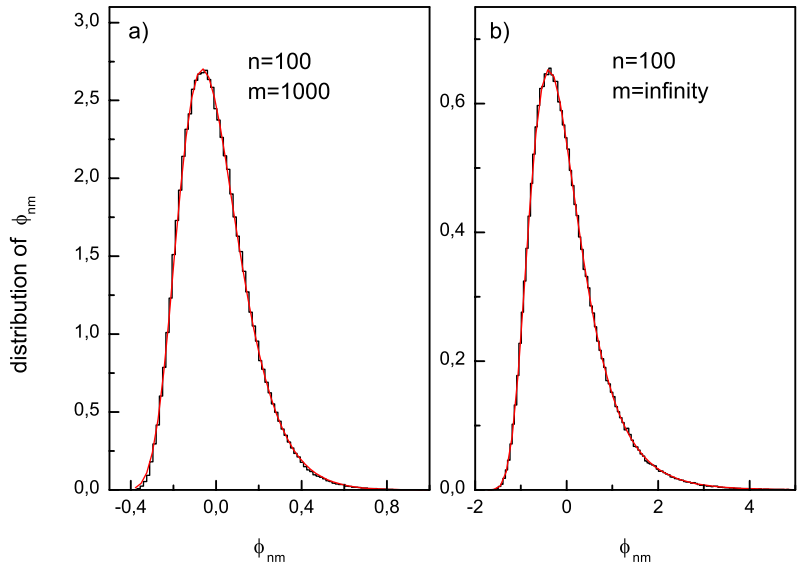

Figure 4.2: Energy distributions for a Gaussian distance function a) and a logarithmic distance function b) and their approximation by generalized extreme value distributions, where the term depending only on f_0 which is independent from the sample is not taken into account.

4.5 The distribution of the Energy test statistic

The distribution of the Energy test statistic depends on the probability distribution function f_0 and on the distance function R. Figure 4.2 shows the energy distributions for uniform f_0 with Gaussian and logarithmic distance functions. The distributions are well described by a generalized extreme value distribution

$$f(x) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi-1}\exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}$$

depending on three parameters, a scale parameter σ, a location parameter μ, and a shape parameter ξ. This class of generalized extreme value distributions is considered, since the energy quantity in GoF testing is in some sense an extreme value. Rather than computing these parameters from the moments of the specific ϕ distributions, we propose to generate the distribution of the test statistic and the quantiles by a Monte Carlo simulation. As a consequence of the dramatic increase of computing power during the last decade, it has become possible to perform the calculations on a simple PC within minutes. There is no need to publish tables of critical values.

4.5.1 Relation to U-statistics

To introduce some basic properties of U-statistics we follow [31] and [32]. For more details on U-statistics, [33] is recommended.

We assume that the c.d.f. F is completely unknown and let $X_1, X_2, \ldots, X_n$ be i.i.d. with F. Consider a 'parametric function' $\theta = \theta(F)$ which may be, for example, the expectation, variance or the median and so on of F.

U-statistics $U(X_1, X_2, \ldots, X_n)$ form a class of unbiased estimators of θ. For instance to estimate

$$\theta(F) = \text{mean of } F = E(X) = \int x dF(x)$$

one will use the sample mean

$$U(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Other statistics for θ may correlate different X_i's, for example

$$\theta = E\varphi(X_1, X_2, \ldots, X_r), \tag{4.16}$$

where we assume that φ, called a *kernel*, is a symmetric function in its r arguments.

A U-statistic $U(X_1, X_2, \ldots, X_n)$ of order r with a kernel φ for estimation of θ is obtained by averaging the kernel φ symmetrically over the observations:

$$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{r}} \sum_{(1 \le i_1 < i_2 < \ldots < i_r \le n)} \cdots \sum \varphi(X_{i_1}, X_{i_2}, \ldots, X_{i_r}). \tag{4.17}$$

U-statistics can be generalized in a natural way to k several samples by

$$U = \frac{1}{\binom{n_1}{r_1} \cdots \binom{n_k}{r_k}} \sum \varphi(X_{1i_1}, X_{1i_2}, \ldots, X_{1i_r}; \ldots; X_{kj_1}, X_{kj_2}, \ldots, X_{kj_s}), \tag{4.18}$$

for example, for $k = 2$ we get

$$U = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum \varphi(X_{i_1}, X_{i_2}, \ldots, X_{i_r}; Y_{j_1}, Y_{j_2}, \ldots, Y_{j_s}), \tag{4.19}$$

where n is the size of the first sample and m is the size of the second sample.

From (4.17) and (4.19) it is clear that the Energy test statistic ϕ_{nm} can be expressed as a sum of three different U-statistics of order 2 with a symmetric kernel

R. Therefore our main concern here is the asymptotic behavior of U-statistics. In the theory of one sample U-statistics it is known that

$$\lim_{n\to\infty} \frac{U-\theta}{\sqrt{Var(U)}} \sim N(0,1)$$

where

$$Var(U) = \left[\sum_{i=1}^{r} \binom{r}{i}\binom{n-r}{r-i}\sigma_i^2\right] \frac{(n-r)!r!}{n!} \tag{4.20}$$

if

$$Var\varphi_i\left(X_1, X_2, \ldots, X_i\right) = \sigma_i^2 < \infty \text{ for all } i = 1, 2, \ldots, r.$$

$\sim$ is read as 'is distributed according to'. A same result holds for U-statistics of two samples, namely a two sample U-statistic is also asymptotically distributed as a Gaussian.

These results might lead to the assumption that the asymptotical distribution of ϕ_{nm} must be Gaussian. But we see from Figure 4.2 that it is a skew distribution and not Gaussian. We have also determined for large sample size, $n = 1000$, the distribution of ϕ_{nm} for uniform f_0 with a logarithmic distance function and we got a skew distribution and not a Gaussian. The reason for that lies in the fact that the three sums in (4.9) are not independent. The first two sums in (4.9) are independent, but they are correlated with the third sum. A theoretical determination of the asymptotical distribution of ϕ_{nm} seems difficult.

4.6 Consistency

One important aspect which a test must have is that, as the sample size increases, it should distinguish better between the hypotheses being tested. A test is termed *consistent* if the power tends to unity as the number of observations increases.

To show that the Energy test is consistent we investigate the variance of the three terms of ϕ_{mn}. The first two terms are one sample U-statistics and we get from (4.20) for $r = 2$

$$\begin{aligned} Var(U) &= \left[\sum_{i=1}^{2} \binom{2}{i}\binom{n-2}{2-i}\sigma_i^2\right] \frac{(n-2)!2!}{n!} \\ &= \frac{4(n-2)}{n(n-1)}\sigma_1^2 + \frac{2}{n(n-1)}\sigma_2^2 \to \frac{4\sigma_1^2}{n}. \end{aligned}$$

If $\sigma_1^2 = Var\left[R(X)\right] < \infty$, which is in most situation given, then the variance of the first term tends to 0 at rate $\frac{4\sigma_1^2}{n}$. The same holds for the second term, since X may be replaced by Y.

If the covariances

$$\begin{aligned}\sigma_{10}^2 &= Cov[\varphi(X_1, X_2, \ldots, X_r; Y_1, \ldots, Y_s),\\ &\quad \varphi(X_1, X_2' \ldots, X_r'; Y_1', \ldots, Y_s')]\\ \sigma_{10}^2 &= Cov[\varphi(X_1, \ldots, X_r; Y_1, Y_2, \ldots, Y_s),\\ &\quad \varphi(X_1', \ldots, X_r'; Y_1, Y_2' \ldots, Y_s')]\end{aligned}$$

are > 0 and if $N = n + m$ and

$$\frac{n}{N} \to \rho, \frac{m}{N} \to 1 - \rho \text{ with } 0 \leq \rho \leq 1,$$

then for the asymptotic variance of the two sample U-statistics holds

$$Var(U) = \frac{1}{N}\left(\frac{r^2}{\rho}\sigma_{10}^2 + \frac{s^2}{1-\rho}\sigma_{01}^2\right), \tag{4.21}$$

see [31].

The asymptotic variance of the last term of ϕ_{nm} is equal to (4.21) with $r = s = 1$ and $\varphi = R$.

We have shown that the variance of the test statistic ϕ_{nm} tends to 0 if the sample sizes of the observed and Monte Carlo samples increase. Therefore the test is consistent.

A test is called *biased*, if for some members of H_1 the probability of rejecting H_0 is smaller when H_0 is false than when it is true. To determine whether the Energy test is biased or not is very difficult. But the consistency of the Energy test implies that it is asymptotically unbiased, see [25].

4.7 A link between ϕ_{nm} and the Bowman-Foster test statistic

The Bowman-Foster (BF) test for multivariate normality can be constructed from a kernel density estimate $\widehat{f}_n(\mathbf{x})$ of the null hypothesis density $f_0(\mathbf{x})$. The test statistic

$$BF = \int \left(\widehat{f}_n(\mathbf{x}) - E\left[\widehat{f}_n(\mathbf{x})\right]\right)^2 d\mathbf{x} \tag{4.22}$$

is an integrated squared error of $\widehat{f}_n(\mathbf{x})$ and its expected value $E\left[\widehat{f}_n(\mathbf{x})\right]$, where

$$E\left[\widehat{f}_n(\mathbf{x})\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[K_h\left(\mathbf{x}, \mathbf{X}_i; h\right)\right] = \int K_h\left(\mathbf{x}, \mathbf{y}; h\right)\, d\mathbf{y},$$

i.e. $E\left[\widehat{f}_n(\mathbf{x})\right]$ is a folding of $f_0(\mathbf{x})$.

$\widehat{f}_n(\mathbf{x})$ is a biased estimator, see [15]. Hence it is inappropriate to consider the L_2 distance between $\widehat{f}_n(\mathbf{x})$ and the hypothesized density $f_0(\mathbf{x})$ as a test statistic. For that reason (4.22) is a suitable test statistic. It is shown in [34] that when the kernel function is a normal density then for testing multivariate normality the integration in (4.22) can be carried out analytically. In concrete applications it is tedious or impossible to compute analytically the integrations involved in (4.22). But they can be replaced by a Monte Carlo integration. To avoid additional statistical uncertainties, the number of simulated observations should be large compared to the sample size. This is not a disadvantage with today's computing power on simple PCs.

The Energy test is related to Bowman-Foster test [34]. To show the link between ϕ_{nm} and BF we put the multivariate version of (3.7) into (4.22) and obtain

$$BF = \int \left[\frac{1}{n} \sum_{i=1}^{n} K_h\left(\mathbf{x}, \mathbf{X}_i; h\right) - \int f_0(\mathbf{y}) K_h\left(\mathbf{x}, \mathbf{y}; h\right)\, d\mathbf{y} \right]^2 d\mathbf{x}, \tag{4.23}$$

By expanding of (4.23) we get

$$\begin{aligned} BF = \frac{1}{n^2} \sum_{i,j=1}^{n} \int K_h\left(\mathbf{x}, \mathbf{X}_i; h\right) K_h\left(\mathbf{x}, \mathbf{X}_j; h\right)\, d\mathbf{x} + \\ - \frac{2}{n} \sum_{i=1}^{n} \int f_0(\mathbf{y}) \left(\int K_h\left(\mathbf{x}, \mathbf{y}; h\right) K_h\left(\mathbf{x}, \mathbf{X}_i; h\right)\, \mathbf{dx} \right) d\mathbf{y} + \\ + \int \left(\int f_0(\mathbf{y}') K_h\left(\mathbf{x}, \mathbf{y}'; h\right)\, d\mathbf{y}' \int f_0(\mathbf{y}) K_h\left(\mathbf{x}, \mathbf{y}; h\right)\, d\mathbf{y} \right) d\mathbf{x}. \end{aligned}$$

We substitute

$$R\left(\mathbf{y}, \mathbf{y}'; h\right) = \int K_h\left(\mathbf{x}, \mathbf{y}; h\right) K_h\left(\mathbf{x}, \mathbf{y}'; h\right)\, \mathbf{dx}, \tag{4.24}$$

where the Fourier transform of $R\left(\mathbf{y}, \mathbf{y}'; h\right)$ is positive which can be seen easily, and obtain

$$\begin{aligned} BF = \frac{1}{n^2} \sum_{i,j=1}^{n} R\left(\mathbf{X}_i, \mathbf{X}_j; h\right) - \frac{2}{n} \sum_{i=1}^{n} \int f_0(\mathbf{x}) R\left(\mathbf{x}, \mathbf{X}_i; h\right)\, d\mathbf{x} + \\ + \int \int f_0(\mathbf{x}) f_0(\mathbf{x}') R(\mathbf{x}, \mathbf{x}'; h)\, d\mathbf{x} d\mathbf{x}'. \end{aligned} \tag{4.25}$$

For a Gaussian kernel K_h we get from the Eq. (4.24) that $R\left(\mathbf{y}, \mathbf{y}'; h\right)$ is also a Gaussian. Hence, we see from the form of (4.25) that the BF test statistic resembles the Energy statistic ϕ_{nm} with a Gaussian distance function R. Therefore the Energy

test statistic with a Gaussian distance function R has an alternative interpretation in terms of L_2 distances between a kernel density estimate $\widehat{f}_n(\mathbf{x})$ of the null hypothesis density $f_0(\mathbf{x})$ and its expected value $E\left[\widehat{f}_n(\mathbf{x})\right]$.

Note that it is impossible to get the kernel K_h from a distance function R with the Eq. (4.24). Therefore the Energy test is more general.

It was pointed out by [35] that the Bowman-Foster statistic is essentially the same as that given by [22] and [20]. In [36] is shown that T_β can be written as

$$BF = \beta^d (2\pi)^{-d/2} T_\beta$$

with $\beta = \frac{1}{h\sqrt{2}}$ and $h = \left(\frac{4}{n(2d+1)}\right)^{1/(4+d)}$. Hence the Bowman-Foster test is a member of the BHEP tests. From this point of view the Energy test statistic with a Gaussian distance function R has also another interpretation in terms of weighted integral of the squared modulus of the difference between the empirical characteristic function and the characteristic function of the proposed distribution.

4.8 Power study

4.8.1 Univariate case

Optimization of the test parameters

To study the dependence of the power of the Energy test on the choice of the distance function and its parameters, we have chosen for H_0 a uniform, univariate p.d.f. f_0 restricted to the unit interval $[0, 1]$. We determined the rejection power with respect to contaminations of f_0 with a linear and two different Gaussian distributions which represent a wide and a more local distortion. These alternative hypothesis densities are

$$f_1(x) = 0.3 + 1.4x \tag{4.26}$$

$$f_2(x) = 0.7 + 0.3\left[c_2 e^{-64(x-0.5)^2}\right] \tag{4.27}$$

$$f_3(x) = 0.8 + 0.2\left[c_3 e^{-256(x-0.5)^2}\right] \tag{4.28}$$

where the c_2 and c_3 are normalization constants for the associated Gaussians.

We required 5% significance level and computed the rejection power which is equal to one minus the probability for an error of the second kind. As a reference, we also computed the power of a χ^2 test with bins of fixed width. The number of bins $B \approx 2n^{2/5}$ was chosen according to the prescription proposed in [37]. The cut-off parameter ϵ for (4.13) and (4.14) was set equal to $1/(4n)$.

