Distribution-free Analysis of Homogeneity

Dissertation

Max Wornowizki

TU Dortmund University

Faculty of Statistics

14.08.2015

Supervisor: Prof. Dr. Roland Fried

Contents

1 Introduction **4**

2 Demixing Empirical Distribution Functions **9**

2.1 Motivation 9

2.2 Problem Definition 12

2.3 Demixing Algorithm 18

2.3.1 Main Algorithm 18

2.3.2 Shrink-Down Algorithm 22

2.3.3 Shrink-Up Algorithm 22

2.3.4 Binary Search Algorithm 24

2.3.5 Normalisation Algorithm 25

2.4 Theoretical Analysis 28

2.4.1 Correctness of the Algorithm 28

2.4.2 Runtime Analysis 40

2.5 Practical Evaluation 41

2.5.1 Performance and Runtime 43

2.5.2 Estimated Shrinkage Factors under the Null Hypothesis . . . 47

2.5.3 Application to Astrophysical Data 48

2.5.4 Application to Bioinformatical Data 51

2.6 Related Methods 54

3 Two-sample Tests based on Divergences **58**

3.1 Divergence Measures 59

3.2 Divergence Estimation 61

3.2.1 Density Ratio Estimation 61
3.2.2 Divergence Estimation using Density Ratio Estimates 65
3.2.3 Comparison of the Divergence Estimators 68
3.3 Testing Homogeneity based on Divergences 71
3.3.1 Divergence-based Tests 71
3.3.2 Comparison of Divergence-based Tests 73
3.3.3 Application to Biometrical Data 79
3.4 Conclusions and Extensions 80

4 Testing Time Series for a Constant Volatility 84

4.1 Framework 84
4.2 Test Statistics for Volatility Changes 86
4.3 Testing for a Global Constant Volatility 90
4.3.1 Testing Procedure 90
4.3.2 Localisation of Presumable Structural Breaks 91
4.3.3 Alternative Methods 92
4.4 Comparison of Volatility Tests 95
4.4.1 Choice of Weighting Scheme and Block Sizes 95
4.4.2 Evaluation of Volatility Tests 97
4.4.3 Number and Position of Estimated Structural Breaks 99
4.4.4 Application to Financial Data 101
4.5 Extension to Structural Breaks in Kurtosis 102
4.6 Conclusions and Outlook 105

5 Summary 107

6 Tables **109**

References **117**

1 Introduction

Modern statistics provides countless tools for the investigation of various types of problems. Despite the myriad of new applications, the two main questions after the data acquisition remain the same: which assumptions on given observations are justifiable and which method relying on these assumptions is most suitable to study the points of interest? Homogeneity is a topic closely related to both of these questions. With regard to the first, homogeneity tests allow to compare several pieces of data checking the assumption of identical distributions. This way, either the dissimilarities of the observations are recognised and treated adequately or the whole data can be combined resulting in increased information. In both cases homogeneity tests are valuable tools to verify assumptions on the data and thereby facilitate an appropriate data analysis. With respect to the second question, homogeneity itself often lies at the heart of the analysis. This holds for example when one is looking for the best of several products, as it is the case with new commodities, medical treatments and also more abstract concepts like teaching methods. Homogeneity is also relevant in itself in the context of temporal data allowing to construct monitoring procedures and to check the effectiveness of conducted interventions such as new laws.

Since applying methods under wrong assumptions frequently leads to incorrect conclusions, it is of great importance to work with universal procedures to achieve reasonable results. Furthermore, highly multidimensional datasets containing quite different types of attributes are ubiquitous nowadays. It is thus rarely possible to state adequate distributional assumptions for each variable. For both these reasons fast distribution-free methods are highly desirable in the context of homogeneity. In this thesis three such procedures are presented. Each of them treats at a different homogeneity problem. The corresponding three chapters of this work are based

on the papers by Wornowizki and Munteanu (2015), Wornowizki and Fried (2014) and Wornowizki et al. (2015).

Chapter 2 is motivated by the fact that a mere rejection of homogeneity is unsatisfactory in many applications. To illustrate this we consider an arbitrary simulation procedure designed to imitate some observable data source. In other words, the simulation should generate artificial data resembling an observed sample. To check the quality of the simulation a statistician typically applies a homogeneity test. In case of rejection the simulation is inappropriate. Unfortunately, it is not clear then which particular data regions are modelled incorrectly. In order to improve the simulation efficiently it is of interest to automatically quantify the regions with too many or not enough observations in the artificial sample. In Chapter 2 an algorithm for this task is proposed. It is based on the classical distribution-free two-sample Kolmogorov-Smirnov test. The test is combined with a fairly general mixture model resulting in a highly flexible method. The algorithm determines a shrinkage factor and a correction distribution function. The first one measures how well the datasets resemble each other. The latter captures all discrepancies between them relevant in the sense of the Kolmogorov-Smirnov test. With regard to our illustrating example, the correction distribution indicates the deficiencies of the current simulation procedure and thus facilitates its improvement. The proposed procedure is illustrated using simulated as well as real datasets from astrophysics and bioinformatics and leads to intuitive results. We also prove its correctness and linear running time when applied to sorted samples. Since our approach is completely distribution-free and fast to compute, it is widely applicable and in particular suited for large multivariate datasets.

Up to now there is not much work on distribution-free density-based methods for testing homogeneity in the two-sample case. Chapter 3 is devoted to this topic. Classical two-sample test procedures such as the method investigated in the second

chapter often rely on distribution functions. Such functions can be estimated in a nonparametric way quite easily by their empirical counterparts, which is certainly one of their appealing properties. However, they cannot be interpreted as intuitively as probability density functions, which are in turn more difficult to estimate in a distribution-free setting. We focus on the concept of f-divergences introduced by Ali and Silvey (1966) in order to develop two-sample homogeneity tests. These distance like measures for pairs of distributions are defined via the corresponding probability density functions. Thus, homogeneity tests relying on f-divergences are not limited to discrepancies in location or scale, but can detect arbitrary types of alternatives. We propose a distribution-free estimation procedure for this class of measures in the case of continuous distributions. It is based on kernel density estimation and spline smoothing. As shown in extensive simulations, the new method performs stable and quite well in comparison to several existing non- and semiparametric divergence estimators. Furthermore, we construct two-sample homogeneity tests relying on various divergence estimators using the permutation principle. Just like for the new estimator, this approach does not require any assumptions on the underlying distributions and is therefore broadly applicable. The new tests are compared to an asymptotic divergence procedure as well as to several traditional parametric and nonparametric tests on data from different distributions under the null hypothesis and several alternatives. The results suggest that divergence-based methods have considerably higher power than traditional methods if the distributions do not primarily differ in location. Therefore, it is advisable to use such tests if changes in scale, skewness, kurtosis or the distribution type are possible while the means of the samples are of comparable size. The methods are thus of great value in many applications as illustrated on ion mobility spectrometry data.

In Chapter 4 we take a step further moving from two-sample problems to the

detection of structural breaks in time series. As in the previous chapters, the approach we prose is distribution-free. It is also highly flexible in two senses: the method can be applied in order to focus on arbitrary features of a time series such as the location, the scale or the skewness. In addition, any test statistic reflecting the feature under study can be incorporated. This is for example quite valuable, if outliers in the data are an issue. In such a case one can simply use the proposed test plugging in a suitable robust estimator. The method is based on a Fourier-type transformation of blockwise estimates of the quantity under study. The blockwise construction allows to handle multiple structural changes, which is often advantageous in real world applications. The Fourier-type transformation also related to characteristic functions leads to nice representations of the test statistics, which makes them easily computable. We introduce the approach testing the null hypothesis that a given sequence of variables has an unknown but constant volatility over time. Under the assumption of independent and piecewise identically distributed zero mean observations, several statistics for this testing problem are proposed. All of them are given in simple explicit formulas. Conducting extensive Monte Carlo experiments the new approach is compared to other tests for constant volatility. It shows a comparatively high power as well as an accurate localisation of the structural break positions in particular in the case of multiple volatility changes. The method also determines reasonable regimes of volatility on real exchange rate data. To illustrate the flexibility of our approach it is modified to test for a constant kurtosis. Its performance on artificial samples suggests that it behaves comparable to its volatility counterpart.

The three main chapters are structured in a similar way: at first the problem under study is motivated. Hereafter a new method solving the problem is introduced and its details are elaborated. Finally, it is evaluated using artificial as well as real data and the main conclusions are presented. The final chapter gives an

overview over the thesis and summarises its main results. The new algorithms and all alternative methods used for comparison are implemented using the statistical software R (R Development Core Team, 2013), version 2.15.1-gcc4.3.5. To run the data experiments in a batch and to distribute the computations to the cores the R package BatchExperiments by Bischl et al. (2013) is applied. The computations are conducted on a 3.00GHz Intel Xeon E5450 machine with 15 GB of available RAM running a SuSE EL 11 SP0 Linux distribution. All test are carried out at a nominal significance level of $\alpha = 0.05$, unless stated otherwise. The work is supported by the collaborative research centers SFB 823 - "Statistical modelling of nonlinear dynamic processes" and SFB 876 - "Providing Information by Resource-Constrained Data Analysis".

2 Demixing Empirical Distribution Functions

In this chapter a new statistical method for the comparison of two samples is presented. The algorithm provides detailed information on the dissimilarities of the datasets and extends the classical Kolmogorov-Smirnov test, cf. (Durbin, 1973). Our aim is motivated in Section 2.1. In Section 2.2 we formalise the setting and propose a general mixture model for the two-sample problem. Hereafter, several desirable properties of the unknown quantities of the model are established. On this basis two optimisation problems allowing to determine them are formulated. An algorithm solving these problems is proposed in Section 2.3. We hereby give detailed explanations for the main method and each subalgorithm. The proofs of the algorithm's correctness and linear running time are conducted in Section 2.4. In Section 2.5 the performance of the procedure is illustrated on real and simulated data examples. Section 2.6 concludes the chapter providing an overview on existing methods for related problems. In particular, we consider alternative procedures based on probability density functions. This part of the thesis has been published before in *Computational Statistics* by Wornowizki and Munteanu (2015). Besides giving the basic ideas, I contributed substantially to all parts of this chapter. My co-author greatly supported the development of the method and in particular proposed embedding the binary search technique in our algorithm.

2.1 Motivation

To introduce the method proposed in the following let us consider an example from astrophysics. The gamma ray detectors MAGIC-I and MAGIC-II are telescopes located at the Roque de los Muchachos on the Canary Island La Palma. For detailed

information on their structure and functionality the interested reader is referred to Cortina et al. (2009) and the MAGIC Collaboration (2014). The telescopes consist of a mirror surface of over 200 square metres each. It allows to measure atmospheric signals induced by the interaction of high energetic photons, called gamma rays, with the atmosphere. Gamma rays do not interact with magnetic fields, since they do not have an electric charge. They thus are able to carry valuable information about their sources in space far away from the detectors. The physicists exploring these sources are interested in gamma rays. They thus utilise the detectors to reconstruct the particles' trajectories, their energies and some related quantities. However, there are other particles generating somewhat similar atmospheric signals. For each gamma ray in the measurements there are about 1 000 observations of so called background events, which are not of interest in the given context. The background events mainly consist of protons, but also contain heavier hadrons and electrons. Classification algorithms relying on characteristics of the measured signals could be applied in order to distinguish between the background and the gamma particles. Unfortunately, these methods cannot be trained on real data, because it is not labelled. Therefore, simulation procedures for gamma rays as well as for protons based on models of particle propagation have been constructed and improved in several steps. The main software generating such simulations is CORSIKA (Heck et al., 1998).

Clearly, it is of major importance to compare simulated proton samples with actually observed data. On the one hand, suitable artificial background data is crucial for the classification analysis. Hence, variables with low agreement of generated background data and the sample must be identified, so that a purposeful improvement of the simulation is possible. On the other hand, small deviations between the simulations and the real data can be caused by gamma ray signals. If one assumes to have a reasonable simulation, variables with comparably high

discrepancies can be quite helpful for the upcoming classification task.

A typical statistical approach to check the similarity of the observed and the simulated data is the application of a homogeneity test. Note that since the datasets include a large number of variables of various types, a distribution-free procedure like for example the two-sample Kolmogorov-Smirnov test must be used. However, a mere rejection of the null hypothesis is not satisfying in our situation. If the simulation is highly inadequate, the data analyst wants to quantify the issues. In other words, the regions with too many or not enough observations in the artificial sample must be identified. Such information can then be used to update CORSIKA using more suitable simulation parameters. It even may give rise to the inclusion of additional simulation steps. If the discrepancies between the samples potentially stem from gamma ray signals, their quantification is necessary as well. It allows to assess and validate the gamma ray simulations in a subsequent step of the analysis.

In this chapter we present a novel approach allowing to gain additional insight into the discrepancies between two samples. It provides useful information for improving simulation procedures and is illustrated from this point of view in the following. The fast algorithm is applicable for small and large datasets. It is however mainly designed for the latter case, since often large amounts of multivariate simulated data are generated. Note that our contribution helps to improve an existing simulation procedure, which is often based on prior domain specific knowledge. We therefore assume that such a simulation procedure exists a priori.

We work with a mixture model linking the distributions of the observed and the simulated samples by a third distribution. The latter is called correction distribution. It represents all discrepancies between the first two distributions and can therefore be used to correct the simulation. Our algorithm determines an empirical distribution function corresponding to this correction distribution along

with a mixing proportion for the mixture model. Both are computed such that the resulting mixture of the simulation and the correction resembles the observed data in the sense of the Kolmogorov-Smirnov distance. The algorithm does not aim at statistical testing during or after the modification of the simulated sample. Thus, the corresponding type I and type II errors are not investigated. The method rather utilises quantiles of the Kolmogorov-Smirnov distribution to obtain intuitive bounds on the distance between empirical distribution functions. The algorithm does not construct a mixture fitting the observed data perfectly, but leads to a reasonably close approximation taking the sample variance into account. The amount of closeness can be regulated by the critical value c_α or equivalently by the significance level α and may be adjusted for a given application. For the sake of brevity, we illustrate the problem focussing on the improvement of a simulation procedure in the following. The method can also be applied to characterise subgroups in the data. For example, the correction distribution provides an approximation to the distribution of the gamma ray signals assuming that the background simulation is correct.

2.2 Problem Definition

In this section the basic notations for this chapter are introduced. We then suggest a general mixture model for the two-sample problem under study. Within this model all deviations between the distributions of the observed and the simulated data are represented by a correction distribution. In order to identify these discrepancies the correction distribution must be determined. For this purpose, the model is transferred to an empirical equivalent. To calculate the unknown quantities of the empirical model we motivate several constraints to it.

Let $x_1, \ldots, x_n \in \mathbb{R}$ denote the observed sample stemming from an unknown continuous distribution P with probability density function p and distribution function F. The underlying data generating process is modelled by a simulation procedure represented by the distribution Q. The corresponding probability density function and distribution function are denoted by q and G, respectively. To evaluate the quality of the simulation m simulated observations $y_1, \ldots, y_m$ are independently drawn from Q. If the simulation procedure works well, G resembles F so that the samples are similar.

To check the equality of P and Q a statistician typically applies a homogeneity test such as the classical two-sample Kolmogorov-Smirnov test, see Durbin (1973). Denote the empirical distribution functions of the samples by F_e and G_e, respectively, and set $\mathcal{N} = \frac{n \cdot m}{n+m}$. Choosing $M = \mathbb{R}$ the null hypothesis $\mathbb{H}_0 : P = Q$ is rejected by the two-sample Kolmogorov-Smirnov test, if the statistic

$$KS_M(F_e, G_e) = \sqrt{\mathcal{N}} \sup_{x \in M} |F_e(x) - G_e(x)|$$

exceeds an appropriately chosen critical value c_α. It is also possible to consider this procedure from a different perspective. Define an upper boundary function U setting $U(x) = \min(1, F_e(x) + \frac{c_\alpha}{\sqrt{\mathcal{N}}})$ for all $x \in \mathbb{R}$. In analogy, define a lower boundary function L by $L(x) = \max(0, F_e(x) - \frac{c_\alpha}{\sqrt{\mathcal{N}}})$ for all $x \in \mathbb{R}$. Using these notations the Kolmogorov-Smirnov test does not reject $\mathbb{H}_0$ if and only if G_e is an element of the set

$$B = \{f : \mathbb{R} \to [0,1] | \forall x \in M : L(x) \leq f(x) \leq U(x)\}$$

called the confidence band. We are interested in the regions of undersampling and oversampling, that is, the regions where G_e violates L or U.

In order to quantify the amount of such violations we work with the fairly general two-component mixture model

$$P = \tilde{s} \cdot Q + (1 - \tilde{s}) \cdot H. \tag{2.1}$$

The so-called mixture proportion or **shrinkage factor** $\tilde{s} \in [0, 1]$ measures the degree of agreement of P and Q. The **correction distribution** H represents all dissimilarities between P and Q. Since P is fully described by Q, $\tilde{s}$ and H, the latter two contain all information helpful for a modification of Q towards P. We thus want to determine them. Setting $\tilde{s} = 0$ and $H = P$ solves equation (2.1). However, this is not an appropriate solution in our application, because the data analyst is interested in correcting and not in discarding the current simulation. A modification of the current procedure, which is often motivated by expert knowledge, may give more insight into the data generating process itself and is thus preferable. For $\tilde{s} = 1$ the simulation is correct and H is irrelevant. However, for any $\tilde{s} \in (0, 1)$ the corresponding H is unique. In this case demixing P, that is, determining $\tilde{s}$ and H, provides useful information for an improvement of the simulation.

Unfortunately, the distributions P and Q are unknown in practice. We thus consider the corresponding empirical distribution functions F_e and G_e. They are consistent estimators for the true distribution functions F and G, which in turn entirely characterise P and Q. Combining the Kolmogorov-Smirnov approach with the mixture model (2.1), we propose to identify an (empirical) shrinkage factor $s \in (0, 1]$ and an (empirical) correction distribution function $\mathcal{H}$ such that the resulting mixture

$$\mathcal{F} = s \cdot G_e + (1 - s) \cdot \mathcal{H} \tag{2.2}$$

lies within the confidence band B around F_e. In other words, the corrected empirical distribution function $\mathcal{F}$ is regarded as close to F_e in the Kolmogorov-Smirnov sense. $\mathcal{H}$ is a distribution function and therefore must lie in the set

$$\mathcal{M} = \left\{ f : \mathbb{R} \to [0,1] \;\middle|\; f \text{ mon. nondecreasing step function}, \lim_{x \to -\infty} f(x) = 0 \right\}.$$

This is a superset of the set of all distribution functions on $\mathbb{R}$.

Since lying in B does not completely determine the structure of $\mathcal{F}$, neither s nor $\mathcal{H}$ are unique up to now. We thus introduce additional constraints on them allowing to determine reasonable solutions. Before addressing this point we propose another simplification of the problem. Since we work with empirical distribution functions, all derived quantities are characterized by their values on the joint sample $x_1, \ldots, x_n, y_1, \ldots, y_m$. Therefore, it is not necessary to consider all functions $\mathcal{H} \in \mathcal{M}$. Instead, we restrict ourselves to those functions which are discontinuous only on $Z = \{z_1, \ldots, z_{n+m}\}$, where $z_1, \ldots, z_{n+m}$ is the ordered joint sample. We denote this set of functions by $\mathcal{M}^* \subset \mathcal{M}$. This restriction is not very strong regardless of the sample sizes. For the Kolmogorov-Smirnov distance it does not make a difference, whether we add observations at the values in Z or at intermediate values. In addition, the position of such an intermediate value would be arbitrary between two of the given data points with respect to our distance. We thus focus on the original observations, thereby also avoiding additional computational costs. Keep in mind that the sample sizes in the applications we aim for is often comparably large. In these cases the restriction to Z is particularly weak, because the observations cover the relevant data regions quite well.

One helpful constraint on the model can be deduced from the fact that the data analyst wants to change the current simulation as little as possible. With regard to our model this means that s should be chosen maximal such that the corrected

distribution $\mathcal{F}$ fits the observed data. This directly implies a minimal weight $(1-s)$ for the correction function $\mathcal{H}$. We thus formulate **Problem 1**:

$$\begin{aligned} \max_{s\in[0,1]} : \quad & s \\ s.t. : \quad & \exists \mathcal{H} \in \mathcal{M}^* : \; s \cdot G_e + (1-s) \cdot \mathcal{H} \in B \end{aligned}$$

Since this maximum is unique, the shrinkage factor is now identifiable. Note that for $0 < s^* = \frac{c_\alpha}{\sqrt{N}}$ and $\mathcal{H}^* = \frac{1}{1-s^*} \cdot L$ the property $s^* \cdot G_e + (1-s^*) \cdot \mathcal{H}^* \in B$ holds. Thus, the sought-after value of s, called s_{opt} in the following, is larger than 0. Hence, the simulated data is always included in the mixture.

After Problem 1 is solved the resulting mixture $\mathcal{F} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}$ lies in B. Since this does not imply the property $\lim\limits_{x\to\infty} \mathcal{F}(x) = 1$, the function $\mathcal{H}$ could be an improper distribution function. Therefore, there might exist several choices of $\mathcal{H}$ solving Problem 1 given s_{opt}. To find a reasonable $\mathcal{H}$ we define the related function $\mathcal{H}_{min}$ via $\mathcal{H}_{min}(z) = \min \mathcal{H}(z) \; \forall z \in Z$. Hereby, the minimisation is taken over the set of all functions $\mathcal{H} \in \mathcal{M}^*$ satisfying $s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H} \in B$. The function $\mathcal{H}_{min}$ is clearly unique. We propose to first identify and then enlarge $\mathcal{H}_{min}$ in a meaningful way, see Section 2.3.5. This means that we construct a distribution function $\mathcal{H}_{opt}$ such that $\mathcal{H}_{opt}(z) \geq \mathcal{H}_{min}(z)$ for all $z \in Z$ and the corresponding mixture $\mathcal{F}_{opt} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{opt}$ is a proper empirical distribution function lying in B.

In most cases $\mathcal{H}_{min}$ should not be enlarged for small $z \in Z$. Due to the restriction on s in Problem 1 there often exists a point in Z, where $\mathcal{F}_{min} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{min}$ intersects the upper boundary U. Enlarging $\mathcal{H}_{min}$ before or in such a point $z_{meq} = \max \{z \in Z | \mathcal{F}_{min}(z) = U(z)\}$ leads to violations of U in z_{meq}. In case of such an intersection the Kolmogorov-Smirnov distance on $M = \mathbb{R}$ between the final mixture and F_e is just the radius of the confidence band for any meaningful

enlargement of $\mathcal{H}_{min}$. On subsets of $\mathbb{R}$, however, the distance can be improved if $\mathcal{H}_{min}$ is enlarged appropriately. Hence, we propose to identify z_{norm}, the smallest value after z_{meq} such that adding mass after z_{norm} minimises the Kolmogorov-Smirnov distance restricted to the set $M_{norm} = \{z \in Z | z \geq z_{norm}\}$. We then add the probability mass in such a way that the minimal distance $KS_{M_{norm}}$ is attained. If there is no intersection between $\mathcal{F}_{min}$ and U, we set $z_{meq} = \min(Z)$ and proceed in the same way. In total, finding a suitable distribution function $\mathcal{H}$ for a given value of s_{opt} can be formalised in **Problem 2**:

$$\begin{aligned} \min_{\mathcal{H} \in \mathcal{M}^*} : \quad & KS_{M_{norm}}(\mathcal{F}, F_e) \\ s.t. : \quad & \mathcal{F} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H} \in B \\ & \mathcal{H} \geq \mathcal{H}_{min} \\ & \lim_{x \to \infty} \mathcal{H}(x) = 1 \end{aligned}$$

An optimal solution to Problem 2 is called $\mathcal{H}_{opt}$. The corresponding final mixture is denoted by

$$\mathcal{F}_{opt} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{opt}. \tag{2.3}$$

Note that $\mathcal{F}_{opt}$ is not unique even with these constraints. Although the shrinkage factor s_{opt} is unique by its maximality property, there may exist several optimal enlargements of $\mathcal{H}_{min}$ equally appropriate in the sense of the restricted Kolmogorov-Smirnov distance.

2.3 Demixing Algorithm

In this section we present an algorithm solving Problems 1 and 2 formulated in Section 2.2. At first the main procedure is described. All subsequent subroutines called within the main algorithm are explained in more detail hereafter. In order to illustrate the algorithm and its subroutines pseudo code is provided.

2.3.1 Main Algorithm

Algorithm 1 on page 20 is our main procedure to solve Problems 1 and 2. It requires two sorted sample vectors $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ and a significance level α. At first it calculates the empirical distribution functions F_e and G_e of the samples and determines the critical value c_α at level α. In fact, c_α is the α-quantile of the distribution of $C = \sup_{t\in[0,1]} |\mathcal{B}(t)|$, where $\mathcal{B}$ is a Brownian bridge, cf. (Durbin, 1973). For the commonly used significance levels $\alpha_1 = 0.05$ and $\alpha_2 = 0.01$ the critical values are $c_{\alpha_1} = 1.358$ and $c_{\alpha_2} = 1.628$, respectively. The algorithm also initialises the values s and $\mathcal{F}$, candidates for the shrinkage factor s_{opt} and the final mixture $\mathcal{F}_{opt}$, and sets the lower bound for the binary search procedure, l_b, to the value of s^*, see page 16. The upper and lower boundary functions of the confidence band around F_e denoted by U and L, respectively, are computed next. These steps can be considered as preprocessing and are carried out in the lines 1 and 2. The two-sample Kolmogorov-Smirnov test does not reject the null hypothesis of equal distributions, if the relation $L \leq G_e \leq U$ holds. In this case, the empirical distribution functions resemble each other and the algorithm stops in line 4.
If the test rejects the null hypothesis, G_e does not completely lie within the confidence band. The algorithm thus carries out certain steps to determine an optimal mixture within the confidence band. To solve Problem 1 the following

operations are applied iteratively in the main loop in lines 5 to 11: if a candidate mixture $\mathcal{F}$ lies above the upper boundary U for any observation $z \in Z$, it has to be multiplied by a factor $s_d \in (0, 1)$ in order to correct the violation of U. This problem is addressed in line 7 in the so called Shrink-Down algorithm. The corresponding correction is illustrated in the upper row of Figure 1.

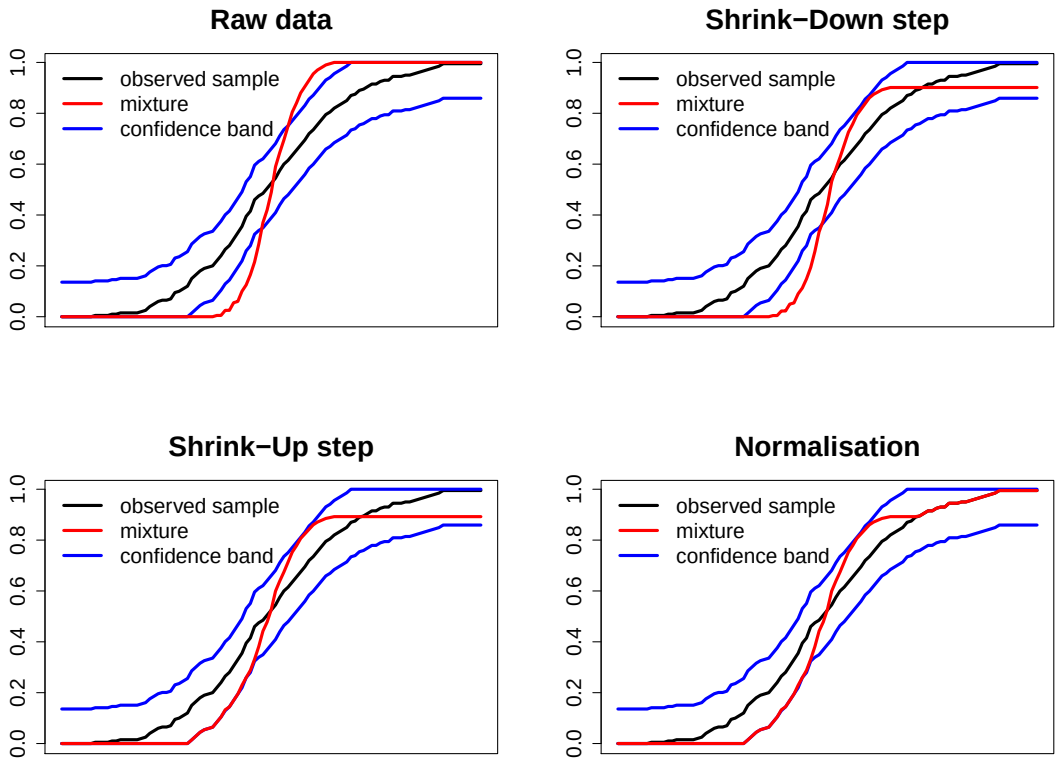

Figure 1: Empirical distribution functions describing the initial data situation (top left), the correction via the Shrink-Down (top right) and subsequent Shrink-Up step (bottom left) as well as the final normalisation (bottom right).

