Technical Report

Two-sample Homogeneity Tests Based on Divergence Measures

Max Wornowizki, Roland Fried

September 10, 2014

technische universität dortmund

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Part of the work on this technical report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C3.

Speaker: Prof. Dr. Katharina Morik
Address: TU Dortmund University
Joseph-von-Fraunhofer-Str. 23
D-44227 Dortmund
Web: http://sfb876.tu-dortmund.de

Abstract

The concept of f-divergences introduced by [2] provides a rich set of distance like measures between pairs of distributions. Divergences do not focus on certain moments of random variables, but rather consider discrepancies between the corresponding probability density functions. Thus, two-sample tests based on these measures can detect arbitrary alternatives when testing the equality of the distributions. We treat the problem of divergence estimation as well as the subsequent testing for the homogeneity of two samples. In particular, we propose a nonparametric estimator for f-divergences in the case of continuous distributions, which is based on kernel density estimation and spline smoothing. As we show in extensive simulations, the new method performs stable and quite well in comparison to several existing non- and semiparametric divergence estimators. Furthermore, 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. According to the results, divergence based methods detect discrepancies between distributions more often than traditional methods if the distributions do not differ in location only. The findings are illustrated on ion mobility spectrometry data.

1 Introduction

Given two distributions P and Q with probability density functions p and q, respectively, the f-divergence from P to Q is defined by

$$D_f(P,Q) = \int f\left(\frac{p(y)}{q(y)}\right) dQ(y) \; = E_Q\left(f\left(\frac{p(Y)}{q(Y)}\right)\right), \tag{1}$$

where f is a given convex function applied to the density ratio $r(x) = \frac{p(x)}{q(x)}$. In order to ensure a well-defined density ratio, P must be dominated by Q. An f-divergence attains its minimal value $f(1)$ if and only if $P = Q$, see [2]. For all common divergences $f(1) = 0$ holds, giving a rather intuitive interpretation to the above property.
We illustrate this class of measures introducing some popular divergences, which will be considered in the following. The choice $f_{aKL}(x) = x \cdot \log(x)$ yields the asymmetric Kullback-Leibler divergence denoted by D_{aKL}, which is closely related to the popular AIC information criterion ([20]) and the classical maximum likelihood estimation ([5]). The measure can be symmetrised using $f_{KL}(x) = (x-1) \cdot \log(x)$. This leads to the symmetric Kullback-Leibler divergence $D_{KL}(X,Y)$ fulfilling $D_{KL}(X,Y) = D_{aKL}(X,Y) + D_{aKL}(Y,X)$. In case of continuous and one-dimensional random variables this measure has the representation

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

Another member of this class is the squared Hellinger distance, also called Hellinger

divergence, which is defined by

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

As suggested by the names, D_H is a metric, while D_H^2 is the f-divergence corresponding to $f_H(x) = \frac{1}{2}\cdot(\sqrt{x}-1)^2$. In contrast to the unbounded Kullback-Leibler divergence, the Hellinger divergence does not exceed 1.
Divergence measures, similar to Kolmogorov-Smirnov type statistics ([12]), reflect general discrepancies of the distributions. They take into account deviations in the mean, the scale, the skewness, the tail behaviour and any other characteristics of the distributions, and weight them implicitly according to the function f. Thus, methods based on divergences can reveal arbitrary dissimilarities between distributions. For this reason, divergence measures and related methods are considered in various estimation and testing problems like contingency tables ([3]), model selection ([20]), survival analysis ([26]) and detection of structural breaks in time series ([16]). Quite often, they yield a good compromise between efficiency and robustness, cf. [6] and [5]. A downside when working with divergence measures in nonparametric settings is their problematic estimation, because the ratio of the involved densities is usually unknown. Hence, the problem is often solved in two steps:

- Estimate the density ratio function $r(x) = \frac{p(x)}{q(x)}$ by $\hat{r}$
- Estimate the divergence given $\hat{r}$