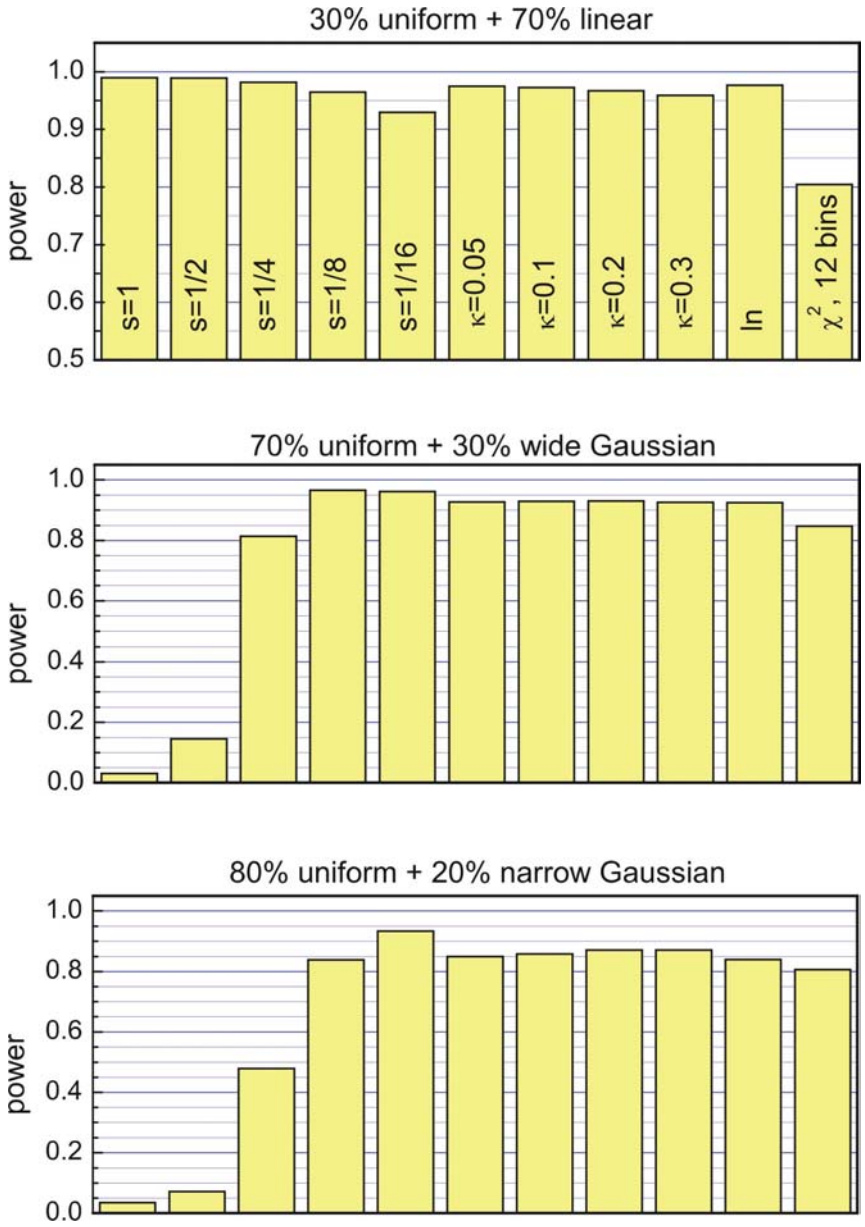

Figure 4.3: Rejection power of different Energy tests for a uniform distribution f_0 with respect to a linear distribution and two Gaussian contaminations, $exp[-64(x-0.5)^2]$ and $exp[-256(x-0.5)^2]$.

In Figure 4.3 we show the results for samples of 100 observations. Five different values of the Gaussian width parameter s, four different power laws and the logarithmic distance function have been studied. As expected, the linear distribution is best discriminated by slowly varying distance functions like the logarithm, low power laws and the wide Gaussians. The p.d.f. f_3 contains a narrow Gaussian contamination and since is better recognized by the narrow distance functions ($s = 1/16, \kappa = 0.3$). Here the two wide Gaussian distance functions fail completely.

Figure 4.4 illustrates the dependence of the rejection power on the sample size for the three different alternative hypothesis densities as in (4.26), (4.27), (4.28) to the uniform distribution. The amount of contamination was reduced with increasing sample size.

We applied the Energy tests ϕ_{nm} with power law $\kappa = 0.3$, the logarithmic and two Gaussian distance functions with fixed width $s = 1/8$ (Gfix) and variable width (Gvar). The variable width was chosen such that the full width at half maximum is equal to the χ^2 bin width, chosen according to the $2n^{2/5}$ law. This allows for a fair comparison between the two methods. As expected the linear distribution is best discriminated by the Energy test with logarithmic distance function. The power of the χ^2 test is considerably worse. The Energy test with variable Gaussian distance function follows the trend of the χ^2 test but performs better in 14 out of the 15 cases.

Comparing the samples with sizes 50 and 100, respectively, we realize one of the caveats of the χ^2 test: For the sample of size 50 there are 9 bins. The central bin coincides favorably with the location of the distortion peak of the background sample and consequently leads to a high rejection power. For the sample of size 100, however, there are 12 bins, thus two bins share the narrow peak and the power is reduced. The Gaussian Energy test is insensitive to the location of the distortion.

Comparison with alternative univariate tests

We have investigated the following GoF tests, which are introduced in this thesis: χ^2 test, Kolmogorov-Smirnov test, Kuiper test, Anderson-Darling test, Watson test, Neyman smooth test, three region test, Energy tests with logarithmic and Gaussian distance functions.

Samples contaminated by different background sources were tested against H_0, corresponding to the uniform distribution. The power of each test for a 5% significance level was evaluated for 100 observations. The following 6 different alternative hypothesis densities are considered to study the power of the tests against uniform distribution on $[0, 1]$ where some of them are illustrated in Figure 4.5.

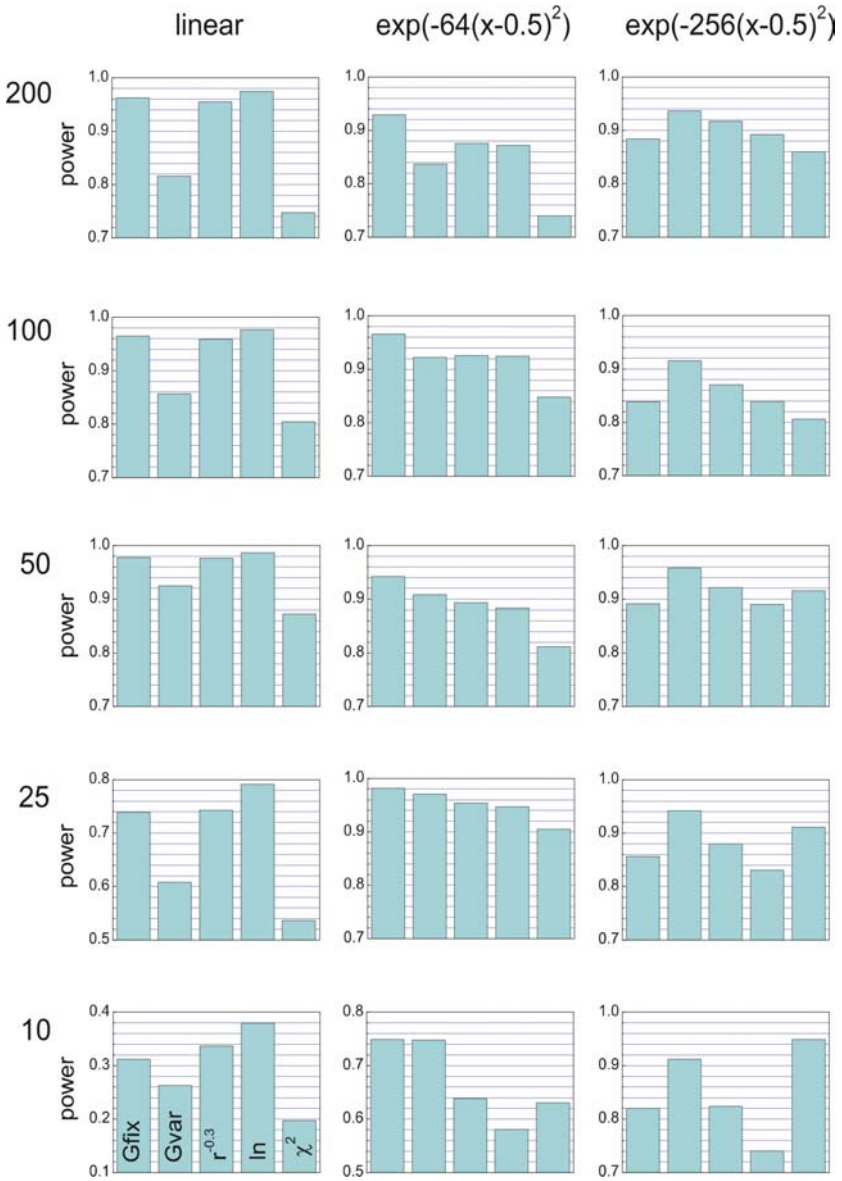

Figure 4.4: Power of tests for 3 different contaminations to uniform distribution and 5 different sample sizes ranging from 10 to 200. The shape of the contamination is displayed on top of the columns. The type of test is indicated in the lowest left hand plot.

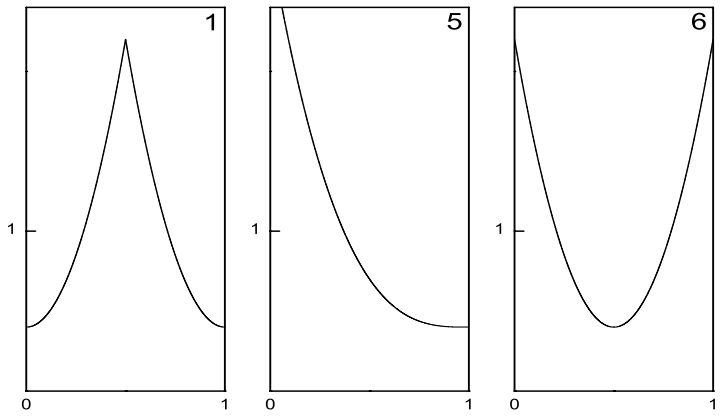

Figure 4.5: Illustration of some alternative distributions to the uniform distribution.

$$
\begin{aligned}
&1 : 0.7 + 0.3 \begin{cases} (k+1)2^k X^k \ , \ 0 \le X \le 0.5 \\ (k+1)2^k (1-X)^k \ , \ 0.5 < X \le 1 \end{cases} \ , \ k = 2 \\
&2 : 0.8 + 0.2N\left(\frac{1}{2}, \frac{1}{24}\right) \\
&3 : 0.8 + 0.2N\left(\frac{1}{2}, \frac{1}{32}\right) \\
&4 : 0.5 + X \ , \ 0 \le X \le 2 \\
&5 : 0.7 + 0.3(k+1)(1-X)^k \ , \ k = 3, \ 0 \le X \le 1 \\
&6 : 0.7 + 0.3(k+1)2^k \left(X - \frac{1}{2}\right)^k \ , \ k = 2, \ 0 \le X \le 1
\end{aligned}
$$

The power of the different tests is presented in Figure 4.6. As expected, none of the tests is optimum for all kind of distortions. The Energy tests are quite powerful independent of the background function.

4.8.2 Bivariate case

The real power of the Energy test manifests itself in multidimensional applications. We compare the Energy test with logarithmic and Gaussian distance functions to tests for multivariate normality which are introduced in third chapter.

The question of how to maximize the power of the BHEP tests in terms of the choice of the smoothing parameter β is still under investigation. For our power study we chose $\beta = 1$.

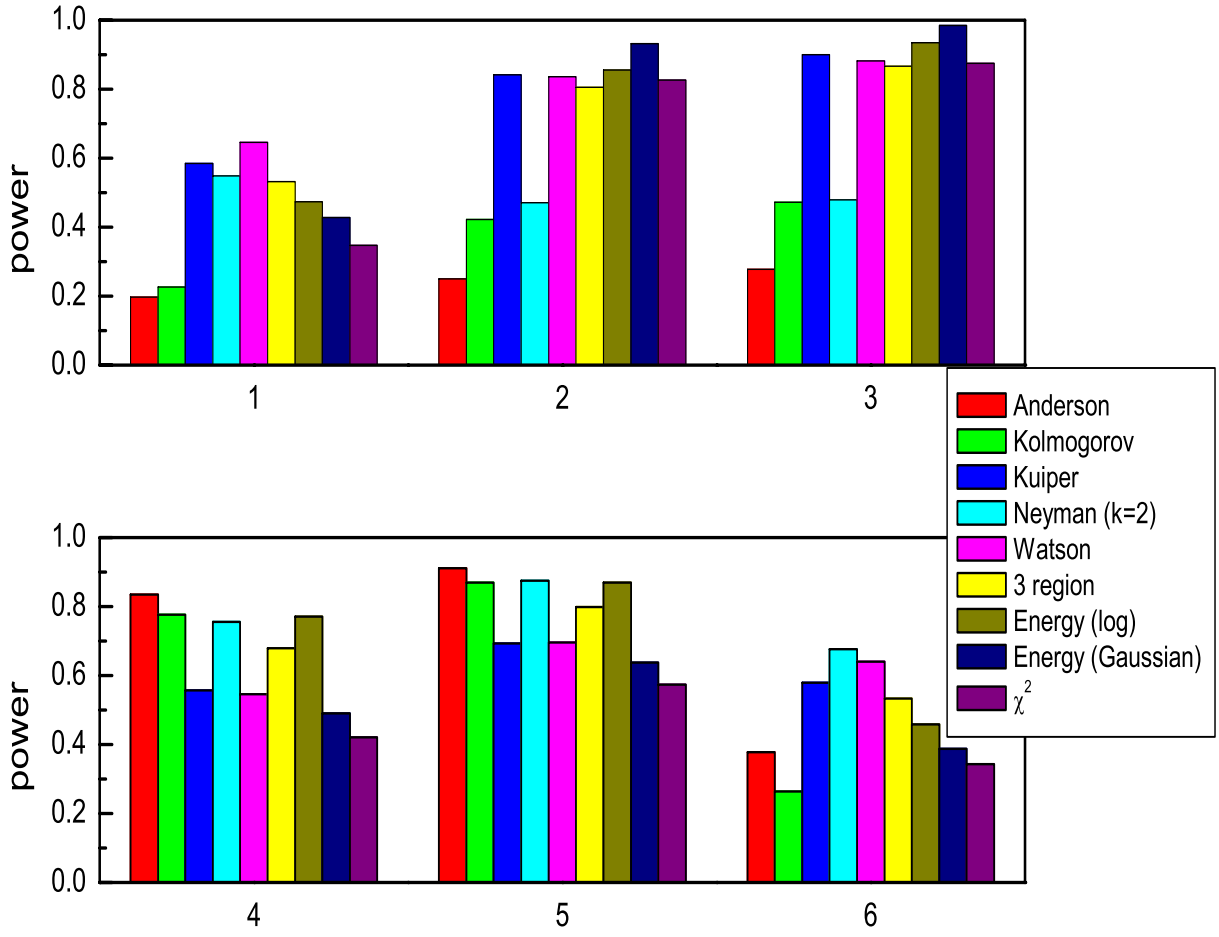

Figure 4.6: Rejection power of tests with respect to different contaminations to the uniform distribution.