Due to the maximal property of the optimal shrinkage factor stated in Problem 1, the mixture candidate intersects U after a Shrink-Down step.

Algorithm 1: Demixing

Input : Sorted observations $x \in \mathbb{R}^n, y \in \mathbb{R}^m$, significance level α
Output : Optimal shrinkage factor s_{opt},
optimal correcting function $\mathcal{H}_{opt} \in \mathcal{M}^*$

$Z \leftarrow \text{sort}((x, y))$; $c_\alpha \leftarrow c(\alpha)$; $\mathcal{N} \leftarrow \frac{n \cdot m}{n+m}$; $l_b \leftarrow \frac{c_\alpha}{\sqrt{\mathcal{N}}}$; $s \leftarrow 1$;
$F_e \leftarrow \text{EmpDistrFun}(x)$; $G_e \leftarrow \text{EmpDistrFun}(y)$; $\mathcal{F} \leftarrow G_e$;
$L \leftarrow \max\left\{0, G_e - \frac{c_\alpha}{\sqrt{\mathcal{N}}}\right\}$;$U \leftarrow \min\left\{1, G_e + \frac{c_\alpha}{\sqrt{\mathcal{N}}}\right\}$;
if $\forall z \in Z : L(z) \leq \mathcal{F}(z) \leq U(z)$ **then**
  return $(s, 0)$
repeat
  if $\exists z \in Z : \mathcal{F}(z) > U(z)$ **then**
    $(s, \mathcal{F}) \leftarrow$ **Shrink-Down**$(s, \mathcal{F})$;
  if $\exists z \in Z : \mathcal{F}(z) < L(z)$ **then**
    $(s, \mathcal{F}) \leftarrow$ **Shrink-Up**$(s, \mathcal{F})$;
  $(l_b, s, \mathcal{F}) \leftarrow$ **BinSearch**$(l_b, s, \mathcal{F})$;
until $\forall z \in Z : L(z) \leq \mathcal{F}(z) \leq U(z)$;
$\mathcal{F} \leftarrow$ **Normalise**$(\mathcal{F})$;
$\mathcal{H} \leftarrow (\mathcal{F} - s \cdot G_e)/(1 - s)$;
return $(s, \mathcal{H})$;

As shown in the upper row of Figure 1, the Shrink-Down algorithm eliminates violations of U, but can create or enhance violations of the lower boundary L. A candidate falling below L must receive additional probability mass in appropriate regions. This is taken achieved in line 9 by applying the Shrink-Up algorithm. Its effect is presented in the lower left part of Figure 1. The two shrink steps are conducted whenever necessary in the presented order. Since they have opposite effects, some data situations require multiple executions of the Shrink-Down and the Shrink-Up step. Their iteration generates a decreasing sequence of upper bounds to s_{opt}. To guarantee the solution of Problem 1 we embed the well-known binary search technique in our demixing algorithm, cf. (Cormen et al., 1990). It is applied in line 10 and bounds s_{opt} from below and above. The method is connected with the Shrink-Down and Shrink-Up step by using the current shrinkage factor s learned from them as an upper bound to s_{opt}. In return, the binary search updates s and $\mathcal{F}$, which are then passed back to the Shrink-Down and Shrink-Up steps. The lower bound for the optimal shrinkage factor, l_b, is updated by the binary search procedure itself.

Once the main loop is terminated, the optimal shrinkage factor s_{opt} and the corresponding minimal correction function $\mathcal{H}_{min}$ introduced on page 16 are determined. Thus Problem 1 is solved and the current candidate lies within the confidence band. The normalisation step in line 12 solves Problem 2 returning an optimal mixture $\mathcal{F}_{opt}$ depicted in the lower right part of Figure 1. Hereafter, $\mathcal{H}_{opt}$ is identified rearranging equation (2.3) in line 13. Finally, it is returned together with the optimal shrinkage factor s_{opt}.

2.3.2 Shrink-Down Algorithm

This procedure is applied whenever a candidate $\mathcal{F}$ exceeds the upper boundary U at some point $z \in Z$. This problem can be solved intuitively exploiting the mixture model (2.2). One simply computes the maximal shrinkage value $s_d \in (0, 1)$ such that $s_d \cdot \mathcal{F}$ does not violate U any more. In other words, $\mathcal{F}$ is shrunk down. The maximal shrinkage factor achieving this is $s_d = \min_{z \in Z} \left\{ \frac{U(z)}{\mathcal{F}(z)} \right\}$, where we set $\frac{v}{0} = \infty$ for any $v > 0$. The Shrink-Down subroutine presented in Algorithm 2 calculates this factor in line 1. It then updates the total shrinkage and the candidate mixture $\mathcal{F}$ accordingly and returns them. The effect of the Shrink-Down step is visualised in the upper row of Figure 1.

Algorithm 2: Shrink-Down

Input : Current mixture $\mathcal{F}$ and shrinkage factor s
Output : Updated mixture $\mathcal{F}$ and shrinkage factor s

1 $s_d \leftarrow \min\limits_{z \in Z} \left\{ \frac{U(z)}{\mathcal{F}(z)} \right\}$;
2 $s \leftarrow s_d \cdot s$;
3 $\mathcal{F} \leftarrow s_d \cdot \mathcal{F}$;
4 **return** $(s, \mathcal{F})$;

2.3.3 Shrink-Up Algorithm

The Shrink-Up step presented in Algorithm 3 is carried out whenever the current candidate mixture $\mathcal{F}$ violates L, the lower boundary of the confidence band. In order to increase the values of the mixture in the problematic regions probability mass is added. This is illustrated in the lower left part of Figure 1. Note that $\mathcal{F}$ may lie below L before $z_{eq} = \min\limits_{z \in Z} \{z | U(z) = \mathcal{F}(z)\}$ as well as after that point. However, these two cases have a crucial difference. Adding probability mass before z_{eq} leads to a new violation of the upper boundary U in z_{eq}. Adding mass after

z_{eq} does not have to imply this problem. To distinguish between these cases the Shrink-Up algorithm first identifies z_{eq} in line 1. The value of z_{eq} is well-defined after initialisation with $\mathcal{F} = G_e$, because $\mathcal{F}(\max(Z)) = G_e(\max(Z)) = 1 = U(\max(Z))$ holds. As we show in Lemma 4, z_{eq} is also well-defined after modifications of $\mathcal{F}$.

Algorithm 3: Shrink-Up

Input : Current mixture $\mathcal{F}$ and shrinkage factor s
Output : Updated mixture $\mathcal{F}$ and shrinkage factor s

1 $z_{eq} \leftarrow \min_{z \in Z}\{z \,|\, U(z) = \mathcal{F}(z)\}$;
2 **if** $\exists z < z_{eq} : \mathcal{F}(z) < L(z)$ **then**
3 $\quad s_u \leftarrow \min_{z < z_{eq}} \left\{ \frac{\mathcal{F}(z_{eq}) - L(z)}{\mathcal{F}(z_{eq}) - \mathcal{F}(z)} \right\}$;
4 $\quad s \leftarrow s_u \cdot s$;
5 $\quad \mathcal{F} \leftarrow s_u \cdot \mathcal{F}$;
6 $\forall z \in Z : d(z) \leftarrow \max\{0, L(z) - \mathcal{F}(z)\}$;
7 $\forall z \in Z : \mathcal{H}(z) \leftarrow \max_{z' \leq z}\{\mathcal{F}(z') - s \cdot G_e(z') + d(z')\}$;
8 $\mathcal{F} \leftarrow s \cdot G_e + \mathcal{H}$;
9 **return** $(s, \mathcal{F})$;

If there are violations of L before z_{eq}, a shrinkage is necessary. Because of Problem 1, we have to shrink minimally. Thus, the largest shrinkage factor s_u must be identified, so that all residuals to L before z_{eq} do not exceed the residual to U in z_{eq} after shrinking. If this property does not hold, adding appropriate probability mass causes a violation of U in z_{eq}. More formally, the shrinkage factor

$$s_u \quad = \max_{s \in [0,1]} \quad \{s \mid \forall z < z_{eq} : L(z) - s \cdot \mathcal{F}(z) \leq U(z_{eq}) - s \cdot \mathcal{F}(z_{eq})\}$$

must be determined. Using basic arithmetic transformations of the constraint and $\mathcal{F}(z_{eq}) = U(z_{eq})$ we get $s_u = \min_{z < z_{eq}} \left\{ \frac{\mathcal{F}(z_{eq}) - L(z)}{\mathcal{F}(z_{eq}) - \mathcal{F}(z)} \right\}$. This value is determined in line 3 of the Shrink-Up algorithm and the shrinkage factor s as well as the candidate

mixture $\mathcal{F}$ are updated.

After the potential shrinkage the algorithm corrects $\mathcal{F}$ by adding probability mass. In order to shift $\mathcal{F}$ appropriately its nonnegative residuals to L, $d(z) = \max\{0, L(z) - \mathcal{F}(z)\}$, are computed for all $z \in Z$. These are the minimal amounts which must be added to $\mathcal{F}$ so that it no longer violates the lower boundary L. They are thus added to the current correction term $\mathcal{F} - s \cdot G_e$ and the sum is minimally monotonised in line 7. The result is added to $s \cdot G_e$ yielding the new candidate mixture $\mathcal{F}$. Note that, in contrast to equation (2.2), we use the notation $\mathcal{H}$ rather than $(1 - s) \cdot \mathcal{H}$ for the properly scaled correction function for the sake of brevity here. Analogical abbreviations are used in the pseudo code of Algorithms 4 and 5.

2.3.4 Binary Search Algorithm

The binary search step presented in Algorithm 4 is called at the end of every iteration in the main loop of Algorithm 1. Its input consists of l_b and u_b, the current lower and upper bound for s_{opt}, respectively. While l_b is derived from previous binary search steps, u_b is set to the current value of s. The algorithm computes the average of the given bounds in line 1. Using this candidate the minimum monotone step function $\mathcal{H}_b$ is determined such that $\mathcal{F}_b = s_b \cdot G_e + \mathcal{H}_b \geq L$ holds in lines 2 and 3. This is done in analogy to the lines 6 and 7 in the Shrink-Up step. If $\mathcal{F}_b$ violates the upper boundary U, then, by minimality of $\mathcal{H}_b$, no monotone step function for the shrinkage factor s_b can exist such that the corresponding mixture lies within the confidence band B. Therefore, as implied by the monotonicity property proved in Lemma 1, it holds that $s > s_b > s_{opt}$. In this case the algorithm updates s to s_b as the new upper bound for s_{opt} and sets the current mixture candidate to $\mathcal{F}_b$ in lines 6 and 7. Otherwise, again by Lemma 1, the relation $s_{opt} \geq s_b > l_b$ must hold, since there exists a monotone step function for the shrinkage factor s_b leading to a

mixture in B. Thus, s_b is a better lower bound to s_{opt} than l_b. In this case l_b is updated to s_b, while all other quantities are kept.

Algorithm 4: Binary Search

Input : l_b and u_b, current lower and upper bounds on s_{opt}
Output: Updated mixture $\mathcal{F}$, shrinkage factor s and lower bound l_b

1 $s_b \leftarrow (l_b + u_b)/2$;
2 $\forall z \in Z : d(z) \leftarrow \max\{0, L(z) - s_b \cdot G_e(z)\}$;
3 $\forall z \in Z : \mathcal{H}_b(z) \leftarrow \max_{z' \leq z}\{d(z')\}$;
4 $\mathcal{F}_b \leftarrow s_b \cdot G_e(z) + \mathcal{H}_b$;
5 **if** $\exists z \in Z : \mathcal{F}_b(z) > U(z)$ **then**
6 $\quad s \leftarrow s_b$;
7 $\quad \mathcal{F} \leftarrow \mathcal{F}_b$;
8 **else**
9 $\quad l_b \leftarrow s_b$;
10 **return** $(l_b, s, \mathcal{F})$;

2.3.5 Normalisation Algorithm

As shown in Theorem 1, Problem 1 is solved when the loop of Algorithm 1 (lines 5 to 11) stops. At this point the current value of s is the optimal shrinkage factor s_{opt}. The current mixture is $\mathcal{F} = s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{min}$ and lies within the confidence band. However, as pointed out in the description of Problem 2, $\mathcal{F}$ may not be a proper distribution function. It may hold that $\lim_{x \to \infty} \mathcal{F}(x) < 1$, as illustrated in the lower left part of Figure 1. This deficiency is corrected by the normalisation step presented in Algorithm 5.

To check whether $\mathcal{F}$ must be enlarged the algorithm computes z_{meq}, the maximal value $z \in Z$ where $\mathcal{F}(z)$ equals $U(z)$. If the last candidate mixture was proposed by the binary search, there may not exist an intersection of $\mathcal{F}$ and U. In this case the algorithm sets $z_{meq} = \min(Z)$. If $z_{meq} = \max(Z)$ is satisfied, the property $\mathcal{F}(\max(Z)) = \mathcal{F}(z_{meq}) = U(z_{meq}) = U(\max(Z)) = 1$ holds, so no further correc-

tions are necessary and $\mathcal{F}$ is returned. Otherwise, as stated in the motivation to Problem 2, adding probability mass before z_{meq} leads to a violation of U in z_{meq}. Since s_{opt} is already determined, such a violation cannot be repaired by further shrinking as in the Shrink-Up step. Thus, probability mass must be added after z_{meq}. In fact, the region where mass should be added can be restricted even further. This holds, because adding mass in regions where $\mathcal{F}$ is above F_e pushes the mixture further apart from F_e. We thus define z_{norm} as the smallest value in Z such that $z_{norm} > z_{meq}$ and $\mathcal{F}(z_{norm}) < F_e(z_{norm})$ holds and modify $\mathcal{F}$ on the set $M_{norm} = \{z \in Z | z \geq z_{norm}\}$ only.

Algorithm 5: Normalisation

Input : Current value of $\mathcal{F}$
Output : Final value of $\mathcal{F}$

1 **if** $\forall z \in Z : \mathcal{F}(z) < U(z)$ **then**
2 $\quad z_{meq} \leftarrow \min(Z)$;
3 **else**
4 $\quad z_{meq} \leftarrow \max_{z \in Z} \{z \,|\, U(z) = \mathcal{F}(z)\}$;
5 **if** $z_{meq} \neq \max(Z)$ **then**
6 $\quad z_{norm} \leftarrow \min_{z > z_{meq}} \{z \,|\, \mathcal{F}(z) < F_e(z)\}$;
7 $\quad \forall z \geq z_{norm} : d(z) \leftarrow \min\{F_e(z) - \mathcal{F}(z), 1 - \mathcal{F}(\max(Z))\}$;
8 $\quad$ **if** $\max_{z \geq z_{norm}} \{-d(z)\} \geq 1 - \mathcal{F}(\max(Z))$ **then**
9 $\quad\quad \tilde{z} \leftarrow \max_{z \geq z_{norm}} \{z \,|\, -d(z) \geq 1 - \mathcal{F}(\max(Z))\}$;
10 $\quad\quad z_{norm} \leftarrow \min_{z > \tilde{z}} \{z \,|\, d(z) > 0\}$;
11 $\quad \forall z \geq z_{norm} : \mathcal{H}_{norm}(z) \leftarrow$
$\quad \max\left\{0, \left(\max_{z_{norm} \leq z' \leq z} \{d(z')\} + \min_{z'' \geq z} \{d(z'')\}\right)/2\right\}$;
12 $\quad \forall z < z_{norm} : \mathcal{H}_{norm}(z) \leftarrow 0$;
13 $\quad \mathcal{F} \leftarrow \mathcal{F} + \mathcal{H}_{norm}$;
14 **return** $(\mathcal{F})$;

The residuals $d(z) = F_e(z) - \mathcal{F}(z)$ are computed for all $z \in M_{norm}$ in line 7. Residuals larger than the remaining mass $1 - \mathcal{F}(\max(Z))$ are decreased to this

value, because more probability mass is not available anyways. Hereafter, the algorithm compares two quantities. The first one is the maximal increase of $\mathcal{F}$ above F_e, the maximum of all negative residuals $-d(z), z \in M_{norm}$. The second one is the imposed maximal decrease of $\mathcal{F}$ below F_e, namely $1 - \mathcal{F}(\max(Z))$. As long as the first is greater or equal to the second one, adding probability mass does not decrease the Kolmogorov-Smirnov distance. Hence, in line 9 the algorithm determines the last position where this inequality holds. It then updates z_{norm} to be greater than this position. This yields an updated set $M_{norm} = \{z \in Z | z \geq z_{norm}\}$ where a reduction of the Kolmogorov-Smirnov distance is possible. At the latest, M_{norm} is the last region where $\mathcal{F}$ lies below F_e.

To determine an appropriate modification of the current candidate $\mathcal{F}$, the residuals d are considered on M_{norm}. Since a monotone function fitting the residuals in the Kolmogorov-Smirnov sense must be determined, we are dealing with an L_∞ isotonic regression problem. Unweighted isotonic regression problems under the L_∞-norm can be efficiently solved in linear time for sorted samples. This can be achieved by a simple approach, which is referred to as Basic by Stout (2012). The method is applied in line 11 of Algorithm 5. For each residual, it computes the maximum of all previous values and the minimum of all subsequent values. The regression value is then determined as the average of these two quantities.

Note that a solution to the isotonic regression problem may in general be negative for some $z \in M_{norm}$. The correction term $\mathcal{H}_{norm}$ however must be nonnegative to guarantee the monotonicity of the mixture and no violations of L. We resolve this issue proving that setting all negative values of $\mathcal{H}_{norm}$ to 0 results in an optimal solution to the isotonic regression problem constraint to nonnegativity, see Lemma 5. Since no correction is applied before z_{norm}, $\mathcal{H}_{norm}$ is set to 0 before z_{norm} in line 12. Finally, $\mathcal{F}$ is updated and returned. The resulting overall mixture is presented in the lower right part of Figure 1.

2.4 Theoretical Analysis

In this section theoretical results for the algorithms presented in Section 2.3 are provided. Among other things, we prove a monotonicity property allowing to apply the binary search technique to Problem 1. Also, the Shrink-Down and Shrink-Up steps are shown to lead to upper bounds on s_{opt}. While the first part of this section deals with the correctness of the demixing algorithm, the second one presents its runtime analysis.

First, let us introduce additional notations used repeatedly in our proofs. The shrinkage factor of G_e, the scaled correction function and the mixture candidate after the k-th iteration of the main loop of Algorithm 1 (lines 5 to 11) are denoted by s_k, $\mathcal{H}_k$ and $\mathcal{F}_k = s_k \cdot G_e + \mathcal{H}_k$, respectively. In order to initialise them, we set $s_0 = 1$, $\mathcal{H}_0 = 0$ and $\mathcal{F}_0 = G_e$. Let $s_{d,k}$ denote the multiplicative update of the shrinkage factor determined in the Shrink-Down step in the k-th iteration. If this update is not computed, we set $s_{d,k} = 1$. The update of the shrinkage factor determined in the Shrink-Up step of the k-th iteration is called $s_{u,k}$ and treated in the same way.

2.4.1 Correctness of the Algorithm

As we show in our first result, the property of lying within the confidence band is monotone in s. In other words, for any $s > s_{opt}$ a corresponding mixture must violate a boundary of B, while for every $s \leq s_{opt}$ it is always possible to find a mixture in B. This fact allows to prove the correctness of our binary search step.

Lemma 1. $(\exists \mathcal{H} \in \mathcal{M}^* : s \cdot G_e + (1-s) \cdot \mathcal{H} \in B) \Leftrightarrow s \in [0, s_{opt}]$.

Proof. First we recall the definition of Problem 1 from page 16:

$$\begin{aligned} \max_{s\in[0,1]} : \quad & s \\ s.t. : \quad & \exists \mathcal{H} \in \mathcal{M}^* : \forall z \in Z : \\ & L(z) \leq \ s \cdot G_e(z) + (1-s) \cdot \mathcal{H}(z) \leq U(z), \end{aligned} \tag{2.4}$$

where $\mathcal{M}^*$ denotes the set of all monotonically nondecreasing step functions discontinuous on Z only and converging to 0 as their argument goes to $-\infty$. We introduce an alternative characterisation of s_{opt} by **Problem A**:

$$\begin{aligned} \max_{s\in[0,1]} : \quad & s \\ s.t. : \quad & \forall z \in Z : s \cdot G_e(z) \leq U(z) && \text{(2.5a)} \\ & \forall z', z'' \in Z, z' < z'' : L(z') - s \cdot G_e(z') \leq U(z'') - s \cdot G_e(z'') && \text{(2.5b)} \end{aligned}$$

Before we proceed with proving the proposition, we show the equivalence of Problem 1 and Problem A. For this sake, choose an arbitrary $s \in [0,1]$ such that (2.4) holds. For all $z \in Z$ it follows that $s \cdot G_e(z) \leq U(z) - (1-s) \cdot \mathcal{H}(z) \leq U(z)$ by nonnegativity of $(1-s) \cdot \mathcal{H}$, which proves inequality (2.5a). Furthermore, choose $z' < z''$ from Z arbitrarily. Then $L(z') - s \cdot G_e(z') \leq (1-s) \cdot \mathcal{H}(z') \leq (1-s) \cdot \mathcal{H}(z'') \leq U(z'') - s \cdot G_e(z'')$ follows by monotonicity of $\mathcal{H}$. Thus, (2.5b) is also respected. For the other direction, let $s \in [0,1]$ respect constraints (2.5a) and (2.5b). From (2.5a) it is clear that $s \cdot G_e(z)$ never exceeds the upper boundary. From (2.5b) we know that correcting any deficiency to the lower boundary L is possible without violating the upper boundary U on subsequent positions. Choosing

$$(1-s) \cdot \mathcal{H}(z) = \max \left\{ 0, \max_{z^* \leq z} \{ L(z^*) - s \cdot G_e(z^*) \} \right\}$$

thus must result in a mixture within the confidence band so that (2.4) holds.
We now make use of the equivalence of Problem 1 and Problem A to prove the proposition:

$$(\exists \mathcal{H} \in \mathcal{M}^* : s \cdot G_e + (1-s) \cdot \mathcal{H} \in B) \Leftrightarrow s \in [0, s_{opt}].$$

For $s \in (s_{opt}, 1]$ the property $s \cdot G_e + (1-s) \cdot \mathcal{H} \notin B$ immediately follows by definition of s_{opt} for any $\mathcal{H} \in \mathcal{M}^*$. So let $s \in [0, s_{opt}]$ be arbitrarily chosen. By the equivalence of Problem 1 and Problem A the constraints (2.5a) and (2.5b) are respected for s_{opt}. From this we deduce that both conditions must also hold for s since $\forall z \in Z : s \cdot G_e(z) \leq s_{opt} \cdot G_e(z) \leq U(z)$ and furthermore for all $z', z'' \in Z$ with $z' < z''$ it follows

$$\begin{aligned}
L(z') - s \cdot G_e(z') &= L(z') - s_{opt} \cdot G_e(z') - (s - s_{opt}) \cdot G_e(z') \\
&\overset{(2.5b)}{\leq} U(z'') - s_{opt} \cdot G_e(z'') \underbrace{-(s - s_{opt})}_{\geq 0} \cdot G_e(z') \\
&\leq U(z'') - s_{opt} \cdot G_e(z'') - (s - s_{opt}) \cdot G_e(z'') \\
&= U(z'') - s \cdot G_e(z'').
\end{aligned}$$

Hence, $L(z') - s \cdot G_e(z') \leq U(z'') - s \cdot G_e(z'')$ holds. As shown before, constraints (2.5a) and (2.5b) are equivalent to constraint (2.4). Therefore, there exists an $\mathcal{H} \in \mathcal{M}^*$ so that $s \cdot G_e + (1-s) \cdot \mathcal{H} \in B$ holds. This completes the proof. □

In the next lemma the correction function $\mathcal{H}_k$ computed in the k-th iteration of the main loop is considered. As we prove, $\mathcal{H}_k$ is indeed the minimal function in $\mathcal{M}^*$ resolving violations of the lower boundary L. This result contributes to the correctness of our construction of $\mathcal{H}_{min}$ and is used in subsequent proofs.

Lemma 2. *$\mathcal{H}_k$ is the pointwise minimal function among all $\mathcal{H} \in \mathcal{M}^*$ satisfying $s_k \cdot G_e + \mathcal{H} \geq L$.*

Proof. Let $\mathcal{H}_{k,min} \in \mathcal{M}^*$ be the minimal function fulfilling $s_k \cdot G_e + \mathcal{H}_{k,min} \geq L$. To prove the claim we show $\mathcal{H}_k = \mathcal{H}_{k,min}$. The correction function $\mathcal{H}_k$ is either computed in the binary search or in the Shrink-Up step. In the first case, the residuals between $s_k \cdot G_e$ and the lower boundary L are determined and then minimally monotonised, cf. lines 2 and 3 of Algorithm 4. This monotonisation is performed considering the maximum of preceding values and is therefore minimal. Hence, this procedure must yield $\mathcal{H}_{k,min}$. In the remainder of this proof we thus treat the second case, namely the computation of $\mathcal{H}_k$ in the Shrink-Up step.
Following the lines 6 to 7 in Algorithm 3 on page 23, let us consider $d_k = \max(0, L - s_{d,k} \cdot s_{u,k} \cdot \mathcal{F}_{k-1})$. These are the positive deficiencies to L after potential shrinking of the last candidate $\mathcal{F}_{k-1}$ in the Shrink-Down and Shrink-Up step of iteration k. Setting $\tilde{\mathcal{F}}_k = s_{d,k} \cdot s_{u,k} \cdot \mathcal{F}_{k-1} + d_k$ the correction function $\mathcal{H}_k$ can be expressed as $\mathcal{H}_k = \text{mon}\left(\tilde{\mathcal{F}}_k - s_k \cdot G_e\right)$. Thereby, $\text{mon}(f)$ denotes the pointwise minimal monotone function such that $\text{mon}(f) \geq f$ for any function f. This monotonisation is performed analogically to the one in the binary search step by considering the maximum of preceding values. Note that the monotonising operator is itself monotone, that is, $\text{mon}(f_1) \leq \text{mon}(f_2)$ holds for two arbitrary functions f_1, f_2 such that $f_1 \leq f_2$. We show the proposition $\mathcal{H}_k = \mathcal{H}_{k,min}$ by induction:

Base case $k = 1$: By assumption $\mathcal{H}_1$ is computed in the Shrink-Up step, so $s_1 = s_{d,1} \cdot s_{u,1}$ holds. In addition, $\mathcal{F}_0$ is defined by $\mathcal{F}_0 = G_e$. Hence, $d_1 = \max(0, L - s_{d,1} \cdot s_{u,1} \cdot \mathcal{F}_0) = \max(0, L - s_1 \cdot G_e) \leq \mathcal{H}_{1,min}$ must hold, since the last inequality holds by definition of $\mathcal{H}_{1,min}$. Because of $\tilde{\mathcal{F}}_1 = s_1 \cdot G_e + d_1$ we

obtain $\mathcal{H}_1 = \text{mon}(\tilde{\mathcal{F}}_1 - s_1 \cdot G_e) = \text{mon}(d_1) \leq \text{mon}(\mathcal{H}_{1,min}) = \mathcal{H}_{1,min}$, where the inequality follows by the monotonicity of the monotonising operator. Thus $\mathcal{H}_1 \leq \mathcal{H}_{1,min}$ is established. To prove the other inequality, note that $\mathcal{H}_1 \in \mathcal{M}^*$ and $\mathcal{H}_1 = \text{mon}(d_1) \geq d_1$. Hence, $\mathcal{H}_{1,min} \leq \mathcal{H}_1$ follows by the definition of $\mathcal{H}_{1,min}$. Altogether, we get $\mathcal{H}_{1,min} = \mathcal{H}_1$.