The remainder of this paper is structured as follows: Section 2 focusses on the problem of the estimation of divergences. We present several existing non- and semiparametric approaches and introduce a new nonparametric divergence estimation technique combining kernel density estimation, smoothing methods and numerical integration. We also transfer the divergence decomposition approach given in [13] to general divergence estimation. In the second part of the section, the performance of the estimation algorithms is evaluated in a simulation study. Section 3 deals with the two-sample homogeneity testing problem. At first, we recapitulate an asymptotic, semiparametric divergence test proposed in [13]. Hereafter, an alternative procedure relying on the permutation technique is described, which can be performed using arbitrary divergence estimators. Both methods are compared to traditional parametric and nonparametric procedures on artificial data paying special attention to model misspecification. The tests performing best are then applied to ion mobility spectrometry data from the field of bioinformatics. Section 4 provides a summary of the results and some recommendations.

2 Divergence estimation

This section is dedicated to the estimation of f-divergences and consists of three parts. The first subsection 2.1 reviews several non- and semiparametric methods for the estimation of the density ratio r. The second part 2.2 utilises the density ratio estimations

introduced before to construct divergence estimators. In addition to stating the standard procedures, we extend the idea of divergence decomposition proposed in [13] to arbitrary density ratio estimators. Hereafter, we introduce a new approach to divergence estimation based on numerical integration and smoothing splines. In subsection 2.3 we report the results of a simulation study involving all divergence estimators presented before and evaluate their performance. The best methods are applied to ion mobility spectrometry data.

In the following we assume that $x_1,\dots,x_n \in \mathbb{R}$ are observations from one-dimensional, continuous, independent and identically distributed random variables $X_1,\dots,X_n$. Each of these follows the distribution P with probability density function p. We make analogous assumptions for the sample $y_1,\dots,y_m$ and the corresponding random variables $Y_1,\dots,Y_m$ with distribution Q and probability density function q. Expectations with respect to P and Q are denoted by E_P and E_Q, respectively.

2.1 Density ratio estimation

A straightforward *naive approach* to nonparametric estimation of the density ratio function $r = \frac{p}{q}$ consists of two steps. At first, a nonparametric estimation of the probability density functions p and q by appropriate estimators $\hat{p}$ and $\hat{q}$, respectively, is performed. Hereafter, $r = \frac{p}{q}$ is approximated by $\hat{r} = \frac{\hat{p}}{\hat{q}}$.
Estimates of the individual probability density functions can be attained by the kernel density procedure ([24]). Given an i.i.d. sample $z_1,\dots,z_l$ generated by an unknown density g, the kernel density estimate of g is

$$\hat{g}(x) = \frac{1}{l \cdot h} \sum_{i=1}^{l} K_h(x, z_i)$$

using for instance the Gaussian kernel function

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

with a bandwidth h. 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. [24]. Standard algorithms for the selection of h are cross validation and the method Sheater and Jones ([21]), which relies on a minimiser of the estimated mean integrated squared error. The latter is used for all computations involving kernel density estimates in this paper. The Gaussian kernel, which is also always applied in the following, ensures $\hat{q}(x) > 0$ for all $x \in \mathbb{R}$ and thus guarantees a well defined density ratio estimate $\hat{r} = \frac{\hat{p}}{\hat{q}}$.

As opposed to the nonparametric approach, semiparametric methods aim at estimating the density ratio directly, omitting the estimation of the underlying densities. The key idea is the introduction of a density ratio model $r(\cdot, \theta)$ assuming that $r(x) = r(x, \theta^*)$ holds for a certain value $\theta^* = (\theta_1^*,\dots,\theta_d^*) \in \mathbb{R}^d$ and all $x \in \mathbb{R}$. Thereby, the identification of

r boils down to the approximation of the parameter θ^* via an estimate $\hat{\theta}$. Since different distributions can result in the same density ratio, the density ratio model does not parametrise the densities completely and thus can be regarded as semiparametric. In the following paragraphs we describe two main techniques of parameter estimation in semiparametric density ratio models.