As null hypothesis we chose two-dimensional Gaussian $N(\mathbf{0}, \mathbf{I}_2)$ and the alternative probability distributions under investigation are the following:

$$
\begin{aligned}
1 &: 0.85N(\mathbf{0}, \mathbf{I}_2) + 0.15U(\mathbf{0}, \mathbf{1}) \\
2 &: 0.7N(\mathbf{0}, \mathbf{I}_2) + 0.3N_{\log}(\mathbf{0}, \mathbf{I}_2) \\
3 &: 0.5N(\mathbf{0}, \mathbf{I}_2) + 0.5U(-\mathbf{1}, \mathbf{1}) \\
4 &: 0.6N(\mathbf{0}, \mathbf{I}_2) + 0.4N\left(\mathbf{0}, \begin{array}{cc} 1 & 0.9 \\ 0.9 & 1 \end{array}\right) \\
5 &: 0.85N(\mathbf{0}, \mathbf{I}_2) + 0.15N\left(\mathbf{0}, \begin{array}{cc} 1/9 & 0.1 \\ 0.1 & 1/9 \end{array}\right) \\
6 &: 0.9N(\mathbf{0}, \mathbf{I}_2) + 0.1N\left(\mathbf{0}, \begin{array}{cc} 0.04 & 0.02 \\ 0.02 & 0.04 \end{array}\right) \\
7 &: 0.95N(\mathbf{0}, \mathbf{I}_2) + 0.05N\left(\mathbf{0}, \begin{array}{cc} 0.0025 & 1/800 \\ 1/800 & 0.0025 \end{array}\right)
\end{aligned}
$$

The distribution denoted by $N_{\log}$ is obtained by the variable transformation $x \to \ln|x|$ applied to each coordinate of a two-dimensional normal distribution and is not to be mixed up with the log-normal distribution. It is extremely asymmetric.

Figure 4.7 displays the results for 200 observations. We were astonished how well the Energy test competes with alternatives especially designed to test normality. We attribute the excellent performance of the Energy test to the fact that it is sensitive to all deviations of two distributions whereas, for example, the Mardia tests are based only on two specific moments.

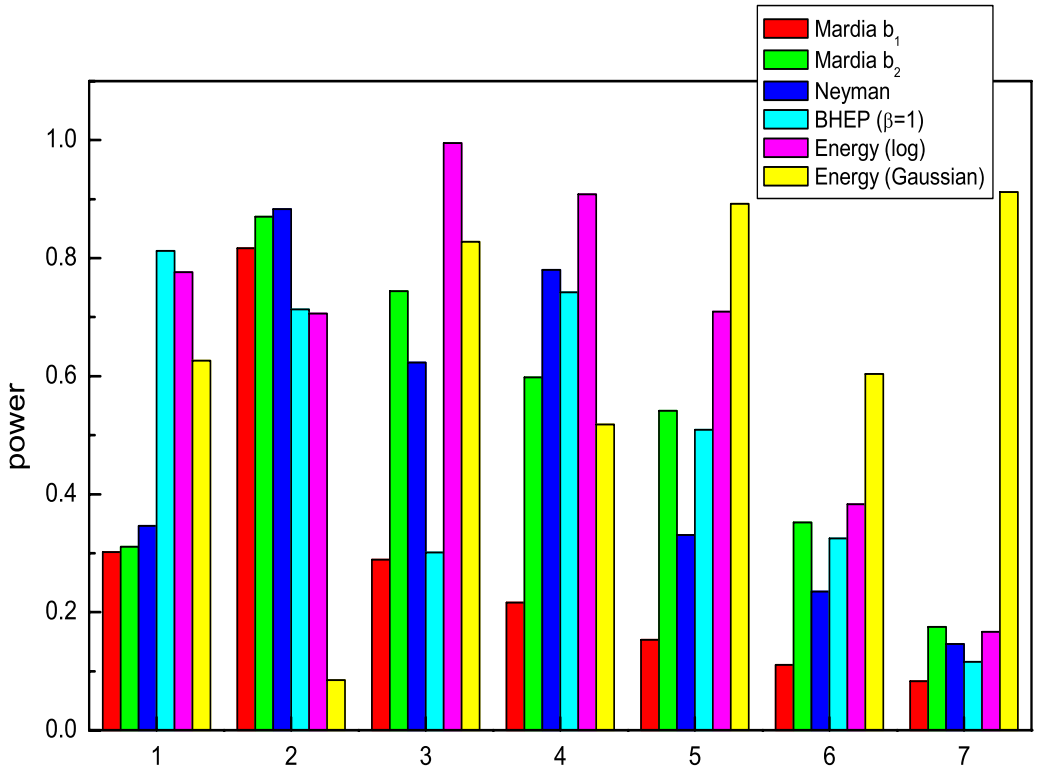

Figure 4.7: Rejection power of tests with respect to different contaminations of a two-dimensional Gaussian.

Chapter 5

Energy for the two-sample problem

In this chapter we consider the classical problem of testing whether two samples of observations are from the same distribution. To develop a multivariate test for the two-sample problem based on the energy approach, we briefly introduce some resampling methods used in statistical analysis. The power of this new test is compared with that of competing tests in different dimensions and the proposed test shows high performance independent of the dimension of the variate space.

5.1 Resampling methods

During the last twenty years, the development of faster and cheaper computers has made Monte Carlo methods more affordable and attractive in statistical analysis. Important developments in this area include the use of resampling methods for improving standard asymptotic approximations. The basic idea of the resampling methods is that, in absence of any other information about the distribution, the observed sample contains all the available information about the underlying distribution, hence resampling the sample is the easiest way to get informations about the underlying distribution.

5.1.1 The bootstrap and permutation principle

The bootstrap and permutation methods are described and explored in detail by [38], and only a brief summary will be given here.

Let $\widehat{\theta}\left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\right)$ denote some sample estimate. Suppose we wish to estimate some feature of this r.v. $\widehat{\theta}$, such as its mean value, its variance, etc. The basic idea of the bootstrap is to construct an empirical distribution function (EDF) F_n of the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and draw a new random sample $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ from F_n. This resample is called a *bootstrap sample.*

The bootstrap algorithm consists of the following steps:

1. Obtain the EDF F_n of the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.
2. Draw a bootstrap sample from F_n. This consists of drawing each $\mathbf{X}_i'$ independently with replacement from the observed sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Using this bootstrap sample, compute $\widehat{\theta}' = \widehat{\theta}\,(\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')$.
3. Repeat step 2 a large number of times, say B times, obtaining bootstrap replications $\widehat{\theta}_1', \widehat{\theta}_2', \ldots, \widehat{\theta}_B'$. Based on this replications, we may compute features of interest of the distribution of $\widehat{\theta}$.

The permutation method proceeds identically as the bootstrap method, except that resampling is done without replacement.

5.1.2 The smoothed bootstrap

Because the EDF F_n is a discrete distribution, samples drawn from F_n, bootstrap samples, are inappropriate for the energy approach. Nearly every bootstrap sample will contain repeated values. The *smoothed bootstrap* [39] is a modification to the bootstrap procedure to avoid samples with this property. The essential idea of the smoothed bootstrap is to perform the repeated sampling not from F_n itself, but from a smoothed version $\widehat{F}_n$ of F_n. For example $\widehat{F}_n$ may be the distribution function of the continuous density estimate $\widehat{f}_n$. If $\widehat{f}_n$ is constructed by the kernel method with kernel K, then it is very easy to find independent realizations from $\widehat{f}_n$. Realizations $\mathbf{X}'$ from $\widehat{f}_n$ can be generated as follows:

1. Choose i uniformly with replacement from $\{1, 2, \ldots, n\}$.
2. Generate α from the kernel K.
3. Set $\mathbf{X}' = \mathbf{X}_i + h\alpha$, where h is a smoothing constant, chosen by trial.

5.2 The two-sample Energy test

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ be two samples of independent random vectors with distributions F and G, respectively. Then as considered in section 2.5 the general two-sample problem consists of testing the hypothesis

$$H_0 : F(\mathbf{x}) = G(\mathbf{x}), \quad \text{for every } \mathbf{x},$$

against the general alternative

$$H_1 : F(\mathbf{x}) \neq G(\mathbf{x}), \quad \text{for at least one } \mathbf{x},$$

where F and G are unknown.

In principle the two-sample problem can be treated with similar methods as used in GoF techniques. The only difference lies in the fact that the underlying distributions F and G of the two samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ are unknown. This difficulty can be relaxed by considering for example a permutational version of the Energy tests which uses the fact that all permutations of the pooled observations $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_{n+m}) := (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$ are equally likely.

The energy concept can also be applied to the two-sample problem. We consider the first sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a system of positive charges and the second sample $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ as a system of negative charges. The two-sample Energy test statistics ϕ_{nm}, which is given below, is based on the fact that the total potential energy of the pooled sample will be minimum if both charge samples have the same distribution.

$$\phi_{nm} = \frac{1}{n(n-1)} \sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \frac{1}{m(m-1)} \sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) + \\ - \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right).$$

To compute the power of the two-sample Energy tests we use the permutation method to evaluate the distribution of ϕ_{nm} under H_0. We merge the $N = m + n$ observations of both samples and draw from the combined sample a subsample of size n without replacement. The remaining m observations represent a second sample. The probability distribution under H_0 of ϕ_{nm} is evaluated by determining the values of ϕ_{nm} of all $\binom{N}{m} = \frac{N!}{n!m!}$ possible permutations. For large N this procedure can become computationally too laborious. Then the probability distribution is estimated from a random sample of all possible permutations.

To determine the distribution of ϕ_{nm} we could also use a smoothed bootstrap sample but we prefer permutation technique, since it is free of arbitrary parameters, as for example smoothing parameters.

5.3 Competing tests

Before we discuss the simulations and their results in the next section we will give a short overview on some tests which are natural competitors to the two-sample Energy test. These tests were used in the simulations to show how the two-sample Energy test perform in comparison with existing tests. The critical value of these tests are determined by a Monte Carlo simulation.

5.3.1 Univariate case

In the literature exists a multitude of two-sample tests. But we select only a few of the proposed tests for a wide variety of functional alternatives. Some of the GoF tests, for example the χ^2 test and the EDF-tests, can be adapted to the two-sample problem. We have seen that as a GoF criterion the EDF-tests compared the empirical distribution function of a sample with a hypothesized distribution. In the two-sample case, the comparison is made between the empirical distribution functions of the two samples. The χ^2 test, Kolmogorov-Smirnov two-sample test, Cramér-von Mises two-sample test, Wilcox test, and the Lepage test will be considered.

Chi-square χ^2 test

χ^2 test requires grouping of observations therefore some informations get lost. But it is the most popular test, which is used in practice, even though it is well known that it has some disadvantages from the point of view of the power. χ^2 statistic is defined by

$$\chi^2 = \sum_i \frac{\sqrt{\frac{m}{n}} n_i - \sqrt{\frac{n}{m}} m_i}{n_i + m_i},$$

where the sum is over all bins and n_i is the number of the observations from the first sample in the ith bin, m_i the number of observations in the same bin i for the second sample. A large value of χ^2 indicates that H_0 is rather unlikely.

Kolmogorov-Smirnov (KS) test

The familiar Kolmogorov-Smirnov two-sample test is defined by

$$KS = \sqrt{\frac{mn}{n+m}} \sup_{1 \le i \le n+m} \left| F_n(Z_{(i)}) - G_m(Z_{(i)}) \right|,$$

where F_n and G_m denote the empirical distribution function for the both samples and the $Z_{(i)}$ are the order statistics of the pooled sample from $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$. The test procedure suggests to reject H_0 if $KS \ge k_{1-\alpha}(n, m)$, where $k_{1-\alpha}$ is the $100\alpha\%$ percentile of the null distribution of KS.

Cramér-von Mises (CvM) test

Another class of test, which uses information from all and not just from the largest deviation is:

$$CvM = \frac{nm}{(n+m)^2} \sum_{i=1}^{n+m} \left(F_n(Z_{(i)}) - G_m(Z_{(i)}) \right)^2.$$

For practical work the alternative form

$$CvM = \frac{1}{n(n+m)}\sum_{i=1}^{n}\left(R(X_{(i)})-i\right)^2+\frac{1}{m(n+m)}\sum_{j=1}^{m}\left(R(Y_{(j)})-j\right)^2+ -\frac{4nm-1}{6(n+m)}$$

is preferable, where $R(X_{(i)})$ is the rank in the combined sample of the ith observation in the first sample and $R(Y_{(j)})$ is the rank in the combined sample of the jth observation in the second sample, respectively . Large values of CvM leads to a rejection of H_0.

Wilcox (W) test

Wilcox test W [40] is a version of KS test, which is based on weighting the Kolmogorov distance between two distributions.

$$W=\sqrt{\frac{nm}{n+m}}\sup_{1\le i\le n+m}\frac{\left|F_n(Z_{(i)})-G_m(Z_{(i)})\right|}{\left(H(Z_{(i)})(1-H(Z_{(i)}))\right)^{\frac{1}{2}}},$$

where

$$H(Z_{(i)})=\lambda F(Z_{(i)})+(1-\lambda)G(Z_{(i)})$$

with $\lambda=\frac{m}{n+m}$.

Wilcox test gives more weights to differences in the tails so that the variance of W remains fairly stable over all possible values. Consequently, W must be more sensitive than KS to differences that occur in the tails.

Lepage (L) test

The Lepage test L [41] is expressed in terms of two independent tests. The Wilcoxon statistic WN

$$WN=\sum_{i=1}^{n}R(X_{(i)})$$

for location alternatives and the Ansari Bradley statistic AB

$$AB=\sum_{i=1}^{n}\frac{n+m+1}{2}-\left|R(X_{(i)})-\frac{n+m+1}{2}\right|$$

for scale alternatives are included in the Lepage statistic L and is given by

$$L=\left(\frac{WN-E(WN)}{\sqrt{Var(WN)}}\right)^2+\left(\frac{AB-E(AB)}{\sqrt{Var(AB)}}\right)^2,$$

where

$$
\begin{aligned}
E(WN) &= \frac{n(n+m+1)}{2} \\
Var(WN) &= \frac{nm(n+m+1)}{12} \\
E(AB) &= \begin{cases} \frac{n(n+m+2)}{4} & \text{if } n+m \text{ is even} \\ \frac{n(n+m+1)^2}{4(n+m)} & \text{if } n+m \text{ is odd} \end{cases} \\
Var(AB) &= \begin{cases} \frac{nm((n+m)^2-4)}{48(n+m-1)} & \text{if } n+m \text{ is even} \\ \frac{nm(n+m+1)((n+m)^2+3)}{48(n+m)^2} & \text{if } n+mm \text{ is odd} \end{cases}
\end{aligned}
$$

The asymptotic distribution of L is under H_0 a χ^2_2 distribution with 2 degrees of freedom, since the statistics WN and AB are independent.