Inductive step $k-1 \Rightarrow k$: The shrink updates $s_{d,k}$ and $s_{u,k}$ are bounded by 1 by construction and thus the inequality $s_k \leq s_{d,k} \cdot s_{u,k} \cdot s_{k-1} \leq s_{k-1}$ holds. Hence, the shrinkage factor s_k does not increase in k. From that we deduce that the corresponding minimal correction function $\mathcal{H}_{k,min}$ does not decrease in k. Consequently, we get $\mathcal{H}_{k,min} \geq \mathcal{H}_{k-1,min} \geq s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1,min}$. The correctness of the $(k-1)$-th step assumed by the induction principle yields $\mathcal{H}_{k-1,min} = \mathcal{H}_{k-1}$ resulting in $\mathcal{H}_{k,min} \geq s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1}$. Next, note that $s_{d,k} \cdot s_{u,k} \cdot \mathcal{F}_{k-1}$ can be rewritten to $s_k \cdot G_e + s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1}$. This allows us to interpret d_k as the minimal function which must be added to $s_k \cdot G_e + s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1}$ so that the lower boundary L of the confidence band is not violated any more. Together with $\mathcal{H}_{k,min} \geq s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1}$ established above this implies $s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1} + d_k \leq \mathcal{H}_{k,min}$. Since in addition d_k is by construction minimally chosen such that $\tilde{\mathcal{F}}_k = s_{d,k} \cdot s_{u,k} \cdot \mathcal{F}_{k-1} + d_k \geq L$ holds, we deduce

$$L - s_k \cdot G_e \leq \tilde{\mathcal{F}}_k - s_k \cdot G_e = s_{d,k} \cdot s_{u,k} \cdot \mathcal{H}_{k-1} + d_k \leq \mathcal{H}_{k,min}.$$

Applying the monotonising operator and exploiting its monotonicity this implies

$$\begin{aligned} L - s_k \cdot G_e \leq \;& \text{mon}\,(L - s_k \cdot G_e) \;\leq \underbrace{\text{mon}\left(\tilde{\mathcal{F}}_k - s_k \cdot G_e\right)}_{=\mathcal{H}_k} \\ \leq \;& \text{mon}\,(\mathcal{H}_{k,min}) \;= \mathcal{H}_{k,min}, \end{aligned} \tag{2.6}$$

and therefore $\mathcal{H}_k \leq \mathcal{H}_{k,min}$. To prove $\mathcal{H}_k \geq \mathcal{H}_{k,min}$ first note that $\mathcal{H}_k \in \mathcal{M}^*$. The inequalities (2.6) imply $L \leq s_k \cdot G_e + \mathcal{H}_k$. So, by definition of $\mathcal{H}_{k,min}$, $\mathcal{H}_k \geq \mathcal{H}_{k,min}$ follows. Thus, overall $\mathcal{H}_k = \mathcal{H}_{k,min}$ holds, which completes the proof. □

The next result shows that the Shrink-Down step always leads to an overall shrinkage factor s not lower than s_{opt}. Therefore, the updated value of s may be used as an improved upper bound for s_{opt} in the binary search procedure.

Lemma 3. *If $s_k > s_{opt}$ is fulfilled, then $s_{d,k+1} \cdot s_k \geq s_{opt}$ must hold.*

Proof. The proposition is trivial for $s_{d,k+1} = 1$ so in the following $s_{d,k+1} < 1$ is assumed. This means that the $(k+1)$-th Shrink-Down step is not skipped but executed. So $\mathcal{F}_k$ must lie above the upper boundary U for some values. Together with the definition of $s_{d,k+1}$ (page 22) this ensures the existence of a $z_{eq} \in Z$ such that $s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) = U(z_{eq})$ holds. In the following we consider the two possible cases for the correction function $\mathcal{H}_k = \mathcal{F}_k - s_k \cdot G_e$:

Case $\mathcal{H}_k(z_{eq}) = 0$: Using the definition of z_{eq} and $\mathcal{F}_k$, we deduce

$$\begin{aligned} U(z_{eq}) &= s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) \\ &= s_{d,k+1} \cdot (s_k \cdot G_e(z_{eq}) + \mathcal{H}_k(z_{eq})) \\ &= s_{d,k+1} \cdot s_k \cdot G_e(z_{eq}) \\ &< G_e(z_{eq}), \end{aligned}$$

where the last inequality follows since $0 < s_{d,k+1} < 1$, $0 < s_k \leq 1$ and $0 < G_e(z_{eq})$. The latter is satisfied, because otherwise $0 = G_e(z_{eq})$ would hold. In this case $\mathcal{H}_k(z_{eq}) = 0$ immediately implies $0 = U(z_{eq})$, which is a contradiction to the positivity of U. The calculations show that the function G_e lies above the upper boundary

U in z_{eq} before any shrinking. However, the first Shrink-Down step solves this problem. Because of $\mathcal{H}_k(z_{eq}) = 0$ there cannot be a new violation of U in z_{eq} in subsequent steps. Hence, $k = 0$ and consequently $s_k = 1$ must hold. The proposition $s_{d,1} = s_k \cdot s_{d,k+1} \geq s_{opt}$ holds in this case, since $s_{d,1}$ is by construction the maximal shrinkage factor avoiding violations of U before adding any correction function.

Case $\mathcal{H}_k(z_{eq}) > 0$: Let $\tilde{\mathcal{H}} \in \mathcal{M}^*$ be the minimal function one must add to $s_{d,k+1} \cdot s_k \cdot G_e$ in order to correct violations of the lower boundary L. Due to $s_{d,k+1} \leq 1$ we get $s_{d,k+1} \cdot s_k \cdot G_e \leq s_k \cdot G_e$. Thus $\tilde{\mathcal{H}} \geq \mathcal{H}_k$ holds by minimality of $\mathcal{H}_k$ shown in Lemma 2. Since by assumption $0 < s_{d,k+1} < 1$ holds, this allows to prove

$$\begin{aligned} U(z_{eq}) &= s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) \\ &= s_{d,k+1} \cdot (s_k \cdot G_e(z_{eq}) + \mathcal{H}_k(z_{eq})) \\ &< s_{d,k+1} \cdot s_k \cdot G_e(z_{eq}) + \mathcal{H}_k(z_{eq}) \\ &\leq s_{d,k+1} \cdot s_k \cdot G_e(z_{eq}) + \tilde{\mathcal{H}}(z_{eq}). \end{aligned}$$

Thus, $s_{d,k+1} \cdot s_k \cdot G_e + \tilde{\mathcal{H}}$ violates the upper boundary of the confidence band and thus does not lie in B. By minimality of $\tilde{\mathcal{H}}$ Lemma 1 yields $s_{d,k+1} \cdot s_k > s_{opt}$, which completes the proof. □

The following proposition concerns the additional shrinkage performed in the Shrink-Up step. Similarly to Lemma 3, it states that a Shrink-Up step cannot lead to shrinkage factors below s_{opt}. The lemma therefore allows to use the updated overall shrinkage factor s as an improved upper bound on s_{opt}.

Lemma 4. *If $s_{d,k+1} \cdot s_k > s_{opt}$ is fulfilled, then $s_{u,k+1} \cdot s_{d,k+1} \cdot s_k \geq s_{opt}$ must hold.*

Proof. The statement is obviously fulfilled for $s_{u,k+1} = 1$. It is also clear in case of $k = 0$ by construction of the shrink update $s_{u,1}$. So let $s_{u,k+1} < 1$ and $k \geq 1$ hold in the following. We prove the proposition by contradiction and thus assume

$$s_{d,k+1} \cdot s_k > s_{opt} > s_{d,k+1} \cdot s_{u,k+1} \cdot s_k. \tag{2.7}$$

Let us consider the preceding candidate mixture $\mathcal{F}_k$. $\mathcal{F}_k \notin B$ must hold, because otherwise the algorithm would have stopped after k iterations. Furthermore, $\mathcal{F}_k \geq L$ is guaranteed by construction of the Shrink-Up and binary search steps. Therefore, $\mathcal{F}_k$ must violate the upper boundary U in the assumed case $k \geq 1$. Thus, a Shrink-Down step was executed before the current Shrink-Up step. Hence, the point

$$z_{eq} = \min \{z \in Z | s_{d,k+1} \cdot \mathcal{F}_k(z) = U(z)\}$$

is well defined. The assumption $s_{u,k+1} < 1$ implies that a Shrink-Up step is carried out and $\exists z \in Z : z < z_{eq}$. By definition of z_{eq}, each $z < z_{eq}$ satisfies $s_{d,k+1} \cdot \mathcal{F}_k(z) < U(z) \leq U(z_{eq})$. From that we deduce

$$\forall z < z_{eq} : s_{d,k+1} \cdot \mathcal{F}_k(z) - U(z_{eq}) < 0. \tag{2.8}$$

Now consider the point

$$z' = \max \left\{ \underset{z<z_{eq}}{\operatorname{argmax}} (L(z) - s_{d,k+1} \cdot s_{u,k+1} \cdot \mathcal{F}_k(z)) \right\}.$$

By the definition of

$$s_{u,k+1} \quad = \max \quad \{s \in [0,1] | \forall z < z_{eq} : L(z) - s \cdot s_{d,k+1} \cdot \mathcal{F}_k(z) \leq U(z_{eq}) \cdot (1-s)\}$$

it follows that

$$L(z') - s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') = U(z_{eq}) \cdot (1 - s_{u,k+1}). \tag{2.9}$$

We also consider another point $z'' = \min\{\operatorname{argmax}_{z \le z_{eq}} \mathcal{H}_k(z)\}$. Using the minimal property of $\mathcal{H}_k$ proved in Lemma 2, for $k \ge 1$ one can deduce $\mathcal{F}_k(z'') = L(z'')$. This implies $z'' < z_{eq}$. Since $\mathcal{F}_k \ge L$ holds by construction of $\mathcal{F}_k$, for all $z \le z''$ we obtain

$$\begin{aligned} L(z) - s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z) &\le \mathcal{F}_k(z) - s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z) \\ &= (1 - s_{u,k+1} \cdot s_{d,k+1}) \cdot \mathcal{F}_k(z) \\ &\le (1 - s_{u,k+1} \cdot s_{d,k+1}) \cdot \mathcal{F}_k(z'') \\ &= \mathcal{F}_k(z'') - s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z'') \\ &= L(z'') - s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z''). \end{aligned}$$

Combining this result with $z'' < z_{eq}$ derived before we get $z' \ge z''$. Using the monotonicity of $\mathcal{H}_k$ and the definition of z'' we deduce

$$\mathcal{H}_k(z') = \mathcal{H}_k(z'') = \mathcal{H}_k(z_{eq}). \tag{2.10}$$

We now combine (2.7), (2.8), (2.9) and (2.10) to prove the proposition. By Lemma 3 and $s_{opt} \ge s^* > 0$ (page 16) the inequality $s_{d,k+1} \cdot s_k > 0$ holds. Thus, $s_{u2} = \frac{s_{opt}}{s_{d,k+1} \cdot s_k}$ is well defined. Inequality (2.7) implies

$$1 \ge s_{u2} > s_{u,k+1}. \tag{2.11}$$

This allows us to show

$$\begin{aligned}
& L(z') - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') \\
= \;& L(z') + (-s_{u,k+1} \cdot s_{d,k+1} + s_{u,k+1} \cdot s_{d,k+1} - s_{u2} \cdot s_{d,k+1}) \cdot \mathcal{F}_k(z') \\
\overset{(2.9)}{=} \;& U(z_{eq}) \cdot (1 - s_{u,k+1}) + s_{u,k+1} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') \\
= \;& U(z_{eq}) - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') + s_{u,k+1} \cdot \underbrace{(s_{d,k+1} \cdot \mathcal{F}_k(z') - U(z_{eq}))}_{< 0 \text{ by (2.8)}} \\
\overset{(2.11)}{>} \;& U(z_{eq}) - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') + s_{u_2} \cdot (s_{d,k+1} \cdot \mathcal{F}_k(z') - U(z_{eq})) \\
= \;& U(z_{eq}) \cdot (1 - s_{u2}).
\end{aligned}$$

We thus get

$$U(z_{eq}) \cdot (1 - s_{u2}) < L(z') - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z'). \tag{2.12}$$

We now use $s_{opt} \cdot G_e + \mathcal{H}_{opt} \geq L$, which holds by definition of $\mathcal{H}_{opt}$, to show

$$\begin{aligned}
U(z_{eq}) \quad = \quad & U(z_{eq}) + s_{u2} \cdot \underbrace{(s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) - U(z_{eq}))}_{= 0 \text{ by definition of } z_{eq}} \\
= \quad & U(z_{eq}) \cdot (1 - s_{u2}) + s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) \\
\overset{(2.12)}{<} \quad & L(z') - s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z') + s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) \\
= \quad & L(z') - s_{u2} \cdot s_{d,k+1} \cdot (s_k \cdot G_e(z') + \mathcal{H}_k(z')) + s_{u2} \cdot s_{d,k+1} \cdot \mathcal{F}_k(z_{eq}) \\
= \quad & L(z') - \underbrace{s_{u2} \cdot s_{d,k+1} \cdot s_k}_{= s_{opt} \text{ by definition of } s_{u2}} \cdot G_e(z') + s_{u2} \cdot s_{d,k+1} \cdot (\mathcal{F}_k(z_{eq}) - \mathcal{H}_k(z')) \\
\leq \quad & \mathcal{H}_{opt}(z') + s_{u2} \cdot s_{d,k+1} \cdot (s_k \cdot G_e(z_{eq}) + \underbrace{\mathcal{H}_k(z_{eq}) - \mathcal{H}_k(z'))}_{= 0 \text{ by (2.10)}} \\
\leq \quad & \mathcal{H}_{opt}(z_{eq}) + s_{opt} \cdot G_e(z_{eq}).
\end{aligned}$$

Thus, the upper boundary U is violated for s_{opt}, which contradicts its definition. Therefore, the proposition follows. □

The next result justifies the way we correct a solution to the unconstrained isotonic regression problem in line 11 of Algorithm 5. As we prove, setting its negative values to zero leads to the same L_∞-distance as in the constrained problem. It therefore yields an optimal solution to the latter. Keep in mind that the unconstrained isotonic regression problem is solved by the Basic approach (Stout, 2012). This algorithm computes the maximum of all previous values and the minimum of all subsequent values for each observation. It then chooses the regression value as the average of these two quantities.

Lemma 5. *Let $x \in \mathbb{R}^d$ be arbitrary. Denote by x_L the optimal solution of the L_∞ isotonic regression of x computed by the Basic approach (Stout, 2012). Define the new vector $x_{L0} = \max(x_L, 0)$ by component-wise comparison to 0. Let x_{Lc} be an optimal solution of the L_∞ isotonic regression of x with the constraint of nonnegativity. Then x_{L0} is also an optimal solution to the constraint problem: $L_\infty(x, x_{Lc}) = L_\infty(x, x_{L0})$.*

Proof. We show the statement considering the two distinct cases $\min(x) \geq 0$ and $\min(x) < 0$ consecutively. At first, assume that $\min(x) \geq 0$ holds. Then, by construction of x_L, we can deduce $x_L \geq 0$. Thus, x_{L0} is equal to x_L. As a nonnegative vector it also satisfies $L_\infty(x, x_{Lc}) \leq L_\infty(x, x_{L0})$. Since introducing constraints to a problem cannot lead to a better value of the objective function in the optimum, it must hold that $L_\infty(x, x_{Lc}) \geq L_\infty(x, x_L) = L_\infty(x, x_{L0})$. Together, this yields the result restricted to the case $\min(x) \geq 0$.
We now consider the case $\min(x) < 0$. The negative values of x_L set to zero in x_{L0} result in a maximal deviation of $-\min(x)$ to x. We thus get $L_\infty(x, x_{L0}) =$

$\max(L_\infty(x, x_L), -\min(x))$. Also, $\min(x) < 0$ and $x_{Lc} \geq 0$ imply $L_\infty(x, x_{Lc}) \geq -\min(x)$, so that we deduce

$$\begin{aligned} L_\infty(x, x_{L0}) &= \max(L_\infty(x, x_L), -\min(x)) \\ &\leq \max(L_\infty(x, x_L), L_\infty(x, x_{Lc})) \\ &= L_\infty(x, x_{Lc}). \end{aligned}$$

The last inequation follows, because a constraint problem cannot lead to a solution with a better value of the objective function compared to the corresponding unconstrained problem. Thus, $L_\infty(x, x_{L0}) \leq L_\infty(x, x_{Lc})$ holds. The converse inequality $L_\infty(x, x_{L0}) \geq L_\infty(x, x_{Lc})$ follows from the definition of x_{Lc}, since $x_{L0} \geq 0$. Both together yield the result restricted to the case $\min(x) < 0$, which completes the proof. □

We now collect all previous results to prove the correctness of our algorithm.

Theorem 1. *Algorithm 1 returns s_{opt} and a corresponding solution $\mathcal{H}_{opt}$ optimal in the sense of Problems 1 and 2, respectively.*

Proof. Lemma 1 shows that for $s > s_{opt}$ no mixture can lie within the confidence band B while for $s \leq s_{opt}$ there always exists a mixture lying in B. By the monotonicity of this property the binary search step converges to s_{opt}. Lemmas 3 and 4 allow to update the upper bound of the binary search by the values of the shrinkage factor after each Shrink-Down and Shrink-Up step. Hence, these steps further reduce the range of possible candidates for s_{opt}, while never excluding s_{opt}. Therefore, the correct s_{opt} is still determined. Lemma 2 implies that the correcting function $\mathcal{H}_k$ after termination of the main loop of Algorithm 1 is the function $\mathcal{H}_{min}$ introduced on page 16, which is required for solving Problem 2. Finally, Lemma 5 allows to correct the solution to the unconstrained L_∞ isotonic regression problem

finding an optimal solution to the constrained problem on the set M_{norm}. Thus a valid solution $\mathcal{H}_{opt}$ is generated, which completes the proof. □

2.4.2 Runtime Analysis

For the runtime analysis of our algorithm we introduce a precision parameter ε. This quantity never appears in our pseudo code or the actual implementation. Instead, think of it as the machine precision, which might depend on the physical architecture, the operating system or the programming environment. The main loop of Algorithm 1 in lines 5 to 11 runs until the mixture $\mathcal{F}$ lies within the confidence band up to an additive deviation of ε. In other words, the loop stops as soon as for all $z \in Z$ the property $L(z) - \varepsilon \leq \mathcal{F}(z) \leq U(z) + \varepsilon$ holds. In the following theorem we prove that this condition is met after a constant number of iterations. This yields an overall running time linear in the input size and logarithmic in $\frac{1}{\varepsilon}$. Hereby, we exclude the $O(n \log n)$ time needed for computing the cumulative distribution functions by assuming sorted input data. We rather focus on the linear running time of the actual analysis.

Theorem 2. *Let $\varepsilon \in (0, 1)$ be a fixed machine precision parameter. On an input of $n + m$ sorted observations, Algorithm 1 runs for at most $O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations. Each iteration can be implemented to run in time $O(n + m)$. The total running time is therefore of order $O((n + m) \log\left(\frac{1}{\varepsilon}\right))$.*

Proof. The Shrink-Down, the Shrink-Up, the binary search step and the normalisation step can be implemented in linear, i.e. $O(n + m)$, time. The solution to the isotonic regression subproblem (line 11 in Algorithm 5) can be computed in linear time as noted by Stout (2012). Therefore, it remains to bound the number of iterations of the loop in lines 5 to 11 of the main algorithm. The search interval

for s is initialized to $[s^*, 1] \subset [0, 1]$. It is halved at the end of every iteration by the binary search step. The Shrink-Down and Shrink-Up steps can only further decrease the upper bound and consequently the size of the search interval. Therefore, after $\lceil \log_2 \left(\frac{2}{\varepsilon}\right) \rceil$ iterations the size of the interval decreases to at most $2^{-\lceil \log_2 \left(\frac{2}{\varepsilon}\right) \rceil} < \frac{\varepsilon}{2}$. So, after $\lceil \log_2 \left(\frac{2}{\varepsilon}\right) \rceil$ iterations every value between the upper and lower boundary lies within additive precision $\frac{\varepsilon}{2}$ to s_{opt}. Consider an $s \in \left[s_{opt} - \frac{\varepsilon}{2}, s_{opt} + \frac{\varepsilon}{2}\right]$ and let $\mathcal{H}_s \in \mathcal{M}^*$ be the minimal function such that $s \cdot G_e + (1 - s) \cdot \mathcal{H}_s \geq L$ holds. Using $s \geq s_{opt} - \frac{\varepsilon}{2}$ we see that $s \cdot G_e \geq \left(s_{opt} - \frac{\varepsilon}{2}\right) \cdot G_e = s_{opt} \cdot G_e - \frac{\varepsilon}{2} \cdot G_e \geq s_{opt} \cdot G_e - \frac{\varepsilon}{2}$ holds. This implies $(1 - s) \cdot \mathcal{H}_s \leq (1 - s_{opt}) \cdot \mathcal{H}_{opt} + \frac{\varepsilon}{2}$ and we deduce

$$\begin{aligned} s \cdot G_e + (1 - s) \cdot \mathcal{H}_s &\leq \left(s_{opt} + \frac{\varepsilon}{2}\right) \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{opt} + \frac{\varepsilon}{2} \\ &\leq s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{opt} + \varepsilon \\ &\leq U + \varepsilon, \end{aligned}$$

because $s_{opt} \cdot G_e + (1 - s_{opt}) \cdot \mathcal{H}_{opt} \leq U$ holds by definition of s_{opt} and $\mathcal{H}_{opt}$. An analogous argument shows $s \cdot G_e + (1 - s) \cdot \mathcal{H}_s \geq L - \varepsilon$. Thus, the stopping criterion $L - \varepsilon \leq \mathcal{F} \leq U + \varepsilon$ is met after $\lceil \log_2 \left(\frac{2}{\varepsilon}\right) \rceil$ iterations and the result follows. □

2.5 Practical Evaluation

In this section the performance of our algorithm is investigated in simulation scenarios as well as on real datasets from astrophysics and bioinformatics. We also illustrate its linear running time and compare it to an alternative demixing procedure. In addition, the method's behaviour in case of false rejections of the null hypothesis is studied.

Our algorithm represents the determined correction distributions by their cumulative distribution functions. However, probability density functions and the first two moments allow to capture the main features of a distribution more intuitively. We thus present our results via estimated densities and empirical moments rather than using the distribution functions. Keep in mind that in applications this approach is not mandatory, because the determined distribution function contains all relevant information available. Improving simulations based on this distribution function directly is perfectly fine in practice. Thus, the additional estimations are not regarded as part of our method and are conducted for the purpose of presentation only. To asses their effect the interested reader is referred to Serfling (1980) and Devroye and Györfi (1985).

In order to attain estimators of the first two moments and the density of a correction distribution, we first determine the empirical density function corresponding to the calculated correction distribution function. This is achieved by considering consecutive differences of $\mathcal{H}_{opt}(z)$ using all $z \in Z$. We then generate 10 000 artificial observations from this density using weighted sampling. Finally, the empirical means and the empirical variances are computed from this artificial data. In addition, kernel density estimation is conducted to estimate the corresponding density. Given an i.i.d. sample $\tilde{x}_1, \ldots, \tilde{x}_l$ generated by an unknown density $\tilde{p}$, the kernel density estimate of $\tilde{p}$ is defined by

$$\hat{p}_h(x) = \frac{1}{l \cdot h} \sum_{i=1}^{l} K_h\left(x, \tilde{x}_i\right) \quad \forall x \in \mathbb{R}.$$

Hereby, K_h is for instance the Gaussian kernel function

$$K_h(x, z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - z}{h}\right)^2\right) \quad \forall x, z \in \mathbb{R} \tag{2.13}$$

with a bandwidth $h > 0$ always applied in the following. It is well known that in most cases the choice of the bandwidth has a much stronger effect on the results than the choice of the kernel function, cf. Devroye and Györfi (1985). Standard algorithms for the selection of h are cross validation and the method of Sheather and Jones (1991). The latter relies on a minimiser of the estimated mean integrated squared error and is used for all computations involving kernel density estimators in this work.

2.5.1 Performance and Runtime

In order to evaluate the algorithm it is applied in the popular setting of finite Gaussian mixtures. For this purpose we generate equally sized dataset pairs for each of the sample sizes $n = m = 100, 500, 1\,000, 5\,000, 10\,000, 50\,000, 100\,000$. In scenario a) for every dataset pair one sample is drawn from a standard Gaussian distribution. The other sample also consists of observations from the standard Gaussian distribution to a fraction of $s = 0.3$. The remaining observations stem from a second Gaussian distribution with mean 3 and standard deviation 1. Our demixing algorithm is therefore supposed to notice the different distributions of the samples, estimate a mixing proportion of about 0.3 and recommend a correction distribution with a mean near 3 and a standard deviation around 1.
In scenario b) the same sample sizes and also the case $n = m$ is investigated. It is more specific and resembles some of the situations encountered in our real data applications. Instead of mixing two Gaussian distributions, the constant value 0 is set for a fraction of 0.7 of the observations in each mixed dataset. The remaining fraction of 0.3 is sampled from the Gaussian distribution with mean 3 and standard deviation 1. The corresponding second sample representing the simulated data consist of observations from the same Gaussian distribution entirely. In this setting

the method is supposed to determine a shrinkage value s_{opt} around 0.3 and propose a correction distribution putting most of its probability mass at 0. Both scenarios are replicated 1 000 times for each of the sample sizes.

Table 1 shows the results for both data cases averaged over the 1 000 replications. In scenario a) we list the determined shrinkage factors s_{opt} as well as the mean and standard deviation of samples of size 10 000 drawn from the determined correction distribution $\mathcal{H}_{opt}$ for each sample pair. The second half of the table corresponds to scenario b). In addition to the determined shrinkage factors s_{opt}, the mean probability assigned to the value 0 by the determined correction distribution functions $\mathcal{H}_{opt}$ is presented. As we see, our demixing leads to an overestimation of the expected mixing proportion 0.3, which decreases in the sample size. This is not surprising, since by definition s_{opt} is the maximal shrinkage factor such that the corresponding mixture lies in the confidence band. Therefore, as the sample size increases, the radius of the confidence band becomes smaller and hence s_{opt} converges towards the true mixture proportion. The estimated mean and standard deviation in data setting a) behave similarly approaching 3 and 1, respectively. In scenario b) even for small sample sizes an overwhelming majority of the probability mass in $\mathcal{H}_{opt}$ is assigned to the value 0. This is correct, since by construction the differences between the sample pairs are caused by the zero values only. Altogether, the correction distributions proposed by the method reflect the discrepancies between the sample pairs quite well in both scenarios.