For the true density ratio function $r = \frac{p}{q}$, the moments $E_P(\eta)$ and $E_Q(\eta \cdot r)$ are equal for an arbitrary moment function η:

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

The *moment matching* method for density ratio estimation is motivated by this equation. Replacing the expectations by appropriate sample means allows to estimate θ^* such that $r(\cdot) = r(\cdot, \theta^*)$ by solving the equation

$$\frac{1}{n}\sum_{i=1}^{n} \eta(x_i, \theta) - \frac{1}{m}\sum_{j=1}^{m} r(y_j, \theta) \cdot \eta(y_j, \theta) = 0 \tag{3}$$

as a function of θ for a given density ratio model $r(\cdot, \theta)$. In other words, the parameter θ is chosen such that the empirical approximations of the considered moments match, which clarifies the name of the method. As shown in [18] the moment function

$$\eta^*(x, \theta) = \frac{1}{1 + \frac{n}{m} \cdot r(x, \theta)} \nabla \log r(x, \theta) \tag{4}$$

is optimal for a given density ratio model $r(\cdot, \theta)$ in the sense that the corresponding estimator induced by the moment matching has the 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.

There are analytic solutions of equation (3) for density ratio models which are linear in θ. Explicit estimators of r in arbitrary density ratio models are only available at the sample points $y_1, \ldots, y_m$. When the problem is not explicitly solvable, it is rephrased via the minimisation of the square of the left-hand side of equation (3) and numerical optimisation is applied. This approach is chosen in the simulations presented in the following, since we apply the popular exponential model

$$r_e(x, \theta) = \exp\left(\theta_1 + \theta_2 \cdot x + \theta_3 \cdot x^2\right) \tag{5}$$

including the case of both P and Q being Gaussian distributions. This model also holds for two exponential distributions, because the density ratio for negative values is of no interest in this case. However, the exponential model r_e is overparametrised by (5), since the quadratic term is redundant.

In the following applications of the moment matching we always use the optimal moment function η^* (4) and the exponential model (5) unless stated otherwise. The optimisation problem is solved using the minimiser of Nelder and Mead proposed in [17], which is implemented in the R-function *optim* and used with default settings. The initialisation

values are derived from the maximum likelihood estimates of the mean and variance under the assumption of Gaussianity. We investigated several other initialisation procedures, which did not improve the estimation performance. Especially the initialisation assuming $r = 1$ provided quite bad results and is not advisable.

The moment matching described above can be conducted using arbitrary density ratio models, as long as the moment function η allows to identify θ^*. 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

$$r_K(x,\theta) = \sum_{i=1}^{d} \theta_i \cdot K_h(x, c_i) \;, \tag{6}$$

where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, K_h is the Gaussian kernel (2) and the $c_1, \ldots, c_d$ are randomly chosen from the first sample $x_1, \ldots, x_n$. This model has a much higher dimension d, which can allow a more flexible density ratio modelling. According to [23] a model dimension $d = \min(100, n)$ is sufficient to guarantee reasonable results together with tolerable computation time in most applications.
In order to estimate the parameter θ^* within this framework, a divergence or a divergence like measure between the true and the modelled density ratio is minimised. One example for that is the Kullback-Leibler importance estimation procedure, abbreviated by KLIEP, which is explained in detail in [23]. The methods relies on the measure $KL(\theta) = -\int \log\left(r_K(x,\theta)\right) \cdot p(x)\, dx$, which is the nonsymmetric 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 of $KL(\theta)$ as a function of θ. Another example for ratio matching is LSIF, the Least-Squares Importance Fitting, also presented in [23]. The method corresponds to the quantity

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

where the constant c is independent of θ. Similar to lasso regression, a penalty term consisting of the weighted L^1-norm of θ is added for regularisation purposes leading to the estimate

$$\theta^* = \arg\min_{\theta} \; \frac{1}{2m}\sum_{j=1}^{m} r_K(y_j,\theta)^2 - \frac{1}{n}\sum_{i=1}^{n} r_K(x_i,\theta) + w\sum_{u=1}^{d} |\theta_d| \;.$$