5.3.2 Multivariate case

For the multivariate two-sample problem with general alternatives there are only a few tests on the market. In the literature tests are proposed which like the Energy test are based on the distance between the observations. One of them is the Friedman-Rafsky test. Its statistic [42] is the number of edges in the *minimal spanning tree* that connect observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ from different samples. Another multidimensional test which we have chosen for a comparison to the Energy test, is the k nearest neighbor test [43].

Friedman-Rafsky test

The Friedman-Rafsky test can be seen as a generalization of the univariate run test [44]. The test statistic in [44] is the run which is defined as a set of adjacent observations which belong to the same sample, where the two samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are combined into a single ordered sequence from smallest to largest, see Figure 5.1. Clearly, small values of the run of the ordered pooled sample lead to a rejection of the null hypothesis.

The problem in generalizing the run test to more than one dimension is that there is no unique sorting scheme for the observations. The minimum spanning tree can be used for this purpose. For independent d-variate random samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ i.i.d. with $f(\mathbf{x})$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ i.i.d. with $g(\mathbf{y})$ a spanning tree on a pooled sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_{n+m}$ is a connected graph with no cycles. A minimal spanning tree is the spanning tree for which the total Euclidean length of the connections is the smallest possible. Clearly, in one dimension the Friedman-Rafsky test is exactly the run test.

The Friedman-Rafsky test proceeds as follows. In the first step the minimal spanning tree of the pooled sample is constructed. In the second step one counts

Figure 5.1: Two samples of observations (5 circles and 4 squares) result in 4 runs.

the connections between observations from different samples. The result is the test statistic R_{nm} and small values of R_{nm} lead to a rejection of H_0. In [45] it is proved that R_{nm} is asymptotically distribution-free under the null hypothesis.

The nearest neighbor test

For independent d-variate random samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ i.i.d. with a p.d.f. f and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ i.i.d. with a p.d.f. g, where the densities f and g are assumed to be continuous, the nearest neighbor test statistic N_{nm} is the sum of the number of vectors $\mathbf{Z}_i$ of the pooled sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_{n+m}$ such that the nearest neighbor of $\mathbf{Z}_i$, denoted by $N(\mathbf{Z}_i)$, is of the same type as $\mathbf{Z}_i$:

$$N_{nm} = \sum_{i=1}^{n+m} I\left(\mathbf{Z}_i \text{ and } N(\mathbf{Z}_i) \text{ belong to the same sample}\right),$$

where I is the indicator function. $N(\mathbf{Z}_i)$ can be determined by a fixed but otherwise arbitrary norm on $\mathbb{R}^d$. We select the Euclidean norm. In [46] it is shown that the limiting distribution of N_{nm} is normal, as $\min(n, m) \to \infty$ and $n/(n+m) \to \tau$ with $0 < \tau < 1$:

$$\sqrt{n+m}\left(\frac{1}{n+m}T_{nm} - \frac{n(n-1)+m(m-1)}{(n+m-1)(n+m)}\right) \to N\left(0, \sigma^2(\tau)\right), \qquad (5.1)$$

where $\sigma^2(\tau)$ is given by

$$\sigma^2(\tau) = 4\tau(1-\tau)\left(\tau(1-\tau)(1+b_1(d)) + (\tau - \frac{1}{2})^2 b_2(d)\right).$$

For dimension $d = 2$ the constants b_1 and b_2 are given by

$$b_1(2) = \frac{6\pi}{8\pi + 3\sqrt{3}} \quad \text{and} \quad b_2(2) \approx 0.633.$$

From (5.1) critical values can be determined for moderate or large sample sizes. Clearly, large values of N_{nm} leads to a rejection of H_0.

Table 5.1: Confidence intervals as a function of the number of permutations for nominal $\alpha = 0.05$.

# of permutations	$CI(95\%)$ for α
100	$[0.006, 0.095]$
300	$[0.025, 0.075]$
500	$[0.031, 0.068]$
1000	$[0.036, 0.063]$

The k nearest neighbor test

As a generalization of the nearest neighbor test statistic, the k nearest neighbor test statistic T_{nm}^k is the number of all k nearest neighbor comparisons in which observations and their neighbors belong to the same sample, i.e.

$$T_{nm}^k = \sum_{i=1}^{n+m} \sum_{r=1}^{k} I_i(r),$$

where

$$I_i(r) = \begin{cases} 1, \text{ if } \mathbf{Z}_i \text{ and } N_r(\mathbf{Z}_i) \text{ belong to the same sample} \\ 0, \text{ otherwise} \end{cases}$$

and $N_r(\mathbf{Z}_i) = \mathbf{Z}_j$ is the rth nearest neighbor to $\mathbf{Z}_i$ satisfying $|\mathbf{Z}_s - \mathbf{Z}_i| < |\mathbf{Z}_j - \mathbf{Z}_i|$, $1 \leq s \leq n+m$, $s \neq i, j$.

Note that for $k = 1$ the test statistic T_{nm}^k reduces to the statistic N_{nm}.

5.4 Power comparisons

The performance of various tests were assessed for finite sample sizes by Monte Carlo simulations in $d = 1$, 2 and 4 dimensions. Also the critical values of all considered tests were calculated by Monte Carlo simulation. We chose a 5% significance level.

For the null hypothesis we determine the distribution of ϕ_{nm} with the permutation technique, as mentioned above. We followed [38] and generated 1000 randomly selected two subsets in each case and determined the critical values ϕ_c of ϕ_{nm}. For the specific case $n = m = 50$ and samples drawn from a uniform distribution we studied the statistical fluctuations. Transforming the confidence interval of ϕ_c into limits for α, we obtain the interval $[0.036, 0.063]$, see Table 5.1. This means that the critical value obtained by 1000 permutations corresponds with 95% confidence to a value α included in this interval.

Even though the Energy test has been designed for multivariate applications, we investigate its power in one dimension because there a comparison with several well established tests is possible . To avoid a personal prejudice we drew the two samples from the probability distributions, which have also been investigated in [47]:

$$
\begin{aligned}
f_1(x) &= \begin{cases} 1 & -\sqrt{3} \le x \le \sqrt{3} \\ 0 & \text{otherwise} \end{cases} \\
f_2(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
f_3(x) &= \frac{1}{2} e^{-|x|} \\
f_4(x) &= \frac{1}{\pi} \frac{1}{1+x^2} \quad , \quad \text{Cauchy} \\
f_5(x) &= e^{-(x+1)} \quad , \quad x \ge -1 \\
f_6(x) &= \chi^2_3 \qquad \begin{array}{c} \chi^2 \text{ with 3 degrees of freedom,} \\ \text{transformed to mean 0, variance 1} \end{array} \\
f_7(x) &= \frac{1}{2} N(1.5, 1) + \frac{1}{2} N(-1.5, 1) \\
f_8(x) &= 0.8 N(0, 1) + 0.2 N(0, 4^2) \\
f_9(x) &= \frac{1}{2} N(1, 2^2) + \frac{1}{2} N(-1, 1)
\end{aligned}
$$

This set f_1 to f_9 of probability distributions covers a variety of cases of short tailed up to very long tailed probability distributions as well as skewed ones.

To evaluate the power of the tests we generated 1000 pairs of samples for small $n = m = 25$, moderate $n = 50$, $m = 40$ and 'large' $n = 100$, $m = 50$, for seven different scenarios. We have transformed the variates $Y_i^* = \theta + \tau Y_j$, $j = 1, \ldots, m$ of the second sample, corresponding to the alternative distribution, with different location parameters θ and scale parameters τ. The power was determined in all cases by counting the number of times a test resulted in a rejection divided by 1000. All tests have a nominal significance level of 0.05.

Table 5.2 shows the estimated power for small sample sizes, $n = 25$, $m = 25$, of the selected tests. The χ^2 test is performed with 5 equal probability classes. Tables 5.3 and 5.4 present the results for $n = 50$, $m = 40$ and $n = 100$, $m = 50$, respectively. For the large sample the number of χ^2 classes was increased to 10. Some of the results from tables 5.2, 5.3 and 5.4 are shown in Figure 5.2.

It is apparent that none of the considered tests performs better than all other tests for all alternatives. The results indicate that the power of the Energy test in most of the cases is larger than that of the well known χ^2 and KS tests and comparable to that of the CvM test. For long tailed distributions, e.g. for combination

Table 5.2: Power of the selected tests for n=m=25, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	θ, τ	KS	CvM	W	L	$\phi_{25,25}$	χ^2
$f_1(x)$	$f_7(x)$	$0.4; 1.4$	0.12	0.18	0.48	0.38	0.24	0.11
		$0.6, 1.6$	0.37	0.41	0.87	0.69	0.54	0.17
		$0.6; 0.8$	0.40	0.55	0.66	0.50	0.45	0.52
		$0.5; 0.5$	0.70	0.70	0.93	0.86	0.85	0.85
$f_7(x)$	$f_2(x)$	$0.4; 1.4$	0.08	0.13	0.22	0.14	0.13	0.08
		$0.6, 1.6$	0.20	0.29	0.46	0.34	0.31	0.14
		$0.6; 0.8$	0.34	0.46	0.57	0.51	0.44	0.45
		$0.5; 0.5$	0.72	0.69	0.93	0.93	0.89	0.88
$f_2(x)$	$f_3(x)$	$0.4; 1.4$	0.17	0.23	0.22	0.19	0.19	0.14
		$0.6, 1.6$	0.33	0.44	0.42	0.37	0.38	0.24
		$0.6; 0.8$	0.64	0.70	0.67	0.66	0.67	0.60
		$0.5; 0.5$	0.74	0.77	0.84	0.91	0.89	0.84
$f_2(x)$	$f_9(x)$	$0.4; 1.4$	0.06	0.09	0.19	0.13	0.12	0.07
		$0.6, 1.6$	0.14	0.22	0.35	0.29	0.21	0.11
		$0.6; 0.8$	0.46	0.59	0.65	0.59	0.45	0.54
		$0.5; 0.5$	0.71	0.72	0.89	0.88	0.82	0.85
$f_6(x)$	$f_5(x)$	$0.4; 1.4$	0.10	0.16	0.15	0.12	0.12	0.12
		$0.6, 1.6$	0.16	0.25	0.24	0.22	0.16	0.20
		$0.6; 0.8$	0.95	0.94	1.00	0.97	0.98	0.97
		$0.5; 0.5$	0.99	0.98	1.00	1.00	1.00	1.00
$f_3(x)$	$f_8(x)$	$0.4; 1.4$	0.26	0.34	0.30	0.25	0.28	0.21
		$0.6, 1.6$	0.55	0.64	0.56	0.53	0.50	0.43
		$0.6; 0.8$	0.85	0.89	0.86	0.84	0.85	0.79
		$0.5; 0.5$	0.90	0.93	0.94	0.97	0.95	0.91
$f_8(x)$	$f_4(x)$	$0.4; 1.4$	0.34	0.42	0.45	0.52	0.54	0.32
		$0.6, 1.6$	0.60	0.67	0.68	0.77	0.80	0.51
		$0.6; 0.8$	0.80	0.85	0.78	0.72	0.82	0.70
		$0.5; 0.5$	0.81	0.84	0.81	0.71	0.84	0.72

Table 5.3: Power of the selected tests for n=50, m=40, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	$\theta; \tau$	KS	CvM	W	L	$\phi_{50,40}$	χ^2
$f_1(x)$	$f_7(x)$	0.3; 1.3	0.22	0.18	0.67	0.41	0.25	0.14
		0.4; 0.8	0.49	0.47	0.67	0.53	0.46	0.62
$f_7(x)$	$f_2(x)$	0.3; 1.3	0.15	0.17	0.34	0.16	0.20	0.10
		0.4; 0.8	0.62	0.56	0.66	0.70	0.58	0.58
$f_2(x)$	$f_3(x)$	0.3; 1.3	0.28	0.29	0.25	0.22	0.26	0.14
		0.4; 0.8	0.67	0.66	0.65	0.70	0.68	0.61
$f_2(x)$	$f_9(x)$	0.3; 1.3	0.07	0.07	0.27	0.12	0.14	0.07
		0.4; 0.8	0.51	0.50	0.66	0.58	0.46	0.61
$f_6(x)$	$f_5(x)$	0.3; 1.3	0.13	0.14	0.18	0.12	0.11	0.14
		0.4; 0.8	1.00	0.97	1.00	0.99	1.00	1.00
$f_3(x)$	$f_8(x)$	0.3; 1.3	0.38	0.39	0.32	0.29	0.31	0.25
		0.4; 0.8	0.84	0.84	0.80	0.78	0.86	0.74
$f_8(x)$	$f_4(x)$	0.3; 1.3	0.50	0.52	0.59	0.66	0.64	0.39
		0.4; 0.8	0.76	0.77	0.70	0.63	0.79	0.59

Table 5.4: Power of the selected tests for n=100, m=50, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	$\theta; \tau$	KS	CvM	W	L	$\phi_{100,50}$	χ^2
$f_1(x)$	$f_7(x)$	0.3; 1.3	0.32	0.33	0.97	0.62	0.44	0.28
		0.4; 0.8	0.68	0.73	0.85	0.76	0.76	0.66
$f_7(x)$	$f_2(x)$	0.3; 1.3	0.21	0.27	0.67	0.28	0.33	0.26
		0.4; 0.8	0.82	0.79	0.82	0.90	0.82	0.64
$f_2(x)$	$f_3(x)$	0.3; 1.3	0.31	0.37	0.41	0.29	0.34	0.17
		0.4; 0.8	0.85	0.86	0.82	0.90	0.89	0.72
$f_2(x)$	$f_9(x)$	0.3; 1.3	0.08	0.10	0.51	0.18	0.21	0.17
		0.4; 0.8	0.65	0.66	0.79	0.77	0.67	0.63
$f_6(x)$	$f_5(x)$	0.3; 1.3	0.13	0.19	0.25	0.18	0.18	0.22
		0.4; 0.8	1.00	1.00	1.00	1.00	1.00	1.00
$f_3(x)$	$f_8(x)$	0.3; 1.3	0.46	0.52	0.44	0.44	0.47	0.25
		0.4; 0.8	0.95	0.97	0.92	0.94	1.00	0.86
$f_8(x)$	$f_4(x)$	0.3; 1.3	0.61	0.68	0.81	0.84	0.86	0.63
		0.4; 0.8	0.90	0.93	0.87	0.86	0.92	0.74