In Figure 2 we illustrate the algorithm output for scenario a) and $n = m = 1\,000$. In the upper row kernel density estimations of the two samples are presented according to their roles in our framework. Demixing the samples using Algorithm 1 leads to the shrinkage factor $s_{opt} = 0.39$, which is a reasonable approximation of the true mixture proportion $s = 0.3$. Using the approach described on page 41, we generate a third sample with 10 000 observations from the correction distribution

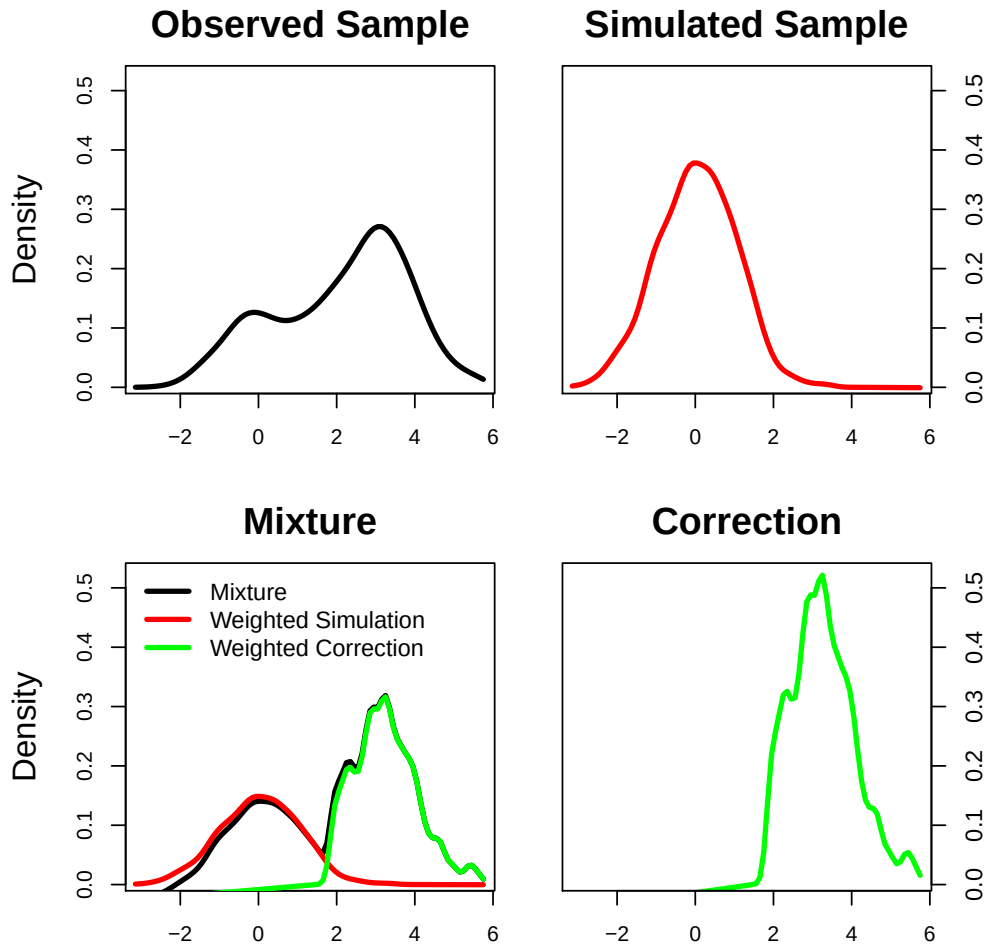

Figure 2: Kernel density estimates for two samples (upper row), the computed mixture (bottom right) and the correction distribution (bottom left) in the Gaussian mixture setup a).

characterised by $\mathcal{H}_{opt}$. Its mean 3.3 and standard deviation 0.81 resemble the desired values 3 and 1, respectively. The corresponding kernel density estimation shown on the right in the lower row is almost symmetrical and unimodal. Hence, the correction distribution represents the Gaussian distribution quite well, which is the difference between the underlying distributions of the first and the second sample. The final mixture distribution proposed by Algorithm 1 is illustrated in the lower left corner. It corresponds to the weighted sum of the distribution of the simulated sample and the correction distribution. The graph resembles the one of the observed sample as desired.

In order to study the running time of our algorithm we again make use of the data scenarios a) and b). For comparison a simpler demixing approach called binary search procedure is considered. It determines the optimal shrinkage factor s_{opt} relying only on the binary search. In contrast to Algorithm 1, the Shrink-Down and Shrink-Up steps are not conducted. Both steps are in principle not necessary to obtain the correct solutions to Problem 1 and 2, but are supposed to accelerate the computation. Thus, the determined s_{opt} and $\mathcal{H}_{opt}$ are identical for both methods, but the running times differ. The corresponding running times for both algorithms in data case a) are shown in Figure 3. Thereby, the time needed for precomputing the empirical distribution functions is not included. For the sake of presentation, the running times for the two largest sample sizes $n = m = 50\,000$ and $n = m = 100\,000$ are not included. These were 1.18 and 2.47 seconds, respectively, for Algorithm 1 and 6.2 and 12.33 seconds, respectively, for the binary search procedure. We also omit the running times for scenario b), which are essentially the same as in a). All results are averages over 1 000 repetitions.

In accordance with Theorem 2, the running time for both algorithms increases linearly in the sample size given sorted input data. It is by a factor of approximately 6 smaller for Algorithm 1 than for the binary search procedure for both data cases.

This shows that the Shrink-Down and Shrink-Up steps lead to huge savings in computation time and are therefore very valuable in particular for large datasets.

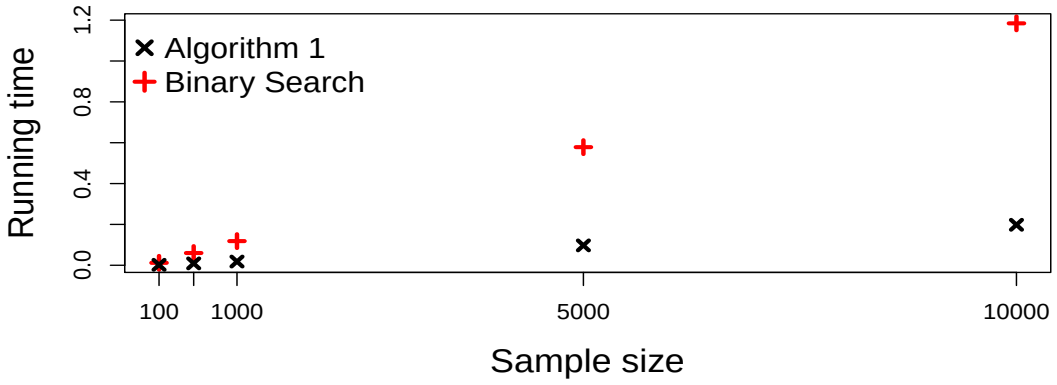

Figure 3: Average running times in seconds for Algorithm 1 (black) and the binary search procedure (red) computed on sorted samples for different sample sizes in the Gaussian mixture case a).

2.5.2 Estimated Shrinkage Factors under the Null Hypothesis

Under $\mathbb{H}_0 : P = Q$ both analysed samples stem from the same distribution. In this situation the Kolmogorov-Smirnov test rejects by mistake in about an α-fraction of the cases, where α is the predefined significance level. A reasonable procedure comparing the samples after a false rejection should recognise their similarity. Thus, a shrinkage factor near 1 is desirable in such cases.

To check the performance of our method under $\mathbb{H}_0$, dataset pairs are generated for the sample sizes $n = m = 100, 500, 1\,000, 5\,000, 10\,000$. All samples stem from the standard Gaussian distribution. Other distribution types like exponential and t-distributions were also considered and led to comparable results. For each sample size, dataset pairs are simulated until the Kolmogorov-Smirnov test rejects 1 000 times. These 1 000 dataset pairs are passed to Algorithm 1. The corresponding

shrinkage factors determined by the method are presented via boxplots in Figure 4. All of them are less than 1 by construction.

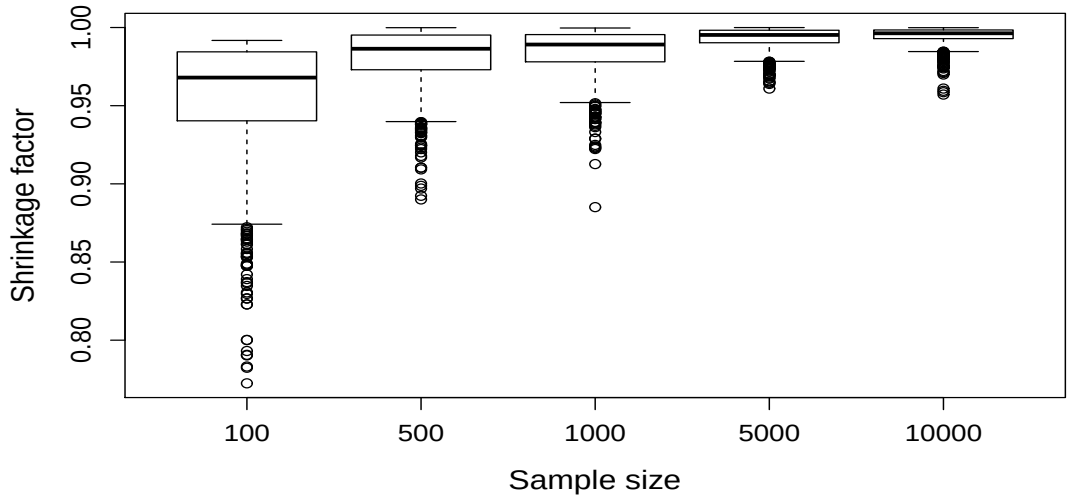

Figure 4: Boxplots of shrinkage factors s_{opt} determined by Algorithm 1 for different sample sizes after a false rejection of the null hypothesis $\mathbb{H}_0 : P = Q$ for data generated from the standard Gaussian distribution.

As suggested by the results, even for small sample sizes the majority of shrinkage values are greater than 0.9. Increasing the sample size further reduces the amount of small shrinkage values. Thus, our method performs as desired: if no modifications are actually necessary, the algorithm proposes to perform none or only small modifications to the current samples.

2.5.3 Application to Astrophysical Data

In this section we apply Algorithm 1 to investigate data from astrophysics motivating our work. The data situation is introduced on page 9. We consider simulated proton data and compare it to observations recorded by the gamma ray detectors MAGIC-

I and MAGIC-II. The latter are almost completely induced by protons. Both datasets consist of 5 000 observations and contain 54 continuous attributes we work with. Among other features, these variables include characteristics of the recorded atmospheric signals and their reconstructed trajectory. The attributes are identical for both datasets. Our method is supposed to determine attributes differing the most for simulated protons and observed data and quantify their discrepancies. This information can subsequently be used to improve the background simulation. The Kolmogorov-Smirnov test comparing the real data and the simulation rejects the null hypothesis for all but two attributes. However, 37 of the 54 attributes have shrinkage factors above 0.85, which indicates a suitable proton simulation overall. The upper row of Figure 5 provides kernel density estimates for the observed and simulated data for the attribute Length1. This variable describes the length of the ellipse fitted to an atmospheric signal measured by the MAGIC-1 detector. The Kolmogorov-Smirnov test for Length1 rejects $\mathbb{H}_0 : P = Q$ and results in a comparably low shrinkage factor of 0.75. Therefore, the simulation of this variable might be inadequate and the corresponding simulation steps seem to be worth inspecting in more detail. In the lower right corner of Figure 5, a kernel density estimation for the corresponding correction distribution characterised by $\mathcal{H}_{opt}$ is presented. It is determined on 10 000 observations generated by the sampling technique introduced on page 41. The plot in the lower left corner shows the density estimates for the simulated and the correction distribution weighted by 0.75 and 0.25, respectively. In addition, the density estimate for the final mixture is included. All plots are presented on the same scale.

The coarse form of the density estimates for the observed data and the simulation in the upper row is quite similar showing one major peak around 25. However, there are some slight discrepancies. Compared to the real data curve, the main peak of the simulation is considerably higher. While the curve for the real data has

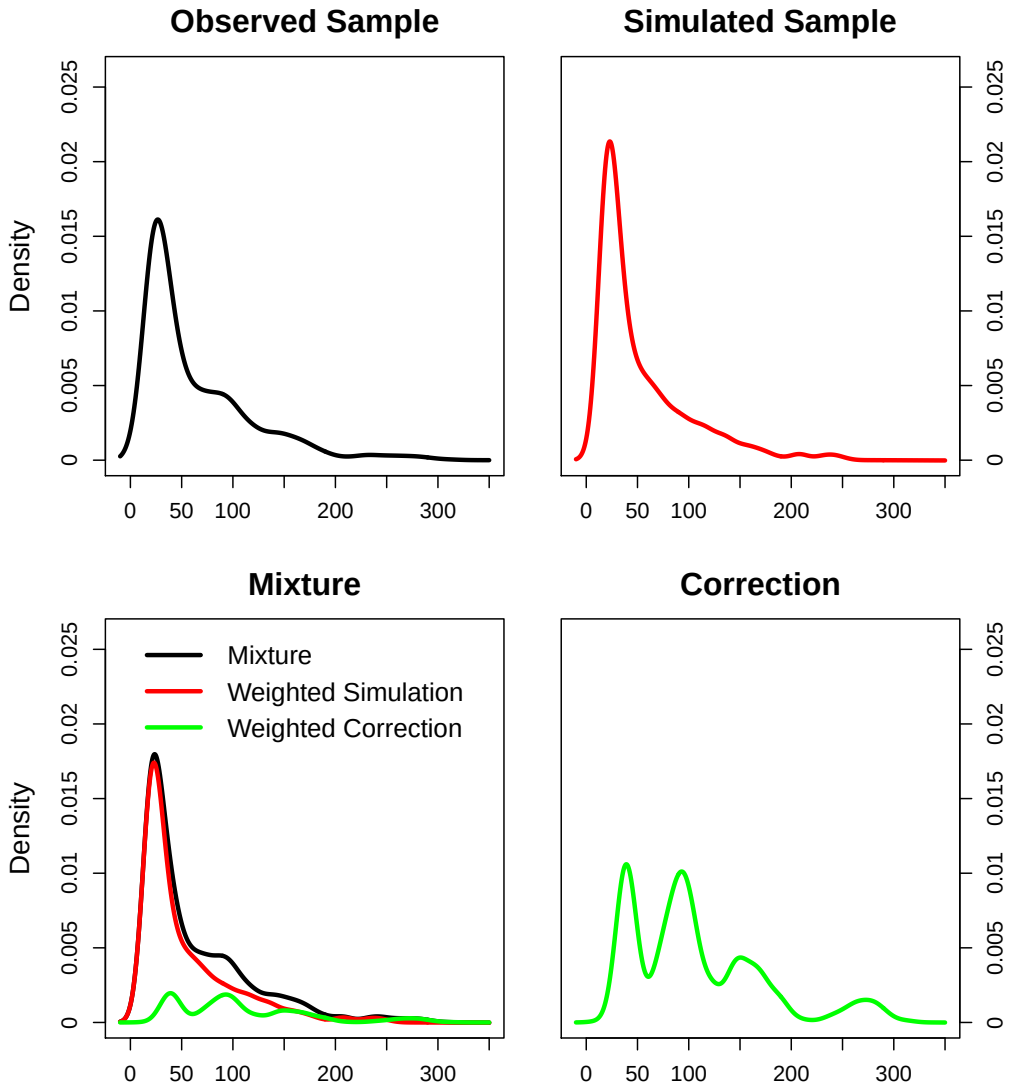

Figure 5: Kernel density estimates for the attribute Length1 based on the observed data (top left), the simulated data (top right), the determined mixture (bottom left) as well as the correction distribution (bottom right).

a plateau around 90, there is a steadily falling curve for the simulation. Although these differences are not very large, it is quite unlikely that they are induced by the sample variance due to the large sample sizes. To assess whether this is realistic or not, we conducted several simulations considering kernel density estimates for a broad class of distributions using 5 000 observations in each sample. The dissimilarities in these simulations were much smaller than for the Length1 attribute. This supports the conjecture that the Kolmogorov-Smirnov test correctly rejects the null hypothesis of equal distributions. To correct the simulated sample one obviously has to generate less observations around 25 and more around 90. Exactly this is proposed by the correction distribution presented in the lower right corner of Figure 5. The corresponding density graph based on the estimated $\mathcal{H}_{opt}$ has a peak near 25, but also another one of comparable height and greater width near 90. Therefore, it gives the region around 90 about as much weight as the one around 25, in contrast to the simulated sample. The combination of the simulated and the correction distributions leads to the density graph of the final mixture presented by the solid black curve in the lower left plot of Figure 5. It resembles the density estimate for the observed data quite well. On the one hand, the height of the main peak is corrected, which is achieved by the shrinking. On the other hand, the required plateau is introduced to the mixture by the correction distribution. As similar but somewhat smaller correction is performed for the plateau around 140.

2.5.4 Application to Bioinformatical Data

Algorithm 1 is also illustrated on so called ion mobility spectrometry (IMS) measurements. IMS data allows to detect volatile organic compounds in the air or in exhaled breath. For the analysis, groups of measurements are summarised in spectrograms, two-dimensional data structures similar to heat-maps. Motivated by

the need to process the measurements in real-time as they arrive one-by-one, it is a usual approach to find and annotate major peaks in the spectrograms. In this way the original information is reduced to the position and shape parameters of the peaks and storage is saved. To automate and speed-up the computations D'Addario et al. (2014) propose to approximate the measurements by finite mixtures of probability density functions. More precisely, both dimensions of the spectrograms are modelled independently by mixtures of inverse Gaussian densities. The corresponding parameters of the densities are estimated using an EM algorithm.

For the evaluation of these models we focus on one of the dimensions and condition on the other. This results in 6 000 spectrograms consisting of 12 500 data points each. They stem from 10 minutes of IMS measurement, cf. Kopczynski et al. (2012). This data consist of 187 groups of spectrograms. Hereby, each spectrogram in a group belongs to the same peak model and the models differ over the groups. Both the spectrograms as well as the bioinformatic mixture models can be regarded as probability density functions up to some normalising constants. This allows us to draw samples of size 1 000 from each spectrogram and the corresponding mixture model. In order to evaluate the bioinformatic models we apply our algorithm to the corresponding pairs of datasets.

In general our algorithm suggests that the models fitted by the bioinformaticians approximate their spectrograms reasonably well, since in 152 of the 187 groups the mean shrinkage factor for the spectrograms is above 0.8. In addition, we identify some interesting groups of spectrograms. The shrinkage factors of two of these are shown in Figure 6. Keep in mind that the spectrogram index represents the second dimension of the data we condition on. In both groups the model in the second dimension consists of a single inverse Gaussian density.

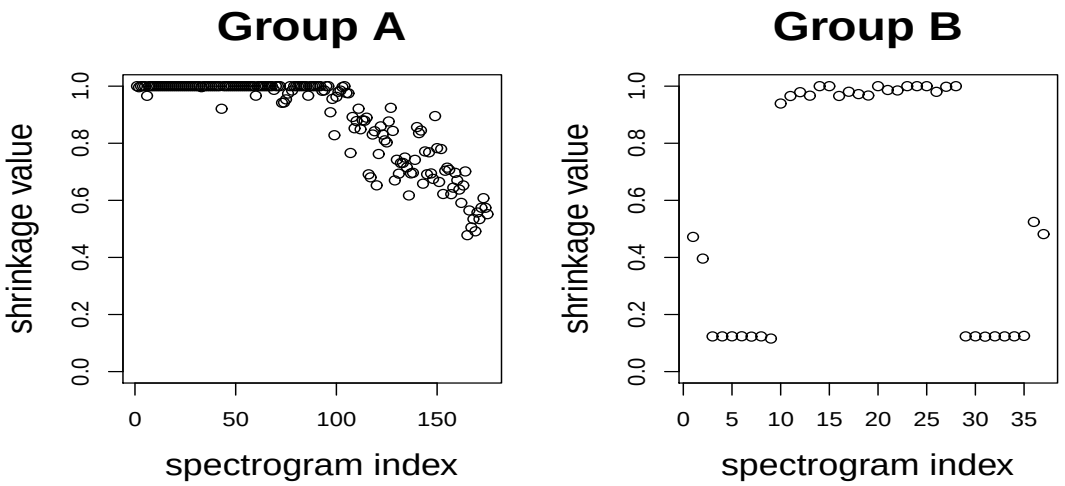

Figure 6: Shrinkage factors s_{opt} for two groups of spectrograms.

Our results for group A suggest that the first half of the measurements are modelled quite well by the bioinformaticians' EM algorithm, but for increasing spectrogram indices the approximation is getting worse. This shows that the bioinformatics model in the second dimension is not appropriate. Instead of a single inverse Gaussian density, two components would probably lead to better approximations. In contrast to that, the shrinkage factors for group B indicate a sufficient number of components used in the second dimension. For the spectrograms in the middle we have shrinkage factors close to one. Thus the corresponding models are close to the observed spectrograms. However, going to the left and right borders, the spectrograms seem to be fitted quite poorly, since the shrinkage factors are lower than 0.2. The two leftmost and two rightmost models are a little closer to their spectrograms with shrinkage values between 0.4 and 0.6. Taking the models of Kopczynski et al. (2012) into account this indicates that their fitted density mixture might be too wide or too narrow in the second dimension. The approximation could be substantially improved by excluding the spectrograms on both margins from this group and treating them by further models.

We also illustrate our procedure using a single spectrogram from the dataset. The upper row of Figure 7 provides a kernel density estimate for the measurement 1157 and its model. Since all four plots are given on the same scale, the two peaks in the model are more narrow and differ much more in height than the ones in the original data. In addition, the peak on the left is not included in the model. Although it looks small in this scale, it appears noteworthy when compared to the other two. In the bottom right part of Figure 7 a kernel estimate for the correction distribution characterised by $\mathcal{H}_{opt}$ is presented. It is determined on 10 000 observations generated by the sampling approach described on page 41. As expected, the correction distribution puts mass on the very right peak in order to fix the height proportions between the peaks on the right. In addition, it generates the left peak missing in the model. The plot in the lower left corner shows the estimates of the modelled and the correction distribution weighted by the determined shrinkage value 0.76 and the remaining mass 0.24, respectively. The kernel estimate for the final mixture, which is the sum of the weighted estimates, is presented here, too. The proposed mixture is still somewhat narrow, but the proportions of the peak heights as well as the small peak are represented more adequately in comparison to the original model.

2.6 Related Methods

The literature offers various suggestions on mixture models. Some of them involve multiple samples and finite mixture models, like for example the Bayesian approach by Kolossiatis et al. (2013). Nevertheless, to the best of our knowledge, there is no literature addressing the two-sample problem investigated in this chapter. Algorithm 1 closes this gap providing a fast distribution-free method to model discrepancies between datasets flexibly, as illustrated in Section 2.5.

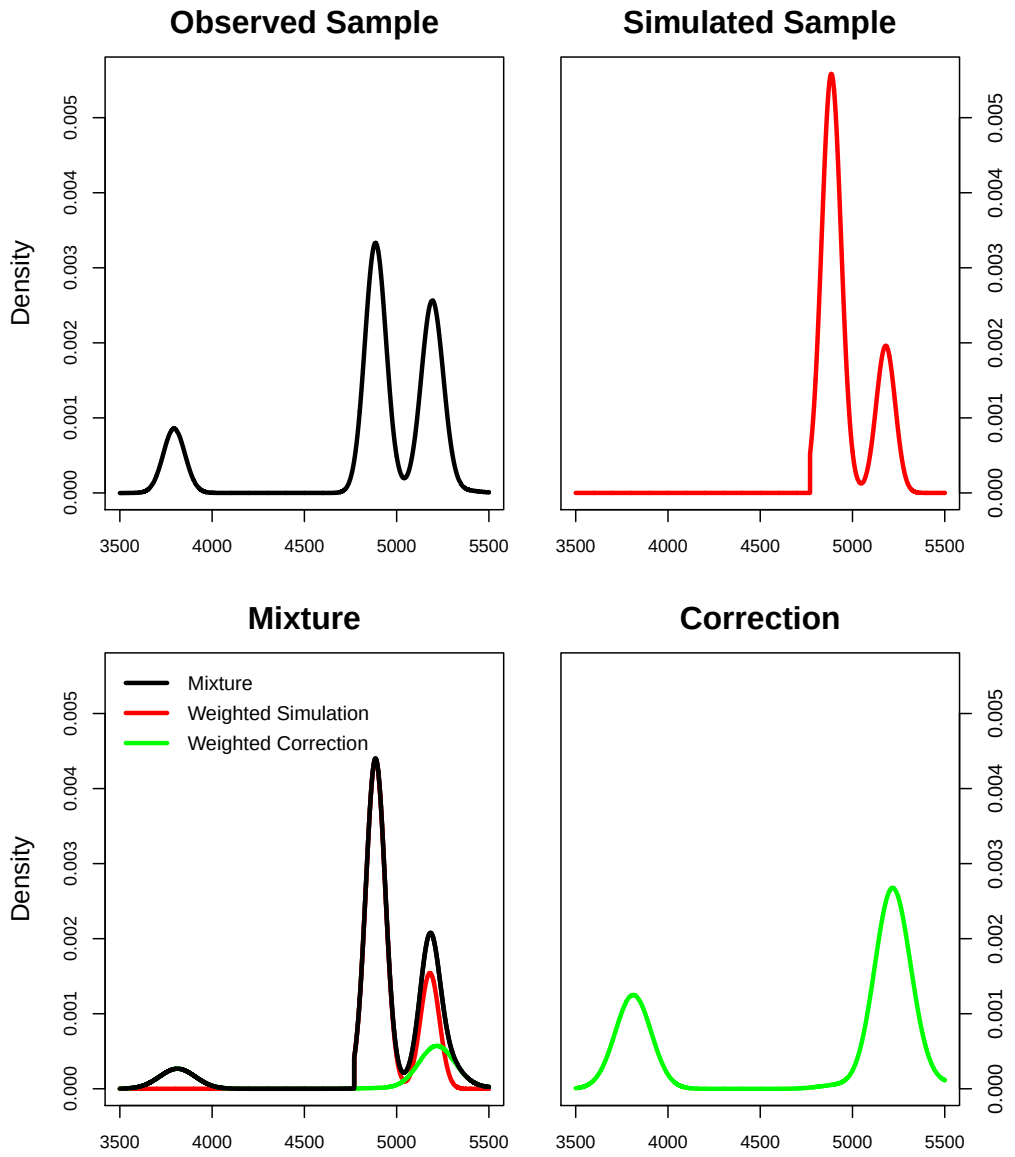

Figure 7: Kernel density estimates for spectrogram 1157 based on the measurements (top left), the corresponding inverse Gaussian model (top right), the determined mixture (bottom left) and the correction distribution (bottom right).

Most of the procedures available in the context of mixture models are tailored for single samples and operate on the level of probability density functions. They estimate the number and shape of components in the specified mixture model and thereby rely on adjusted EM algorithms (Pilla and Lindsay, 2001) or Newton type optimisers (Wang, 2010; Schellhase and Kauermann, 2012). As pointed out on page 41, probability density functions are easier to interpret than distribution functions. It is thus worth discussing, whether the algorithms available can be modified to solve the two-sample problem investigated in this chapter in a straightforward way. Many procedures available like for example Schellhase and Kauermann (2012) use convex combinations of basis functions $p_i, i = 1, \ldots, n_b$ to model the unknown density p:

$$\hat{p}(x) = \sum_{i=1}^{n_b} a_i p_i(x) \quad \forall x \in \mathbb{R}. \tag{2.14}$$

The natural way to exploit such models in our situation is the following one: first, the density q corresponding to the simulated sample is estimated via (2.14). In a second step this estimate $\hat{q}$ is fixed as one of the basis functions for the estimation of the density p corresponding to the observed sample. In short, one fits the model $\hat{p} = a_1\hat{q} + \sum_{i=2}^{n_b} a_i p_i$. This strategy is straightforward, but has a crucial drawback. In general one cannot guarantee that the coefficient a_1 properly reflects the importance of the simulated data. Often a_1 is determined way too small as long as the remaining basis functions are not chosen appropriately, because they also contribute to the region of values modelled by $\hat{q}$. Thus, in terms of a better fit, it is often correct to choose a_1 small. However, this corresponds to discarding the simulation almost completely, which is not desirable in our application. Unfortunately, choosing the remaining basis functions in an adequate way is a highly nontrivial and open problem. Therefore, the obvious adjustment of one-sample density-based approaches does

not lead to satisfactory results. Also, their optimisation is often computationally costly making them less suitable especially in medium to large sample cases. Certainly, much work is necessary to make density-based procedures viable for the problem studied in this chapter. Before going in this direction it is thus interesting to know, whether such methods are advantageous at all in the context of homogeneity. A good starting point to address this question are two-sample tests relying on probability density functions studied in the next chapter.

3 Two-sample Tests based on Divergences

In this chapter we construct and evaluate distribution-free two-sample homogeneity tests based on probability density functions. In particular, we focus on the concept of f-divergences introduced by Ali and Silvey (1966), which provides a rich set of distance like measures between pairs of distributions. The corresponding tests can be applied to detect arbitrary deviations of distributions and are not restricted to location or scale alternatives. This part of the work is based on the manuscript Wornowizki and Fried (2014). The discussions with my thesis advisor R. Fried were of great help to improve the methods as well as their presentation.

The f-divergences are defined in Section 3.1. We then present a new nonparametric divergence estimation technique combining kernel density estimation and spline smoothing in Section 3.2. As we show in extensive simulations, the algorithm performs stable and quite well in comparison to several existing non- and semiparametric divergence estimators. In Section 3.3 we tackle the two-sample homogeneity problem using permutation tests based on various divergence estimators. The methods are compared to an asymptotic divergence test as well as to several traditional parametric and nonparametric procedures under different distributional assumptions and alternatives in simulations. It turns out that divergence-based procedures detect discrepancies more often than traditional methods, if the samples do not predominantly differ in location. The tests performing best are applied to the ion mobility spectrometry data considered before in Section 2.5.4. Section 3.4 concludes the chapter giving some final thoughts on divergence-based testing. Furthermore, potential extensions of the concept are pointed out.

