UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Implementation of Goodness of Fit Methods as an R-Package

Mahmoud Abu Azoum

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Olivier THAS

2021
2022

Faculty of Sciences
School for Information Technology

Master of Statistics and Data Science

Master's thesis

Implementation of Goodness of Fit Methods as an R-Package

Mahmoud Abu Azoum

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Olivier THAS

Acknowledgements

Firstly, I'd like to thank my supervisors, Dr. Prof. ***Olivier Thas***, for his motivation, continual support, and guidance throughout this research project. You assisted me by providing counsel, and they have kindly offered their time for frequent, meaningful meetings and detailed explanations when I am confused. I am delighted to be one of your students.

Second, I'd want to express my appreciation to all of my lecturers at Hasselt University. Thank you for sharing your vast knowledge and wonderful experiences with us. Following that, I'd like to express my heartfelt gratitude to everyone of my classmates for their unwavering support over the last two years.

Lastly, but not least, my deepest gratitude goes to my beloved family members, my parents, my wife and my young princesses, and relatives for your support, advice, sacrifice, and all the challenges you faced while raising me physically and intellectually. All of my daily activities would be meaningless without your well-wishes and fervent prayers. To all of my friends who provided support.

Thank you

Mahmoud ABU AZOUM

Antwerpen; June, 2022

Abstract

In goodness-of-fit tests, the null hypothesis expresses that the true hypothesized distribution and the observed distribution are the same. Accepting this null hypothesis will provide some confidence in using the estimated model. Otherwise, the rejection of such a hypothesis will often not provide any indication of the true distribution. The main reason for such a result is that traditional goodness-of-fit tests like the tests based on the Empirical Distribution Function (EDF) or ***Pearson*** χ^2 goodness-of-fit tests are not good at identifying the nature or sources of departure from the tested hypothesis. Smooth tests and generalized smooth tests can be applied as tests with power that is almost always competitive with alternative tests.

The main objective of this study was to implement smooth goodness-of-fit methods as an **R**-Package. Under this objective, a top-level function **stGOF** was designed, programmed, and validated. This function allows for many approaches like fixed order, bootstrap, and data-driven. In addition, four datasets were added to the **stGOF** package.

As a result, the **stGOF** package has been built. A comprehensive and in-depth explanation was provided in this thesis. In addition, arguments of the functions in the package were presented as well as validated examples.

Keywords: Smooth test, Goodness-of-fit, **stGOF** Package

Contents

Chapter 1

Introduction

1.1 Background

One of the main aspects of statistical inference that is of great importance to researchers in variety fields is the validity of the assumed distribution of the data under investigation. Although the hypothesized distribution of the data can be suggested, it is frequently different from the actual or observed distribution. For instance, to test the validity of the Student's t distribution as a hypothesized distribution, numerous existing techniques could be used like the ***Kolmogorov-Smirnov*** (K-S) or the ***Cramer- von Mises*** classes of omnibus tests of goodness-of-fit [3, 16]. These tests have been widely used to assess the equality of distributions with or without estimated parameters in one-sample and two-sample issues [4]. The null hypothesis is based on the assumption that the predicted and observed distributions are same. Acceptance of this null hypothesis will provide some confidence in using the estimated model. Otherwise, the rejection of such a hypothesis will not provide any indication of the true distribution. The main reason for such a result is that traditional goodness-of-fit tests like the tests based on the Empirical Distribution Function (EDF) or ***Pearson*** χ^2 goodness-of-fit tests are not good enough to identify the nature or sources of departure from the tested hypothesis [3].

Pearson's χ^2-test (1900), was really developed to test the goodness-of-fit of discrete distributions. While when it be used for continuous data, an arbitrarily classification into k groups is applied. Which means it has limitations [14]. In an attempt to take advantage of the properties of a continuous distribution, ***Neyman*** (1937) observed that if a probability integral transform based on the probability density function (PDF) is used under the null hypothesis, all goodness-of-fit tests can be converted to one type of hypothesis test. Apart from an ANOVA-like decomposition of the goodness-of-fit test statistic, the smooth test has the advantage of being able to discover the nature and sources of departure from the null hypothesis for "smooth" alternatives (alternatives that are extremely near to the null hypothesis). The smooth test may be used as both an omnibus and a more directed test for more specific alternatives. This allows, among other things, the smooth test to have higher power than traditional goodness-of-fit tests.

In 1937, ***Neyman*** proposed the smooth test to evaluate a simple null hypothesis declaring that observations follow a well-known continuous distribution function F. Smooth tests are based on embedding the hypothesised distribution g within a large family of distributions, g_k, which is indexed by a k-dimensional parameters vector $\theta^t = (\theta_1, \ldots, \theta_k)$- in such a way that $\theta = 0$ corresponds to the hypothesised distribution. Application of the probability integral transformation means that the distribution for which we test can always be taken to be uniform. The Legrendre polynomials are used in his tests. And the components of the smooth test statistic are used to detect the type of deviation from the hypothesized distribution as a diagnostic. Such as, the first detects a mean shift, the second a variance shift, the third a change in skewness and the fourth a change in kurtosis. If the orthonormal set was chosen differently, then different alternatives would be detected by the components. So, the orthonormal set should be chosen with the alternatives one wishes to most powerfully detect in mind.

In addition, ***Neyman*** didn't use score tests [19]. Instead, he required that the tests derived for the smooth model should be locally uniformly most powerful symmetric, unbiased and of size α. Unfortunately, only asymptotically are those constraints realised for the tests suggested. ***Neyman***'s work has been extended by ***Barton*** (1956) and his model. A little progress was made with testing for composite hypotheses.

Rayner & ***Best*** (1990)[19] also extended the work of ***Neyman***. The main difference between ***Rayner*** & ***Best*** and the others, is that they use orthonormal functions, while the others used powers of the cumulative distribution function. The tests of the others require tables of constants to implement and do not have other advantages of the orthonormal formulation. These advantages include that the components are often identifiable with known moment-type statistics used in other tests of fit, that the components are asymptotically independent and that the components have a convenient asymptotic distribution: the standard normal. For these reasons, the smooth tests based on orthonormal functions are preferred over other tests [23].

Finally, **smooth tests** of goodness-of-fit or smooth order k alternative can be defind as a test which is used to assess the fit of data to a given probability density function within a class of alternatives that differ 'smoothly' from the null model. These alternatives are characterized by their order. The greater the order the richer the class of alternatives [21].

1.2 Objectives

In the current study, the main objective was:

- To implement smooth Goodness of Fit Methods as `R`-Package.

And under this objective, there are several goals to be done

- Design, program, and validate a top-level function `stGOF`, which can take into account all approaches like fixed order, bootstrap, and data-driven.
- Add datasets to the `R` package, such as `Pseudo-Random Generator (PRG)` , `PCB Concentration`, `Pulse Rate` , and `Cultivars` [23].
- Add `help` files for all functions and datasets.

Chapter 2

Methodology

2.1 Introduction

The history of statistical hypothesis testing is extensive. It was traced back to Bayes (1763) to Neyman and Pearson (1933). However, it wasn't until Pearson's goodness-of-fit test was published in 1900 that hypothesis testing became widely used. This was the most significant scientific breakthrough of the twentieth century [6], even after 100 years. Simply stated, Pearson's (1900) test statistic is calculated as follows:

$$P_{\chi^2} = \Sigma_{j=1}^{q} \frac{(O_j - E_j)^2}{E_j},$$

where O_j denotes the observed frequency and E_j is the (expected) frequency that would be obtained under the distribution of the null hypothesis, for the j th class, $j = 1, 2, \ldots, q$. The most common outcome of this goodness of fit test has been to give a $\boldsymbol{p}$-value that reflects the consistency of the data with the hypothesized distribution.