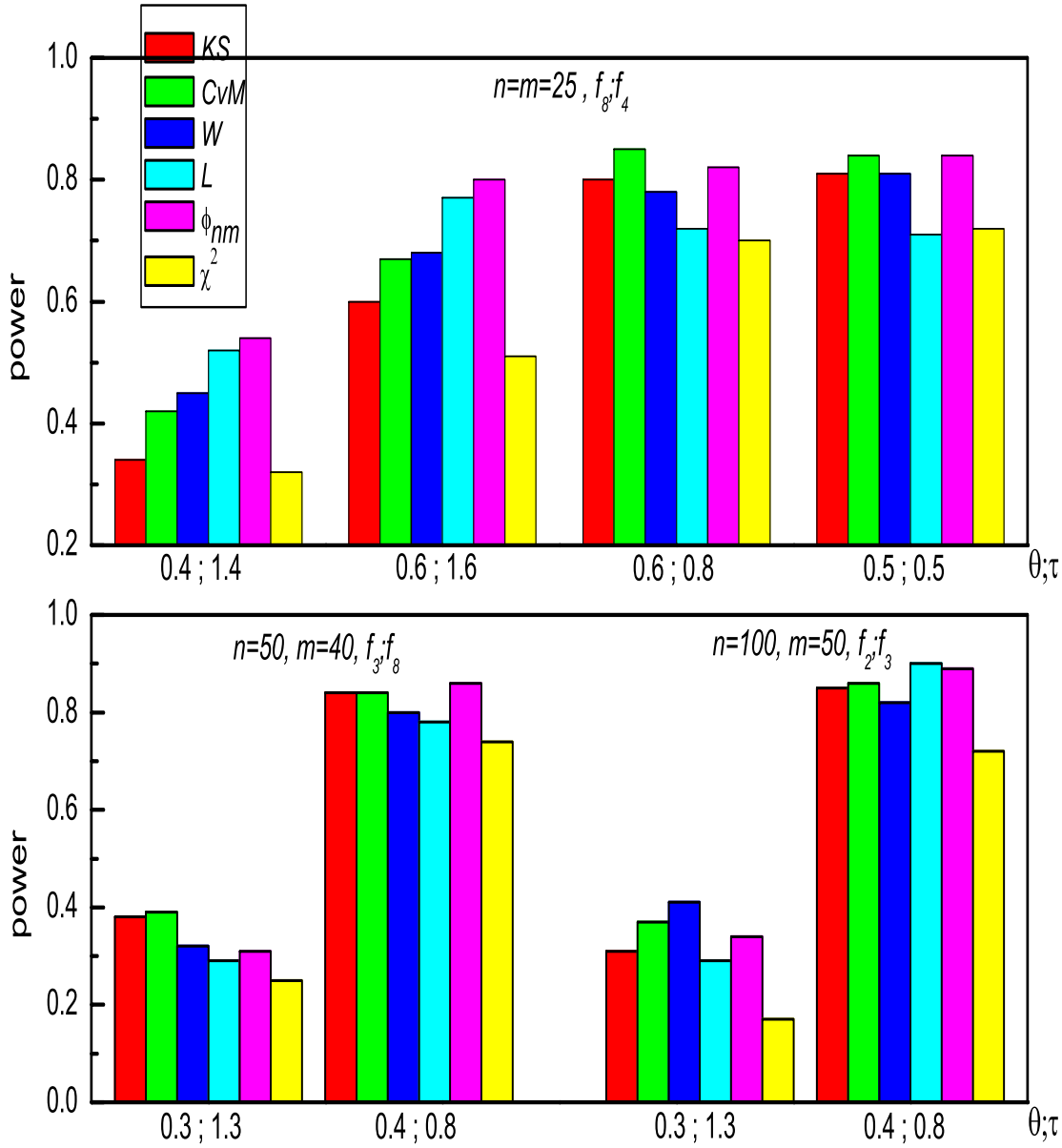

Figure 5.2: Rejection power of six two-sample tests for some selected different alternatives.

(f_8, f_4), the Energy test is the best among those which have been investigated. This is not unexpected since $R(x) = -\ln(x)$ is long range. Lepage and Wilcox tests are powerful tests for all combinations and sample sizes considered, however, the Lepage test is based on the first two moments of the null distribution and therefore specifically adapted to the type of study presented here.

In order to investigate how the performance of the tests using ϕ_{nm}, R_{nm} and T^k_{nm} changes with the dimension, we have considered problems in dimensions $d = 2$ and 4. In Table 5.5 and Table 5.6 we summarize the alternative probability distributions P^X and P^Y from which we drew the two samples. The first sample was drawn either from $N(\mathbf{0}, \mathbf{I})$ or from $U(\mathbf{0}, \mathbf{1})$ where $N(\boldsymbol{\mu}, \mathbf{V})$ is a multivariate normal probability distribution with the indicated mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ and $U(\mathbf{0}, \mathbf{1})$ is the multivariate uniform probability distribution in the unit cube. The parent distributions of the second sample were the Cauchy distribution C, the $N_{\log}$ distribution, correlated normal distributions, the Student's distributions t_2 and t_4 and Cook-Johnson $CJ(a)$ distributions [48] with correlation parameter $a > 0$. $CJ(a)$ converges for $a \to \infty$ to the independent multivariate uniform distribution and $a \to 0$ corresponds to the totally correlated case $X_{i1} = X_{i2} = ... = X_{id}, i = 1, ..., n$. We generated the random vectors from $CJ(a)$ via the standard gamma distribution with shape parameter a, following the prescription proposed by [49]. It is extremely asymmetric, see Figure 5.3

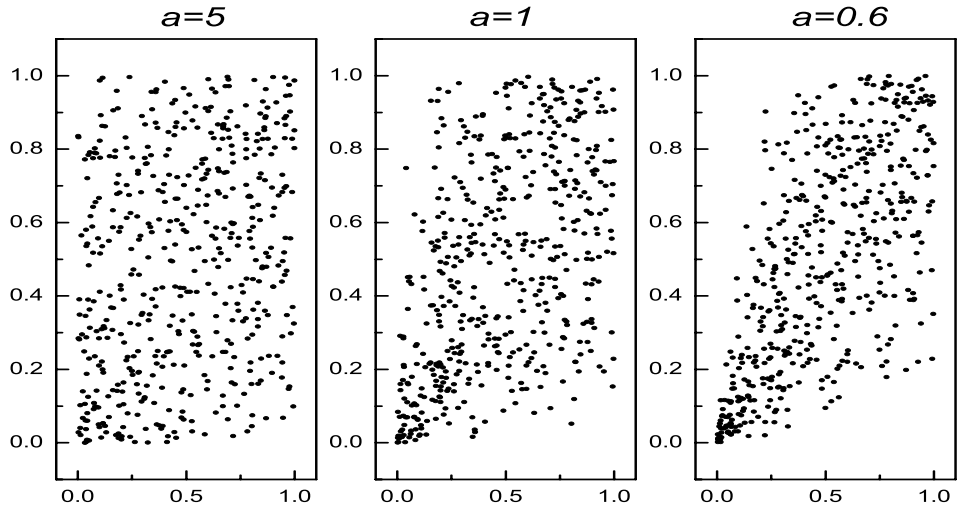

Figure 5.3: Cook-Johnson $CJ(a)$ distribution for different parameter a.

Some of the considered probability densities have also been used in a power study in [50].

Table 5.5: Two dimensional distributions used to generate the samples.

case	P^X	P^Y
1	$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$
2	$N(\mathbf{0},\mathbf{I})$	$N_{\log}(0,\mathbf{I})$
3	$N(\mathbf{0},\mathbf{I})$	$N\left(\mathbf{0}, \begin{matrix} 1 & 0.6 \\ 0.6 & 1 \end{matrix}\right)$
4	$N(\mathbf{0},\mathbf{I})$	$N\left(\mathbf{0}, \begin{matrix} 1 & 0.9 \\ 0.9 & 1 \end{matrix}\right)$
5	$N(\mathbf{0},\mathbf{I})$	Student's t_2
6	$N(\mathbf{0},\mathbf{I})$	Student's t_4
7	$U(\mathbf{0},\mathbf{1})$	$CJ(10)$
8	$U(\mathbf{0},\mathbf{1})$	$CJ(5)$
9	$U(\mathbf{0},\mathbf{1})$	$CJ(2)$
10	$U(\mathbf{0},\mathbf{1})$	$CJ(1)$
11	$U(\mathbf{0},\mathbf{1})$	$CJ(0.8)$
12	$U(\mathbf{0},\mathbf{1})$	$CJ(0.6)$
13	$U(\mathbf{0},\mathbf{1})$	$80\%U(\mathbf{0},\mathbf{1}) + 20\%N(\mathbf{0.5},0.05^2\mathbf{I})$
14	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1}) + 50\%N(\mathbf{0.5},0.2^2\mathbf{I})$

Table 5.6: Four dimensional distributions used to generate the samples.

case	P^X	P^Y
1	$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$
2	$N(\mathbf{0},\mathbf{I})$	$N_{\log}(\mathbf{0},\mathbf{I})$
3	$N(\mathbf{0},\mathbf{I})$	$80\%N(\mathbf{0},\mathbf{I})+20\%N(\mathbf{0},0.2^2\mathbf{I})$
4	$N(\mathbf{0},\mathbf{I})$	$50\%N(\mathbf{0},\mathbf{I})+50\%N\left(\mathbf{0}, \begin{matrix} 1 & 0.4 & 0.5 & 0.7 \\ 0.4 & 1 & 0.6 & 0.8 \\ 0.5 & 0.6 & 1 & 0.9 \\ 0.7 & 0.8 & 0.9 & 1 \end{matrix}\right)$
5	$N(\mathbf{0},\mathbf{I})$	Student's t_2
6	$N(\mathbf{0},\mathbf{I})$	Student's t_4
7	$U(\mathbf{0},\mathbf{1})$	$CJ(10)$
8	$U(\mathbf{0},\mathbf{1})$	$CJ(5)$
9	$U(\mathbf{0},\mathbf{1})$	$CJ(2)$
10	$U(\mathbf{0},\mathbf{1})$	$CJ(1)$
11	$U(\mathbf{0},\mathbf{1})$	$CJ(0.8)$
12	$U(\mathbf{0},\mathbf{1})$	$CJ(0.6)$
13	$U(\mathbf{0},\mathbf{1})$	$80\%U(\mathbf{0},\mathbf{1}) + 20\%N(\mathbf{0.5},0.05^2\mathbf{I})$
14	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1}) + 50\%N(\mathbf{0.5},0.2^2\mathbf{I})$

The various combinations emphasize different types of deviations between the distributions. These include location and scale shifts, differences in skewness and kurtosis as well as differences in the correlation of the variates. The test statistics ϕ_{nm}, R_{nm}, N_{nm} and T^k_{nm} were evaluated.

The power was again computed for 5% significance level and samples of equal size $n = m = 30$, 50, and 100 (small, moderate and large) in two and four dimensions. Table 5.7 and Table 5.8 illustrate the power of the considered tests calculated from 1000 replications.

The Friedman-Rafsky and the nearest neighbor tests show very similar rejection power. For all three sample sizes and dimensions the Energy test performed better than the Friedman-Rafsky and the nearest neighbor tests in almost all considered alternatives. This is astonishing because the logarithmic distance function is long range and the probability distributions in the cases 11 and 12 have a sharp peak in one corner of a d dimensional unit cube and in case 13 a sharp peak in the middle of this unit cube. The multivariate student distribution represents very mild departures from normality, but nevertheless the rejection rate of the Energy test is high. The power of the k nearest neighbor test T^k_{nm} increased considerably for $k \geq 2$ relative to $k = 1$ and almost reached the power of the Energy test for some optimal k in the considered examples. The value of the free parameter k has to be chosen independently from the result and there is no prescription on how to choose k [51]. In Figure 5.4 the power of the considered two-sample tests is shown for some selected combination of p.d.f.s from the Tables 5.5, 5.6.

To study the power of the k nearest neighbor test T^k_{nm} and Energy test ϕ_{nm} for sample sizes $n >> m$, for example $n = 30$ and $m = 3$, we selected combination of four dimensional p.d.f.s (P^X and P^Y) which are strong differently, see Table 5.9. The distribution denoted by N_{corr} is the same as in the case 4 of the Table 5.6, i.e. $N_{corr} = N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.4 & 0.5 & 0.7 \\ 0.4 & 1 & 0.6 & 0.8 \\ 0.5 & 0.6 & 1 & 0.9 \\ 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}\right)$. As result from Table 5.9 we obtain that the power of the k nearest neighbor test remained relatively low for alternatives, which include correlation between variables, for that case the Energy test is recommended. For alternatives which include shift in skewness or kurtosis the k nearest neighbor test should be used.

5.5 An example from high energy physics

We compare observed J/ψ particle production to a Monte Carlo simulation, where the observed sample is taken from the HERA-B detector [52] during the 2000 run.

Table 5.7: Power at significance level α =0.05, calculated from 1000 repetitions, $n = m = 30$, $n = m = 50$ and $n = m = 100$, $d = 2$

case	$R_{30,30}$	$N_{30,30}$	$T^2_{30,30}$	$T^3_{30,30}$	$T^4_{30,30}$	$T^5_{30,30}$	$\phi_{30,30}$
1	0.41	0.38	0.43	0.48	0.48	0.47	0.90
2	0.33	0.30	0.42	0.49	0.51	0.49	0.58
3	0.14	0.12	0.15	0.17	0.20	0.19	0.13
4	0.63	0.57	0.72	0.79	0.81	0.80	0.55
5	0.14	0.14	0.15	0.17	0.16	0.14	0.32
6	0.07	0.07	0.06	0.06	0.05	0.05	0.11
7	0.04	0.07	0.05	0.05	0.05	0.05	0.05
8	0.05	0.06	0.06	0.08	0.07	0.07	0.08
9	0.08	0.08	0.08	0.10	0.10	0.11	0.10
10	0.13	0.12	0.13	0.16	0.18	0.18	0.15
11	0.15	0.14	0.18	0.21	0.22	0.22	0.18
12	0.20	0.20	0.25	0.30	0.34	0.34	0.26
13	0.11	0.10	0.12	0.17	0.18	0.17	0.14
14	0.09	0.09	0.08	0.08	0.09	0.08	0.14

case	$R_{50,50}$	$N_{50,50}$	$T^4_{50,50}$	$T^5_{50,50}$	$T^8_{50,50}$	$T^{10}_{50,50}$	$\phi_{50,50}$
1	0.62	0.55	0.79	0.79	0.76	0.75	1.00
2	0.53	0.50	0.73	0.74	0.77	0.77	0.89
3	0.17	0.20	0.29	0.28	0.31	0.34	0.21
4	0.87	0.83	0.97	0.98	0.98	0.98	0.88
5	0.18	0.20	0.25	0.22	0.21	0.22	0.49
6	0.08	0.08	0.07	0.07	0.08	0.08	0.13
7	0.04	0.07	0.05	0.04	0.05	0.04	0.05
8	0.05	0.08	0.05	0.05	0.06	0.06	0.07
9	0.08	0.10	0.11	0.11	0.13	0.12	0.14
10	0.18	0.18	0.24	0.25	0.28	0.30	0.23
11	0.23	0.22	0.32	0.32	0.37	0.38	0.30
12	0.33	0.31	0.49	0.52	0.59	0.59	0.46
13	0.16	0.15	0.25	0.26	0.36	0.39	0.33
14	0.12	0.11	0.05	0.07	0.07	0.06	0.22

case	$R_{100,100}$	$N_{100,100}$	$T^5_{100,100}$	$T^8_{100,100}$	$T^{13}_{100,100}$	$T^{20}_{100,100}$	$\phi_{100,100}$
1	0.91	0.85	1.00	1.00	1.00	0.99	1.00
2	0.82	0.74	1.00	1.00	1.00	0.99	1.00
3	0.31	0.28	0.58	0.66	0.70	0.70	0.47
4	0.99	0.97	1.00	1.00	1.00	1.00	1.00
5	0.34	0.29	0.57	0.58	0.55	0.47	0.86
6	0.10	0.11	0.14	0.14	0.14	0.12	0.24
7	0.04	0.05	0.06	0.07	0.07	0.07	0.10
8	0.05	0.05	0.09	0.10	0.11	0.10	0.09
9	0.10	0.11	0.19	0.23	0.26	0.28	0.24
10	0.25	0.23	0.45	0.52	0.58	0.59	0.52
11	0.32	0.29	0.64	0.72	0.77	0.77	0.66
12	0.56	0.48	0.87	0.91	0.94	0.93	0.90
13	0.23	0.19	0.55	0.68	0.82	0.87	0.78
14	0.16	0.16	0.08	0.09	0.10	0.13	0.56