3.1 Divergence Measures

As before, let us consider two distributions P and Q with corresponding probability density functions p and q. For any convex function $f : \mathbb{R} \to \mathbb{R}$ the f-**divergence** from P to Q is defined by

$$D_f(P,Q) = \int f\left(\frac{p(y)}{q(y)}\right) dQ(y) \; = E_Q\left(f\left(\frac{p\,(Y)}{q\,(Y)}\right)\right). \tag{3.1}$$

To ensure a well-defined density ratio $r = \frac{p}{q}$ the distribution P must be dominated by Q. An f-divergence attains its minimal value $f(1)$ if, and only if, $P = Q$ (Ali and Silvey, 1966). For all common divergences $f(1) = 0$ holds, giving a rather intuitive interpretation of the minimal property. Note that divergences do not need to be symmetric, that is, $D_f(P,Q) = D_f(Q,P)$ may not hold. Choosing the function f as $f_{aKL}(x) = x \cdot \log(x)$ for all $x \in \mathbb{R}$ yields the asymmetric Kullback-Leibler divergence denoted by D_{aKL}. This measure is closely related to the popular AIC information criterion, see Seghouane and Amari (2007). It also has a central role among the divergences, since minimizing it in the context of parameter estimation corresponds to the classical maximum likelihood approach, cf. Basu et al. (1998). D_{aKL} can be symmetrised using $f_{KL}(x) = (x-1) \cdot \log(x), x \in \mathbb{R}$. This leads to the symmetric Kullback-Leibler divergence D_{KL} fulfilling $D_{KL}(P,Q) = D_{aKL}(P,Q) + D_{aKL}(Q,P)$. In case of continuous and one-dimensional random variables this measure can be represented by

$$D_{KL}(P,Q) = \int [p(x) - q(x)] \cdot [\log(p(x)) - \log(q(x))]\, dx.$$

Another member of this class is the squared Hellinger distance, also called Hellinger divergence. For continuous random variables it is defined by

$$D_H(P,Q) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \, dx = 1 - \int \sqrt{p(x)} \cdot \sqrt{q(x)} dx.$$

As suggested by the names, $\sqrt{D_H}$ is a metric, while D_H is the f-divergence corresponding to $f_H(x) = \frac{1}{2} \cdot \left(\sqrt{x} - 1\right)^2, x \in \mathbb{R}$. In contrast to the unbounded Kullback-Leibler divergence, the Hellinger divergence does not exceed 1. Along with the Kullback-Leibler measure, it is one of the standard divergences investigated in the literature. In particular, it allows to construct robust and first-order efficient parameter estimators (Lindsay, 1994) and is frequently used for asymptotical considerations as for example in Liese and Miescke (2008).

Divergence measures, similar to Kolmogorov-Smirnov type statistics, take into account deviations in the location, the scale, the skewness and any other characteristics of the distributions and weight them implicitly according to the function f. Thus, corresponding methods are able to detect arbitrary heterogeneities. For this reason, divergence measures and related quantities are applied in various estimation and testing problems like contingency tables (Alin and Kurt, 2008), model selection (Seghouane and Amari, 2007), survival analysis (Zhu et al., 2013) and detection of structural breaks of distribution parameters in time series (Lee and Na, 2005). They often yield a good compromise between efficiency and robustness, cf. Beran (1977) and Basu et al. (1998). A downside of divergence measures is the necessity of density ratio estimation. Therefore, the problem is often divided into two steps:

1. Estimate the density ratio function $r = \frac{p}{q}$ by $\hat{r}$.
2. Estimate the divergence $D_f(P,Q) = E_Q\left(f(r(Y))\right)$ given $\hat{r}$.

Several approaches to both steps are discussed in the next section.

3.2 Divergence Estimation

This section is dedicated to the estimation of f-divergences and consists of three parts. Section 3.2.1 reviews several non- and semiparametric methods for the estimation of the density ratio. These procedures are utilised in Section 3.2.2 to construct divergence estimators. In addition to reviewing the standard approaches, we propose a new algorithm for divergence estimation based on spline smoothing. All divergence estimators are evaluated in a simulation study in Section 3.2.3.
As in Chapter 2, we assume that $x_1, \ldots, x_n \in \mathbb{R}$ are observations from continuous, independent and identically distributed random variables $X_1, \ldots, X_n$. Each of them follows a distribution P with probability density function p. We make analogous assumptions for the sample $y_1, \ldots, y_m$ and the corresponding random variables $Y_1, \ldots, Y_m$ with distribution Q and probability density function q.

3.2.1 Density Ratio Estimation

The **direct approach** is an intuitive way to estimate the density ratio function $r = \frac{p}{q}$ without imposing distributional assumptions. Hereby, the probability density functions p and q are estimated nonparametrically by $\hat{p}$ and $\hat{q}$ first. Hereafter, $r = \frac{p}{q}$ is simply approximated by the ratio $\hat{r} = \frac{\hat{p}}{\hat{q}}$. Estimates of the individual probability density functions can be attained by the kernel density procedure (Devroye and Györfi, 1985), cf. page 42. For implementation we use the bandwidth choice of Sheather and Jones (1991) and the Gaussian kernel. The latter ensures strictly positive density estimates, which results in a well defined density ratio estimate $\hat{r} = \frac{\hat{p}}{\hat{q}}$.
In contrast to the nonparametric approach, semiparametric methods estimate the density ratio itself instead of the individual densities. The key idea is to introduce a

density ratio model $r(\cdot, \theta)$, which should fulfill $r(x) = r(x, \theta)$ for a certain parameter $\theta^* = (\theta_1^*, \ldots, \theta_d^*) \in \mathbb{R}^d$ and all $x \in \mathbb{R}$. Thereby, the identification of r boils down to the approximation of the parameter θ^* by an estimate $\hat{\theta}$. Since different distributions can result in the same density ratio, the density ratio model does not parametrise the densities completely. It thus can be regarded as semiparametric. In the following, we describe two parameter estimation techniques for semiparametric density ratio models.

The **moment matching** technique (Qin, 1998) is motivated by the equation

$$E_P(\eta) = \int \eta(x) \cdot p(x)\, dx = \int \eta(x) \cdot \frac{p(x)}{q(x)} \cdot q(x)\, dx = E_Q(\eta \cdot r),$$

which holds for the true density ratio function $r = \frac{p}{q}$ and for an arbitrary moment function η. As we see, the moments $E_P(\eta)$ and $E_Q(\eta \cdot r)$ are equal for the correct density ratio. Replacing these moments by appropriate sample means allows to estimate θ^* by solving the equation

$$\frac{1}{n} \sum_{i=1}^{n} \eta(x_i, \theta) - \frac{1}{m} \sum_{j=1}^{m} \eta(y_j, \theta) \cdot r(y_j, \theta) = 0 \tag{3.2}$$

in θ for any density ratio model. In other words, the parameter θ is chosen such that the empirical approximations of the considered moments match. As shown by Qin (1998) the moment function

$$\eta^*(x, \theta) = \frac{1}{1 + \frac{n}{m} \cdot r(x, \theta)} \nabla \log r(x, \theta) \quad \forall x \in \mathbb{R}, \tag{3.3}$$

is optimal in the sense that the corresponding estimator induced by the moment matching has minimal asymptotic variance. Hereby, $\nabla \log r(x, \theta)$ denotes the gradient column vector of the function $\log r(x, \theta)$ with respect to θ for all $x \in \mathbb{R}$. There are analytic solutions of equation (3.2) for density ratio models linear in

θ. Explicit estimators of r in arbitrary density ratio models are only available at the sample points $y_1, \ldots, y_m$. If the problem is not explicitly solvable, it is rephrased via the minimisation of the square of the left-hand side of equation (3.2) and numerical optimisation is applied. This is the case in the simulations presented in the following, because we focus on the popular exponential model

$$r_e(x, \theta) = \exp\left(\theta_1 + \theta_2 \cdot x + \theta_3 \cdot x^2\right). \tag{3.4}$$

This model includes the case of two Gaussian distributions, but also holds for two exponential distributions. In the latter case it is overparametrised, because the quadratic term is redundant. In our applications of the moment matching we always use the optimal moment function η^* introduced in (3.3) and the exponential model presented in (3.4) unless stated otherwise. The minimisation problem is solved using the optimiser of Nelder and Mead (1965) implemented in the R-function *optim* with default settings. The initialisation for the parameter θ is derived from the maximum likelihood estimates of the mean and variance under the assumption of Gaussianity. Several other initialisation procedures were investigated, but did not improve the estimation performance. Especially the initialisation assuming $r = 1$ does not provide good results and is therefore not advisable.

Moment matching can be conducted using arbitrary density ratio models. Typically, models with a low dimension are used and thus relatively strong assumptions on the density ratio are made. In contrast to that, the density ratio model in the **ratio matching** approach is fixed to

$$r_K(x, \theta) = \sum_{i=1}^{d} \theta_i \cdot K_h(x, x_i^*) \quad \forall x \in \mathbb{R}. \tag{3.5}$$

Hereby, $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ is the parameter vector of weights, K_h is the Gaussian kernel defined in (2.13) and the $x_1^*, \ldots, x_d^*$ are observations randomly chosen from the sample $x_1, \ldots, x_n$. According to Sugiyama et al. (2009) a model dimension $d = \min(100, n)$ is sufficient to guarantee reasonable results together with a tolerable computation time in most applications. This density ratio model typically has more parameters than the ones used in the moment matching, which leads to a more flexible estimation. To estimate the parameter θ^* via ratio matching, distance-like measures between the true and the modelled density ratio are minimised. One example for this is the Kullback-Leibler importance estimation procedure (**KLIEP**), which is presented in detail in Sugiyama et al. (2009). The method relies on the measure $KL(\theta) = -\int \log\left(r_K(x,\theta)\right) \cdot p(x)\ dx$. This quantity is the asymmetric Kullback-Leibler divergence from p to the implicitly modelled $p(x,\theta) = r_K(x,\theta) \cdot q(x)$ up to a constant independent of θ. Since divergence measures attain their minimal value only for equal distributions, the KLIEP procedure estimates θ by minimising an empirical equivalent to $KL(\theta)$ in θ. Another example for ratio matching is the Least-Squares Importance Fitting (**LSIF**) also presented in Sugiyama et al. (2009). The approach corresponds to the function

$$\begin{aligned} LS(\theta) &= \frac{1}{2}\int r_K(x,\theta)^2 \cdot q(x)\ dx - \int r_K(x,\theta) \cdot p(x)\ dx \\ &= \frac{1}{2}\int \left(r_K(x,\theta) - r(x)\right)^2 \cdot q(x)\ dx + c, \end{aligned}$$

where the constant c is again independent of θ. It thus essentially reflects a squared distance between the density ratio and its model. To obtain sparse solutions a weighted penalty term consisting of the L_1-norm of θ is added to the empirical equivalent of $LS(\theta)$. This leads to the estimator

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min}\ \frac{1}{2m}\sum_{j=1}^{m} r_K(y_j,\theta)^2 - \frac{1}{n}\sum_{k=1}^{n} r_K(x_k,\theta) + w^* \sum_{u=1}^{d} |\theta_d| \ .$$

Both KLIEP and LSIF require constraint optimisation procedures, since $\hat{\theta} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_d\right)$ must not be negative to ensure a nonnegative density ratio estimator. Unfortunately, high-dimensional constrained optimisation tasks are hard to solve efficiently. Therefore, Sugiyama et al. propose to drop the nonnegativity restriction initially and replace the L_1 penalty term by L_2 regularisation. Since the density ratio model (3.5) is linear in θ, the unconstrained minimisation with the L_2 penalty is analytically solvable. The negative entries of its solution $\hat{\theta}$ are set to zero, so that a nonnegative density ratio estimate is ensured. This procedure is called the unconstrained Least-Squares Importance Fitting (**uLSIF**). In addition to resulting in an analytically solvable optimisation problem, uLSIF has another advantage: the score of the leave one out cross-validation can be computed efficiently and stably. This is quite useful for obtaining a suitable bandwidth h and a regularisation weight w^*. In our computations we make use of the R implementation of this algorithm by Takafumi Kanamori with default settings. It is available at `http://www.ms.k.u-tokyo.ac.jp/software.html`.

3.2.2 Divergence Estimation using Density Ratio Estimates

In this subsection we introduce several possibilities for estimating an arbitrary divergence D_f given $\hat{r}$, an estimate of the density ratio function $r = \frac{p}{q}$. By its definition in (3.1) a divergence is nothing but the moment $E_Q\left(f\left(r\left(Y\right)\right)\right)$. Hence, straightforward application of the strong law of large numbers allows to estimate D_f by the **natural estimator**

$$\hat{D}_f = \frac{1}{m} \sum_{j=1}^{m} f\left(\hat{r}(y_j)\right).$$

As a simple mean, this estimator is easy to implement and fast to compute. However, the procedure is asymmetric in the sense that the first sample affects the divergence estimation only implicitly via the density ratio estimation in contrast to the second sample. As we see in the simulations in Section 3.3.2, this does affect the performance of corresponding tests.

Kanamori et al. (2012) expanded $\hat{D}_f$ explicitly including both samples in the final divergence estimation. The authors decompose the convex function f characterising a divergence measure via $f(x) = f_1(x) + x \cdot f_2(x)$ for all $x \in \mathbb{R}$. Given such a pair f_1 and f_2 each f-divergence can be estimated by the **decomposed estimator**

$$\hat{D}_f^D = \frac{1}{m}\sum_{j=1}^{m} f_1\left(\hat{r}(y_j)\right) + \frac{1}{n}\sum_{k=1}^{n} f_2\left(\hat{r}(x_k)\right),$$

because $D_f(P,Q) = E_Q\left(f\left(r\right)\right) = E_Q\left(f_1\left(r\right)\right) + E_P\left(f_2\left(r\right)\right)$ holds. For the moment matching method based on the moment function η^* specified in (3.3) Kanamori et al. prove that the decomposition into

$$f_1^*(x) = \frac{f(x)}{1 + \frac{n}{m} \cdot r(x,\theta)} \text{ and } f_2^*(x) = \frac{\frac{n}{m} \cdot f(x)}{1 + \frac{n}{m} \cdot r(x,\theta)} \quad \forall x \in \mathbb{R} \tag{3.6}$$

leads to an estimator with minimal asymptotic variance under fairly weak and verifiable conditions, cf. Kanamori et al. (2012). Even though the decomposed estimator is introduced for the moment matching density ratio estimation, it is applicable for any density ratio estimation procedure.

We now propose an alternative estimator of f-divergences, which makes use of cubic splines (Green and Silverman, 1994). Cubic splines are piecewise polynomial functions with continuous second derivatives. They are quite appealing for regres-

sion, since given some observations $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_l, \tilde{y}_l)$ and a fixed penalty factor $\tilde{w} > 0$ there is a unique cubic spline minimising

$$S(g) = \sum_{j=1}^{l} \left(\tilde{y}_j - g(\tilde{x}_j)\right)^2 + \tilde{w} \int_{\tilde{x}_1}^{\tilde{x}_l} g''(x)^2 dx \tag{3.7}$$

over all functions g with continuous second derivatives. Hereby, g'' denotes the second derivative of g. The measure S represents a trade-off between the goodness-of-fit (left term) and the roughness (right term) of the regression function g. For $\tilde{w} \to \infty$ its minimiser converges to the linear least squares fit, for $\tilde{w} \to 0$ the solution is a spline interpolating the observations. The minimiser of S is computable in linear time, which is advantageous when dealing with large datasets.

In order to derive our divergence estimation technique, let us regard a divergence as an integral involving the known convex function f, the unknown density ratio r and the unknown density q. The unknown quantities can be estimated following the direct approach to density ratio estimation. The only problem left to solve then is the integration. Since the direct density ratio estimator is quite sensitive to distortions of its denominator, we propose to smooth the integrand via cubic splines. This also allows us to solve the integration problem, because splines are piecewise polynomial and thus can be integrated analytically quite easily. In summary, we propose the following algorithm to obtain a **smoothed estimator** $\hat{D}_f^S$:

1. Compute the kernel density estimates $\hat{p}$ and $\hat{q}$ and set $\hat{r} = \frac{\hat{p}}{\hat{q}}$.
2. Smooth the function $f\left(\hat{r}(\cdot)\right)\hat{q}(\cdot)$ using cubic splines.
3. Integrate the spline analytically over the range of all observations.

Our implementation of this algorithm determines the bandwidth of the kernel density estimations via the method of Sheather and Jones (1991) using the Gaussian

kernel (2.13). The spline smoothing is performed using the routine *smooth.spline* with default settings available in the *stats* package. More detailed information in particular with regard to the proper choice of $\tilde{w}$ are provided on to the corresponding help page and in the references given therein.

3.2.3 Comparison of the Divergence Estimators

To investigate their performance the estimation techniques presented before are applied to artificial data. Hereby, we restrict ourselves to distribution pairs with explicit representations of the corresponding divergence measure. The true divergence values can thus be calculated in an easy way allowing us to validate the estimates. Therefore, we work with exponential, Laplacian and Gaussian data for equal sample sizes $m = n = 50, 100, 300$. The results are reported for the Gaussian data and $m = n = 300$ only, since the findings for the other distributions and sample sizes are essentially the same. In the Gaussian setting the first sample is generated from the standard Gaussian distribution. The second one is drawn from the Gaussian distribution with mean μ and variance σ^2. The cases considered are:

1) $\mu = 0,\ \sigma^2 = 1$ ($\mathbb{H}_0$)

2) $\mu = 3,\ \sigma^2 = 1$ (location alternative)

3) $\mu = 0,\ \sigma^2 = 2$ (scale alternative)

4) $\mu = 3,\ \sigma^2 = 2$ (location and scale alternative)

For each of them 500 sample pairs are generated. The Kullback-Leibler and the Hellinger divergence are estimated on each of them. Two Gaussian distributions

with means μ and ν and variances σ^2 and τ^2, respectively, result in a symmetric Kullback-Leibler divergence given by

$$D_{KL}(P,Q) = \frac{(\sigma^2-\tau^2)^2}{2\sigma^2\tau^2} + \frac{(\mu-\nu)^2}{2} \cdot \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)$$

and a Hellinger divergence of

$$D_H(P,Q) = 1 - \sqrt{\frac{2\sigma\tau}{\sigma^2+\tau^2}} \cdot \exp\left(-\frac{1}{4}\frac{(\mu-\nu)^2}{\sigma^2+\tau^2}\right).$$

In the data cases 1) to 4) this yields 0, 9, 1.125 and 6.75 for the Kullback-Leibler criterion and 0, 0.82, 0.17 and 0.74 for the Hellinger divergence, respectively. The density ratio estimation is conducted using the direct kernel density approach, the moment matching technique and the uLSIF algorithm. Since its computational demand is quite high and its performance is not outstanding (Sugiyama et al., 2009), we do not take the KLIEP procedure into account. Each density ratio estimate is passed to the natural divergence estimator $\hat{D}_f$ as well as to its decomposed version $\hat{D}_f^D$. The latter one relies on the decomposition presented in equation (3.6), which is optimal for the moment matching density ratio estimation. Furthermore, the smoothed estimator $\hat{D}_f^S$ is applied for both divergence measures.

To assess the performance of the estimators the empirical mean squared error (MSE) is computed over the 500 replications of each data case. We make use of this absolute error measure rather than a relative one, since under the null hypothesis the true divergence values are both zero. The results for $n = m = 300$ are given in Tables 2 and 3 in the appendix. The estimated errors for the Hellinger divergence are presented on a 10^{-4} scale, because they are much smaller than those for the Kullback-Leibler divergence. This is caused by the boundedness of the Hellinger divergence. As for the measure itself, the corresponding estimates typically lie within $[0, 1]$ leading to a small empirical MSE in comparison to the unbounded

Kullback-Leibler measure.

According to the results, higher divergence values are more difficult to estimate than smaller ones. This becomes in particular clear focussing on data case 2). Situation 4) seems more difficult than 2) at first glance, since more variability is introduced by a higher variance of the second distribution. However, the estimated MSEs here are mostly lower than in case 2), which leads to the highest errors overall. Note that higher divergence values indicate a density ratio with more high and more low values. Thus, the density ratio estimation is more difficult and larger errors in the divergence estimation become more likely.

Among the density ratio estimators, the moment matching algorithm unsurprisingly leads to the best results overall. Since the correct density ratio model is specified, this semiparametric approach makes use of addition information in comparison to its competitors. In contrast, the uLSIF algorithm leads to extreme estimations and hence achieves the worst results. Sugiyama et al. (2009) stressed its good performance for multidimensional problems, but in the univariate case we find the other methods to estimate the true divergence values better. Among the nonparametric procedures in case of the Kullback-Leibler divergence, the smoothed estimator $\hat{D}_f^S$ outperforms both divergence estimators relying on the direct kernel density approach. The latter lead to huge overestimations, while $\hat{D}_f^S$ behaves more stable. In the most realistic sample case 4) it even attains the smallest MSE of all methods considered. For the bounded Hellinger divergence the decomposed estimator using the direct kernel density estimation leads to slightly better results than the smoothed estimator, which performs quite well overall. In general decomposing drastically improves the estimators' performance in the majority of the cases and never leads to huge increases of the MSE. This holds for all methods and not just for the moment matching it was proposed for.

3.3 Testing Homogeneity based on Divergences

In this section we study two-sample tests using divergence measures to test $\mathbb{H}_0 : P = Q$. At first, we review an asymptotic method by Kanamori et al. (2012), which relies on a semiparametric divergence estimator. Hereafter, we construct alternative tests via the permutation technique. They can be conducted using arbitrary divergence estimators. In the second part of the section, all tests procedures introduced so far are compared to some parametric and nonparametric competitors in a broad simulation study. The methods performing best are applied to ion mobility spectrometry data also investigated in Section 2.5.4.

3.3.1 Divergence-based Tests

Kanamori et al. (2012) propose an asymptotic test for the two-sample homogeneity problem relying on divergence measures. They estimate the density ratio via the semiparametric moment matching. The divergence of choice is then estimated in the subsequent step by the decomposed estimator $\hat{D}_f^D$. Hereby, they rely on the moment function η^* and the decomposition functions f_1^* and f_2^* presented in (3.3) and (3.6), respectively. The authors prove that under the null hypothesis $\mathbb{H}_0 : P = Q$ the test statistic

$$T = \frac{2 \cdot n \cdot m}{(n+m) \cdot f''(1)} \cdot \hat{D}_f^D$$

is asymptotically chi-square distributed with $(d-1)$ degrees of freedom for any divergence measure D_f. Hereby, d denotes the dimension of the parameter vector θ in the density ratio model $r(\cdot, \theta)$ and f'' is the second derivative of the convex function f specifying the divergence. The corresponding homogeneity test is referred

to as the Kanamori test from here on.

Research in various fields shows that the permutation principle introduced by Fisher (1935) and its extensions can lead to quite powerful tests, cf. Cardot et al. (2007), Sohn et al. (2012) and Zeileis and Hothorn (2013) and the references given therein. Motivated by these facts, we propose the following distribution-free procedure to test the null hypothesis: given the original sample pair $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, first generate $n_p \geq 500$ new sample pairs from the original data. For this purpose, draw n of the $n+m$ observations from the original joint sample at random without replacement yielding a new first sample. The remaining m observations form the new second sample. In other words, the group labels of the original samples are permuted at random. Repeating this procedure n_p times results in n_p new sample pairs. Next, the divergence of choice is estimated on each of the n_p sample pairs as well as on the original data always using the same estimator. This leads to $n_p + 1$ divergence estimates. Under $\mathbb{H}_0 : P = Q$ all of them stem from identically distributed random variables. Recall that the true divergence value for any convex function f is minimal under the null hypothesis. Thus, a permutation test based on a divergence estimator rejects the null hypothesis, if the divergence estimate on the original data exceeds the empirical $(1-\alpha)$-quantile of the $n_p + 1$ divergence estimates, where α is the predefined significance level. In contrast to the Kanamori test, the permutation procedure leads to a valid testing procedure for all sample sizes and also does not impose any distributional assumptions on the data. In addition, it allows to use arbitrary estimators for testing. As we show in the next section, these advantages lead to results superior to the combination of the moment matching and the decomposed estimator in several settings.

3.3.2 Comparison of Divergence-based Tests

Since the Kanamori test is an asymptotical procedure, we assess the minimum sample size it requires to hold the nominal significance level of $\alpha = 5\%$ prior to a comparison of the homogeneity tests. For this purpose, the method is applied to 500 pairs of equally sized samples drawn from the standard Gaussian distribution for different sample sizes and both divergence measures. An analogue simulation is conducted with data generated from the exponential distribution with mean 1. Kanamori et al. (2012) also check the convergence rate of their method in a similar simulation, but focus on the multidimensional rather than the univariate setting. As mentioned before, the exponential density ratio model given in (3.4) is adequate for two Gaussian or two exponential distributions. However, it is overparametrised in the latter case. In order to quantify the effect of this overparametrisation equivalent tests are performed using the reduced exponential model

$$r_{re}(x, \theta) = \exp(\theta_1 + \theta_2 \cdot x) \tag{3.8}$$

for the exponential data. In comparison to the exponential model it lacks a quadratic term. To analyse the behaviour of the Kanamori test we consider its empirical size as a function of the sample size in Table 4. The results illustrate the strong impact of the density ratio model quite well. For the correctly specified models presented in rows a) and c) 150 or at most 250 observations are sufficient to ensure a proper test procedure. The rejection rate for the overparametrised exponential model in line b) converges much slower to $\alpha = 5\%$ and leads to too many false rejections. The results also indicate a faster convergence of the test using the Hellinger distance compared to the Kullback-Leibler version.

We now compare the empirical power of several homogeneity tests in different simulation scenarios proceeding as follows: at first, a certain data setting such as "location alternatives for Gaussian distributions" is chosen. The distribution P is fixed, while the parameters of the second distribution Q vary on a grid reflecting different degrees of discrepancy. For each of these parameter constellations 500 sample pairs of size $m = n = 50$ are generated from the respective P and Q. Then, several tests are applied to each of the sample pairs. Finally, the empirical rejection rate is computed for each test and each parameter constellation. The tests include six divergence-based permutation tests as well as other parametric and nonparametric competitors from the literature. As illustrated at the beginning of this section, the asymptotic Kanamori test is not a valid procedure for small sample sizes. We thus repeat the simulation for $m = n = 300$ and hereby apply the Kanamori test for both divergence measures. In exchange, some of the permutation tests performing similar to others for $m = n = 50$ are excluded in the large sample case. Due to the huge amount of results we do not list all rejection rates for the small and the large sample case in all settings. Instead, we give some representative examples for qualitatively similar results and summarize the main conclusions.

Before going into detail with regard to the data settings, we specify the tests investigated. The testing via the permutation approach is conducted using six different divergence estimators and $n_p = 500$ permutations. The divergence estimators considered are the smoothed estimators $\hat{D}^S_{KL}$ and $\hat{D}^S_H$, the natural estimators $\hat{D}_{KL}$ and $\hat{D}_H$ as well as the decomposed estimators $\hat{D}^D_{KL}$ and $\hat{D}^D_H$. The corresponding density ratios are estimated by the direct kernel density approach. The semiparametric uLSIF algorithm is omitted due to its high computational demand and its modest results in Section 3.2.3. We also do not consider permutation tests based on the density ratio estimation by moment matching due to the large amount of simulations. Nevertheless, this technique is investigated, because it

is applied in the Kanamori test studied in the large sample case. In addition to the six permutation tests, we apply the nonparametric Wilcoxon rank-sum test (Wilcoxon, 1945), the Anderson-Darling test (Anderson and Darling, 1952) and the Kolmogorov-Smirnov test, cf. page 13. The first primarily detects location alternatives. The other two reveal arbitrary deviations from the null hypothesis and are motivated by distribution functions. If appropriate, we also include optimal distribution specific tests like the F-test and the t-test. In particular when dealing with exponential distributions, a two-sided parametric test is considered. It is based on two one-sided tests and rejects the null hypothesis $\mathbb{H}_0 : P = Q$ if and only if one of the one-sided tests rejects $\mathbb{H}_0$. The one-sided tests are optimal for testing $\lambda_P > \lambda_Q$ and $\lambda_P < \lambda_Q$, respectively, whereby λ_P and λ_Q denote the parameters of the exponential distributions P and Q. The statistic of the tests is the ratio of the sample means, which follows an F-distribution under $\mathbb{H}_0$, see Lee et al. (1975). Both one-sided tests are carried out at a significance level of 2.5% to ensure the global significance level of $\alpha = 5\%$. This method as well as the Kanamori test are implemented by the authors. All other competitors of the permutation procedures are conducted using the implementations in the R packages *stats* and *adk*.