In addition, other powerful goodness of fit tests are commonly used, such as the Anderson-Darling, Cramer-von Mises, and Shapiro-Wilk tests.

Angus Data is a very good example introduced by Rayner, Thas, and Best (2011) in their article[20] to explain the inconsistency with exponentiality when Pearson's χ^2, Anderson-Darling, Cramer-von Mises, and Shapiro-Wilk tests were applied. Angus[1] gives 20 operational lifetimes in hours:

6278, 3113, 5236, 11584, 12628, 7725, 8604, 14266, 6125, 9350, 3212, 9003, 3523, 12888, 9460, 13431, 17809, 2812, 11825, 2398.

In testing the null hypothesis that the data follow the exponential distribution, the Pearson–Fisher test with four classes constructed as in Figure 1, the Anderson-Darling, Cramer-von Mises, and Shapiro-Wilk tests give bootstrap p-values 0.0126, 0.0065, 0.0058, and 0.0009, respectively. It would appear that the data are not consistent with exponentiality. They return to the analysis of this data set several times throughout their article [20].

One issue with traditional goodness-of-fit tests was that the outcome was either confirm or reject the model. Furthermore, there were no diagnoses, and the tests did not recommend other models. The smooth test doesn't have the same issue. The smooth test approaches can give diagnostics as well as model selection techniques that produce smooth alternative models.

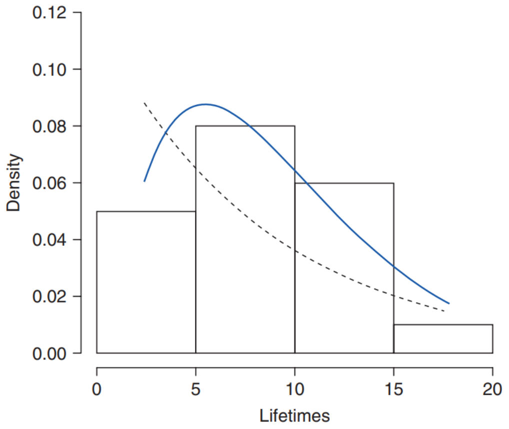


Figure 2.1: Histogram and density estimates of the hypothesized exponential distribution (dashed line) and the improved density (full line) for the Angus lifetimes data. Lifetimes are given in thousands of hours. The counts in the four classes are 5, 8, 6, and 1.

2.2 Smooth Test Structure

Let $X_1, \ldots, X_n$ donate an *i.i.d.* random sample from a distribution with density function $f(x;\beta)$, where β is a $q \times 1$ vector of nuisance parameters. In most cases, the X_i will be simply discrete or continuous. If the distribution is fully specified, such as the continuous uniform distribution on (0,1), or an exponential distribution with known rate parameter λ_0, then β is empty, and may be omitted. It is more usual for the distribution to be loosely specified, such as the normal distribution with unspecified mean μ and unspecified standard deviation σ; in which case $\beta = (\mu, \sigma)^T$. To construct a smooth test $f(x;\beta)$ is first nested in an alternative of order k:

$$g_k^N(x;\theta,\beta) = C(\theta,\beta)exp\left\{\sum_{i=1}^{k} \theta_i h_i(x;\beta)\right\} f(x;\beta),$$

where $C(\theta,\beta)$ is a normalizing constant that is assumed to exist, and $\{h_i(x;\beta)\}$ is a set of orthonormal functions on $f(x;\beta)$ with $h_0(x;\beta) = 1$ for all x. If E_0 means expectation when the model that generated the data is $f(x;\beta)$, then orthonormality means $E_0[h_r(X;\beta)h_s(X,\beta)] = \delta_{rs}$ for r and $s = 0, 1, \ldots$. Testing for $f(x;\beta)$ is achieved by assessing if the θ_i are consistent with zero: that is, by testing $H_0 : \theta = (\theta_i) = 0$ against $K : \theta \neq 0$.

This approach is a generalization of Neyman's vision [15]: he considered testing for the continuous uniform distribution and used the Legendre polynomials for the orthonormal system. The N in $g_k^N(x;\theta,\beta)$ is to acknowledge Neyman's contribution.

To utilize the orthonormality under the null hypothesis the score test of $H_0 : \theta = 0$ against $K : \theta \neq 0$ is used. Information about score tests applied to this setting is given by Rayner (1997) [18], and Rayner, Thas, and Best (2011) [20]. This leads to a quadratic form $S_k = V^T \Sigma^{-1} V$ in which V is the score vector and Σ is the asymptotic covariance matrix of V under the null hypothesis. Using orthonormal functions in $g_k^N(x;\theta,\beta)$ results, in certain circumstances, in components that under the null hypothesis are asymptotically independent and asymptotically standard normal. Hence Σ is the identity matrix, the score test statistic S_k is a sum of squares of the elements of V and has null distribution χ^2_k . These circumstances include when testing for distributions without nuisance parameters and in testing for distributions from exponential families.

In the fully specified case use of the score test leads to the same result as Neyman obtained. However, Neyman[15] used what has become known as ***the generalized fundamental lemma of Neyman and Pearson*** to show that the optimal test is approximately that which rejects for large values of $S_k = V_1^2 + \cdots + V_k^2$ in which the orthonormal functions $h_r(X)$ are the normalized Legendre polynomials and the *components* V_r are defined as $V_r = \Sigma_{j=1}^n h_i(X_j)/\sqrt{n}$.

When there are nuisance parameters an immediate application of this approach often finds that the asymptotic covariance matrix in the score test is singular. This can be interpreted as indicating that some of the θ_i are playing the same role as some of the nuisance parameters. If $\theta_1, \ldots, \theta_q$ are redundant, a sensible solution is to remove these parameters from the model, by modifying $g_N^k(x;\theta,\beta)$ so that it involves the k $h_i(x;\beta)$ after the first q:

$$g_k^N(x;\theta,\beta) = C(\theta,\beta)exp\left\{\sum_{i=q+1}^{q+k} \theta_i h_i(x;\beta)\right\}f(x;\beta),$$

If $\hat{\beta}_0$ is the maximum likelihood estimator of β under the null hypothesis and if $\hat{V}_r = \Sigma_{j=1}^n h_r(X_j;\hat{\beta}_0)/\sqrt{n}$ then for some distributions including the binomial, Poisson, exponential, univariate normal, and geometric, the likelihood equations are equivalent to $\hat{V}_1 = \cdots = \hat{V}_q = 0$ and the score test statistic simplifies from a general quadratic form to

$$\hat{S}_k = \hat{V}_{q+1}^2 + \cdots + \hat{V}_{q+k}^2$$

The main difference between Rayner, Thas,& Best and the others is that they frequently use orthonormal functions, while the others used powers of the cumulative distribution function. The tests of the others require tables of constants to implement and do not have other advantages of the orthonormal formulation. So, for example, when $\left\{h_i(x;\beta)\right\}$ is a set of Hermite polynomials orthonormal on the normal distribution, $\hat{V}_3$ and $\hat{V}_4$ are standardized versions of the skewness and kurtosis coefficients. When using the orthonormal polynomials for the univariate Poisson and binomial distributions, the order two components are standardized versions of the indices of dispersion. In these cases, the higher order components are thus generalizations of well-established tests. These advantages include that the components are often identifiable with known moment-type statistics used in other tests of fit, that the components are asymptotically independent, and that the components have a convenient asymptotic distribution: the standard normal. For these reasons, the smooth tests based on orthonormal functions are preferred over other tests [19, 20].