Table 5.8: Power at significance level α =0.05, calculated from 1000 repetitions, $n = m = 30$, $n = m = 50$ and $n = m = 100$, $d = 4$

case	$R_{30,30}$	$N_{30,30}$	$T^2_{30,30}$	$T^3_{30,30}$	$T^4_{30,30}$	$T^5_{30,30}$	$\phi_{30,30}$
1	0.76	0.70	0.77	0.81	0.80	0.76	1.00
2	0.54	0.51	0.61	0.66	0.67	0.66	0.90
3	0.14	0.13	0.15	0.19	0.22	0.23	0.23
4	0.17	0.17	0.18	0.20	0.20	0.19	0.13
5	0.24	0.21	0.20	0.22	0.22	0.21	0.73
6	0.07	0.07	0.05	0.08	0.07	0.06	0.18
7	0.06	0.06	0.06	0.06	0.05	0.05	0.06
8	0.06	0.07	0.06	0.07	0.07	0.06	0.08
9	0.12	0.12	0.15	0.16	0.18	0.19	0.18
10	0.26	0.26	0.31	0.36	0.39	0.39	0.30
11	0.42	0.37	0.41	0.49	0.53	0.53	0.47
12	0.55	0.51	0.64	0.71	0.75	0.74	0.66
13	0.16	0.16	0.16	0.22	0.27	0.29	0.27
14	0.15	0.13	0.10	0.12	0.14	0.13	0.17

case	$R_{50,50}$	$N_{50,50}$	$T^4_{50,50}$	$T^5_{50,50}$	$T^8_{50,50}$	$T^{10}_{50,50}$	$\phi_{50,50}$
1	0.94	0.93	0.99	0.99	0.98	0.97	1.00
2	0.80	0.78	0.92	0.93	0.95	0.95	1.00
3	0.16	0.17	0.30	0.32	0.43	0.44	0.47
4	0.25	0.26	0.31	0.30	0.31	0.30	0.18
5	0.37	0.35	0.40	0.38	0.36	0.36	0.95
6	0.09	0.10	0.08	0.07	0.07	0.07	0.31
7	0.06	0.06	0.06	0.06	0.07	0.08	0.07
8	0.08	0.09	0.07	0.08	0.09	0.09	0.07
9	0.18	0.20	0.29	0.29	0.30	0.31	0.30
10	0.45	0.42	0.59	0.62	0.66	0.68	0.62
11	0.61	0.58	0.76	0.78	0.83	0.83	0.77
12	0.79	0.76	0.94	0.95	0.96	0.96	0.93
13	0.19	0.20	0.39	0.42	0.58	0.57	0.62
14	0.19	0.18	0.21	0.21	0.26	0.27	0.31

case	$R_{100,100}$	$N_{100,100}$	$T^5_{100,100}$	$T^8_{100,100}$	$T^{13}_{100,100}$	$T^{20}_{100,100}$	$\phi_{100,100}$
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.98	0.98	1.00	1.00	1.00	1.00	1.00
3	0.31	0.25	0.71	0.83	0.92	0.94	0.94
4	0.52	0.46	0.77	0.81	0.79	0.75	0.49
5	0.70	0.63	0.87	0.87	0.83	0.79	1.00
6	0.16	0.16	0.22	0.20	0.18	0.16	0.63
7	0.06	0.08	0.08	0.08	0.08	0.07	0.10
8	0.09	0.09	0.15	0.16	0.17	0.17	0.12
9	0.31	0.29	0.54	0.60	0.65	0.67	0.60
10	0.72	0.66	0.94	0.96	0.97	0.98	0.97
11	0.88	0.84	0.99	1.00	1.00	1.00	0.99
12	0.98	0.96	1.00	1.00	1.00	1.00	1.00
13	0.35	0.29	0.79	0.92	0.98	0.99	0.99
14	0.28	0.23	0.50	0.57	0.63	0.65	0.68

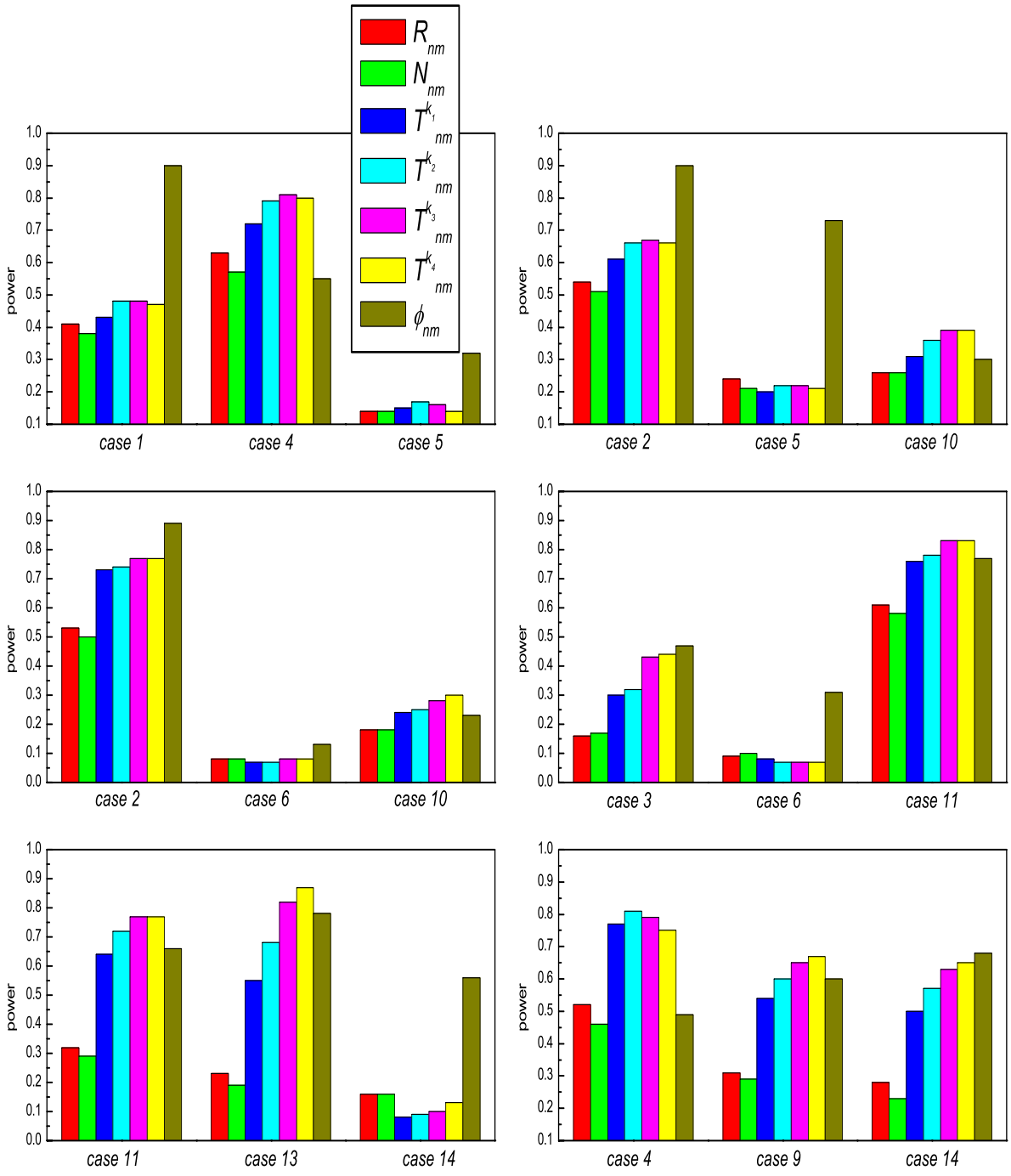

Figure 5.4: Power comparison for selected two-sample tests shown as histograms for $d = 2$ (left) and $d = 4$ (right)

Table 5.9: Power at significance level α =0.05, calculated from 1000 repetitions, $n = 30$, $m = 3$, $d = 4$

P^X	P^Y	$T^2_{30,3}$	$T^4_{30,3}$	$T^6_{30,3}$	$T^{10}_{30,3}$	$T^{15}_{30,3}$	$\phi_{30,3}$
$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$	0.39	0.56	0.63	0.73	0.78	0.42
$N(\mathbf{0},\mathbf{I})$	$N_{\log}(\mathbf{0},\mathbf{I})$	0.08	0.13	0.16	0.23	0.23	0.16
$N(\mathbf{0},\mathbf{I})$	N_{corr}	0.04	0.04	0.05	0.08	0.08	0.11
$N(\mathbf{0},\mathbf{I})$	t_2	0.17	0.31	0.37	0.49	0.55	0.11
$U(\mathbf{0},\mathbf{1})$	$CJ(0.5)$	0.08	0.08	0.09	0.09	0.09	0.15
$U(\mathbf{0},\mathbf{1})$	$CJ(0.3)$	0.11	0.09	0.10	0.09	0.08	0.20
$U(\mathbf{0},\mathbf{1})$	$CJ(0.1)$	0.13	0.11	0.11	0.09	0.07	0.30

The detector was designed to exploit the collisions of the HERA proton beam with a wire target. In proton-nucleus interaction $b\bar{b}$ quark pairs are created and hadronize to B mesons. The $b\bar{b}$ production cross section in proton-nucleus collisions is not well measured. The $b\bar{b}$ events are identified via the decay $B \to J/\psi$, where the J/ψ mesons decay to $\mu^+\mu^-$ or e^+e^-. In the selected data set the J/ψ mesons decay to $\mu^+\mu^-$. The B events are separated from directly produced J/ψ's by requiring that the J/ψ decay vertex does not coincide with the primary vertex, i.e. the $B \to J/\psi$ events are selected using the impact parameter of the muon tracks with respect to the wire and the decay distance, projected on the beam direction, between the secondary vertex and the target wire.

The original intention was to perform a detailed comparison of the observed J/ψ particle production to a Monte Carlo simulation. However, we found that even for small samples the agreement was not satisfactory and could not easily be improved. As a consequence all considered tests with large samples and high dimensions led to p-values much below 10^{-3}. Nevertheless we show a comparison in two dimensions for a restricted sample of the detached J/ψ events. The two similar samples of the detached J/ψ events, an observed sample of size $n = 32$ and a simulated sample of size $m = 800$, were compared using only the two variables, momentum of μ^+ and momentum of μ^-, see Figure 5.5. We have chosen only these two variables because then the distribution can be visualized.

The compatibility of the observed sample with the simulated sample is determined by the two-sample Energy test statistic ϕ_{nm} with the normalization of each components as described in subsection 4.2.4 and by the k nearest neighbor test statistic T^k_{nm}. The p-values for ϕ_{nm} and T^k_{nm} with $n = 32$, $m = 800$, $k = 5$, were found to be 0.07 and 0.6, respectively. Whereas the Energy test detects a significant difference between experimental data and simulation which supports the visiual impression of the Figure 5.5, the k nearest neighbor test is insensitive to it.

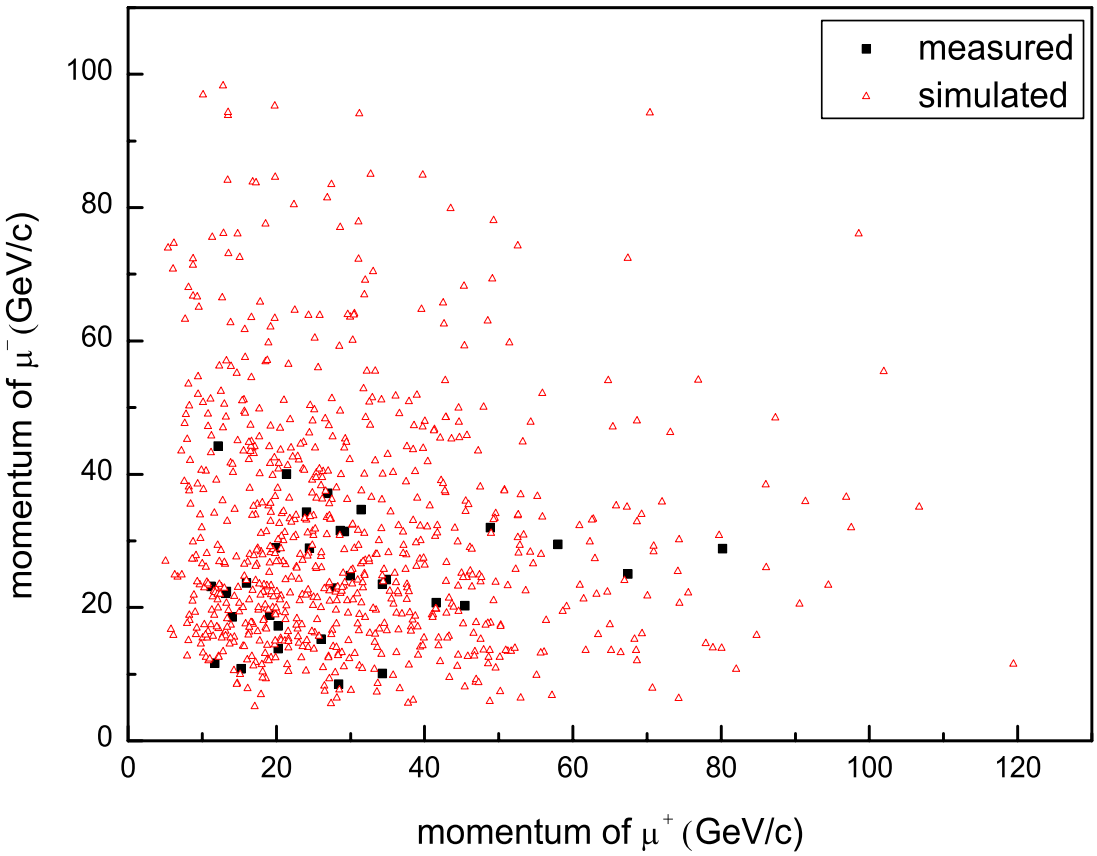

Figure 5.5: Momentum of $\mu^+\mu^-$ from $J/\psi \rightarrow \mu^+\mu^-$ for measured and simulated events.

Chapter 6

Deconvolution

This chapter concerns the estimation of p.d.f. in cases where no parametric form is available, and where the measurements (observations) are subject to additional random fluctuations due to the limited resolution of the measuring device. In high energy physics the procedure of correcting for these distortions is usually called *deconvolution* or *unfolding*. An introduction to unfolding and the related statistical problems can be found in [53].

6.1 The problem

A standard task in high energy physics and other fields of science is the measurement of some p.d.f. $f(\mathbf{x})$ using a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ where the r.v. $\mathbf{X}$ may be for example energy, decay times, etc. or some other quantity of one or more dimensions. Up to now in this thesis we have considered the observed values of the r.v.s under the assumption that they are free of error. But in practice measured values are affected by the *limited resolution* of the measuring device as expressed by a resolution function t and by the *acceptance* a of the measuring device which is the probability to obtain a given observation is less than 1. It may also be the case that t and a are not known analytically and must be obtained approximately by using Monte Carlo simulation, which will be another source of error.