Both KLIEP and LSIF require constraint optimisation procedures, since the estimate $\hat{\theta} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_d\right)$ must not be negative so that a nonnegative density ratio estimation is ensured. In order to avoid the numerical problems of high-dimensional constrained optimisation tasks, 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 (6) is linear in θ,

the unconstrained optimisation minimisation with the L^2 penalty is analytically solvable. To guarantee a reasonable density ratio estimate, the negative entries of the resulting estimate $\hat{\theta}$ are set to zero. This procedure is called the unconstrained Least-Squares Importance Fitting, abbreviated by uLSIF. In addition to leading to an analytically solvable optimisation problem, the score of the leave one out cross-validation in uLSIF can be computed efficiently and stably, which is quite useful for obtaining a suitable bandwidth h and a regularisation weight w. An implementation of this algorithm, which is utilized in the following computations with default settings, is available at `http://sugiyama-www.cs.titech.ac.jp/~sugi/software/`.

2.2 Divergence estimation using density ratio estimates

In this subsection, we introduce several possibilities for estimating an arbitrary divergence D_f assuming that $\hat{r}$, an estimate of the density ratio function $r = \frac{p}{q}$, is available. Recalling the definition of D_f in (1), note that a divergence is nothing but the expectation $E_Q(f(r(Y)))$. 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 second sample is used for density ratio estimation and for divergence estimation, while the first one only affects the density ratio estimation. As we will see in the simulations in Section 3, this can lead to tests with asymmetric performance even for symmetrical divergence measures.

Working on the moment matching approach introduced in 2.1, $\hat{D}_f$ is expanded in [13] by including the second sample in the divergence estimation. The convex function f, which characterises a divergence measure, is decomposed via $f = f_1 + r \cdot f_2$. Given such a pair f_1 and f_2, each f-divergence can be estimated by the *decomposed estimator*

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

because $D_f = E_Q(f(r)) = E_Q(f_1(r)) + E_P(f_2(r))$ holds. For the moment matching method based on the moment function $\eta = \eta^*$ given in (4) Kanamori et al. proved that the decomposition into

$$f_1^*(x) = \frac{f(x)}{1 + \frac{n}{m}\cdot r(x,\theta)} \text{ and } f_2^*(x) = \frac{\frac{n}{m}\cdot f(x)}{1 + \frac{n}{m}\cdot r(x,\theta)} \tag{7}$$

leads to an estimator with minimal asymptotic variance under fairly weak and verifiable conditions, cf. [13].
Even though the decomposed estimator was introduced for the moment matching density

ratio estimation, it is applicable for any density ratio estimation procedure.

We now introduce an alternative numerical procedure for the estimation of f-divergences. For the moment we do not consider a divergence measure as an expectation of a random variable, but rather understand it as an integral involving the known convex function f, the unknown density ratio r and the unknown density q. The unknown functions can be estimated following the naive approach of density ratio estimation. The only problem left to solve then is the integration process, which can be tackled by numerical integration. Preliminary investigations not reported in this work indicate that the performance of this approach can be improved by smoothing the integrand before the numerical approximation. We use cubic splines ([11]) for this purpose. In summary, we propose the following algorithm to obtain a *numerical estimator* $\hat{D}_f^N$:

1. Compute the kernel density estimates $\hat{p}$ and $\hat{q}$ and set $\hat{r} = \frac{\hat{p}}{\hat{q}}$.
2. Smooth the function $f(\hat{r}(x)) \cdot \hat{q}(x)$ via cubic splines.
3. Integrate the smoothed function of step (2) numerically.

The numerical estimation is implemented using the statistical software R ([19]). The kernel density estimation relies on the method of Sheater and Jones for bandwidth optimisation and the Gaussian kernel function (2) ensuring a well-defined density ratio estimate. Spline smoothing is performed using the routine *smooth.spline* with default settings and numerical integration is carried out via the function *integrate* for 500 subdivisions. All these functions are available in the *stats* package. For more detailed information we refer to the corresponding help pages and the references provided therein.