In the construction of the smooth tests the orthonormal functions $\left\{h_r(x,\beta)\right\}$ and the order k must be specified. These two define the class of smooth alternatives $\left\{g_k^N(x;\theta,\beta)\right\}$.

In testing for the continuous uniform distribution Neyman [15] recommended $k = 4$. If there is no additional information, this is a bit of sound advice. However, it is possible to use the data in a data-driven manner.

The orthonormal functions $\Big\{h_r(x,\beta)\Big\}$ should be chosen to best detect the alternatives of interest. If the null hypothesis specifies the continuous uniform distribution and $\Big\{h_r(x,\beta)\Big\}$ is considered to be the Legendre polynomials, then the null hypothesis specifies the probability density function to be a polynomial of degree zero, while the order k alternative is a polynomial of degree k. However, if a periodic rather than a polynomial alternative is more compatible with the scenario under which the data are collected then a more appropriate set of orthonormal functions could be $\sqrt{2}cos(j\pi x)$. In choosing the orthonormal functions the objective is to specify the alternatives of interest with smallest possible order to improve the power.

The smooth test construction requires a set of functions $h_i(x;\beta)$ that are orthonormal. Subsequently it will be shown that using the orthonormal polynomials often results in convenient and interpretable components [20].

2.3 Interpretation of Components

One of the reasons that explain the popularity of smooth tests is the interpretability of the components when orthonormal polynomials are used for their construction. The question is whether or not the components $\hat{V}_r$ of the smooth test statistic can be used to identify the cause of the failure of the null hypothesis. The interpretation of significant components is important for both formal hypothesis testing and informal data analysis.

$\hat{V}_r^2$ is considered to be the score test statistic for testing $H_r : \theta_r = 0$ VS $K_r : \theta_r \neq 0$. And it gives a random sample from the model $C(\theta_r, \beta)\ \exp\{\theta_r h_r(x,\beta)\}f(x;\beta)$. This indicates that $\hat{V}_r^2$ is a detector of θ_r. The significant large $\hat{V}_r^2$ not only indicates the unacceptability of the model $f(x;\beta)$ but also, it diagnoses the reason of the failure in that model. This failure occurs due at least to a departure of the data from the hypothesized probability density function in θ_r. Unfortunately this argument fails because it is routine to demonstrate that for virtually any $f(x;\beta)$ a given $\hat{V}_r^2$ has some power to detect parameters other than θ_r. Nevertheless a significant $\hat{V}_r^2$ gives an indication of what might be called an order r deviation from the null model [21].

When testing for distributions from exponential families if the score functions are linearly related to the orthonormal functions, then maximum likelihood estimation and method of moments estimation coincide. In these cases for the smooth tests, estimation of the nuisance parameters is via method of moments estimation, and, recalling that β is $q \times 1$, the data agrees with the hypothesized distribution in moments up to the q^{th}. In general, the first significant component beyond the q^{th} would appear to be diagnostic, as all moments of the data and the hypothesized distribution up to and including the q^{th} are consistent. However, even this conclusion is suspect: higher order moments could still be the cause.

All of these issues would be resolved if, in the score test statistics, the asymptotic covariance matrix was replaced by one that estimated the component variances and covariances consistently, under both the null and alternative hypotheses. Henze and Klar [9, 8] worked in this vein, in which the smooth testing problem is treated in a semiparametric framework.

Unfortunately the simulation studies show that convergence of these 'rescaled' components to their asymptotic limits is extremely slow, with samples as large as 10,000 required to achieve satisfactory results. Thus, rescaling does not create diagnostic components and inference. Nevertheless, although in the finite samples that occur in practice the rescaled components may be somewhat 'tainted' by higher order moments, it is reasonable to say they are more diagnostic than the raw, unscaled components.

In the power studies, powers based on the rescaled components are often less than those based on the unscaled components, and where power gains are achieved it is often at the expense of power loss for alternatives elsewhere in the parameter space[20].

2.4 Generalized Smooth Tests

A smooth model's score test statistic is a quadratic form in the components. If the null hypothesis specifies a probability density function from an exponential family of distributions, the score test statistic $\hat{S_k}$ often has the appealing form of being a sum of squares of components that are asymptotically independent and asymptotically standard normal under the null hypothesis.

When the distribution is not from the exponential families, this convenient form cannot be expected. Distributions like the zero-inflated Poisson, extreme-value, negative binomial, and generalized Pareto distributions are examples of that. As the components are not even asymptotically uncorrelated, the significance of one component may be associated with the significance of others. And this is the difficulty for these distributions. Generally, it is not helpful to use, for example, the Gram-Schmidt transformation to diagonalize the asymptotic covariance matrix; the interpretation of those components would be usually impossible.

An alternative approach is to use the generalized score test can be applied when the score test does not produce a test statistic that is a sum of squares of components. The generalized smooth tests use the generalized score tests, with the $q \times 1$ nuisance parameter β being estimated by solving $V_r = 0$, for $r = 1, \dots, q$. The solution $\hat{\beta}_0$ is a method of moments estimator. These estimators are not usually fully efficient, and their use often means estimating efficiency is sacrificed to gain interpretable components.

A generalized score test statistic is of the form $\tilde{V}^T\tilde{\Sigma}^{-1}\tilde{V}$, where $\tilde{\Sigma}$ is a consistent estimator of the asymptotic covariance matrix of the score $\tilde{V}$. A ***Cholesky decomposition*** of $\tilde{\Sigma}^{-1}$ gives $\tilde{\Sigma}^{-1} = MM^T$, where M is upper triangular. Putting $\tilde{V}^* = M^T\tilde{\Sigma} = (\tilde{V}^*_{q+1}, \dots, \tilde{V}^*_{q+k})^T$ gives $\tilde{V}^T\tilde{\Sigma}^{-1}\tilde{V} = \tilde{V}^T\tilde{V}$. Thus, the generalized score test statistic is of the form $(\tilde{V}^*_{q+1}, \dots, \tilde{V}^*_{q+k})^2$. As for most distributions of interest a multivariate central limit theorem applies and $\tilde{V}^*$ has asymptotic covariance matrix the identity, the elements $\tilde{V}^*_r$ of $\tilde{V}^*$ are asymptotically independent and asymptotically standard normal [20].

As M^T is lower triangular, $\tilde{V}^*_r$ is the sum of the first r elements of $\tilde{V}^*$. It follows that the discussion about whether or not the components $\tilde{V}^*_r$ in the smooth test are diagnostic applies equally to the components $\tilde{V}^*_r$ in the generalized smooth test.

The conclusion is that when testing for any distribution, the generalized score test and Cholesky decomposition together yield components that are equally as convenient as those resulting from the score test when testing for distributions from exponential families. Moreover, because of their construction using the Cholesky decomposition, a significant Cholesky component $\tilde{V}^*_r$ suggests the data and the model agree in moments up to the r^{th}, although the significance may be due to moments up to the $2r^{th}$.

The only caveat on this approach is that the orthonormal polynomial of order r requires the existence of the first $2r$ moments of the null distribution.

Thus, for example, using this approach it is not possible to directly test for the Cauchy, which has no moments of any order. One option would be to use the probability integral transformation to essentially test for the continuous uniform distribution on (0, 1). However, the resulting components are then difficult to interpret in terms of the original null hypothesis.

2.5 Data-Driven Smooth Test

The determination of the order is one of the difficulties associated with the alternative of order k defined previously. Even if there are higher order inconsistencies, the hypothesized distribution may be suitable for most purposes. This can be reflected roughly in the first four moments if the data should correspond with the hypothesized distribution in four parameters.

If an order four smooth test for normality accepts the null hypothesis there may be some non-normality that a higher order test may have detected. However, given the known robustness of the analysis of variance, whatever slight non-normality there may be in data is almost certainly of no consequence.