Each observed value is characterized by a true measured value $\mathbf{x}$ and a measured value $\mathbf{x}'$. Suppose that the true observation $\mathbf{X}$ is distributed according to f. Mathematically the true p.d.f. $f(\mathbf{x})$ and the measured p.d.f. $f'(\mathbf{x}')$ are related by an equation of the type

$$f'(\mathbf{x}') = \int_{\mathbb{R}^d} k\left(\mathbf{x}', \mathbf{x}\right) f(\mathbf{x})\, d\mathbf{x}, \tag{6.1}$$

where $k\left(\mathbf{x}', \mathbf{x}\right) := t\left(\mathbf{x}', \mathbf{x}\right) a(\mathbf{x})$ is called *response function*, which gives the probability to obtain $\mathbf{x}'$, including the effect of acceptance, given that the true measured value was $\mathbf{x}$. The resolution function $t\left(\mathbf{x}', \mathbf{x}\right)$ is the conditional probability for the measured value $\mathbf{x}'$ given that the true value was $\mathbf{x}$. In what follows, for simplicity

we assume the acceptance to be unity. The acceptance correction can be separated from the unfolding problem.

Eq. (6.1) is an integral equation of the first kind where f is unknown. Solving Eq. (6.1) is known as unfolding or deconvolution.

If we have a priori information about the form of the p.d.f. f, for example f belongs to a parametric family $\mathcal{P} = \{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ then standard techniques for parameter estimation can be used to obtain estimators $\widehat{\boldsymbol{\theta}}$, where the hypothesized p.d.f. will be denoted by $\widehat{f}\left(\mathbf{x}, \widehat{\boldsymbol{\theta}}\right)$. In this case unfolding can be avoided. We can apply a GoF test by computing

$$\widehat{f'}(\mathbf{x}') = \int_{\mathbb{R}^d} t(\mathbf{x}', \mathbf{x})\, \widehat{f}(\mathbf{x})\, d\mathbf{x}$$

and by comparison $\widehat{f'}$ with the measured p.d.f. f'. This procedure is considerably simpler than unfolding.

As we will see unfolding is not a simple, straightforward procedure. One is well advised to ask in each problem if it can be avoided. But without unfolding many problems in practice cannot be solved, for example the results of experiments cannot be compared, since each experiment will have in general a different resolution.

6.2 Unfolding methods

It is not our intention to list all existing unfolding methods in the literature. Most of the important methods are discussed in [53].

6.2.1 Matrix inversion

For the numerical solution of Eq. (6.1) the p.d.f.s of f and f' have to be estimated by histograms. Then we obtain from Eq. (6.1)

$$d'_i = E[n_i] = \sum_{j=1}^{r} t_{ij} d_j \tag{6.2}$$

where $n_i \in \mathbb{N}$, $i = 1, 2, \ldots, s$, is the actual number of entries in bin i of the histogram of $\mathbf{X}'$ (measured histogram), $d_j \in \mathbb{R}_{\geq 0}$, $j = 1, 2, \ldots, r$, is the expectation value for the jth bin contents of the histogram of $\mathbf{X}$ and t_{ij} is the resolution matrix, which is the discrete version of $t(\mathbf{x}', \mathbf{x})$ and has the simple interpretation as a conditional probability:

$$t_{ij} = P(\text{observed in bin } i \mid \text{true value in bin } j).$$

Under the condition that the resolution matrix can be inverted, we obtain from Eq. (6.2)

$$d_i = \sum_{j=1}^{s} t_{ij}^{-1} E\left[n_j\right] = \sum_{j=1}^{s} t_{ij}^{-1} d'_j$$

where t_{ij}^{-1} is the matrix element of the inverted resolution matrix.

d'_j can be estimated by n_j and as estimator for d_i we will obtain

$$\widehat{d_i} = \sum_{j=1}^{s} t_{ij}^{-1} n_j. \tag{6.3}$$

It is well known that this procedure of matrix inversion produces bin-to-bin oscillations in the unfolded histogram.

Example 3 *Suppose $N = M = 2$ and let be the resolution matrix given by*

$$(t_{ij}) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},$$

i.e. if the true value lies in the first bin, it has a 60% chance of lying in the first bin of the histogram of X' and 40% for the second bin. The inverse matrix is

$$\left(t_{ij}^{-1}\right) = \begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix}$$

and if we measure $\begin{pmatrix} 20 \\ 30 \end{pmatrix}$ unfolding gives $\begin{pmatrix} 0 \\ 50 \end{pmatrix}$. Unfolding $\begin{pmatrix} 15 \\ 35 \end{pmatrix}$ gives $\begin{pmatrix} -25 \\ 75 \end{pmatrix}$. The bin contents tend to vary strongly from bin to bin, even matrix inversion method can produce unphysical negative bin contents.

6.2.2 Iterative unfolding

Statistically not significant bin-to-bin oscillations are damped by introducing in some way a measure of smoothness to $\widehat{d_i}$. This approach is known as regularization of the unfolded distribution, see [53]. Matrix inversion can be done iteratively. If the matrix inversion is stopped after some iterations, bin-to-bin oscillations are suppressed. Hence the choice of the number of iterations playing the role of the regularization.

To illustrate the idea of the iterative method we consider the Figure 6.1, which is taken from [53]. The first histogram is the input histogram, which should be smooth. It is obtained by grouping the simulated observation which is denoted by $\mathbf{X}^{MC}$, MC (Monte Carlo) refer to simulation of the true distribution. The bin contents d_j^{MC} of the simulated input histogram is an estimator for d_j. If no information about

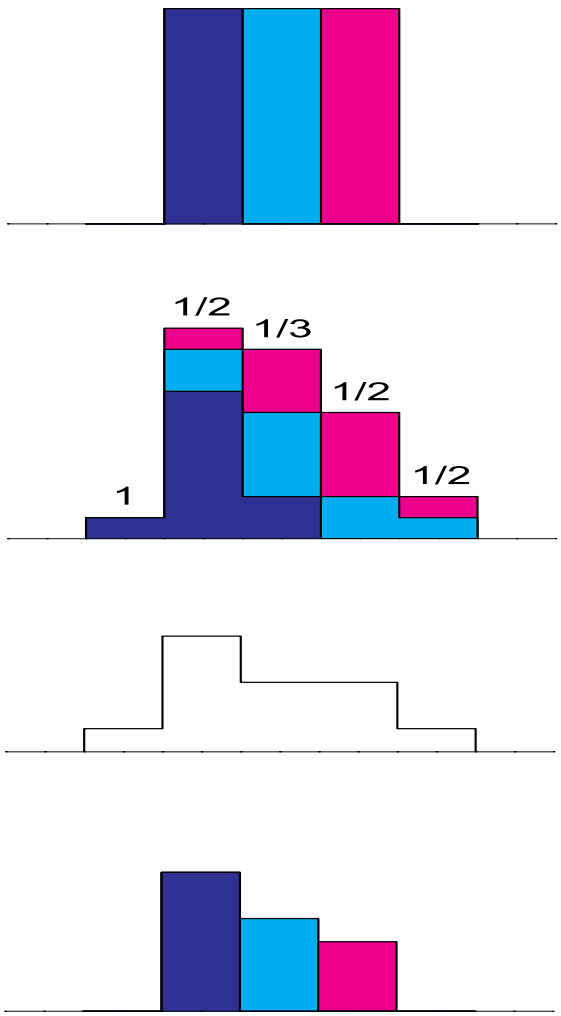

Figure 6.1: Illustration of iterative unfolding method.

the true histogram is available then it is recommended to start with a uniform distribution [53]. The second histogram, where the number of entries in ith bin is denoted by $d_i'^{MC}$, is obtained by folding the first histogram. The third histogram is the measured histogram. To get an agreement with the measured histogram in the iterative unfolding method the bin contents of the second histogram is weighted with the factor $\frac{n_i}{d_i'^{MC}}$. In the Figure 6.1 these factors are displayed on top of the bins. These weights are then propagated back into the input histogram. As result we get a first updated estimator $d_j^{(1)^{MC}}$ for d_j. This procedure of iterative unfolding can be expressed by following equations:

$$d_j^{(k+1)^{MC}} = \sum_{i=1}^{s} t_{ij} d_j^{(k)^{MC}} \frac{n_i}{d_i'^{(k)^{MC}}}$$

with

$$d_i'^{(k)^{MC}} = \sum_{j=1}^{r} t_{ij} d_j^{(k)^{MC}},\ k = 0, 1, \ldots$$

and $d_j^{(0)^{MC}} = d_j^{MC}$, $d_i'^{(0)^{MC}} = d_i'^{MC}$. The statistical uncertainties of the simulation are not taken into account. Therefore the size of the simulated true random sample should be much larger than the total number of measured values. In the limit $k \to \infty$ this procedure is equivalent to matrix inversion method, provided that the latter does not contain negative bin contents, i.e. it yields

$$\lim_{k\to\infty} d_j^{(k+1)^{MC}} = d_j.$$

6.3 A new binning-free unfolding approach

Binning of observations is always linked with loss of information and should be avoided where possible. Consequently we should expect unfolding methods based on binning to be inferior to methods based on each observation. In addition unfolding without binning offers several advantages:

- Arbitrary bin boundaries are avoided.
- Variable transformations are possible after unfolding.
- Arbitrary histogramming is possible after unfolding.
- Low statistic of the measurement can be handled in arbitrarily high dimensions where histogramming is problematic.

The iterative unfolding approach can be easily generalized to a binning-free method which is proposed in [54]. The difference is that instead of weighting the bin contents $d_i'^{MC}$ each individual observation $\mathbf{X}'^{MC}$ is weighted according to the ratio

of the 'local' densities of $\mathbf{X}'$ and $\mathbf{X}'^{MC}$. The support of these local densities is of the order of the resolution and the densities can be estimated with standard techniques like kth nearest neighbor method, see [15]. To suppress the statistical uncertainties of the estimation of the local densities the binning-free iterative unfolding method requires high sample sizes n. Often in high energy physics experiments the statistic of the measurements is not high. Therefore a binning-free unfolding method is required which can also be applied to low sample sizes in arbitrarily high dimensions.

The new method proposed in [55] is based on the migration of the simulated observations $\mathbf{X}_i^{MC}$. $\mathbf{X}^{MC}$ migrates in the true variate space until the two samples, the simulated sample and the observed sample, coincide. The new unfolding approach can be performed as follows:

1. Choose the size of the simulated sample of $\mathbf{X}^{MC}$ equal to the size of the observed sample of $\mathbf{X}'$, $n = m$.
2. Generate for each $\mathbf{X}_i^{MC}$ k observations $\mathbf{Y}_r^s$, $r = 1, 2, \ldots, n$, $s = 1, 2, \ldots, k$, from a p.d.f. which estimates the resolution function. k is typically of the order of 20.
3. Compute the energy of the pooled sample
   $\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n; \mathbf{Y}_1^1, \mathbf{Y}_1^2, \ldots, \mathbf{Y}_1^k, \mathbf{Y}_2^1, \mathbf{Y}_2^2, \ldots, \mathbf{Y}_2^k, \ldots, \mathbf{Y}_n^1, \mathbf{Y}_n^2, \ldots, \mathbf{Y}_n^k$.
4. Select randomly one $\mathbf{X}_i^{MC}$, set $\mathbf{X}_i^{MC} = \mathbf{X}_i^{MC} + \mathbf{\Delta}$, where $\mathbf{\Delta}$ is distributed according to a arbitrary p.d.f. with a variance of order of the FWHM of the resolution function and regenerate at this new position the k accompanying observations $\mathbf{Y}_r^s$, $s = 1, 2, \ldots, k$.
5. Compute the change of energy ϕ, see Eq.(6.4). The migration of $\mathbf{X}_i^{MC}$ is accepted if energy ϕ has decreased, otherwise it is rejected.
6. Step 4 and Step 5 is repeated until the minimum of the energy ϕ is reached.

The computation of the energy ϕ follows the relation:

$$\begin{aligned}\phi \;=\; & \frac{1}{n(n-1)} \sum_{i<j}^{n} R\left(\left|\mathbf{x}'_i - \mathbf{x}'_j\right|\right) + \frac{1}{nk(k-1)} \sum_{i=1}^{n} \sum_{l<s}^{k} R\left(\left|\mathbf{y}_i^l - \mathbf{y}_i^s\right|\right) + \\ & + \frac{1}{k^2 n(n-1)} \sum_{l,s}^{k} \sum_{i<j}^{n} R\left(\left|\mathbf{y}_i^l - \mathbf{y}_j^s\right|\right) - \frac{1}{kn^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{k} R\left(\left|\mathbf{x}_i - \mathbf{y}_j^l\right|\right). \end{aligned} \tag{6.4}$$

Also this unfolding technique produces oscillating solutions unless a regularization is introduced, which can be done in two different ways:

1. The migration process can be stopped before the oscillations become intolerable.

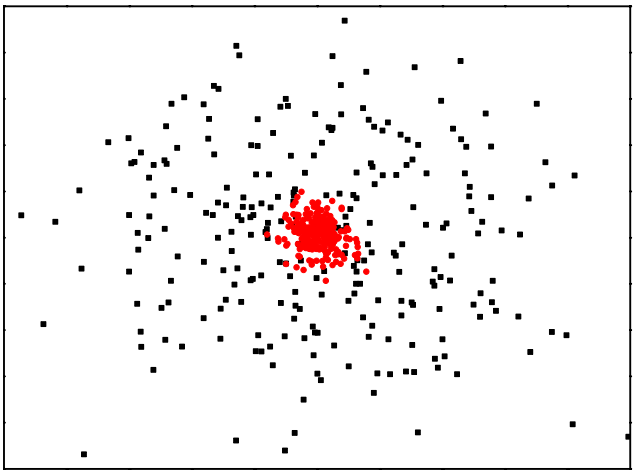

Figure 6.2: Unfolding of a point source with the new method.

2. The value of the parameter k can be used. Large values provide high resolution but introduce oscillation.

Here we present first experiences with the new binning-free unfolding method. A detailed study is beyond the scope of this work. A preliminary study of unfolding a distribution generated for a point source with Gaussian resolution $\sigma = 1$ is shown in Figure 6.2. The black dots represent the observed point source, the red points are result of unfolding.

The dependence of the point resolution on the parameter k is shown in Figure 6.3 as a function of the number of observations and of the value of k. Each measurement is an average of 40 simulations. The results show that the resolution converges with increasing of k and n.