We now explain the data scenarios under study. Hereby, we always list the parameters considered in the small sample simulation and give the corresponding values for the large sample simulation in brackets. At first, we simulate choosing both P and Q as Gaussian distributions. While P is the standard Gaussian distribution, random variables with distribution Q have mean μ and variance σ^2. For location alternatives we fix $\sigma^2 = 1$ and vary $\mu = -1, -0.9, \ldots, 0.9, 1$ ($\mu = -0.5, -0.45, \ldots, 0.45, 0.5$). Scale alternatives are studied setting $\mu = 0$ and changing the values of $\sigma^2 = 0.1, 0.2, \ldots, 1.9, 2$ ($\sigma^2 = 0.5, 0.55, \ldots, 1.45, 1.5$). In order to investigate simultaneous discrepancies in location and scale, the mean and variance are linked using $\mu = \theta - 1$ and $\sigma = \theta$ for $\theta = 0.1, 0.2, \ldots, 1.9, 2$

($\theta = 0.5, 0.55, \ldots, 1.45, 1.5$). Analogous simulations are performed for the family of scaled t-distributions with 5 and 20 degrees of freedom, respectively. Some representative rejection rates for these settings are given in Tables 6 and 7. In addition, Table 5 provides the empirical sizes for the large sample case. Since the null hypothesis is included in the location, the scale and the location and scale design and each of them is replicated 500 times, the results are based on 1500 replications.

In a second step, we evaluate the performance of the methods in case of skewness alternatives making use of the skewed Gaussian distribution class (Azzalini, 1985). The skewness of the corresponding random variables is regulated by the parameter $\tilde{\lambda}$. For $\tilde{\lambda} = 0$ the skewed Gaussian distribution coincides with the standard Gaussian. For negative (positive) values of $\tilde{\lambda}$ it is left-skewed (right-skewed). Note that a skewed Gaussian random variable does not have mean 0 and variance 1 for $\tilde{\lambda} \neq 0$. We therefore always generate data from a standardised skewed Gaussian distribution Q for $\tilde{\lambda} = -50, -40, \ldots, 40, 50$ ($\tilde{\lambda} = -5, -4, \ldots, 4, 5$) and compare it to observations drawn from the standard Gaussian distribution P. The results for the large sample case in this scenario are presented in Table 8.

Next, we investigate the methods' capability of detecting departures from the Gaussian distribution in terms of heavy tails. P is again set to the standard Gaussian distribution. Q is chosen as a t-distribution with a number of degrees of freedom ν varying from 3 to 10 for $m = n = 50$ and $m = n = 300$. As for the skewness, we draw data from a standardised version of Q, so that P and Q neither differ in location nor in scale. The corresponding results are listed in Table 9.

As shown in Lindsay (1994), the Hellinger divergence allows to construct a robust and first-order efficient parameter estimator. Motivated by this fact we investigate the robustness of the tests with respect to outliers. For this purpose, we set P to the standard Gaussian distribution. Q is chosen as the mixture of the standard

Gaussian distribution and another Gaussian distribution with mean μ and variance 1 to a proportion of ε^* and $1-\varepsilon^*$, respectively. The parameter μ is set to 0.5 (0.1). We consider $\varepsilon^* = 0, 0.05, 0.1, 0.2, 0.3$ to illustrate the effect of outliers under the null hypothesis and $\varepsilon^* = 1, 0.95, 0.9, 0.8, 0.7$ to assess the effect of outliers under the alternative. The corresponding results are given in Tables 10 and 11.

Finally, we analyse the case of two exponential distributions. P is fixed to have mean $\lambda_P = 1$. The mean of Q, $\frac{1}{\lambda_Q}$, is chosen by $\lambda_Q = 0.2, 0.3, \ldots, 1.7, 1.8$ ($\lambda_Q = 0.6, 0.7, \ldots, 1.5, 1.4$). Representative rejection rates for the exponential data scenario are summarised in Table 12.

According to the rejection rates for the parametric methods, the t- and F-test, as expected, perform best under Gaussianity for discrepancies in location and scale, respectively. However, they reject $\mathbb{H}_0$ quite rarely if their specific alternative is not met, cf. Tables 8 and 9. As illustrated in Table 5, the F-test is more affected by an incorrect distributional assumption and does not hold the significance level for non-gaussian data. As opposed to that, the t-test becomes conservative when applied to data generated by a t-distribution. In the exponential setting, the parametric test consisting of two one-sided optimal tests proposed by Lee et al. (1975) also attains the highest rejection rates.

Among the classical nonparametric procedures, the Anderson-Darling test achieves better results than the Kolmogorov-Smirnov test in almost every case investigated. Although both asymptotic tests are applicable for samples of size 50 already, they still reject $\mathbb{H}_0$ in less than 5% of the cases for the t-distribution with 5 degrees of freedom even for $m = n = 300$. Both of them detect various kinds of discrepancies between the distributions in contrast to the Wilcoxon test, which mainly reveals location alternatives. The latter is solely superior to the Anderson-Darling test if the samples differ in location only.

With regard to the permutation tests, the ones based on the Hellinger divergence

perform better than their Kullback-Leibler counterparts in most cases. However, the discrepancies are not large even in the outlier scenarios (Table 10 and 11). This is somewhat surprising, because the Hellinger divergence leads to more robust parameter estimators than the maximum likelihood estimator corresponding to the Kullback-Leibler divergence (Lindsay, 1994). The divergence estimation technique appears to be much more crucial for the performance of the tests. The methods using the smoothed estimator or the decomposed estimator lead to similar and stable results. The ones relying on the natural estimators $\hat{D}_H$ and $\hat{D}_{KL}$ perform quite differently, see Tables 6 and 12. For example in the setting of different scales, they detect departures from the null hypothesis more often if the variance of the second sample, 1.5, exceeds the one of the first, which is 1. However, they reject rarely compared to other methods in the opposite case, where the variances are 0.5 and 1. This appears counterintuitive due to the relative size of the variances and could be caused by the asymmetry of the estimation procedure discussed on page 66. Overall, the decomposed and smoothed estimators lead to higher rejection rates in most of the cases under study.

All in all, permutation tests using divergence estimators detect discrepancies between distributions less often than the Wilcoxon test, the Kolmogorov-Smirnov test and the Anderson-Darling test, if the corresponding samples differ primarily in location. More precisely, the nonparametric procedures outperform the divergence tests only for the location and the exponential setting. In all other cases studied the tests based on the smoothed divergence estimator and the decomposed estimator attain at least competitive and often considerably higher empirical powers. Especially in situations where the means of the distributions are equal the advantages of the divergence procedures are striking. This holds for the scale and the skewness setting as well as for the comparison of Gaussian to t-distributed random variables. This behaviour results in less proneness to outliers under the null hypothesis (Table

10), but lower rejection rates under contaminated location alternatives (Table 11) in comparison to the Anderson-Darling test, the overall best classical procedure. The asymptotic Kanamori test shows even better results as long as the exponential density ratio model is correct. However, if the model is inadequate, it does not hold the nominal significance level and leads to considerably worse results than the permutation tests, cf. Table 5, 8 and 9.
Since the two best permutation tests using $\hat{D}_H^S$ and $\hat{D}_H^D$ lead to quite similar results, they are evaluated in terms of running time. We apply them to equally large samples of varying size $n = m = 50, 100, 200 \ldots, 1\,000$ and determine the mean computation time over 200 replications for each sample size. All tests are conducted using 1 000 permutations and data stemming from the standard Gaussian distribution in both samples. The runtime in the case of different Gaussian distributions is also investigated and is essentially the same. According to the results given in Table 13, the test based on $\hat{D}_H^S$ is always considerably faster than the decomposed estimator $\hat{D}_H^D$. Its runtime also increases notably slower in the sample size. Since both methods lead to comparable rejection rates in our simulations, we recommend the smoothed divergence estimator for applications.

3.3.3 Application to Biometrical Data

To assess how our tests perform on real data we consider the ion mobility spectrometry (IMS) measurements studied in Section 2.5.4. The homogeneity tests are applied to compare the peak modelling proposed by D'Addario et al. (2014) to the corresponding datasets. As before, we focus on one of the dimensions and condition on the other. This time we investigate 500 spectrograms and generate 500 observations from each spectrogram. In addition, we sample an equal amount of data from each mixture modelling a corresponding spectrogram. The permutation

test based on the smoothed divergence estimator $\hat{D}_H^S$ and the Anderson-Darling test are applied to each of these 500 dataset pairs.

In general, the results for both tests suggest that the inverse Gaussian models fit the spectrograms quite well. The null hypothesis of equal distribution is rejected for only 62 and 51 of the 500 spectrograms, respectively. For 91 spectrograms the tests come to different conclusions. We illustrate two of these 91 situations by looking at kernel density estimates associated with the spectrograms and the corresponding mixture model in Figure 8.

Most of the 91 cases are unimodal or almost unimodal like spectrogram A. In some of them the Anderson-Darling test rejects the null hypothesis, while the divergence test does not, and vice versa. Presumably, most of them are false rejections of one or the other test. For all of the few multimodal situations similar to spectrogram B the Anderson-Darling test does not reject $\mathbb{H}_0$ in contrast to the divergence test. Since the discrepancies between the densities in spectrogram B look notably larger than in spectrogram A, the test based on $\hat{D}_H^S$ seems preferable to the Anderson-Darling test. These results also go well with our impressions based on the simulation study. The Anderson-Darling test has problems, if the samples differ in shape but not in location, while the divergence-based test detects such discrepancies more often.

3.4 Conclusions and Extensions

Finding out whether density-based procedures are beneficial in the context of homogeneity tests is the key motivation to the work presented in this chapter. Given the results in Sections 3.3.2 and 3.3.3 we can state quite surely this is the case. Permutation tests relying on stable estimators of f-divergences do not require any assumptions on the underlying distributions and are therefore widely applicable.

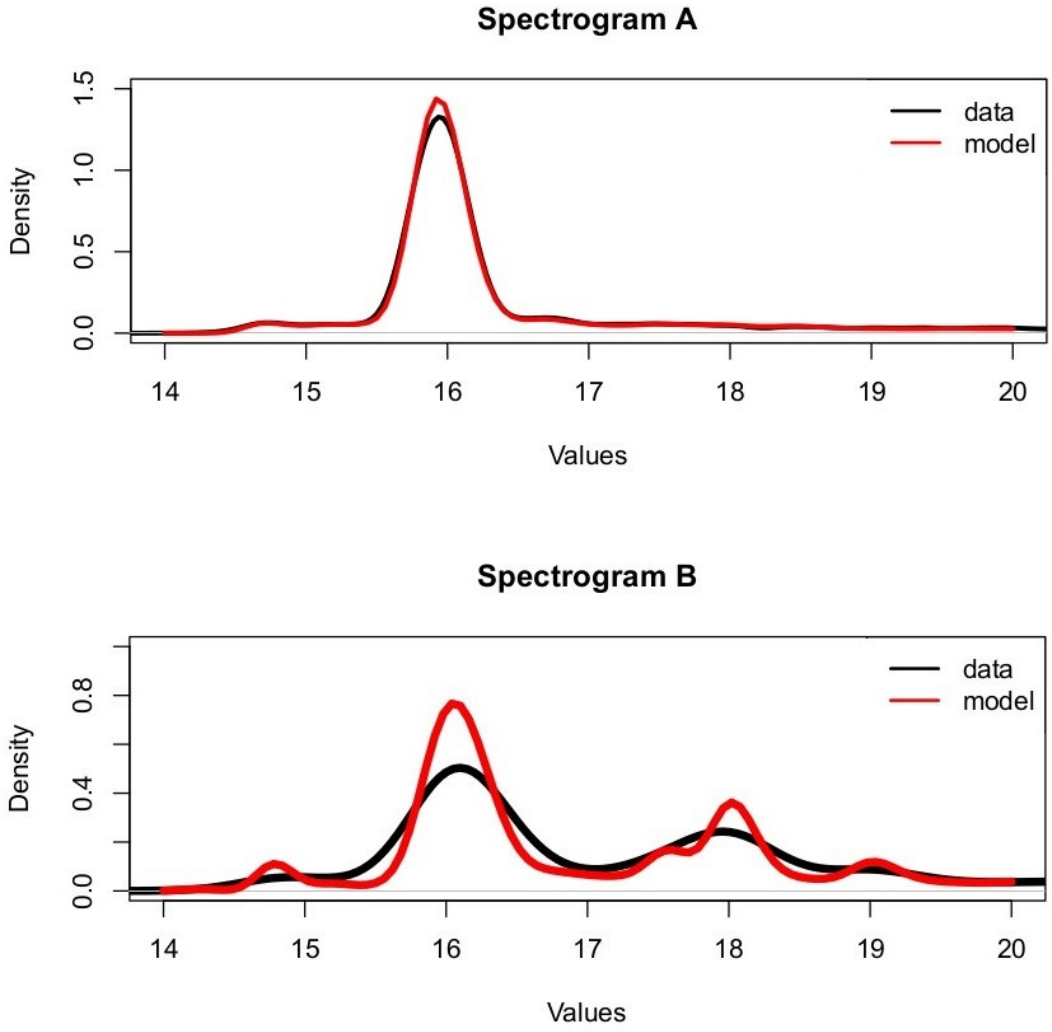

Figure 8: Kernel density estimates based on the measurements and the corresponding fitted model for two spectrograms.

As long as the samples under study do not predominantly differ in location, they clearly outperform classical homogeneity tests like the Kolmogorov-Smirnov and the Anderson-Darling test. If this assumption is justifiable in an application, we heavily recommend to use the permutation test based on the stable smoothed estimator of the Hellinger divergence. This is in particular the case, if it is not clear what kind of discrepancies can be expected or the samples differ in several ways. A combination of our method and the Anderson-Darling test via a Bonferroni correction (Bonferroni, 1937) might also be a reasonable option.

There are several possibilities to extend the approach presented in this chapter. First, one could consider other divergences or density-based dissimilarity measures. In the context of parameter estimation the Hellinger divergence has better robustness properties than the Kullback-Leibler divergence. However, in the nonparametric setting studied here both measures lead to comparable results. Nevertheless the choice of divergence is expected to have some influence of the performance of the corresponding test in general. A natural first idea is to investigate other divergences with beneficial properties in the case of parameter estimation. One candidate is the class of blended weighted Hellinger divergences (Basu and Lindsay, 1994) extending the Hellinger measure. A second way to expand this work is to adapt the procedures to the multidimensional case. According to the studies of Sugiyama et al. (2009) the estimation of divergences via the natural approach and kernel density estimators performs poorly for multiple dimensions. However, multivariate smoothing techniques or combinations of numeric integration with the uLSIF algorithm might be worth considering. We studied a first version of the latter proposal conducting density ratio estimation via uLSIF, kernel density estimation of the density q and numerical integration of the resulting function. Unfortunately, in the univariate case this estimator gave poor results even for a correctly specified density ratio model and has therefore been omitted in this work.

Another possible point of improvement is the robustification of the estimators presented here. A robustified divergence estimation procedure can be constructed using robust kernel density estimators as for example proposed by Kim and Scott (2012). It might also be helpful to develop a sophisticated approach for determining the region of integration or find a suitable weighting of the observations for the decomposed estimator.

The concepts introduced in this chapter could also be transfered to the detection of structural breaks in time series. Temporal data plays a huge role in many different fields of application such as biometry (vital signs, disease progression, outbreak of epidemics), quality control (survival analysis) and finance (exchange rates, returns). Therefore, there is a strong need to develop new and improve old methods in the context of time series. We do not further pursue this goal using density-based methods. Instead, we turn to procedures closely related to Fourier transform and characteristic functions in the next chapter. Just like probability density functions, characteristic functions uniquely characterise the corresponding distribution and do not focus on particular moments such as the mean or the variance. They are thus flexible tools allowing to construct distribution-free tests with competitive power.

4 Testing Time Series for a Constant Volatility

In this chapter we develop a new procedure for testing the null hypothesis that a given sequence of variables has constant volatility over time. First, we sketch the framework in Section 4.1. Several test statistics resulting from different weight functions are proposed in Section 4.2. They are based on a Fourier-type transformation of the volatility process, which is assumed to be piecewise constant. The corresponding testing approach as well as several competitors are introduced in Section 4.3. They are compared conducting extensive Monte Carlo experiments in Section 4.4. As it turns out, our proposals have high power in particular in the case of multiple changes of the volatility and locate structural break positions adequately. The best methods are applied to exchange rate data. In Section 4.5 the construction principle used in Section 4.2 is transfered to obtain a test for structural breaks in kurtosis. The performance of the new kurtosis tests is also investigated in a simulation study. Section 4.6 highlights the main results of the chapter and provides an outlook on possible extensions. This chapter is based on the manuscript Wornowizki et al. (2015). The basic idea for the approach presented in the following was proposed by S. Meintanis and refined and investigated by me and my thesis advisor R. Fried.

4.1 Framework

Let us consider a real-valued stochastic process $X(\cdot)$ at times $t = 1, \ldots, n$ for some $n \in \mathbb{N}$. Denote the associated volatility process by $\sigma^2(\cdot)$, where $\sigma^2(t) = \mathrm{Var}(X(t))$

is the variance of $X(t)$ at each of the times $t = 1, \ldots, n$. We are interested in testing whether $\sigma^2(\cdot)$ is constant as expressed by the hypotheses

$$\mathbb{H}_0 : \forall t, t' = 1, \ldots, n : \sigma^2(t) = \sigma^2(t') \text{ vs. } \mathbb{H}_1 : \exists t, t' = 1, \ldots, n : \sigma^2(t) \neq \sigma^2(t'). \quad (4.1)$$

Often the volatility is thought to vary permanently, for instance according to a GARCH or another stochastic volatility model. Since this assumption is controversial, we adopt an alternative idea of a blockwise constant volatility, which has drawn attention in Mercurio and Spokoiny (2004), Stărică and Granger (2005), Vassiliou and Demetriou (2005), Spokoiny (2009), Davies et al. (2012) and Fried (2012), among others. Within this concept, some specified time points $0 = t_0 < t_1 < \ldots < t_N = n$ are understood as potential change point positions. If there is no external knowledge allowing to choose them appropriately, one can set $t_0, \ldots, t_N$ equidistantly. The potential change point positions correspond to important events, which may trigger an upward or downward change in the volatility. The values of the volatility process $\sigma^2(\cdot)$ are thus allowed to differ for some of the time blocks $B_j = \{t_{j-1} + 1, t_{j-1} + 2, \ldots, t_j\}$, $j = 1, ..., N$. However, within each time block the volatility is assumed to be constant. In addition, we work with independent zero mean random variables $X(1), \ldots, X(n)$, which are identical distributed up to scale. The latter means that for some unknown but fixed distribution function F the relation $P(X(t) \leq x) = F(x/\sqrt{\sigma^2(t)})$ holds for all $t = 1, \ldots, n$ and all $x \in \mathbb{R}$. The centered Gaussian distribution is thus contained in our framework as a special case. Heavy tails, for example often encountered in financial applications, are also included. The zero mean assumption is justifiable when dealing with returns or comparable data, which is obtained from differences of consecutive observations. In other cases it can be relaxed to blockwise constant means with known block structure, so that zero mean data results from prepro-

cessing. For a detailed discussion of the other assumptions in comparison to the classical GARCH framework see Spokoiny (2009) and the references given therein.

4.2 Test Statistics for Volatility Changes

In this section a new class of test statistics allowing to test the constancy of the volatility process is introduced. We then derive explicit formulas for some representatives of this class and discuss alternative class members.

To test $\mathbb{H}_0$ specified in (4.1) a reasonable first step is the estimation of the volatility in each time block B_j, $j = 1, \ldots, N$. Since all random variables are assumed to have zero mean, a natural volatility estimator for the j-th block is given by

$$\widehat{\sigma}_j^2 = \frac{1}{\tau_j} \sum_{t \in B_j} X(t)^2 \quad \forall j = 1, \ldots, N. \tag{4.2}$$

Hereby, $\tau_j = |B_j|$ denotes the number of observations with observation times in the time block B_j, $j = 1, \ldots, N$. To ensure volatility estimates with reasonable accuracy a sufficient number of observations should be included in each block.
Instead of considering the estimated volatilities themselves, we rather work with their logarithms in the following. As demonstrated later, this allows us to construct scale independent test procedures. Let i denote the imaginary number defined by $i^2 = -1$. Under the null hypothesis specified in (4.1) the logarithmised volatility process $\log(\sigma^2(\cdot))$ is constant. Thus, under $\mathbb{H}_0$ the function $\varphi : \mathbb{R} \times \{1, \ldots, n\} \to \mathbb{C}$

defined by $\varphi(u,t) = e^{iu\log(\sigma^2(t))}$ does not depend on t. Hence, for any $t = 1,\dots,n$ it can be estimated by

$$\widehat{\varphi}(u) = \sum_{j=1}^{N} \frac{\tau_j}{n} e^{iu\log\left(\widehat{\sigma}_j^2\right)}$$

in a straightforward way. Hereby, each $\tau_j/n = \tau_j/(\sum_{i=1}^{N}\tau_i)$ weights the corresponding term derived from the j-th block according to the number of observations in the block, $j = 1,\dots,N$. If all blocks contain the same number of observations, all weights are equal to $\frac{1}{N}$. The transformation of the blockwise estimators induced by φ is closely related to the Fourier transformation and to characteristic functions. In this situation it has the nice and intuitive behaviour illustrated in Figure 9.

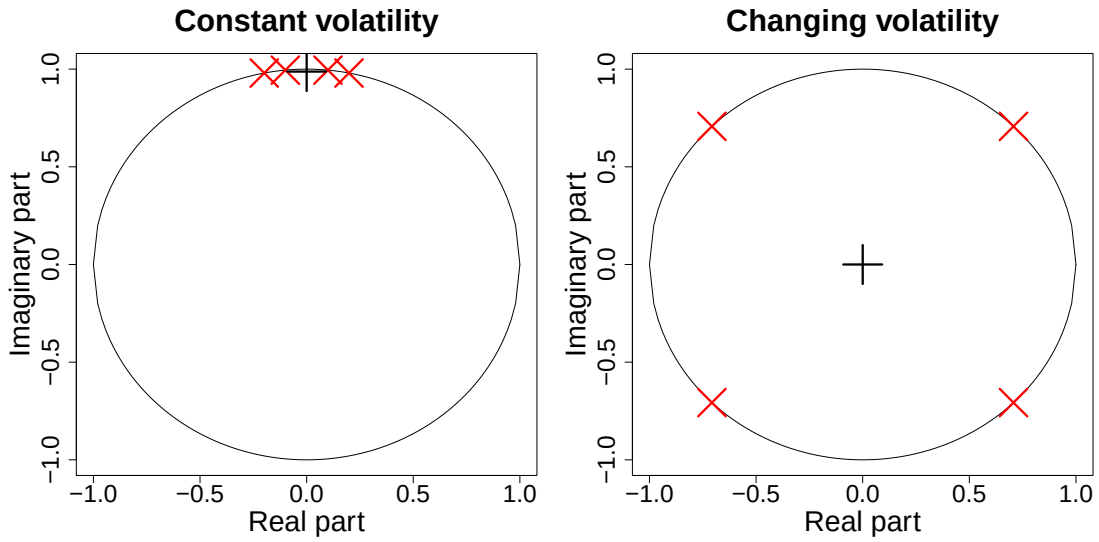

Figure 9: Behaviour of $\hat{\varphi}(u)$ for some fixed $u \in \mathbb{R}$ under the null hypothesis of constant volatility (left) and under an alternative (right), cf. Section 4.2. The red crosses represent the blockwise volatility estimates mapped to the unit circle. The black crosses mark the corresponding values of $\hat{\varphi}(u)$.

Under the null hypothesis of global constant volatility the blockwise estimates are about the same. Because of that, the function $f : \mathbb{R} \to \mathbb{C}, f(x) = e^{iux}$ maps them to the red points on the unit circle close to each other, as presented in the left part

of the figure. Consequently, their weighted mean $\widehat{\varphi}(u)$ depicted by the black cross lies near the unit circle for every $u \in \mathbb{R}$ and has a modulus close to one. Under an alternative some of the logarithmised blockwise estimates differ. They are thus mapped to distant points on the unit circle for most $u \in \mathbb{R}$, see the right part of Figure 9. Hence, for most $u \in \mathbb{R}$ their weighted mean $\widehat{\varphi}(u)$ is closer to the origin than under the null hypothesis. In these cases $\widehat{\varphi}(u)$ has a comparatively small modulus. In view of this fact, we consider test statistics of the form

$$V = \int \left(1 - |\widehat{\varphi}(u)|^2\right) w(u)du$$

to test for a global constant volatility. V is nonnegative because of $|\widehat{\varphi}(u)|^2 \leq 1$. The latter obviously holds, since a weighted mean of points at the unit circle cannot lie outside the unit circle. The weight function $w : \mathbb{R} \to \mathbb{R}_0^+$ is chosen such that a finite test statistic is ensured.

In order to handle the integral in our test statistic we define $W = \int w(u)du < \infty$ for an integrable weight function w. In this case V can be rewritten to

$$V = W - \int |\widehat{\varphi}(u)|^2 w(u)du. \tag{4.3}$$

Since W is independent of the data, it can be dropped. Using the definition of $\widehat{\varphi}$ the integral in (4.3) reduces to

$$T_{Four} = \frac{1}{n^2} \sum_{j,k=1}^{N} \tau_j \tau_k I_w \left(\log\left(\widehat{\sigma}_j^2\right) - \log\left(\widehat{\sigma}_k^2\right)\right), \tag{4.4}$$

where

$$I_w(x) = \int \cos(ux) w(u)du.$$

Since small values of V support the null hypothesis, $\mathbb{H}_0$ is rejected for small values of T_{Four}. This test statistic depends on the data only via the terms $\log\left(\widehat{\sigma}_j^2\right) - \log\left(\widehat{\sigma}_k^2\right) = \log\left(\widehat{\sigma}_j^2/\widehat{\sigma}_k^2\right)$ for $1 \leq j < k \leq N$. Therefore, thanks to the logarithm, any scale factor is canceled out and T_{Four} is scale invariant. This is the reason why we logarithmise the estimated volatilities.

The function I_w can be expressed explicitly for several standard choices of w presented in Hušková and Meintanis (2006a). These are the uniform, the Laplace and the Gaussian weighting with corresponding weight functions

$$w_U(u) = \mathbf{1}_{(-a,a)}(u), \quad w_L(u) = e^{-a|u|} \quad \text{and} \quad w_G(u) = e^{-au^2},$$

respectively. All of them are integrable and depend on a parameter $a > 0$. Straightforward computations lead to

$$I_{w_U}(x) = \frac{2\sin(ax)}{x}, \quad I_{w_L}(x) = \frac{2a}{a^2 + x^2} \quad \text{and} \quad I_{w_G}(x) = \sqrt{\frac{\pi}{a}} \exp\left(\frac{-x^2}{4a}\right)$$

for the uniform, Laplace and the Gaussian weight function, respectively. There also exist alternative choices for w. One example is the data adaptive weighting scheme proposed by Meintanis et al. (2014) for goodness-of-fit testing. Another weight function is studied in Matteson and James (2014) in the context of multivariate nonparametric detection of general distributional changes. These two types of weight functions were also considered in the simulations. We do not provide the corresponding results in the following, because the weight function does not have a large impact on the performance of the proposed tests. This is illustrated in Section 4.4.1 for the three standard weighting schemes w_U, w_L and w_G and also holds for these weight functions. In our simulations the data adaptive weighting leads to slightly worse and the weighting proposed by Matteson and James to essentially the same rejection rates as the standard weight functions.

4.3 Testing for a Global Constant Volatility

In this section we show how time series can be tested for a global constant volatility. First, a simple approach using the statistics defined in Section 4.2 is presented. Hereafter, a natural estimator of the structural break position in case of a rejection is defined. The procedure allows to locate multiple presumable structural break positions, which is quite valuable in applications. The section closes by introducing alternative methods for the testing problem taken from or inspired by the literature.