In the case of finite samples, the order k can have a significant impact on the test's power. The test will provide effective protection against a wide variety of alternatives if k is set to a large value. However, as k increases, the ability to identify any particular alternative in the parameter space will be decreased. Contrariwise, if k is chosen to be small, the test will be fairly directional, providing strong protection against a small set of alternatives but none against the rest. As a result, it's essential to understand which alternatives are the most crucial.

Inglot, Kallenberg, & Ledwina (1997) [10] let the data make the decision about the order k. If L_k is the likelihood of a random sample of size n from a distribution that is an alternative of order k it would be natural to maximize L_k by choice of k in some set, say $1, 2, \ldots, d$, where d is specified before sighting the data. Here d is the maximum order one is prepared to accept. However, this procedure would simply choose $k = d$. A penalty term is needed to discourage complexity. The Bayesian information criterion (BIC)

$$BIC_K = -2\text{log } L_k + k \text{ log } n$$

is proposed as a model selection rule to determine the order. The optimal order, say K, is taken to be the smallest order that maximizes BIC_k. The test statistic is then chosen to be the sum of squares of the first K components, as previously defined. As the order is no longer a predetermined constant but a random variable, the test statistic $\hat{S}_k$ is no longer asymptotically χ^2 distributed. However, asymptotically the selected order converges to one in probability under the null hypothesis, and thus $\hat{S}_k$ asymptotically has a χ^2_1 null distribution. The procedure of Ledwina and co-workers allows the maximal order to depend on the sample size n. When d increases with n at a certain rate their construction makes the data-driven test omnibus consistent against every fixed alternative. As the convergence is rather slow, critical values and p-values are best determined using re-sampling methods.

These data-driven smooth tests are powerful competitors. It would be absurd to expect them to have more power than all given order smooth tests since they safeguard against mis-specifying the order.

It is recognized that it is computationally easier to work with a modified BIC_k, $\hat{S}_K + k$ log n.

An alternative penalty term is Aikaike's information criterion (AIC)

$$AIC_K = -2\log\, L_k + 2k$$

or its modified form $\hat{S}_K + 2k$. BIC_k penalizes complex models more heavily than AIC_k. The point is that many different model selection rules are possible. In different circumstances different rules will be appropriate. The approach of Ledwina and co-workers requires technical proofs for each selection criterion, and the rate of convergence of the maximal order also depends on the criterion, the class of distributions and the orthonormal functions. A more flexible method, and one of wider applicability, but one which does not result in omnibus consistent tests, is done by Inglot and Ledwina [11]. They fix the maximal order, and they also consider subset selection. This means that for a given maximal order, say d, the model selection criterion can select any subset of indexes from $1, \ldots, d$ and the data-driven test statistic is then built from the corresponding components V_j with j in the selected index set.

2.6 Model selection

Suppose now that the order k of a smooth alternative has been determined, and a smooth or generalized smooth test of this order applied. The moment interpretation of the components of these tests may not provide helpful insight. The insignificant $\hat{\theta}_r$ in the order k alternative could be replaced by zero, significant $\hat{\theta}_r$ by their method of moments estimators $\hat{\theta}_r$, and β by its method of moments estimator under the null hypothesis, $\tilde{\beta}_0$. This 'plug-in' estimator seems intuitively reasonable, but there are better options.

Two approaches were described by Rayner and co-workers [21]. Model selection through hypothesis testing and model selection using model selection criteria. To describe these, first suppose $S_h = \{1, 2, \ldots, d\}$ is an index set called the horizon. Consider Neyman smooth models of the form

$$g_k^N(x;\theta,\beta) = C(\theta_S,\beta)exp\left\{\sum_{i\in S}\theta_i h_i(x;\beta)\right\}f(x;\beta),$$

in which $S \subset S_h, \theta_S = \{\theta_i : i \in S\}$ and in which, as before, $C(\theta_S;\beta)$ is a normalizing constant and $h_i(x;\beta)$ is a set of functions orthonormal on $f(x;\beta)$. Clearly $g_{S_h}^N(x;\theta_{S_h},\beta)$ is the most complex model that is prepared to accept. A horizon of four clearly limits the maximal order more than 1 of 44 does. The probability density functions here are non-negative but require the determination of $C(\theta_S;\beta)$, if indeed it exists. In iterative work it is far more convenient to work with the Barton smooth models

$$g_k^N(x;\theta,\beta) = \left\{1 + \sum_{i\in S}\theta_i h_i(x;\beta)\right\}f(x;\beta),$$

Working with Barton models, and adjusting for non-negativity using the methods proposed is the favorite approach.

The first approach to be described is similar to the familiar forward selection and backward elimination techniques used in regression analysis. In forward selection at the u^{th} step the model is $g_S^B(x; \theta_S, \beta)$ and we consider whether or not to add a single θ_i term to the model, where $\theta_i \in \theta_{S_u}$, for every possible θ_i.

A slightly modified score test is derived to test each of these hypotheses. If any are significant at a predetermined level, then the θ_i corresponding to the most significant test is added to the model. Backward elimination is similar, with the least significant θ_i being eliminated from successively reduced models until only significant terms remain in the model. The score test statistics change at each iteration, and each requires the consistent variance estimate mentioned in the previous section [20].

Chapter 3

The stGOF package overview

The **stGOF** package provides functions for performing smooth tests of goodness of fit, and for computing orthonormal polynomials. It also provides functions for calculating the maximum likelihood estimates **MLE** and method of moments estimates **MME** of distributions. The top-level function **stGOF** can perform the smooth test with fixed order and the data-driven smooth test that are described in *Rayner et al. (2009).*

The latest (under development) version of the **stGOF** package is also available and can be installed in **R(>=3.1)** from the github repository of the project as follows:

```
  #Package devtools must be installed
    install.packages("devtools")
    devtools::install_github("krakla/stGOF")
    library(stGOF)
```

The function is structured around one top-level function, **stGOF** and some other functions like the function which is used in calculating the test statistic, **test_stat**, and the **orth_poly** function which generates polynomials that are orthonormal to the density function of a specified distribution. Some distributions are indexed by nuisance parameters which may be estimated from sample observations. Both maximum likelihood estimation and method of moments estimation are implemented.

3.1 stGOF function

The **stGOF** function performs the smooth test for the one-sample goodness-of-fit problem as described by *Rayner et al.(2009)*. Both simple and composite null hypotheses can be tested. The maximum likelihood (**MLE**) and the method of moments (**MME**) methods for nuisance parameter estimation are implemented. In addition, the function can be applied in case of fixed order or in the case of the data-driven version. The function is considered to be as the following:

```
stGOF <- function(formula, data = NA, order = NULL, method = "MLE",
                  rescale = FALSE, B = NULL, max.order = 0,
                  horizon = "", criterion = "", output = TRUE)
```

There are 10 arguments in **stGOF**. Four arguments must to be provided:

- **formula**: A formula is used to specify the data vector and the hypothesized distribution. And it should be written as **data** $\sim$ **distr**, where **distr** the distribution to be tested, and must be one of "**unif**", "**pois**" "**exp**", "**norm**", or "**logis**".

- **data**: A numeric vector of sample observations.
- **method**: Indicates the method for parameter estimation and must be one of "MLE" (default), or "MME".
- **output**: Logical item; if **TRUE** (default) an extensive output of the smooth test is given, otherwise, no output is given.