2.3 Evaluation of the divergence estimators

In this subsection the estimators presented above are applied to artificial data. In order to investigate their performance, we restrict ourselves to distribution pairs with explicit representations of the corresponding divergence measure. This allows us to calculate the true divergence, which should be estimated by the methods. Therefore, we study exponential, Laplacian and Gaussian random variables, but report the results for the latter only, since the findings for the distributions were essentially the same. We work with equally sized samples and $m = n \in \{50, 100, 300\}$. While the first sample is drawn 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 :
(i) $\mu = 0$, $\sigma^2 = 1$, (ii) $\mu = 3$, $\sigma^2 = 1$, (iii) $\mu = 0$, $\sigma^2 = 2$, (iv) $\mu = 3$, $\sigma^2 = 2$.
In each of these four data settings, 500 sample pairs are generated, where the Kullback-Leibler and the Hellinger divergence should be estimated. We investigate the naive kernel density approach, the moment matching technique and the uLSIF algorithm for the density ratio estimation. Due to its high computational demand mentioned in ([23]), we do not take the KLIEP procedure into account. Each of these three methods is passed to the natural estimators $\hat{D}_f$ and its decomposed version $\hat{D}_f^D$. The latter one relies on

the decomposition presented in equation (7), which is optimal for the moment matching density ratio estimation. Furthermore, the numerical estimator $\hat{D}_f^N$ introduced at the end of Section 2.2 is computed for both divergence measures.
For two Gaussian distributions with means μ and ν and variances σ^2 and τ^2, respectively, the symmetric Kullback-Leibler divergence is 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 the squared Hellinger distance can be expressed as

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

These explicit representations of the divergence measures allow us to compute the true divergence values. In the settings (i) to (iv) 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. This allows us to assess the performance of the estimators by the empirical mean squared error (MSE). For our evaluation we use the statistical software R ([19]), version 2.15.1-gcc4.3.5. The R-package BatchExperiments ([7]) is applied to run the experiments in a batch and to distribute the computations to the cores of the computer. All computations presented in the following are conducted on a 3.00GHz Intel Xeon E5450 machine with 15GB of available RAM running a SuSE EL 11 SP0 Linux distribution.
Since we get analogous results for different sample sizes, for both divergence measures we only present the empirical MSEs in the case of $n = m = 300$ in Tables 1 and 2 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 could be caused by the boundedness of the Hellinger divergence. As for the measure itself, the estimates for the Hellinger divergence will typically lie within $(0,1)$ causing a small empirical MSE compared to the estimates of the unbounded Kullback-Leibler measure.
According to the results in general higher divergence values are more difficult to estimate than smaller ones. This becomes in particular clear focussing on data setting (ii). Although situation (iv) seems more difficult than (ii) at first glance, since more variability is introduced by a higher variance of the second distribution, the estimated MSEs are mostly lower than in case (ii). In fact, case (ii) leads to the highest errors overall. 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.
A comparison of the density ratio estimators shows that the moment matching algorithm leads overall to the best results. This semiparametric approach makes use of more information than the nonparametric methods, because the correct density ratio model is specified. In contrast, the uLSIF algorithm leads to extreme estimations and hence achieves the worst results. Sugiyama et al. stressed its good performance for multidimensional problems, but in the univariate case we find the other methods to estimate the true divergence value better. Among the nonparametric methods in case of the Kullback-Leibler divergence, the numerical estimator $\hat{D}_f^N$ outperforms the naive kernel density approach, which leads to some huge overestimations, while the numerical alternative seems more

stable. In the most realistic sample case (iv) $\hat{D}_f^N$ even attains the smallest MSE of all methods considered. For the bounded Hellinger divergence, the decomposed estimator based on the naive kernel density estimation gives slightly better results than the numerical estimator, which performs quite well overall. In addition, the results suggest that decomposing dramatically decreases the MSE in the majority of the cases for all methods and not just for the moment matching.