4.3.1 Testing Procedure

The distribution of the test statistic T_{Four} heavily depends on the distribution of the random variables $X(1), \ldots, X(n)$. Getting critical values without imposing distributional assumptions is thus impossible at least for small sample sizes. As shown for the divergence estimators in Chapter 3, the permutation principle introduced by Fisher (1935) can be of great help in such situations. As an additional motivation, note that Hušková and Meintanis (2006b) successfully make use of it testing for general structural changes of the distribution in temporal data using characteristic functions, a topic closely related to our procedure. Since under the null hypothesis the observations stem from identically distributed random variables, the approach described on page 72 can be adapted to our framework. We thus determine the test statistic T_{Four} on the original sample as well as on each of its n_p permutations always assuming that each sample is observed in the given order. Thereby, the same parameters N, w, a and block lengths $\tau_1, \ldots, \tau_N$ are used for all computations. The permutation procedure rejects $\mathbb{H}_0$ at the predefined significance level α, if the test statistic determined on the original sample falls below the empirical α-quantile of all $n_p + 1$ test statistics.

4.3.2 Localisation of Presumable Structural Breaks

If the tests procedure proposed in 4.3.1 rejects $\mathbb{H}_0$, we are interested in locating the first presumable change point. A rough approximation for this position is t_{j^*}, where $j^* = \text{argmax} \left|\log\left(\widehat{\sigma}_j^2\right) - \log\left(\widehat{\sigma}_{j+1}^2\right)\right|$ and the maximisation is performed over the blocks $j = 1, \ldots, N-1$. Unfortunately, the resolution of this estimator is limited by the block lengths. This is in particular problematic, if the potential change point positions $t_1, \ldots, t_N$ are not determined by a priori knowledge. In order to alleviate this problem the presumable change point position can be fine tuned if desired. To achieve this we work with the union of the two blocks around the rough estimate t_{j^*}, $B = \{t_{j^*-1} + 1, t_{j^*-1} + 2, \ldots, t_{j^*}, \ldots, t_{j^*+1}\}$ for $j^* = 1, \ldots, N-1$. For each $t \in B$ the volatility before and after t is estimated analogously to (4.2) using observations from B only. Let us denote them by $\widehat{\sigma}_1^2(t)$ and $\widehat{\sigma}_2^2(t)$. The position of the presumable structural break is then estimated by

$$\underset{t}{\text{argmax}} |\log(\widehat{\sigma}_1^2(t)) - \log(\widehat{\sigma}_2^2(t))|.$$

Hereby, the maximisation is performed over all $t \in B$ far enough away from the bounds of B so that a meaningful estimations of the local volatility is ensured. In our implementation of the method we always leave out the five observations closest to each bound of B.

Multiple structural break positions are located in a recursive manner in the spirit of Vostrikova (1981). After identifying the first presumable change point as described above, the sample is split into two parts at that point. The test procedure is then repeated on each of the subsamples large enough to ensure reasonable estimations. In case of new rejections, corresponding presumable change points are determined and the splitting continues. As soon as no splitting is performed

anymore, the current data blocks seem homogeneous and the method stops. This testing procedure attains a predefined significance level α under the null hypothesis, since under $\mathbb{H}_0$ the permutation test conducted on the full sample rejects in α percent of the cases.

4.3.3 Alternative Methods

The idea of using Fourier-type transforms and characteristic functions in change point detection is not new. Work in this context is often related to similar issues such as the two–sample problem (Meintanis, 2005) and the k–sample problem (Hušková and Meintanis, 2008) and empirical characteristic functions of the observations themselves are used to test for general deviations of distributions. Papers on the general change point problem also include Hušková and Meintanis (2006a,b) for change point detection with independent observations and Hlávka et al. (2012) for sequential testing in the context of autoregressive models. In all of these it is shown that methods using transformations to the complex plane are convenient from the computational point of view and lead to theoretically sound asymptotics. However, to the authors' knowledge up to now there are no methods based on such concepts specifically tailored for testing the constancy of the volatility or other features of time series explicitly.

In the following, we sketch several other approaches for this task derived from the literature. All four methods presented in this section reject the hypothesis of a global constant volatility for large values of the corresponding test statistic. The CUSUM procedure is a standard tool in the detection of structural breaks and

a lot of work is available on it. We choose the method proposed by Wied et al. (2012) as a representative for this class of tests. It relies on the CUSUM statistic

$$T_{CUS} = \max_{1\leq t\leq n} \left| \widehat{D} \frac{t}{\sqrt{n}} \left(\widehat{\sigma}^2_{1:t} - \widehat{\sigma}^2_{1:n} \right) \right|,$$

where $\widehat{\sigma}^2_{1:l}$ denotes the empirical variance of the first l observations for $l = 1, \ldots, n$. The normalising scalar $\widehat{D}$ allows to attain the asymptotic distribution. The CUSUM approach compares the discrepancies between the estimated volatility on the whole sample to all volatilities estimated on proper subsamples. It then determines the maximal deviation signaling a possible structural break. The test is designed to detect at most one change in volatility and critical values are derived from asymptotics.

As opposed to the CUSUM strategy, Peña (2005) also compares variances estimated on subsamples to a measure of volatility estimated on the complete sample. He proposes the test statistic

$$T_{Log} = n \log \left(\sum_{t=1}^{n} X(t)^2 \right) - \sum_{j=1}^{N} \tau_j \log(\widehat{\sigma}^2_j) \tag{4.5}$$

built in a blockwise manner. As for T_{Four}, its distribution under the null hypothesis also heavily depends on the data. To construct a level α test we therefore again apply the permutation principle.

Another approach testing for a global constant volatility is given by Ross (2013). It is motivated by the classical distribution-free procedure proposed by Mood (1954). The method first determines $r(1), \ldots, r(n)$, the ranks of the random variables $X(1), \ldots, X(n)$. It then splits the time series into two parts $X(1), \ldots, X(t)$ and $X(t+1), \ldots, X(n)$ for each possible split position $t = 1, \ldots, n-1$. For each of these splittings the standardised test statistic of the Mood test is calculated using

the ranks. The expected value $\mu_t = t(n^2-1)/12$ and the standard deviation $\sigma_t = \sqrt{t(n-t)(n+1)(n^2-4)/180}$ used hereby are derived under the null hypothesis. Taking the maximum over the possible split positions $t = 1, \ldots, n-1$ results in

$$T_{Mood} = \max_{t=1,\ldots,n-1} \frac{\left| \sum_{h=1}^{t} \left(r(h) - \frac{n+1}{2} \right)^2 - \mu_t \right|}{\sigma_t}.$$

Since only the ranks of the random variables contribute to the test statistic, the procedure is distribution-free. Appropriate critical values depend solely on the sample size n and can be derived by simulations. For several critical values and more details we refer to Ross (2013).

In addition, we consider the monitoring procedure based on characteristic functions introduced by Steland and Rafajłowicz (2014). Their statistics have the advantage that changes in the location process do not affect the monitoring of the volatility and vice versa. According to the authors,

$$S_j = \int \left[\left(\widehat{U}_j(u) \right)^2 + \left(\widehat{V}_j(u) \right)^2 \right] w(u) du$$

is a good estimator in the context of characteristic functions, which reflects the volatility in the j-th block, $j = 1, \ldots, N$. Hereby, w is a weight function as before. $\widehat{U}_j$ and $\widehat{V}_j$ denote the natural estimators of the real and imaginary part of the characteristic function of the random variables in the j-th block for $j = 1, \ldots, N$. They are defined by

$$\widehat{U}_j(u) = \frac{1}{\tau_j} \sum_{t \in B_j} \cos\left(u \cdot X(t)\right) \text{ and } \widehat{V}_j(u) = \frac{1}{\tau_j} \sum_{t \in B_j} \sin\left(u \cdot X(t)\right) \quad \forall j = 1, \ldots, N.$$

We adopt the monitoring procedure to the retrospective case in the following way: since the null hypothesis should be rejected for substantially different volatilities in two blocks, we propose the quantity

$$T_{cf} = \max_{1 \leq j < k \leq N} |S_j - S_k| \tag{4.6}$$

as a test statistic for the testing problem under study. The testing is carried out via the permutation principle. Note that for any $j = 1, \ldots, N$ one can rewrite S_j to

$$S_j = \frac{1}{\tau_j^2} \sum_{t,t' \in B_j} I_w \left(X(t) - X(t') \right).$$

Therefore, using (4.4) our statistic T_{Four} can be interpreted as a weighted version of S_j computed on the pseudo observations $\log\left(\widehat{\sigma}_1^2\right), \ldots, \log\left(\widehat{\sigma}_N^2\right)$.

4.4 Comparison of Volatility Tests

In this section the new Fourier-type tests are compared to the competitors described in Section 4.3.3. This is achieved by determining the rejection rates of all methods in different data scenarios. We thereby address the choice of parameters and weighting functions. The best methods are applied to exchange rate data, along with a GARCH approach.

4.4.1 Choice of Weighting Scheme and Block Sizes

As a first step of the analysis, we asses the influence of the weight function w, its parameter a and the number of the blocks N on the two tests using weight

functions. These are the ones based on the Fourier-type statistic T_{Four} and the statistic T_{cf} derived from characteristic functions defined, see (4.4) and (4.6). Since both methods are constructed using the permutation principle, they attain a predefined significance level α under the null hypothesis of global constant volatility. Therefore, their empirical powers under alternatives are adequate performance measures. We thus generate 1 000 datasets consisting of 200 observations each. The first half of each sample is drawn from the standard Gaussian distribution. The second 100 observations are sampled from the Gaussian distribution with increased standard deviation 1.5 and mean 0. Appropriate values of the weight parameter $a > 0$ are chosen from the literature on empirical characteristic functions, which are comparable to our quantity $\widehat{\varphi}_N$. Since characteristic functions contain the most information around the origin (Epps, 1993), the weight functions used in this context are decreasing in the modulus of their argument. In accordance with this prior experience, for all three weighting functions investigated we choose the values $a = 0.5,\ 1,\ 1.5$ for the parameter a regulating how fast a weight function decreases to zero. The number of equidistant blocks is set to $N = 5$ and $N = 10$ and 2 000 permutations are conducted for both tests. The corresponding rejection rates are given in Table 14. Comparable results not listed here were attained for the weighting proposed by Matteson and James (2014) introduced on page 89.
According to the rejection rates the choice of the weight function w and its parameter a does not have a large influence on the performance of the test using T_{Four}. The second method seems more affected by them and in particular does not show the same behaviour in a for each choice of w. Unsurprisingly, both tests heavily depend on the initial number of blocks, because a few large blocks in general allow better estimations of the blockwise volatilities. For that reason, the methods lead to lower rejection rates for $N = 10$ in comparison to $N = 5$. This is the case despite of the fact that for $N = 5$ the true structural break lies in the middle of one block, which

certainly has a negative effect on the power of the test. We also observe that the test based on T_{Four} considerably outperforms the one using T_{cf} for all parameter constellations.

4.4.2 Evaluation of Volatility Tests

In the following we apply all tests introduced in Section 4.3 in five data scenarios. To present the settings in a clear and compact way, let $|n, \sigma = \tilde{a}|\, m, \sigma = \tilde{b}|$ denote n observations with standard deviation $\tilde{a}$ followed by m observations with standard deviation $\tilde{b}$. The data cases under study are:

1) $|200, \sigma = 1|$ ($\mathbb{H}_0$)

2) $|100, \sigma = 1|\, 100, \sigma = 1.5|$ (one structural break)

3) $|100, \sigma = 1|\, 100, \sigma = 1.5\, |100, \sigma = 1|$ (two structural breaks)

4) $|100, \sigma = 1|\, 100, \sigma = 1.4\, |100, \sigma = 1|\, 100, \sigma = 1.4\, |100, \sigma = 1|$
(four structural breaks)

5) $|100, \sigma = 1|\, 50, \sigma = 1.6\, |100, \sigma = 1|\, 150, \sigma = 1.2\, |100, \sigma = 1|$
(four nonequidistant structural breaks)

For each of these five data cases three different distributions are considered as prototypes. These are the standard Gaussian distribution (G), the t-distribution with 5 degrees of freedom (t5) and the exponential distribution with parameter $\lambda = 1$ (exp) shifted to zero mean. We make use of scaling to obtain the desired standard deviations from these prototypes. For each of the five data cases and each distribution 10 000 time series are generated. All methods introduced in Section

4.3 are applied to them. The permutation tests are executed using $n_p = 2\ 000$ permutations. For the blockwise procedures the data is divided into $N = 10$ equidistant blocks. In case of rejection we proceed in the same way on subsamples. To reduce the computational burden the two tests relying on weight functions are only carried out for $a = 1.5$ and the Gaussian weighting based on the analysis in the previous section. The results are given in Table 15.

Apparently, heavier tails make the detection of the structural breaks more difficult, since the rejection rates for data generated from the t-distribution are always lower than the corresponding ones for Gaussian data. Under the null hypothesis all methods keep the significance level of 5%. For the non-gaussian data the rejection rates of the CUSUM test are quite low under the null reflecting an inaccurate approximation. As expected, the method leads to the best results for Gaussian data with one volatility change, but loses a considerable amount of power in presence of multiple structural breaks due to masking effects. The Mood-type test clearly outperforms its competitors in the case of exponential data in all but the fifth data scenario. The procedure therefore might have problems in nonequidistant settings. For data from the Gaussian and the t-distribution it suffers a similar loss of efficiency as the CUSUM test. The problem is quite similar, because the procedure proposed by Ross is based on a two-sample test. Much like the CUSUM approach, it therefore implicitly anticipates one structural break at a time and thus always divides the sample into two parts. The tests relying on T_{Four} and T_{Log} lead to competitive results overall. In particular, they suffer considerably less from multiple structural breaks in comparison to the CUSUM and the Mood-type test. As a consequence, they clearly outperform the CUSUM test for all distributions under study and the test proposed by Ross for the symmetric distributions in case of more than one volatility change. Among the two, T_{Log} leads to slightly higher rejection rates under alternatives. The procedure using T_{cf} performs similarly

to T_{Four} and T_{Log}, but is inferior to both in all considered scenarios. This is in accordance with the results in Section 4.4.1.

4.4.3 Number and Position of Estimated Structural Breaks

A good test for structural breaks should have high rejection rates under various alternatives. However, it also must determine the causes of the heterogeneity adequately after a rejection. Otherwise, it connects the correct rejection with an irrelevant event leading to false conclusions. We therefore take a closer look at the number and location of the presumable structural breaks estimated by the methods. The blockwise procedures based on the statistics T_{Four} and T_{Log} performed best in terms of power for the Gaussian and the t-distribution with multiple structural breaks. We therefore focus on the corresponding tests, since the assumption of Gaussianity is often encountered in practice due to the central limit theorem and multiple structural breaks as well as heavy tails are also realistic scenarios for example in financial applications. Both tests are conducted in the recursive manner explained in Section 4.3.2. In Table 16 their mean number of presumable structural breaks is listed for all five scenarios studied in the previous section and data stemming form the t-distribution.

On average both methods do not detect all structural breaks, particularly if several volatility changes are present. It seems especially problematic to detect the four equidistant volatility changes in setting 4). Now, each data case is constructed such that all rejection rates are below 1 in order to make the tests comparable. Therefore, the structural breaks are simply not that obvious to all procedures under study and are thus not always detected. The test via T_{Log} rejects more often and thus unsurprisingly finds more structural breaks on average. The differences are quite small though.

Next, the replications where the tests reject are considered. As indicated by the results given in brackets in Table 16, both methods determine a reasonable number of presumable structural breaks given a rejection. The test based on T_{Four} estimates the number of structural breaks more adequately under the null hypothesis and in presence of four structural breaks, but the results are again of comparable size.

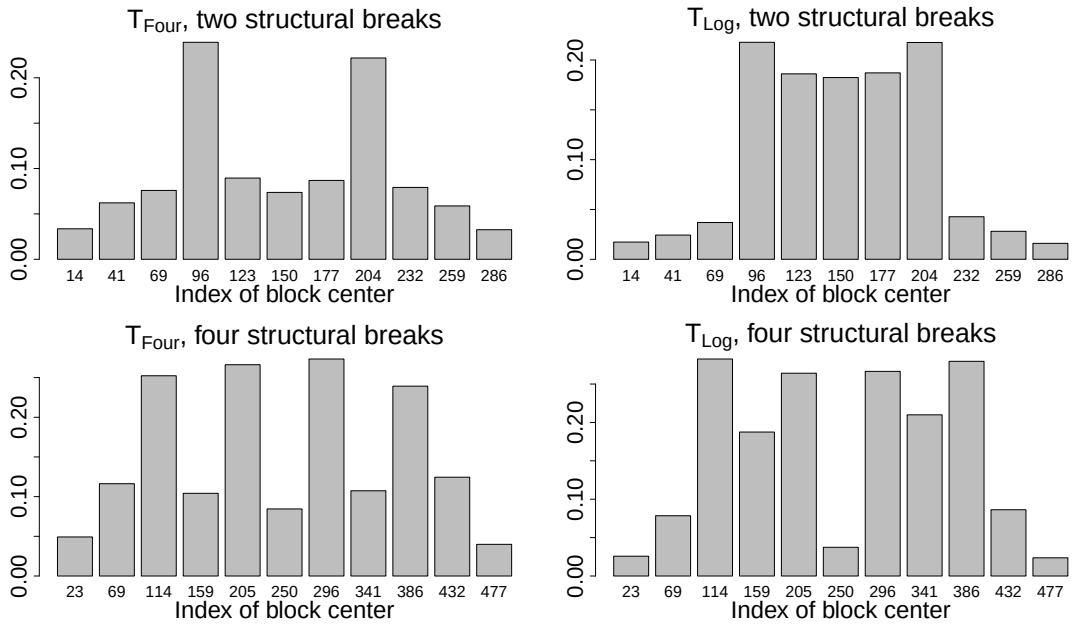

Figure 10: Mean number of the change points estimated by the permutation procedure based on T_{Four} (left column) and T_{Log} (right column) grouped by the presumable change point position. The data is generated using scaled t-distributions and contains two (upper row) and four (lower row) equidistant structural changes.

The presumable change point positions for the case of two and four equidistant structural breaks are presented in Figure 4.4.3 for both methods. Since both tests locate presumable change points in the same way, the different results are mainly a consequence of additional rejections on subsamples by the method using T_{Log}. These additional presumable change point positions do not coincide with the true structural break positions for the most part. Hence, the tests based on

T_{Four} locates the true change point positions much more precisely. This suggests that tackling the detection of structural breaks via Fourier-type transformations is indeed advantageous in comparison to the blockwise approach of T_{Log}. The price paid by a somewhat smaller rejection rate is outweighed by a more exact location of the change position in case the latter is relevant at all.

4.4.4 Application to Financial Data

For further illustration of the methods we consider the daily exchange rates of the Chinese yuan to the U.S. dollar from the 1st of January 2006 to the 1st of January 2015. The data is available on the web page of the US Federal Reserve (`http://www.federalreserve.gov`). In accordance with Ross (2013) we study the logarithmised daily exchange rate differences. For this purpose, we apply the permutation tests relying on T_{Four} and T_{Log} with the same settings as in Section 4.4.2 at a significance level of 5%. Both of them are conducted in the recursive manner described in Section 4.3.2. In Figure 4.4.4 we present the data as well as twice the estimated standard deviations in the blocks derived from the presumable changepoints for both methods. In addition, we fit a GARCH(1,1) model based on t-distributions using the R package fGarch by Wuertz and Chalabi (2013).
The data shows several regimes with considerably different magnitudes of volatility. These can be associated with events such as the financial crisis starting in the summer of 2007 and the bankruptcy of Lehman Brothers in September 2008. Apparently, both permutation tests manage to detect the regime changes quite well and lead to similar time intervals of constant volatility. Both volatility estimates can be regarded as a smoothed version of the GARCH(1,1) prediction, which is far more wiggly. Among the two tests, the block arrangement obtained using T_{Log} seems to be too fine for the summer of 2006 and more sensitive to single

observations around 2011. As opposed to that, the blocks for the T_{Four} statistics give a clear overview of the behaviour of the volatility for the whole time series.

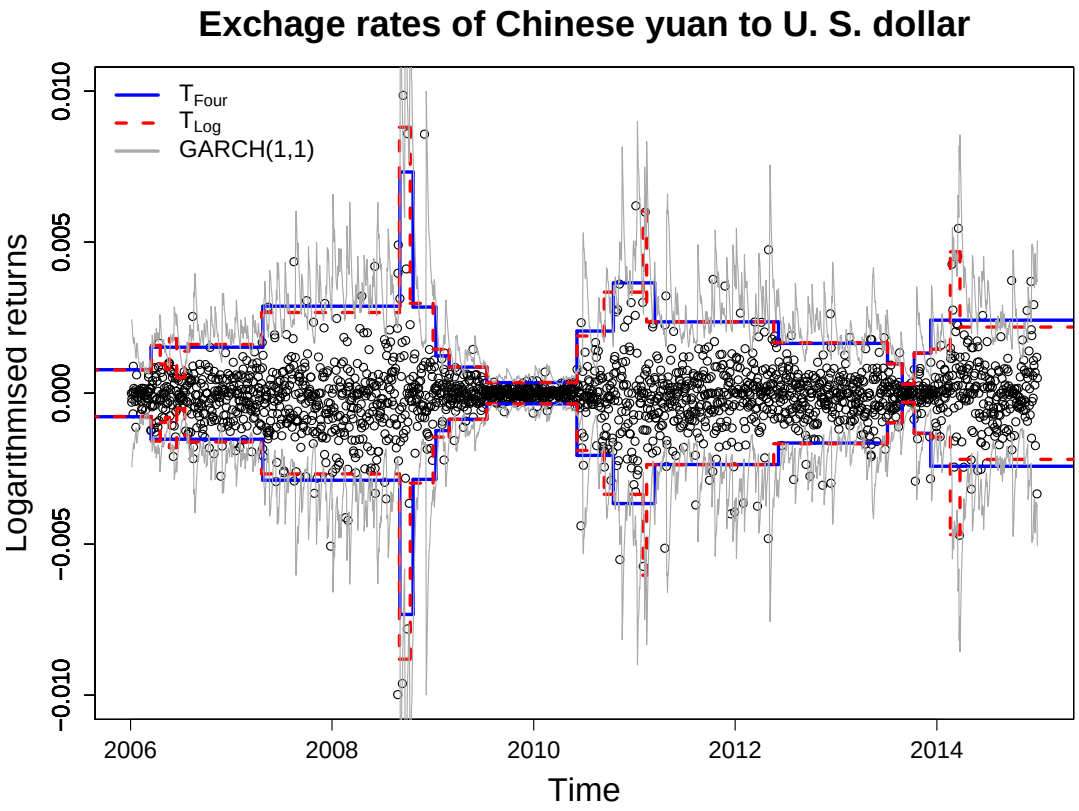

Figure 11: Logarithmised daily exchange rate differences of the Chinese yuan and the U.S. dollar. The lines represent twice the estimated local volatility derived from the presumable structural breaks obtained via the permutation test based on T_{Four} (blue) and T_{Log} (red) as well as the a GARCH(1,1) model (gray).

4.5 Extension to Structural Breaks in Kurtosis

The concept introduced in Section 4.2 as well as the procedure proposed in 4.3.1 are not restricted to testing for a constant volatility. They can easily be adapted to test the constancy of any desired feature of the data as long as reasonable

estimators for this feature are available. This can be achieved in a straightforward way: one simply substitutes the estimator (4.2) by another measure reflecting the quantity of choice. To illustrate the procedure and give a first impression on its performance we consider structural changes in the kurtosis. Kurtosis has recently gained additional attention in financial applications and is increasingly regarded as an alternative risk measure, see for example Bertram (2013) and the references given therein. In analogy to Section 4.4.2, we consider 1 000 replications of each of the four data cases

1) $|6\,000, N(0,1)|$ $(\mathbb{H}_0)$

2) $|3\,000, N(0,1)|\,3\,000, t_{10}|$ (one structural break)

3) $|3\,000, N(0,1)|\,3\,000, t_{10}\,|3\,000, N(0,1)|$ (two structural breaks)

4) $|3\,000, N(0,1)|\,3\,000, t_{10}\,|3\,000, N(0,1)|\,3\,000, t_{10}\,|3\,000, N(0,1)|$
(four structural breaks)

Hereby, $N(0,1)$ denotes standard Gaussian data, while t_{10} stands for observations drawn from a t-distribution with 10 degrees of freedom. To eliminate the effect of different volatilities the data from the t-distribution is standardised by the corresponding theoretical standard deviation. We work with considerably larger sample sizes than for the volatility, because the kurtosis is much harder to estimate. Four tests are applied to the datasets. The following three of them rely on the permutation principle. The first one is the proposed adaption to the procedure motivated in Section 4.2. Its test statistic is

$$\tilde{T}_{Four} = \frac{1}{n^2} \sum_{j,k=1}^{N} \tau_j \tau_k I_w(\log(\widehat{\kappa}_j) - \log(\widehat{\kappa}_k)),$$

where

$$\widehat{\kappa}_j = \frac{1}{\tau_j} \sum_{t \in B_j} X(t)^4 \quad \forall j = 1, \ldots, N$$

is the natural estimator of kurtosis in the j-th block for zero mean random variables with unit variance. $\tilde{T}_{Four}$ resembles the statistic T_{Four} in (4.4), but the volatility estimators $\widehat{\sigma}_j^2$ are replaced by the kurtosis estimators $\widehat{\kappa}_j$ for all $j = 1, \ldots, N$. The weighting is conducted using the Gaussian weight function with parameter $a = 1.5$. Additional simulations not reported here show that both the weighting scheme as well as the parameter a do not affect the results much as in the volatility case. In analogy to that we adopt the statistics T_{Log} introduced in (4.5) resulting in

$$\tilde{T}_{Log} = n \log \left(\sum_{t=1}^{n} X(t)^4 \right) - \sum_{j=1}^{N} \tau_j \log(\widehat{\kappa}_j).$$

The third statistic considered via the permutation approach is

$$\tilde{T}_{Max} = \max_{1 \leq j,k \leq N} |\widehat{\kappa}_j - \widehat{\kappa}_k|.$$

It is an intuitive measure to capture changes in the kurtosis process. In addition to the three permutation tests, an asymptotical CUSUM test for the kurtosis is conducted. Following Bertram (2013) we define

$$\tilde{T}_{CUS} = \max_{1 \leq h \leq n} \sqrt{n} \left| \frac{\widehat{\kappa}_{1:h}}{\widehat{\kappa}_{1:n}} - \frac{h}{n} \right|$$

and derive corresponding critical values from its asymptotics. Hereby, $\widehat{\kappa}_{1:h} = \frac{1}{h} \sum_{t=1}^{h} X(t)^4$ denotes the kurtosis estimator on the first h observations, $1 \leq h \leq n$. The results in Table 17 lead to similar conclusions as for the volatility. The CUSUM

procedure attains the highest rejection rates as long as the data contains only a few structural breaks. With increasing number of changes in kurtosis the method's performance worsens in comparison to the other tests, although this happens not as fast as in the volatility case. The tests using $\tilde{T}_{Four}$ and $\tilde{T}_{Log}$ on the contrary reject more often with increasing number of structural breaks. As before, their results are quite close. Both tests outperform the method using $\tilde{T}_{Max}$ in all settings considered. In analogy to Section 4.4.3 we examined the positions of the structural breaks determined by these two tests. As for the volatility, the test based on $\tilde{T}_{Four}$ determines the location of the true structural breaks more accurately, cf. page 101.

4.6 Conclusions and Outlook

In this chapter we construct statistics allowing to test whether the volatility of a time series is constant over time. For this purpose, we make use of a Fourier-type transformation and blockwise estimates. The method does not assume a specific distribution type, but is based on independent and blockwise identically distributed data. Our studies suggest that it has a competitive power for symmetric distributions in particular when several structural changes are present. In case of rejection it locates the structural break positions adequately.

The procedure can be extended to test for arbitrary quantities, if appropriate estimators are used. We demonstrate this in Section 4.5 for the kurtosis and obtain results comparable to the volatility case. The concept also enables the data analyst to substitute the volatility notion introduced in (4.2) by any other measure more suitable in a given context. For example, robust estimators of scale may be preferable if outliers are an issue. Deriving asymptotics for the test statistic T_{Four} is one goal for future research. Under the null hypothesis T_{Four} is nothing but a mean of identically distributed, but not independent, random variables. Therefore, the

formulation of a law of large numbers as well as a central limit theorem is an obvious task partially solved already. Upon completion, the results will allow to speed up the computations significantly especially for large sample sizes. New algorithms for an adequate choice of blocks can greatly contribute to an improvement of the method's performance and are thus desirable as well. Finally, the procedure could also be extended to the multivariate case.