Other arguments are required depending on the case in which the smooth test is to be carried out. For instance, in the case of performing the smooth test with the Henze and Klar rescaling of the components (bootstrap version with fixed order), then three more arguments should be exist. And they are:

- **order**: The order of the test
- **rescale**: If **rescale**=TRUE the empirical variance is used to rescale the components with the Henze and Klar rescaling method.
- B: The number of bootstrap runs/iterations for p-value calculation

If **rescale=FALSE** (default), then the theoretical variance under the null hypothesis is used. This is case of the regular smooth test with a fixed order. If B=NULL then the asymptotic chi-squared distribution is used for p-value calculation.

If the order is not specified, the data-driven smooth test is applied. And other arguments are needed here like:

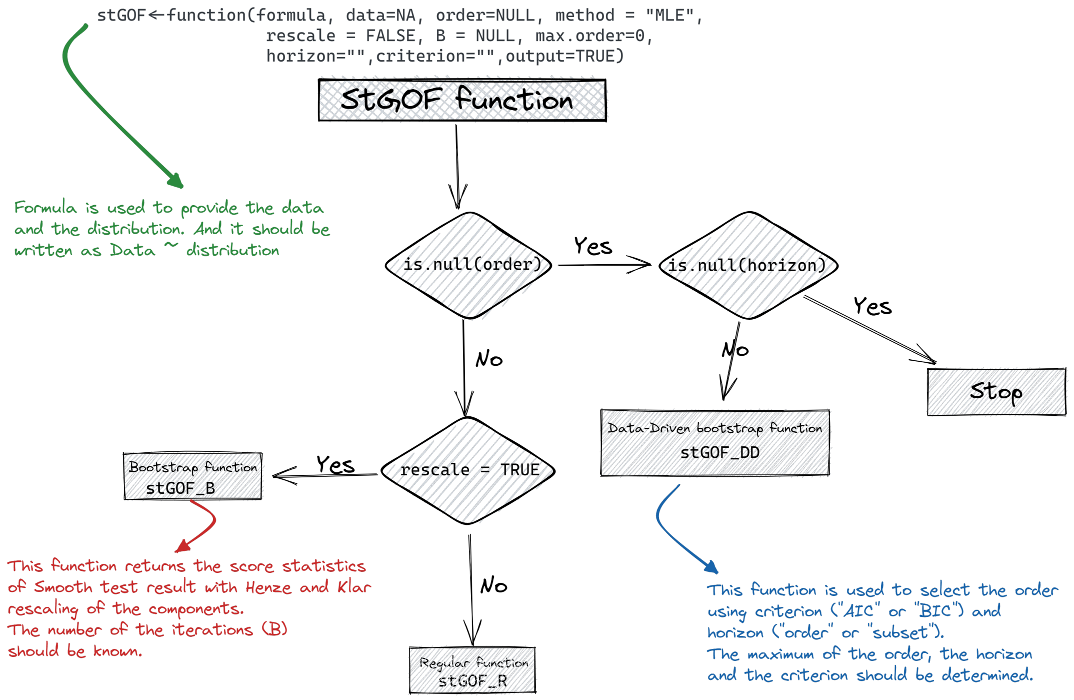
- **max.order**: The maximum order of the test
- **horizon**: **horizon**="order" (default) refers to order-selection, and **horizon**="subset" refers to subset selection.
- **criterion**: A character specifying the model selection criterion: "AIC" (default), or "BIC".

The workflow diagram in Figure 3.1 of the **stGOF** function describes how the function works. The first step is to check if the order is fixed or not. If the order is not fixed, then the data-driven function **stGOF_DD** will be run. Otherwise, if the order is fixed, there are two options. To perform the regular smooth test which means the theoretical variance under the null hypothesis is used. Or rescale the components by using the empirical variance as described in Henze and Klar rescaling method [9, 8].

For more understanding of the running sequence of the functions under stGOF function work, we present the workflow of the regular **stGOF** function as the example in Figure 3.2. Suppose the funcation is runing to perform the smooth test for **PRG** data with a fixed order = 4 and **"MLE"** as the estimation method of the parameters. Then the function will be as the following:

```
stGOF(PRG ~ unif, PRG, order = 4, method = "MLE")
```

The workflow schema shows that the regular **stGOF_R** function will start to embed under the top-level function **stGOF**. Then the **unif_MLE** function runs to return the **MLE** estimators for the uniform distribution. These estimators will be pluged-in the **orth_poly** function which contains all the sub-functions that calculate the orthonormal polynomial for the chosen distribution by calling its function (Here, it calls the **unif_orth**). After calculating the orthonormal polynomial, we use it in the **test_stat** to return the results of test statistics.



```
stGOF<-function(formula, data=NA, order=NULL, method = "MLE",
                        rescale = FALSE, B = NULL, max.order=0,
                        horizon="",criterion="",output=TRUE)
```

Formula is used to provide the data and the distribution. And it should be written as Data ~ distribution

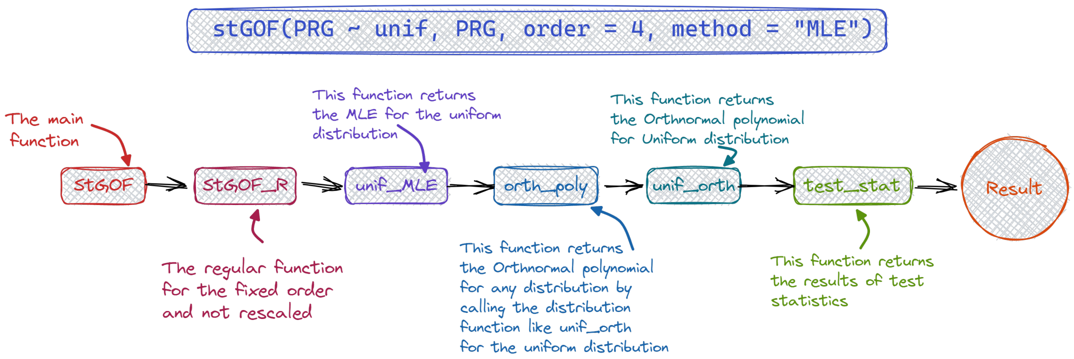
This function returns the score statistics of Smooth test result with Henze and Klar rescaling of the components.
The number of the iterations (B) should be known.

This function is used to select the order using criterion ("AIC" or "BIC") and horizon ("order" or "subset").
The maximum of the order, the horizon and the criterion should be determined.

Figure 3.1: **`stGOF` function workflow**

This scheme is in order to generalize the top-level function and implement the methods in a generic way. And that means there is no need to program the smooth test from scratch for a new distribution when we aim for further extending of the R package with other distributions. The only functions that will be added or modified are the functions related to the suggested distribution like the estimation function and the orthonormal polynomial functions.