3 Homogeneity tests based on divergences

In this section we study two sample procedures testing $H_0 : P = Q$ based on divergence measures. At first, we review the asymptotic test given in [13], which relies on a semiparametric divergence estimator. Hereafter, we propose alternative tests based on arbitrary divergence estimators based on the permutation technique. In the second part of the section, the tests presented before are compared to some parametric and nonparametric tests in a broad simulation study. Finally, the methods performing best are applied to ion mobility spectrometry data.

3.1 Divergence based tests

In [13], an asymptotic test for the two-sample homogeneity problem based on divergence measures is developed. The authors estimate the divergence via the semiparametric moment matching introduced in Section 2.1. Hereby, they used the moment function η^* and the decomposition functions f_1^* and f_2^* presented in (4) and (7), respectively. The authors proved that the chi-square distribution with $d-1$ degrees of freedom is the asymptotic distribution of the test statistic

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

under the null hypothesis $H_0 : P = Q$ for any divergence measure D_f, where d denotes the length of the parameter vector θ in the density ratio model and f'' is the second derivative of the convex function f specifying the chosen divergence. In the following, we refer to this method as the Kanamori test.

Recent research in various fields shows that the permutation principle ([10]) and its extensions can lead to quite powerful tests, cf. [25], [22] and [8] and the references given therein. Motivated by these results, we propose the following distribution free procedure to test H_0 : Given the original sample pair $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, we generate b new sample pairs from the original data. For this purpose, we draw n of the $n+m$ observations from the sample at random without replacement yielding a new first sample. The remaining m observations form the new second sample. In other words, the sample labels are permuted at random. In this way, we obtain b sample pairs. Next, the divergence of choice is estimated on each of the b sample pairs as well as on the original data using the same estimator leading to $b+1$ estimates. Under the null hypothesis $H_0 : P = Q$, these

stem from identically distributed random variables. Since the true divergence value for any convex function f is minimal under H_0, a permutation test based on a divergence estmator rejects the null hypothesis, if the divergence estimate on the original data exceeds the empirical $(1-\alpha)-$quantile of the $b+1$ divergence estimates, where α is the predefined significance level.
Note that, in contrast to the Kanamori test, the permutation procedure itself is constructed without any restrictions of the sizes of the samples and also does not impose any distributional assumptions. Thus, it does not need a correctly specified density ratio model to keep the significance level under the null hypothesis. Another advantage is the possibility to use arbitrary divergence estimators, which can be superior to the moment matching in certain settings.

3.2 Evaluation of the divergence tests

Before comparing the tests described above with parametric and nonparametric alternatives, we assess the minimum sample size required for the Kanamori test to keep the nominal significance level of $\alpha = 5\%$. For this purpose the Kanamori tests based on the Kullback-Leibler and the Hellinger divergence are applied to 500 pairs of equally sized samples drawn from the standard Gaussian distribution for different sample sizes. An analogue computation is performed with data generated from the exponential distribution with mean 1. Kanamori et al. also checked the convergence rate of their method in a similar simulation, but focussed on the multidimensional rather than the univariate setting. As mentioned before, the exponential density ratio model given in (5) is adequate if two Gaussian or two exponential distributions are considered. However, it is overparametrised in the latter case. In order to examine the way this affects the results, we perform the same tests in the case of exponential distributions using the reduced exponential model

$$r_{re}(x,\theta) = \exp(\theta_1 + \theta_2 \cdot x) \ , \tag{8}$$

dropping the quadratic term.
We analyse the convergence of the method considering the empirical size of the test given in Table 3 as functions of the sample size. The exponential distribution scenarios illustrate the strong impact of the density ratio model. While for the correctly specified models 250 observations seem sufficient to ensure a proper test procedure, the rejection rate for the overparametrised exponential model converges much slower to $\alpha = 5\%$ and leads to increased 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 scenarios proceeding as follows: At first, we choose a certain data setting like for example location alternatives for Gaussian distributions. The distribution P is fixed, while the parameters of the second distribution Q vary on a chosen grid reflecting different degrees of discrepancy. For each of these parameter constellations, we generate 500 sample pairs of size $m = n = 50$ from the respective P and Q. We then apply several tests on each of the