We present now a simple example in two dimensions shown in Figure 6.4. This example of course is not very typical for a physics application but it shows how an observed distribution can be unfolded even when the original distribution is rather peaked and when the statistics is low. The true picture of the face, displayed in Figure 6.4 a), contains 500 observations, which has been blurred with a Gaussian resolution function, Figure 6.4 b). It would have been quite difficult to apply standard unfolding techniques to the picture of the Figure 6.4 b), which are based on histogramming. The binning-free iterative unfolding method to Figure 6.4 b) would also fail, since the two-dimension sample size of $n = 500$ observations is not large enough. We know of no other unfolding methods which are able to perform this

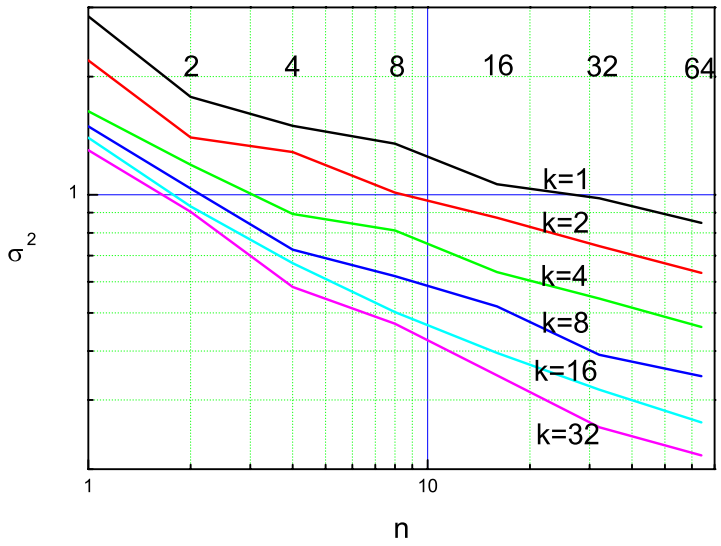

Figure 6.3: Illustration of the dependence of the new method of the parameter k.

job. Figure 6.4 c) shows the result of the new unfolding method with a Gaussian distance function and $k = 20$, which is obtained after 40000 trials of random moves.

6.3.1 Some remarks

- We propose to use Gaussian with width similar to the resolution or logarithmic distance functions.
- The average migration steps should be larger than the resolution. We propose to generate $\mathbf{\Delta}$ from a uniform distribution.
- Only the combinations which contain the moving observations lead to a change of the energy.
- If acceptance and resolution are independent of the location, the k accompanying observations can migrate together with the Monte Carlo observation.

The new approach opens the possibility to solve problems which are not accessible with conventional unfolding methods. It is especially powerful in multidimensional applications with sharp structures. If the statistic of the measurement is high, histogramming methods are preferable because they are faster.

More work is needed in order to study the effect of local minima of ϕ and to optimize the migration process.

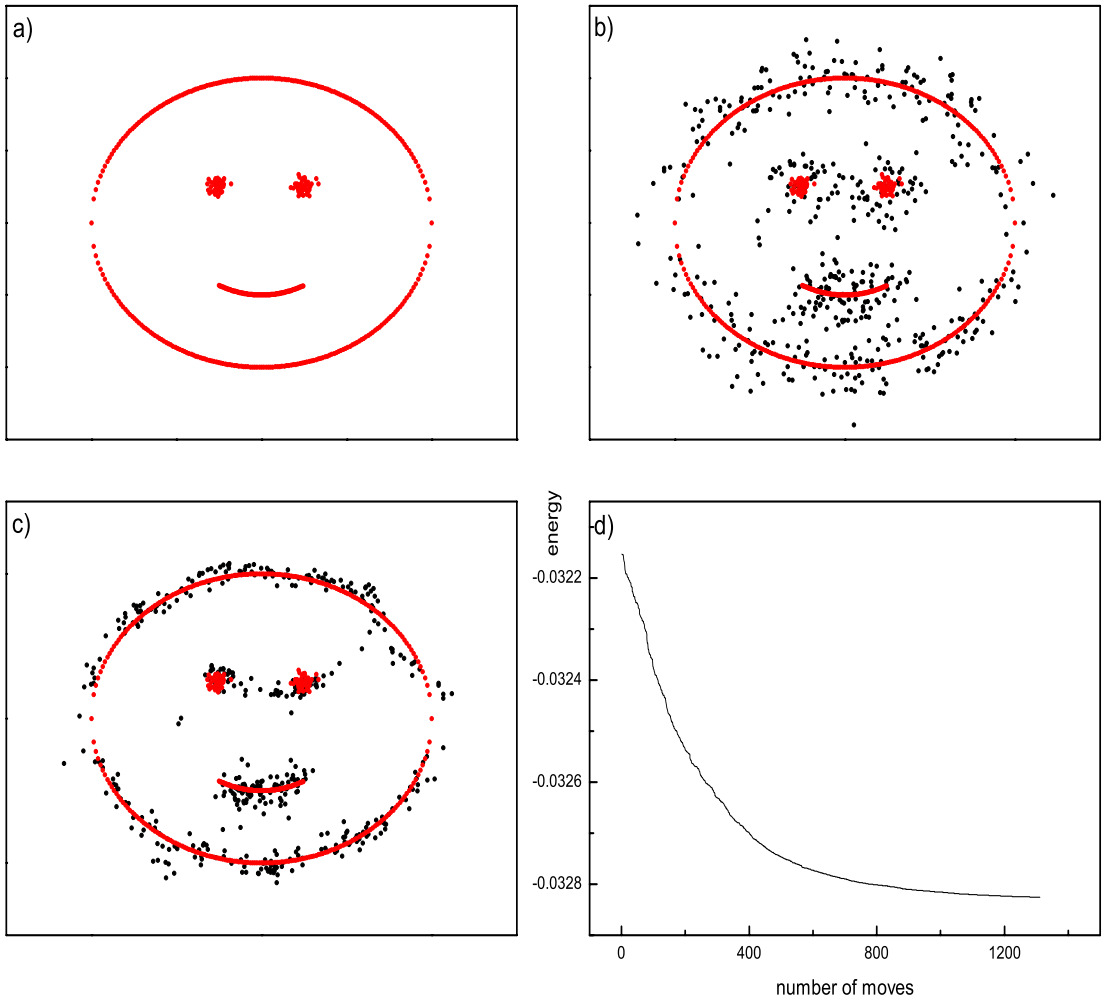

Figure 6.4: Unfolding of a simple picture: a) true face, b) observed face, blurred with a Gaussian resolution function, c) the energy unfolded face, d) convergence of energy

Chapter 7

Summary

The concept of physics energy is introduced into statistics. It provides very general tools to compare two samples to each other. The energy of samples, which are drawn from statistical distributions, is defined in a similar way as for discrete charge density distributions in electrostatics. It is computationally simple, nonparametric, and avoids arbitrary binning.

A system of two sets of point charges with opposite sign is in a state of minimum energy if they are equally distributed. This property is used to construct new nonparametric, multivariate Goodness-of-Fit tests, to check whether two samples belong to the same parent distribution and to deconvolute distributions distorted by measurement.

The Energy tests are powerful nonparametric, multivariate omnibus tests. Under many alternatives investigated in the simulation study, the Energy tests give much better results than the existing relevant tests in all considered dimensions. It is especially powerful in multidimensional applications.

The statistical minimum energy configuration does not depend on the application of the one-over-distance power law of the electrostatic potential. To increase the power of the new approach other monotonic decreasing distance functions may be chosen. We proved that the new energy technique is applicable to all distance functions which have positive Fourier transforms.

Multivariate, binning-free unfolding based on energy is straightforward. The new approach opens the possibility to solve problems which are not accessible with the conventional unfolding methods. The new method of unfolding that has been proposed in this study looks promising for future research.

Bibliography

[1] S. Brandt, Data analysis: Statistical and computational methods for scientists and engineers, third edition, Springer Verlag (1999).

[2] P.H. Garthwaite, I.T. Jolliffe and B. Jones, Statistical inference, Prentice Hall, London (1995).

[3] J.C.W. Rayner and D.J. Best, Smooth tests of goodness of fit, Oxford University Press, Oxford (1989).

[4] B.S. Duran, A survey of nonparametric tests for scale, Communications in Statistics: Theory and Methods **5** (1976) 1287.

[5] W.J. Conover, M.E. Johnson and M.M. Johnson, A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data, Technometrics **23** (1981) 351.

[6] H. Buening, Robuste und adaptive Tests, De Gruyter, Berlin (1991).

[7] R.B. D'Agostino and M.A. Stephens, Goodness-of-fit techniques, Marcel Dekker, New York (1986).

[8] B. Aslan and G. Zech, Comparison of different goodness-of-fit tests, Proceedings of Conf. Advanced Statistical Techniques in Particle Physics, ed. M.R. Whalley and L. Lyons, Durham (2002).

[9] R. Fisher, The conditions under which χ^2 measures the discrepancy between observations and hypothesis, Journal of the Royal Statistical Society **87** (1924) 442.

[10] T.R.C. Read and N.A.C. Cressie, Goodness-of-fit statistics for discrete multivariate data, Springer, New York (1988).

[11] T.R.C. Read, Small sample comparisons for the power divergence goodness-of-fit statistics, Journal of the American Statistical Assosiation **79** (1984) 929.

[12] G. Shorack and J. Weller, Empirical processes with applications to statistics, Wiley, New York (1986).

[13] O. Vasicek, A test for normality based on sample entropy, Journal of the Royal Statistical Society Ser. B **38** (1976) 54.

[14] E.J. Dudewicz and E.C. van der Meulen, Entropy-based tests of uniformity, Journal of the American Statistical Assosiation **76** (1981) 376, 967.

[15] B.W. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, London (1986).

[16] E. Parzen, On estimation of a probability density function and mode, Annals of Mathematical Statistics **33** (1962) 1065.

[17] P.J. Bickel and M. Rosenblatt, On some global measures of the deviations of density function estimates, Annals of Statistics **1** (1973) 1071.

[18] M. Rosenblatt, A quadratic measure of deviation of two dimensional density estimates and a test of independence, Annals of Statistics **3** (1975) 1-14.

[19] K.V. Mardia, Measures of multivariate skewness and kurtosis with applications, Biometrika **57** (1970) 519.

[20] N. Henze and B. Zirkler, A class of invariant and consistent tests for multivariate normality, Communications in Statistics: Theory and Methods **19** (1990) 3595.

[21] T.W. Epps and L.B. Pulley, A test for normality based on the empirical characteristic function, Biometrika **70** (1983) 723.

[22] L. Baringhaus and N. Henze, A consistent test for multivariate normality based on the empirical characteristic function, Metrika **35** (1998) 339.

[23] S. Csörgö, Consistency of some tests for multivariate normality, Metrika **36** (1989) 107.

[24] D.W. Scott, Multivariate density estimation: Theory, practice and visualisation, Wiley, New York (1992).

[25] M.G. Kendall and A. Stuart, The advanced theory of Statistics, Volume 2, 4th Edn., Charles Griffin, London (1960).

[26] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, Statistical methods in experimental physics, North-Holland, Amsterdam (1971).

[27] D. Morgenstern, Proof of a conjecture by Walter Deuber concerning the distances between points of two types in R^d, Discrete Mathematics **226** (2001) 347.

[28] W. Deuber, 2. Problem of W. Deuber, Problem 303, Discrete Mathematics **192** (1998) 348.

[29] D.C. Champeney, Fourier transforms and their physical applications, Academic Press, London (1973).

[30] I.M. Gel'fand and G.E. Shelov, Generalized functions, Vol.1: Properties and operations, Academic Press, New York (1964).

[31] E.L. Lehmann, Elements of large-sample theory, Springer, New York (1998).

[32] R.J. Serfling, Approximation theorems of mathematical statistics, Wiley, New York (1980).

[33] A.J. Lee, U-statistics, Marcel Dekker, New York (1990).

[34] A. Bowman and P. Foster, Adaptive smoothing and density-based tests of multivariate normality, Journal of the American Statistical Assosiation **88** (1993) 529.

[35] N. Henze and T. Wagner, A new approach to the BHEP tests for multivariate normality, Journal of Multivariate Analysis **62** (1997) 1.

[36] N. Gürtler, Asymptotische Untersuchung zur Klasse der BHEP-Tests auf multivariate Normalverteilung mit festem und variablem Glättungsparameter, Ph.D. thesis, University of Karlsruhe (2000).

[37] D.S. Moore, Tests of chi squared type, In Goodness-of-Fit Techniques (eds R.B. D'Agostino and M.A. Stephens), pp. 63-95, Marcel Dekker, New York (1986).

[38] B. Efron and R. Tibshirani, An introduction to the bootstrap, Chapman and Hall, New York (1993).

[39] B. Efron, Bootstrap methods; another look at the jackknife, Annals of Statistics **7** (1979) 1.

[40] R.R. Wilcox, Introduction to robust estimation and hypothesis testing, Academic Press, San Diego (1997).

[41] Y.A. Lepage, A combination of Wilcoxon's and Ansari-Bradley's statistics. Biometrica **58** (1971) 213.

[42] J.H. Friedman and L.C. Rafsky, Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests, Annals of Statistics **7** (1979) 697.

[43] N. Henze, A multivariate two-sample test based on the number of nearest neighbor type coincidences, Annals of Statistics **16** (1988) 772.

[44] A. Wald and J. Wolfowitz, On a test whether two samples are from the same population, Annals of Mathematical Statistics **11** (1940) 147.

[45] N. Henze and M.D. Penrose, On the multivariate runs test, Annals of Statistics **27** (1998) 290.

[46] N. Henze, Über die Anzahl von Zufallspunkten mit typ-gleichem nächsten Nachbarn und einen multivariaten Zwei-Stichproben-Test, Metrika **31** (1984) 259.

[47] H. Buening, Kolmogorov-Smirnov- and Cramèr-von Mises type two-sample tests with various weight functions, Communications in Statistics-Simulation and Computation **30** (2001) 847.

[48] L. Devroye, Non-Uniform Random Variate Generation, Springer-Verlag, New York (1986).

[49] J.H. Ahrens and U. Dieter, Pseudo-Random Numbers, Wiley, New York (1977).

[50] R. Bahr, A new test for the multi-dimensional two-sample problem under general alternative, (German) Ph.D. thesis, University of Hannover (1996).

[51] M.F. Schilling, Multivariate two-sample tests based on nearest neighbors, Journal of the American Statistical Assosiation **81** (1986) 799.

[52] The HERA-B collaboration, report on status and prospects, DESY-PRC 00/04 (2000).

[53] G. Zech, Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding, Desy 95-113 (1995).

[54] L. Lindemann and G. Zech, Unfolding by weighting Monte Carlo events, Nuclear Instruments and Methods in Physics Research A**354** (1995) 516.

[55] G. Zech and B. Aslan, Binning-free unfolding based on Monte Carlo migration, Proceedings of Conf. Statistical Problems in Particle Physics, Astrophysics and Cosmology, Stanford (2003).

Danksagung

Vorrangig gilt mein Dank Herrn Prof. Dr. Günter Zech. Als kompetenter Ansprechpartner auf fachlicher wie auch privater Ebene stand er mir stets freundlich zur Verfügung. Durch viele anregende Diskussionen hat er wesentliche Teile meiner Arbeit geprägt. Die wesentlichen Techniken des wissenschaftlichen Arbeitens konnte ich unter seiner Anleitung erlernen.

Mein Dank gilt auch Herrn Dr. Ulrich Werthenbach und Herrn Dr. Torsten Zeuner, die über die Jahre immer hilfsbereiter Kollegen waren. So manche Diskussionen über Physikalisches und Alltägliches schaffte ein angenehmes Arbeitsklima.

Schließlich möchte ich meiner Mutter für ihre vielseitige Unterstützung während der Promotion danken.