In addition, there are no major changes to the function if the components are rescaled using the empirical variance or if the data-driven approach is applied. Only one additional function runs before the statistic calculation function to return the bootstrap results.



```
stGOF(PRG ~ unif, PRG, order = 4, method = "MLE")
```

The main function

This function returns the MLE for the uniform distribution

This function returns the Orthnormal polynomial for Uniform distribution

The regular function for the fixed order and not rescaled

This function returns the Orthnormal polynomial for any distribution by calling the distribution function like unif_orth for the uniform distribution

This function returns the results of test statistics

Figure 3.2: Regular **`stGOF` function workflow**

3.2 data

Besides the top-level function **stGOF**, the package contains four databases that will be utilized later to build, develop, and validate functions in this package. To explore the list of the names of the datasets in the **stGOF** package

```
#names of data sets in the stGOF package
(data(package = "stGOF"))$results[, "Item"]
```

```
[1] "cultivars" "PCB"       "PRG"       "pulse"
```

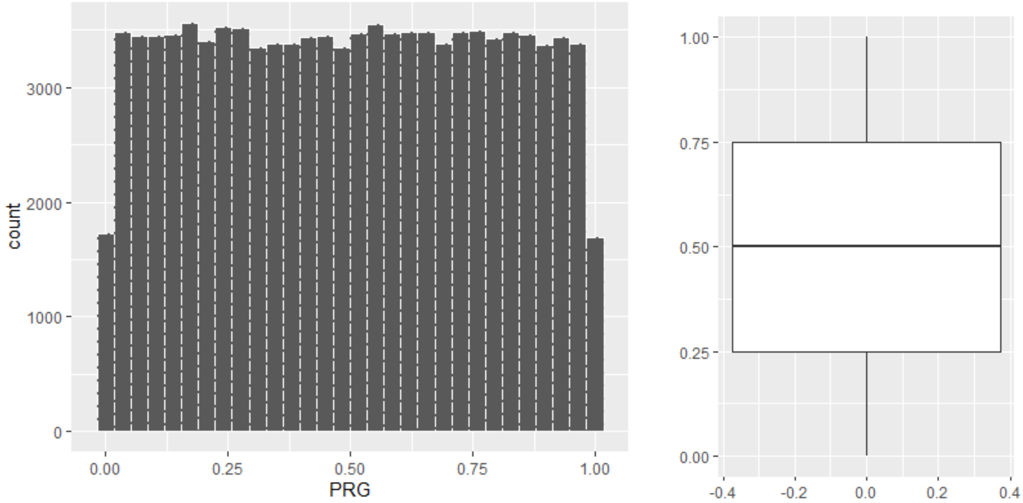
3.2.1 Pseudo-Random Generator Data - PRG

The generation of random numbers is important in many areas. For instance, in modern cryptographic algorithms 'good' random numbers are needed. Good random number generators are also essential in many sciences, *e.g.*, in physics and, of course, in statistics, where it is common practice today to assess empirically the validity of theoretical distribution theory by means of a simulation experiment in which statistics are calculated on repeatedly generated random samples from a given distribution. A device that generates true random numbers is hard to achieve. A true random generator is, for instance, based on a radioactive source, but it is unrealistic to have this built into every computer. Therefore, computer scientists, mathematicians, and engineers have created algorithms that generate pseudo-random numbers. These algorithms are based on a sound mathematical theory, and despite their deterministic nature they generate sequences of numbers that come close to true random number sequences. Apart from having as much randomness in the sequence as possible, pseudo-random generators 'sample' the numbers from a particular distribution. Often this is the uniform distribution over [0, 1]. Whenever a new pseudo-random generator is developed, it should be tested. Using the terminology of Knuth (1969) [13], two types of tests exist: theoretical and empirical tests. The former are based on algorithmic properties and their application does not need to let the algorithm generate sequences of pseudo-random numbers. The result of the test is a score of the randomness. The empirical tests, on the other hand, are basically statistical goodness-of-fit tests that should be applied to a generated sequence. These tests are used to test the null hypothesis that the generated numbers are indeed sampled from a uniform distribution over [0, 1] [2]. Atkinson (1980) is a reference in the statistical literature describing the problem. A nice reference in the computer science literature in which goodness-of-fit tests are applied to several pseudo-random generators, is Entacher and Leeb(1995) [5].

As an example we examine the quality of the uniform pseudo-random generator, the histogram and the boxplot in Figure 3.3 present the PRG dataset because it would be quite useless to list all 100,000 numbers in the dataset.

3.2.2 PCB Concentration Data

In a study on the effect of environmental pollutants on animals, Risebrough (1972) gives data on the concentration of several chemicals in the yolk lipids of pelican eggs. The data considered here are the PCB (polychlorinated biphenyl) concentrations for 65 Anacapa birds. In the original study the mean PCB concentration in Anacapa eggs was compared to the mean concentration in eggs of other birds. Here we concentrate on the Anacapa eggs [22]. A histogram and a boxplot are presented in Figure 3.4



(a) (b)

Figure 3.3: The histogram (left) and the boxplot (right) of the pseudo-random generator data

3.2.3 Pulse Rate Data

At a hospital the pulse rates of 50 patients were measured in beats per minute. The data are taken from Hand *et al.* (1994) [7]. Figure 3.5 shows the histogram and the boxplot of the pulse rate data.

3.2.4 Cultivars Data

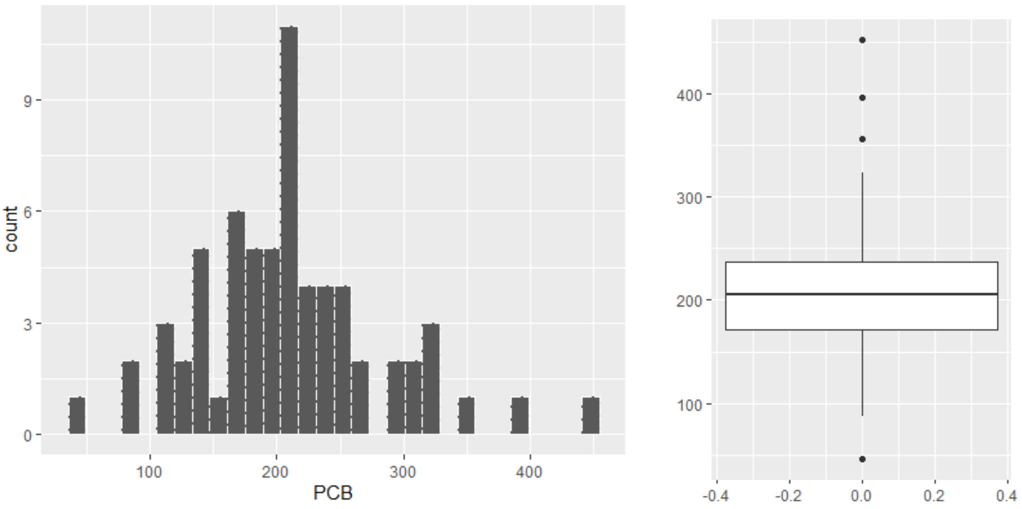
The cultivars dataset is taken from Karpenstein-Machen et al. (1994) and Karpenstein-Machan and Maschka (1996) [12]. It has also been analyzed by Piepho (2000) [17]. The dataset contains the yields (in tons per hectare) of two triticale cultivars: Alamo and Modus. Yields on both cultivars are obtained in 19 different environments. For each environment, a fertility score ("Ackerzahl" (AZ)) was recorded. A histogram and boxplot of the data are shown in Figure 3.6

3.3 Examples

In this section we test the functions of the **stGOF** package. In the next examples we test the composite null hypothesis that the PCB concentration data come from a normal distribution. We test this hypothesis first with a traditional smooth test based on the efficient scores. Because the normal distribution belongs to the exponential family, and **MME** and **MLE** coincide, it does not matter which $\sqrt{n}$-consistent estimation scheme we choose. The output below shows the **R**-code and the results of two smooth tests with fixed orders $k = 6$ and $k = 7$. All p-values are obtained from the asymptotic χ^2 approximation, but the results based on the simulated null distribution give the same conclusions.

```
stGOF(PCB ~ norm, PCB, order = 3, method = "MLE")
```

```
##  Results of the Smooth test
##  Ho: Normal against 3 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 210 72.26383  ( mean sd )
```



(a)

(b)