5 Summary

In this dissertation three problems strongly connected to the topic of homogeneity are considered. For each of them a distribution-free approach is motivated and investigated using simulated as well as real data.

The first method is based on the classical nonparametric two-sample Kolmogorov-Smirnov test, cf. Durbin (1973). In case of a rejection by this test, the proposed algorithm quantifies the discrepancies between the corresponding samples. These dissimilarities are represented by the so called shrinkage factor and the correction distribution. The former measures the degree of discrepancy between the two samples. The latter contains information with regard to the over- and undersampled regions when comparing one sample to the other in the Kolmogorov-Smirnov sense. To the best of our knowledge our proposal is the first attempt to measure the dissimilarities between two datasets in a general distribution-free framework. We prove the correctness of the algorithm as well as its linear running time when applied to sorted samples. As illustrated in various data settings, the fast method leads to adequate and intuitive results.

The second topic investigated in this work is a new class of two-sample homogeneity tests. Classical nonparametric procedures such as the Kolmogorov-Smirnov test and the Anderson-Darling test (Anderson and Darling, 1952) rely on distribution functions, which can be estimated comparatively easily. The estimation of probability density functions is not that straightforward, if no particular distribution type is assumed. Nevertheless, two-sample homogeneity tests using density-based dissimilarity measures lead to much higher power in certain data scenarios as shown in the analysis in Chapter 3. In particular, they perform considerably better than classical procedures, if the samples under study do not predominantly differ in location. We thus highly recommend them for testing scale alternatives, skewness

alternatives or general unspecified discrepancies of datasets with almost equal means. In addition to proposing and evaluating new tests, we introduce a novel estimation technique for f-divergences the tests can rely on. The procedure is fast, does not require strong assumptions on the data and competes well with other estimators in terms of mean squared error.

In Chapter 4 we deal with structural breaks in time series. The method introduced there is motivated by characteristic functions and Fourier-type transforms. It is highly flexible in several ways: firstly, it allows to test for the constancy of an arbitrary feature of a time series such as location, scale or skewness. It is thus applicable in various problems. Secondly, the method makes use of arbitrary estimators of the feature under investigation. Hence, a robustification of the approach or other modifications are straightforward. We demonstrate the testing procedure focussing on volatility as well as on kurtosis. In both cases it leads to reasonable rejection rates for symmetric distributions. In particular the test shines in presence of multiple structural breaks, because its test statistic is constructed in a blockwise manner. The position and number of the presumable change points located by the new procedure also correspond to the true ones quite well. The method is thus well suited for many applications as illustrated on exchange rate data.

6 Tables

In the tables below, we use the abbreviations given in brackets: Gaussian distribution (G), t-distribution with 5 degrees of freedom (t5), t-distribution with 20 degrees of freedom (t20), exponential distribution with mean 1 (exp), t-test (t), F-test (F), Wilcoxon test (Wil), Kolmogorov-Smirnov test (KS), Anderson-Darling test (AD), Kanamori test based on the Kullback-Leibler divergence (Kan_{KL}), Kanamori test based on the Hellinger divergence (Kan_{H}), parametric test for two exponential distributions (Exp), natural density ratio estimator (Nat), uLSIF density ratio estimator (uLSIF) and moment matching density ratio estimator (MM). The tests not listed to far are denoted by their test statistic. All rejection rates are given in percent.

		100	500	1 000	5 000	10 000	50 000	100 000
a)	s_{opt}	0.516	0.409	0.380	0.341	0.331	0.317	0.313
	Mean	3.504	3.282	3.217	3.121	3.096	3.055	3.043
	SD	0.693	0.795	0.832	0.892	0.910	0.944	0.954
b)	s_{opt}	0.481	0.381	0.358	0.326	0.318	0.308	0.306
	$P_{\mathcal{H}_{opt}}(0)$	0.979	0.993	0.995	0.998	0.999	0.999	1.000

Table 1: Results for the the Gaussian mixture case a) and the the zero mixture b) for different sample sizes averaged over 1 000 replications, cf. page 43. For a) the determined shrinkage factors s_{opt} and the estimations of the mean and the standard deviation of the correction distribution $\mathcal{H}_{opt}$ are given. For b) the determined shrinkage factors s_{opt} and estimated probability mass assigned to 0 by $\mathcal{H}_{opt}$ denoted by $P_{\mathcal{H}_{opt}}(0)$ are presented.

Table 2: Empirical mean square errors for estimators of the Kullback-Leibler divergence in situations 1) to 4), cf. page 68. The estimators are grouped by their density ratio estimators.

	Nat		uLSIF		MM		
	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}^S_{KL}$
1)	0.0014	0.0019	0.0027	0.0041	0.0004	0.0004	0.0015
2)	8.2009	17.8172	91.3088	7.1588	10.9003	1.1715	3.9507
3)	0.9449	0.9276	0.3728	0.4041	0.0611	0.0610	0.1801
4)	77.2846	79.2853	9.6543	10.0414	0.8303	0.8276	0.6996

Table 3: Empirical mean square errors for estimators of the Hellinger divergence in situations 1) to 4) multiplied by 10^4, cf. page 68. The estimators are grouped by their density ratio estimators.

	Nat		uLSIF		MM		
	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}^S_H$
1)	0.20	0.24	0.37	0.56	0.07	0.07	0.22
2)	685.34	15.79	7699.41	161.73	466.80	11.05	18.03
3)	3.05	3.02	6.31	7.56	3.35	3.34	3.53
4)	18.13	11.78	27.89	25.79	10.05	8.88	11.99

Table 4: Rejection rates of the Kanamori test under $\mathbb{H}_0 : P = Q$ for different sample sizes, cf. page 71. Distributions: a) standard Gaussian, b) and c) mean 1 exponential. Density ratio models: a) and b) exponential, c) reduced exponential.

	n	10	30	50	75	100	150	200	250	300	400	500
a)	KLD	16.8	8.6	7.6	7.4	6.2	5.2	5.8	5.4	4.4	4.4	6.0
	Hell	11.4	7.6	7.0	6.6	5.6	5.0	5.8	5.2	4.2	4.2	6.0
b)	KLD	23.2	17.8	9.0	8.6	11.6	8.4	8.2	7.2	7.0	6.6	6.2
	Hell	11.8	12.8	5.4	6.4	9.4	6.4	6.0	5.8	6.6	6.0	5.4
c)	KLD	8.6	9.0	4.2	5.8	6.4	4.6	6.6	5.2	5.2	5.6	5.4
	Hell	7.6	8.2	4.2	5.4	6.2	4.6	6.4	5.2	5.2	5.6	5.4

Table 5: Rejection rates of several homogeneity tests under $\mathbb{H}_0 : P = Q$ for $m = n = 300$, cf. page 76.

	t	F	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_{KL}$	$\hat{D}^S_H$
G	5.2	5.2	5.0	5.0	5.6	4.2	3.8	4.8	5.0
t5	4.0	22.4	5.2	3.6	4.0	7.6	7.0	4.9	4.8
t20	4.4	6.4	5.6	5.2	4.8	5.0	4.8	5.9	5.7

Table 6: Rejection rates of some homogeneity tests under several alternatives for $m = n = 50$. The parameters of the distribution Q are $\mu_1 = -0.5$, $\mu_2 = 0.5$ for location alternatives, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1.5$ for scale alternatives and $\theta_1 = 0.6$, $\theta_2 = 1.4$ for alternatives in both location and scale simultaneously, cf. page 76.

	Location						Scale						Location and Scale					
	G		t20		t5		G		t20		t5		G		t20		t5	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
t	72	68	68	69	71	73	5	4	4	5	4	4	67	40	70.2	35	67	38
F	5	5	7	7	20	20	100	79	100	80	97	73	92	61	92	59	85	60
Wil	68	68	73	69	80	80	5	6	5	4	4	5	63	36	68	33	75	44
KS	58	52	53	54	67	69	36	11	36	13	29	10	73	36	78	36	85	43
AD	70	63	68	66	77	78	75	23	72	23	57	20	86	50	89	45	89	56
$\hat{D}_{KL}$	46	40	42	40	43	49	33	68	26	67	12	54	42	64	41	61	40	58
$\hat{D}_H$	46	41	43	43	49	52	40	65	35	65	22	56	47	63	45	59	47	60
$\hat{D}_{KL}^D$	50	45	46	45	48	54	94	52	93	51	75	34	89	54	90	48	85	43
$\hat{D}_H^D$	50	45	47	46	52	56	94	52	94	51	84	37	89	54	90	46	89	46
$\hat{D}_{KL}^S$	43	48	48	41	49	46	95	50	95	45	87	34	91	54	91	49	89	44
$\hat{D}_H^S$	45	49	48	43	55	54	95	49	95	42	88	33	91	52	92	49	91	49

Table 7: Rejection rates under several alternatives for $m = n = 300$. The parameters of the distribution Q are $\mu_1 = -0.2$, $\mu_2 = 0.2$ for location alternatives, $\sigma_1^2 = 0.8$, $\sigma_2^2 = 1.2$ for scale alternatives and $\theta_1 = 0.8$, $\theta_2 = 1.2$ for alternatives in both location and scale simultaneously, cf. page 76.

	Location						Scale						Location and Scale					
	G		t20		t5		G		t20		t5		G		t20		t5	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
t	66	71	69	69	69	69	5	6	4	4	4	4	78	64	77	62	76	62
F	5	5	6	6	22	22	98	89	96	86	89	78	98	89	96	86	89	78
Wil	67	66	67	69	78	77	6	5	6	5	6	5	75	57	75	60	85	69
KS	51	57	56	57	71	70	26	17	25	16	22	16	85	69	86	67	90	73
AD	63	70	66	69	78	75	60	37	57	32	41	27	95	83	96	79	95	83
Kan_{KL}	55	63	57	62	63	62	96	80	92	75	68	49	99	94	99	90	93	80
Kan_{H}	55	63	57	61	62	62	96	80	92	75	67	48	99	94	99	90	92	79
$\hat{D}_{KL}$	46	43	45	45	34	34	87	68	80	64	52	38	97	86	97	83	83	66
$\hat{D}_H^S$	46	44	49	46	47	45	87	68	80	64	61	45	97	86	97	84	89	76

Table 8: Rejection rates for testing the equality of the standard Gaussian and a skewed Gaussian distribution with skewness parameter $\tilde{\lambda}$ for $m = n = 300$, cf. page 76.

$\tilde{\lambda}$	t	F	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_{KL}$	$\hat{D}_H^S$
-5	6.8	7.8	13.6	35.0	51.8	7.4	7.0	96.6	97.0
-3	6.6	7.2	11.0	21.2	23.8	6.2	5.8	66.6	66.4
-1	7.4	5.0	7.0	5.6	7.0	6.4	6.4	6.2	6.2
0	6.0	5.2	5.6	4.2	5.4	4.4	4.4	6.0	5.0
1	4.8	4.2	4.4	4.4	5.8	6.0	6.0	5.6	5.4
3	3.0	6.4	9.8	19.0	20.4	5.0	5.0	66.0	66.4
5	3.8	7.2	13.8	31.8	46.8	5.4	5.2	95.8	95.0

Table 9: Rejection rates for testing the equality of the standard Gaussian and a standardised t-distribution for varying degrees of freedom and $m = n = 300$, cf. page 76.

d.o.f.	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_H$	$\hat{D}_H^S$
3	4.8	74.2	90.8	24.0	23.2	97.6	99.4
4	5.0	25.2	35.8	9.8	9.6	65.0	72.2
5	5.6	14.0	16.4	7.6	7.0	41.1	46.8
10	4.8	5.2	6.8	6.6	6.4	10.2	11.0

Table 10: Rejection rates for comparing the standard Gaussian to an asymmetrically contaminated standard Gaussian distribution with different contamination levels ε^* for $m = n = 50$, cf. page 77.

ε^*	t	F	Wil	KS	AD	$\hat{D}_{KL}$	$\hat{D}_H$	$\hat{D}_{KL}^D$	$\hat{D}_H^D$	$\hat{D}_{KL}^S$	$\hat{D}_H^S$
0	4.2	6.0	4.8	3.6	5.0	4.2	4.4	3.6	4.6	4.2	4.4
0.05	5.0	4.4	4.6	4.8	5.0	5.8	5.8	5.2	4.2	5.0	4.4
0.1	6.8	4.6	6.8	4.6	6.6	7.0	6.6	5.2	6.0	5.2	5.8
0.2	15.4	7.2	12.8	8.6	13.0	11.6	11.2	8.6	8.6	9.6	8.8
0.3	28.8	10.0	26.0	16.8	25.6	21.4	20.2	17.2	17.6	18.6	17.4

Table 11: Rejection rates for comparing the standard Gaussian distribution to a mixture of the standard Gaussian distribution and the Gaussian distribution with mean 0.1 with different levels of mixture proportion ε^* for $m = n = 300$, cf. page 77.

ε^*	t	F	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_{KL}$	$\hat{D}_H^S$
1	94.6	5.0	94.4	84.8	94.2	90.6	90.6	81.2	82.0
0.95	93.0	4.8	92.4	84.2	92.2	89.2	89.0	76.0	77.0
0.9	89.2	4.8	87.2	79.2	88.0	84.2	84.2	72.0	73.2
0.8	82.6	4.8	80.2	69.4	81.2	75.2	75.2	61.0	61.6
0.7	73.4	5.2	70.6	57.6	70.0	62.4	61.6	48.2	48.6

Table 12: Rejection rates for testing the equality of two exponential distributions with parameters $\lambda_P = 1$ and varying λ_Q for $m = n = 300$, cf. page 77.

λ_Q	Exp	Wil	KS	AD	$\hat{D}_{KL}$	$\hat{D}_H$	$\hat{D}_{KL}^D$	$\hat{D}_H^D$	$\hat{D}_{KL}^S$	$\hat{D}_H^S$
0.7	98.8	96.0	93.8	97.4	83.0	95.4	60.0	85.8	80.8	87.4
0.8	79.4	67.2	56.4	71.6	42.6	61.6	23.0	41.2	36.2	42.2
0.9	26.2	22.4	15.2	22.6	17.4	20.6	9.6	12.6	12.0	13.6
1	6.6	5.4	6.2	6.2	6.6	4.8	6.0	4.4	5.6	4.6
1.1	19.8	15.6	14.4	16.6	1.4	1.8	6.6	8.4	9.6	8.8
1.2	60.2	48.4	37.8	49.6	1.4	5.8	16.6	25.0	26.6	27.6
1.3	88.4	77.4	68.4	80.0	3.4	12.6	31.4	54.6	50.6	57.4

Table 13: Runtimes of the permutation tests using the estimators $\hat{D}_H^D$ and $\hat{D}_H^S$ on standard Gaussian data in seconds for different sample sizes, cf. page 79.

n	50	100	150	200	250	300	350	400	450	500
$\hat{D}_H^D$	16.9	23.9	31.4	39.1	46.6	54.2	62.4	70.3	77.8	85.6
$\hat{D}_H^S$	11.4	11.4	11.8	12.0	12.3	12.8	13.2	13.7	14.3	15.0

Table 14: Rejection rates in case of one volatility change for the permutation tests based on T_{Four} and T_{cf} for different weight functions w, the parameter values $a = 0.5, 1, 1.5$ and two numbers of initial equidistant blocks N, cf. Section 4.4.1.

		w_U			w_L			w_G		
	a	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
N=5	T_{Four}	80	80	81	81	80	80	80	80	81
	T_{cf}	68	69	65	60	65	68	66	68	69
N=10	T_{Four}	72	72	72	71	72	71	72	72	71
	T_{cf}	47	50	42	34	42	46	43	48	49

Table 15: Rejection rates for the volatility tests in five different scenarios for Gaussian distributions (G), t-distributions with 5 degrees of freedom (t5) and exponential distributions (exp). The tests are denoted by their test statistics, cf. Section 4.3.3 and 4.4.2.

		T_{Four}	T_{CUS}	T_{Mood}	T_{cf}	T_{Log}
$\mathbb{H}_0$	G	5.2	4.9	5.3	5.0	5.4
	t5	5.0	3.5	5.1	5.3	5.1
	exp	5.0	3.0	4.8	5.2	5.1
1 break	G	71.0	89.3	78.4	44.1	74.6
	t5	47.2	51.1	65.8	35.1	47.4
	exp	49.1	31.8	99.0	26.3	41.4
2 breaks	G	78.5	2.6	38.1	55.2	82.7
	t5	50.6	1.9	29.0	42.9	52.3
	exp	44.3	1.4	83.6	31.0	41.1
4 breaks	G	77.7	0.7	14.6	55.6	80.1
	t5	70.1	0.7	20.4	64.8	69.7
	exp	56.1	0.9	70.4	43.7	53.7
4 noneq. breaks	G	88.9	0.3	14.6	80.8	92.0
	t5	56.2	0.7	12.6	65.5	60.8
	exp	44.7	0.8	42.6	43.3	46.4

Table 16: Mean number of structural breaks estimated by the permutation tests using the blockwise statistics T_{Four} and T_{Log}. The data is generated by the t_5-distribution and five data cases are considered, cf. Section 4.4.3. In brackets the mean number of estimated structural breaks among the samples with rejection is given.

	$\mathbb{H}_0$	1 break	2 breaks	4 breaks	4 noneq. breaks
T_{Four}	0.07 (1.4)	0.70 (1.48)	1.05 (2.08)	1.66 (2.37)	2.16 (3.84)
T_{Log}	0.08 (1.56)	0.76 (1.60)	1.15 (2.20)	1.74 (2.48)	2.28 (3.75)

Table 17: Rejection rates for tests presented in Section 4.5 in the four different kurtosis scenarios introduced in Section 4.4.4.

	$\tilde{T}_{Four}$	$\tilde{T}_{CUS}$	$\tilde{T}_{Log}$	$\tilde{T}_{Max}$
$\mathbb{H}_0$	4.7	5.1	4.6	4.9
1 break	22.2	81.1	22.7	16.0
2 breaks	28.1	56.6	29.5	22.0
4 breaks	65.1	56.5	65.7	43.6

References

Ali, S. M. and Silvey, S. D. (1966): A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B*, 131–142.

Alin, A. and Kurt, S. (2008): Ordinary and Penalized Minimum Power-divergence Estimators in Two-way Contingency Tables. *Computational Statistics*, 23 (3), 455–468.

Anderson, T. W. and Darling, D. A. (1952): Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23 (2), 193–212.

Azzalini, A. (1985): A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12 (2), 171–178.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998): Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika*, 85 (3), 549–559.

Basu, A. and Lindsay, B. G. (1994): Minimum Disparity Estimation for Continuous Models: Efficiency, Distributions and Robustness. *Annals of the Institute of Statistical Mathematics*, 46 (4), 683–705.

Beran, R. (1977): Minimum Hellinger Distance Estimates for Parametric Models. *The Annals of Statistics*, 5 (3), 445–463.

Bertram, P. (2013): Testing for Breaks in Kurtosis. *Unpublished manuscript*. URL `http://www.wiwi.uni-hannover.de/fileadmin/statistik/papers/CUSQ.pdf`.

Bischl, B., Lang, M., and Mersmann, O. (2013): *BatchExperiments: Statistical Experiments on Batch Computing Clusters*. URL `http://CRAN.R-project.org/package=BatchExperiments`. R package version 1.0-968.

Bonferroni, C. (1937): Teoria statistica della classi e calcolo delle probabilit. *Volume in onore di Riccardo Dalla Volta*, 1–62.

Cardot, H., Prchal, L., and Sarda, P. (2007): No Effect and Lack-of-Fit Permutation Tests for Functional Regression. *Computational Statistics*, 22 (3), 371–390.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (1990): *Introduction to Alorithms*. MIT Press.

Cortina, J., Goebel, F., Schweizer, T., and for the MAGIC Collaboration (2009): Technical Performance of the MAGIC Telescopes. *arXiv:0907.1211*.

D'Addario, M., Kopczynski, D., Baumbach, J. I., and Rahmann, S. (2014): A Modular Computational Framework for Automated Peak Extraction from Ion Mobility Spectra. *BMC Bioinformatics*, 15 (1), 25.

Davies, L., Höhenrieder, C., and Krämer, W. (2012): Recursive Computation of Piecewise Constant Volatilities. *Computational Statistics & Data Analysis*, 56 (11), 3623–3631.

Devroye, L. and Györfi, L. (1985): *Nonparametric Density Estimation: The L_1 View*. New York: Wiley & Sons.

Durbin, J. (1973): *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia: SIAM. Regional Conference Series in Applied Mathematics 9.

Epps, T. W. (1993): Characteristic Functions and Their Empirical Counterparts: Geometrical Interpretations and Applications to Statistical Inference. *The American Statistician*, 47 (1), 33–38.

Fisher, R. A. (1935): *The Design of Experiments*. Oxford: Oliver & Boyd.

Fried, R. (2012): On the Online Estimation of Local Constant Volatilities. *Computational Statistics & Data Analysis*, 56 (11), 3080–3090.

Green, P. J. and Silverman, B. W. (1994): *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. New York: Chapman & Hall. Monographs on Statistics and Applied Probability 58.

Heck, D., Knapp, J., Capdevielle, J. N., Schatz, G., Thouw, T., et al. (1998): *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*. Forschungszentrum Karlsruhe FZKA 6019.

Hlávka, Z., Hušková, M., Kirch, C., and Meintanis, S. G. (2012): Monitoring Changes in the Error Distribution of Autoregressive Models Based on Fourier Methods. *Test*, 21 (4), 605–634.

Hušková, M. and Meintanis, S. G. (2006a): Change Point Analysis Based on Empirical Characteristic Functions. *Metrika*, 63 (2), 145–168.

Hušková, M. and Meintanis, S. G. (2006b): Change-point Analysis Based on Empirical Characteristic Functions of Ranks. *Sequential Analysis*, 25 (4), 421–436.

Hušková, M. and Meintanis, S. G. (2008): Tests for the Multivariate k-sample Problem Based on the Empirical Characteristic Function. *Journal of Nonparametric Statistics*, 20 (3), 263–277.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012): F-Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models. *IEEE Transactions on Information Theory*, 58 (2), 708–720.

Kim, J. S. and Scott, C. D. (2012): Robust Kernel Density Estimation. *The Journal of Machine Learning Research*, 13 (1), 2529–2565.

Kolossiatis, M., Griffin, J. E., and Steel, M. F. J. (2013): On Bayesian Nonparametric Modelling of Two Correlated Distributions. *Statistics and Computing*, 23 (1), 1–15.

Kopczynski, D., Baumbach, J. I., and Rahmann, S. (2012): Peak Modeling for Ion Mobility Spectrometry Measurements. In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 1801–1805. IEEE.

Lee, E. T., Desu, M. M., and Gehan, E. A. (1975): A Monte Carlo Study of the Power of Some Two-Sample Tests. *Biometrika*, 62 (2), 425–432.

Lee, S. and Na, O. (2005): Test for Parameter Change Based on the Estimator Minimizing Density-based Divergence Measures. *Annals of the Institute of Statistical Mathematics*, 57 (3), 553–573.

Liese, F. and Miescke, K.-J. (2008): *Statistical Decision Theory: Estimation, Testing, and Selection*. New York: Springer.

Lindsay, B. G. (1994): Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods. *The Annals of Statistics*, 22 (2), 1081–1114.

MAGIC Collaboration (2014): The MAGIC Telescopes. URL `https://magic.mpp.mpg.de/`.

Matteson, D. S. and James, N. A. (2014): A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, 109 (505), 334–345.

Meintanis, S. G. (2005): Permutation Tests for Homogeneity Based on the Empirical Characteristic Function. *Nonparametric Statistics*, 17 (5), 583–592.

Meintanis, S. G., Swanepoel, J., and Allison, J. (2014): The Probability Weighted Characteristic Function and Goodness-of-Fit Testing. *Journal of Statistical Planning and Inference*, 146, 122–132.

Mercurio, D. and Spokoiny, V. (2004): Statistical Inference for Time-inhomogeneous Volatility Models. *The Annals of Statistics*, 32 (2), 577–602.

Mood, A. M. (1954): On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests. *The Annals of Mathematical Statistics*, 25 (3), 514–522.

Nelder, J. A. and Mead, R. (1965): A Simplex Method for Function Minimization. *The Computer Journal*, 7 (4), 308–313.

Peña, D. (2005): *Analisis de Series Temporales*. Madrid: Alianza Editorial, 319.

Pilla, R. S. and Lindsay, B. G. (2001): Alternative EM Methods for Nonparametric Finite Mixture Models. *Biometrika*, 88 (2), 535–550.

Qin, J. (1998): Inferences for Case-control and Semiparametric Two-sample Density Ratio Models. *Biometrika*, 85 (3), 619–630.

R Development Core Team (2013): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Ross, G. J. (2013): Modelling Financial Volatility in the Presence of Abrupt Changes. *Physica A: Statistical Mechanics and its Applications*, 392 (2), 350–360.

Schellhase, C. and Kauermann, G. (2012): Density Estimation and Comparison with a Penalized Mixture Approach. *Computational Statistics*, 27 (4), 757–777.

Seghouane, A.-K. and Amari, S.-I. (2007): The AIC Criterion and Symmetrizing the Kullback-Leibler Divergence. *IEEE Transactions on Neural Networks*, 18 (1), 97–106.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Sheather, S. J. and Jones, M. C. (1991): A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society: Series B*, 53 (3), 683–690.

Sohn, S., Jung, B. C., and Jhun, M. (2012): Permutation Tests Using Least Distance Estimator in the Multivariate Regression Model. *Computational Statistics*, 27 (2), 191–201.

Spokoiny, V. (2009): Multiscale Local Change Point Detection with Applications to Value-at-Risk. *The Annals of Statistics*, 37 (3), 1405–1436.

Stărică, C. and Granger, C. (2005): Nonstationarities in Stock Returns. *Review of Economics and Statistics*, 87 (3), 503–522.

Steland, A. and Rafajłowicz, E. (2014): Decoupling Change-point Detection Based on Characteristic Functions: Methodology, Asymptotics, Subsampling and Application. *Journal of Statistical Planning and Inference*, 145, 49–73.

Stout, Q. F. (2012): Strict L_∞ Isotonic Regression. *Journal of Optimization Theory and Applications*, 152 (1), 121–135.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I., and Wang, L. (2009): A Density-ratio Framework for Statistical Data Processing. *IPSJ Transactions on Computer Vision and Applications*, 1, 183–208.

Vassiliou, E. and Demetriou, I. C. (2005): An Adaptive Algorithm for Least Squares Piecewise Monotonic Data Fitting. *Computational Statistics & Data Analysis*, 49 (2), 591–609.

Vostrikova, L. J. (1981): Detecting Disorder in Multidimensional Random Processes. *Soviet Mathematics Doklady*, 24, 55–59.

Wang, Y. (2010): Maximum Likelihood Computation for Fitting Semiparametric Mixture Models. *Statistics and Computing*, 20 (1), 75–86.

Wied, D., Arnold, M., Bissantz, N., and Ziggel, D. (2012): A New Fluctuation Test for Constant Variances with Applications to Finance. *Metrika*, 75 (8), 1111–1127.

Wilcoxon, F. (1945): Individual Comparisons by Ranking Methods . *Biometrics Bulletin*, 1 (6), 80–83.

Wornowizki, M. and Fried, R. (2014): Two-sample Homogeneity Tests Based on Divergence Measures. *Under revision in Computational Statistics*.

Wornowizki, M., Meintanis, S., and Fried, R. (2015): Fourier methods for analysing piecewise constant volatilities. *Submitted to Computational Statistics & Data Analysis*.

Wornowizki, M. and Munteanu, A. (2015): Correcting Statistical Models via Empirical Distribution Functions. *Computational Statistics*. DOI 10.1007/s00180-015-0607-5.

Wuertz, D. and Chalabi, Y. (2013): *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. URL `http://127.0.0.1:19747/library/fGarch/html/00fGarch-package.html`. R package version 270.73.

Zeileis, A. and Hothorn, T. (2013): A Toolbox of Permutation Tests for Structural Change. *Statistical Papers*, 54 (4), 931–954.

Zhu, Y., Wu, J., and Lu, X. (2013): Minimum Hellinger Distance Estimation for a Two-sample Semiparametric Cure Rate Model with Censored Survival Data. *Computational Statistics*, 28 (6), 2495–2518.