Figure 3.4: The histogram (left) and the boxplot (right) of the PCB data

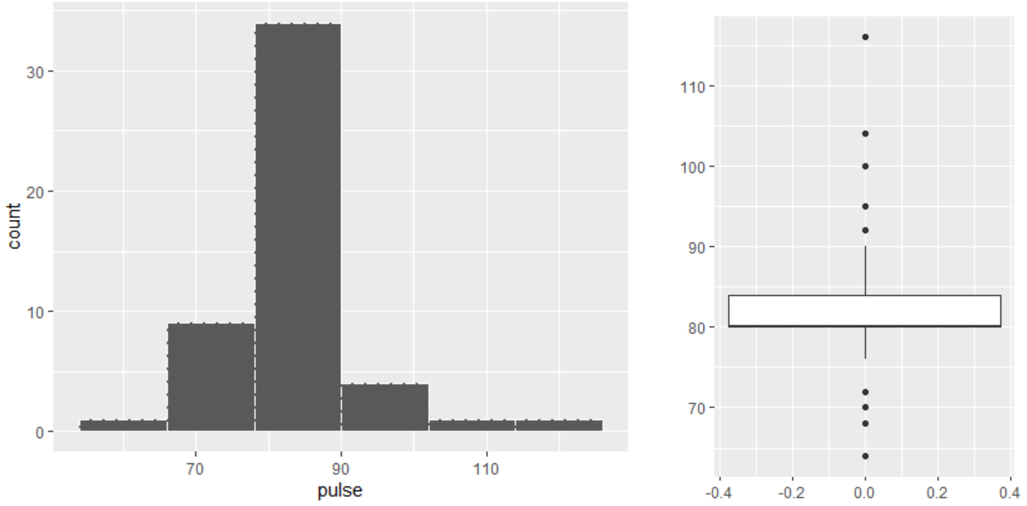
```
##
##  All p-values are obtained by the asymptotical chi-square approximation
##
##     Smooth test statistic S_k : 5.4369
##                p-value        : 0.01972
##
##  3 th component V_k = 2.33172  p-value = 0.01972
```

The function returns several outputs, such as the parameter estimates, the smooth test statistic, the p-value, and the component. As presented, the first two components are exactly zero because `MME` and `MLE` coincide. We read $p = 0.01972$ for $k = 3^{th}$-order smooth test, and conclude at the 5% level of significance that the null hypothesis of normality is rejected.

```
stGOF(PCB ~ norm, PCB, order = 6, method = "MLE")
```

```
##  Results of the Smooth test
##  Ho: Normal against 6 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 210 72.26383  ( mean sd )
##
##  All p-values are obtained by the asymptotical chi-square approximation
##
##     Smooth test statistic S_k : 10.1826
##                p-value        : 0.03746
##
##  3 th component V_k = 2.33172  p-value = 0.01972
##  4 th component V_k = 2.03024  p-value = 0.04233
##  5 th component V_k = 0.43434  p-value = 0.66404
##  6 th component V_k = -0.65966  p-value = 0.50947
```

The same conclusion when the order become $k = 6$. The null hypothesis of normality at the 5% level of significance is rejected. However, if $k = 7$ were chosen, as in the next example, then the smooth test would have p-value equal to 0.060 which does not imply the rejection of the null hypothesis at the 5% level of significance. The reason



(a) (b)

Figure 3.5: The histogram (left) and the boxplot (right) of the pulse rate data

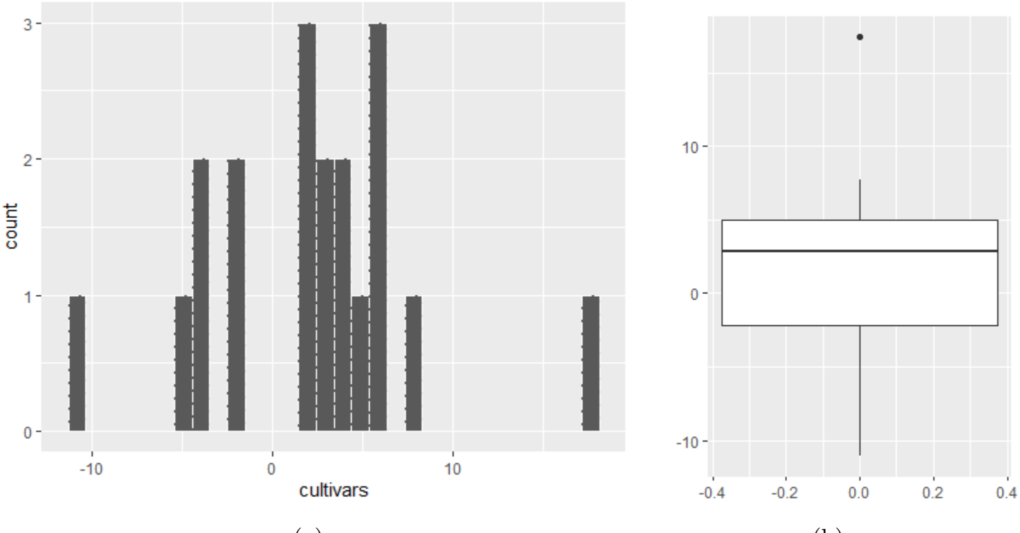
may be found by looking at the p-values of the individual component tests. The third- and the fourth-order component tests have small p-values, but as the order increases, the p-values increase too. This is a typical illustration of the ***dilution effect***.

```
stGOF(PCB ~ norm, PCB, order = 7, method = "MLE")
```

```
##  Results of the Smooth test
##  Ho: Normal against 7 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 210 72.26383  ( mean sd )
##
##  All p-values are obtained by the asymptotical chi-square approximation
##
##     Smooth test statistic S_k : 10.5948
##                p-value        : 0.06003
##
##  3 th component V_k = 2.33172  p-value = 0.01972
##  4 th component V_k = 2.03024  p-value = 0.04233
##  5 th component V_k = 0.43434  p-value = 0.66404
##  6 th component V_k = -0.65966  p-value = 0.50947
##  7 th component V_k = -0.642  p-value = 0.52087
```

We previously used the p-values of the individual component tests, but we argued extensively that the components should be rescaled to recover their full diagnostic property. The method of Henze and Klar has been explained in Section 2.3 to represent the nuisance parameters. The following code and output concern these rescaled component tests (using the `rescale=TRUE` option in the `stSOF` function by the bootstrap with 1000 runs).

The output shows that first, the p-value of order k smooth test equal 0.029. Then in the next lines the components are shown and the p-values. Whereas we previously concluded that the third and the fourth order component tests gave significant results. Now, we must conclude that they are not significant at the 5% level. This may look like a contradiction. There are two possible explanations.



(a)

(b)

Figure 3.6: The histogram (left) and the boxplot (right) of the cultivars data

```
set.seed(11)
  stGOF(PCB ~ norm, PCB, order = 6, method = "MLE", B = 1000,
      rescale = TRUE)
```

```
##  Results of the Smooth test
##  Ho: Normal against 6 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 210 72.26383  ( mean sd )
##
##  All p-values are obtained by the bootstrap with 1000 runs
##
##
##     Smooth test statistic S_k : 10.1826
##                p-value        : 0.029
##
##  3 th component V_k = 1.49321  p-value = 0.154
##  4 th component V_k = 1.21281  p-value = 0.088
##  5 th component V_k = 0.35025  p-value = 0.704
##  6 th component V_k = -0.97439  p-value = 0.31
```

The first is that the skewness and the kurtosis of the `PCB` concentration distribution agree with those of the normal distribution, and that it was falsely suggested by the non rescaled component tests due to an incorrect standardisation of the components. A second explanation might be that the use of the empirical variance estimator in the rescaled component test introduces additional variance, which further implies a loss in power. Thus maybe the large p-values of the rescaled component tests are a consequence of a smaller power. Which one of the two arguments is correct is still not clear at this point.

Another problem still left unanswered. Which analysis should be trusted, the smooth test with $k < 7$ or with $k = 7$? There is no preferable choice. The best answer is to let the data select the order k. The adaptive smooth test (data-driven approach) can be applied with `max.order=7`, `criterion = "BIC"` and the order selection rule `horizon="order"`

```
set.seed(11)
  stGOF(PCB ~ norm, PCB, method = "MLE", B = 1000, max.order = 7,
      horizon="order", criterion="BIC")
```

```
##  Results of the Smooth test
##  Result for Data-Driven Smooth goodness-of-fit test
##  Null hypothesis: norm against 7 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 210 72.26383  ( mean sd )
##
##      Horizon: order
##      Selection criterion: BIC
##
##  All p-values are obtained by the bootstrap with 1000 runs
##
##    Data-Driven Smooth test statistic S_k = 5.43692 p-value = 0.034
##     Selected model: 3
```

The *p*-value of this data-driven smooth test is 0.034. Based on this adaptive test we decide to reject the null hypothesis of normality at the 5% level of significance. The BIC criterion selected only the 3^{rd}-order term. Although the test statistic that was used here is not properly scaled to guarantee the diagnostic property, we may at least have trust in the overall conclusion: rejection of the null hypothesis of normality.

Finally, to illustrate the smooth test for a discrete distribution, we test the null hypothesis that the **pulse** rate data of Subsection 3.2.3 comes from a Poisson distribution. Note that for the Poisson distribution the **MLE** and **MME** coincide. The null hypothesis is tested by means of a smooth test of order $k = 6$, and the first component is exactly zero by the estimation process. The computation of the *p*-value the asymptotic χ^2 approximation is chosen.

```
stGOF(pulse ~ pois, pulse, order = 6, method = "MLE")
```

```
##  Results of the Smooth test
##  Results of the Smooth test
##  Ho: Poisson against 6 th order alternative
##  Parameter estimation method: MLE
##  Parameter estimates: 82.3  ( lambda )
##
##
##     All p-values are obtained by the asymptotical chi-square approximation
##
##     Smooth test statistic S_k : 20.9846
##                p-value        : 0.00082
##
##  2 th component V_k = -0.24605  p-value = 0.80564
##  3 th component V_k = 3.04171  p-value = 0.00235
##  4 th component V_k = 3.24207  p-value = 0.00119
##  5 th component V_k = 0.66246  p-value = 0.50768
##  6 th component V_k = -0.84982  p-value = 0.39543
```

From the output we read that the *p*-value of the order k smooth test equals $p = 0.0008 < 0.05$, and therefore we conclude at the 5% level of significance that the observations do not come from a Poisson distribution. A closer look at the individual components may shed some light on how the distribution differs from the Poisson distribution. Here the third- and the fourth-order components show very large values.

This suggests that the pulse rate distribution has a different skewness and a different kurtosis from a Poisson distribution with mean equal to 82.3.

3.4 Validation & Improvement

To achieve the best performance of the functions in the **stGOF**, the top-level function had to be redesigned several times. In addition, many conditions were modified or dropped besides improving the bootstrap part in the rescale and data-driven approaches. Also, reducing the time was taken into account in the process of improvement.

To validate the functions, each function was checked individually. As well as taking into account the examination of functions to work together in a consistent and compatible with the workflow of the package. The results mentioned in chapter 4 of the book *"Comparing Distributions"* by Thas (2010) [23] were adopted as a reference to compare with the results produced by the **stGOF** function. We found insignificant differences between the results in the book and the results produced by our function. And the reason behind that is the bootstrap functions in the book were without reference/seed.

Chapter 4

Conclusion

Smooth and generalized smooth goodness of fit tests are a good choice as tests with power that is nearly always comparable to alternative tests. `stGOF` is a package which is build, developed to perform the smooth test. This package contains functions for performing smooth tests of goodness of fit and for computing orthonormal polynomials. Smooth tests with a fixed order are performed as well as the data-driven versions. Structure of the smooth test, and the interpretation of the components were discussed in chapter 2. In addition, Data-driven approach was presented to explain the order selection.

A comprehensive explanation of all arguments of the top-level function with validated and compared practical examples. This work can be developed and improved by adding functions for other distributions to be test the null hypothesis. Also, it can be uploaded and published to be available on CRAN mirrors.

Bibliography

[1] John E Angus. "Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation". In: *Journal of Statistical Planning and Inference* 6.3 (1982), pp. 241–251.

[2] AC Atkinson. "Tests of pseudo-random numbers". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.2 (1980), pp. 164–171.

[3] Ralph B D'Agostino and Michael A Stephens. *Goodness-fo-fit-techniques*. Tech. rep. 1986.

[4] James Durbin. *Distribution theory for tests based on the sample distribution function*. SIAM, 1973.

[5] Karl Entacher and Hannes Leeb. "Inversive pseudorandom number generators: empirical results". In: *Proceedings of the Conference Parallel Numerics*. Vol. 95. 1995, pp. 15–27.

[6] Ian Hacking. "Trial by number; Karl Pearson's chi-square test...." In: *Science'84* 5 (1984), pp. 69–71.

[7] David J Hand et al. *A handbook of small data sets*. cRc Press, 1993.

[8] Norbert Henze. "Do components of smooth tests of fit have diagnostic properties?" In: *Metrika* 45.1 (1997), pp. 121–130.

[9] Norbert Henze and Bernhard Klar. "Properly rescaled components of smooth tests of fit are diagnostic". In: *Australian journal of statistics* 38.1 (1996), pp. 61–74.

[10] Tadeusz Inglot, Wilbert CM Kallenberg, and Teresa Ledwina. "Data driven smooth tests for composite hypotheses". In: *The Annals of Statistics* 25.3 (1997), pp. 1222–1250.

[11] Tadeusz Inglot and Teresa Ledwina. "Towards data driven selection of a penalty function for data driven Neyman tests". In: *Linear algebra and its applications* 417.1 (2006), pp. 124–133.

[12] M Karpenstein-Machan and R Maschka. "Investigations on yield structure and local adaptability of Triticale, Hybrid-Rye and Population-Rye based on data of regional variety trails". In: *Agribiological Research* 49 (1996), pp. 130–143.

[13] Donald E Knuth. *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional, 2014.

[14] A Maydeu-Olivares and C Garcı a-Forero. "Goodness-of-Fit Testing [PDF]". In: *Elsevier Ltd. Retrieved June* 1 (2010), p. 2016.

[15] Jerzy Neyman. "» Smooth test» for goodness of fit". In: *Scandinavian Actuarial Journal* 1937.3-4 (1937), pp. 149–199.

[16] Anthony N. Pettitt and Michael A. Stephens. "The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data". In: *Technometrics* 19.2 (1977), pp. 205–210.

[17] HP Piepho. "Exact confidence limits for covariate-dependent risk in cultivar trials". In: *Journal of agricultural, biological, and environmental statistics* (2000), pp. 202–213.

[18] JCW Rayner. "The asymptotically optimal tests". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 46.3 (1997), pp. 337–345.

[19] JCW Rayner and DJ Best. "Smooth tests of goodness of fit: an overview". In: *International Statistical Review/Revue Internationale de Statistique* (1990), pp. 9–17.

[20] JCW Rayner, Olivier Thas, and Donald John Best. "Smooth tests of goodness of fit". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 3.5 (2011), pp. 397–406.

[21] John CW Rayner, Olivier Thas, and Donald John Best. *Smooth tests of goodness of fit: using R*. John Wiley & Sons, 2009.

[22] Robert W Risebrough. "Effects of environmental pollutants upon animals other than man". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 6: Effects of Pollution on Health*. University of California Press. 1972, pp. 443–463.

[23] Olivier Thas. *Comparing distributions*. Vol. 233. Springer, 2010.