sample pairs and compute the empirical rejection rate for each test and each parameter constellation. These tests include six divergence based permutation tests as well as other parametric and nonparametric tests known from the literature. Since the asymptotic Kanamori test is not applicable for such small sample sizes as is illustrated above, we repeat the simulation for samples of size 300. In exchange to running the Kanamori test for both divergence measures in the large sample case, we exclude some of the permutation tests here, which perform similar to others for $m = n = 50$. 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 based on the permutation procedure is conducted in six versions based on different divergence estimators always permuting the sample labels $b = 500$ times. As divergence estimators we consider the numerical estimators $\hat{D}^N_{KL}$ and $\hat{D}^N_H$, 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 naive kernel density approach. The semiparametric uLSIF algorithm is omitted due to its high computational demand and its modest results in Section 2.3. We also leave out the asymptotic Kanamori test, because it does not attain the nominal significance level for such small samples as shown in the previous subsection. In addition to the six permutation tests, we apply the nonparametric Wilcoxon rank-sum, the Kolmogorov-Smirnov and the Anderson-Darling test ([12]). While the first primarily detects location alternatives, the other two reveal arbitrary deviations from the null hypothesis and are based on differences between the empirical 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 $H_0 : P = Q$ if and only if one of the one-sided tests rejects H_0. The one-sided tests are optimal for the comparison of exponential distributions and are constructed to detect the alternatives $\lambda_P > \lambda_Q$ and $\lambda_P < \lambda_Q$, respectively ([15]). Their test statistic is the ratio of the sample means, which follows an F-distribution under H_0. Both one-sided tests are carried out at a significance level of 2.5% to ensure the global significance level of 5%. This method was implemented by the authors, while all other parametric and nonparametric tests are conducted using the implementations in the R packages *stats* and *adk*. All tests are carried out at a nominal significance level of 5%.

In this paragraph, we extensively explain the data settings the test are applied to. We always list the parameter values for the small sample simulation and give the corresponding quantities for the large sample case in brackets.
At first, we consider the case of two 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. A selection of representative rejection rates for these settings is given in Tables 5 and 6. In addition, Table 4 provides the rejection rates under the null hypothesis for the large sample case, which is included in the location, the scale and the linked setting. Consequently, Table 4 is 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 ([4]). The skewness of corresponding random variables is regulated by the parameter λ. For $\lambda = 0$ the skewed Gaussian distribution coincides with the standard Gaussian, while for negative (positive) values of λ it is left-skewed (right-skewed). Note that a skewed Gaussian random variable does not have mean 0 and variance 1 for $\lambda \neq 0$. Therefore, we always generate data from a standardised skewed Gaussian distribution Q for $\lambda = -50, -40, \ldots, 40, 50$ ($\lambda = -5, -4, \ldots, 4, 5$) and compare it to observations drawn from the standard Gaussian distribution P. The results for the large sample case in this setting are presented in Table 7.
Next, we investigate the methods' capability of detecting departures from the Gaussian distribution class in terms of heavy tails. Hereby, P is again chosen as the standard Gaussian distribution, while Q is a t-distribution with a number of degrees of freedom ν varying between 3 and 10. In the same manner as with the skewness, we draw data from a standardised version of Q, so that P and Q neither differ in location nor in scale. The rejection rates for this setting are listed in Table 8.
Finally, we generate data from two exponential distributions. While P is fixed to have parameter value $\lambda_P = 1$, the parameter of Q is chosen as $\lambda_Q = 0.2, 0.3, \ldots, 1.7, 1.8$ ($\lambda_Q = 0.6, 0.7, \ldots, 1.5, 1.4$). The corresponding small sample results are given in Table 9.

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 H_0 quite rarely if their specific alternative is not met, cf. Tables 7 and 8. As illustrated in Table 4, the F-test is more affected by an incorrect distributional assumption and does not hold the significance level, whereas 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 attains the highest rejection rates, too.
Among the traditional 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 H_0 more rarely than in 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.
Comparing the permutation tests to each other we see that the ones based on the Hellinger divergence perform most of the time somewhat better than their Kullback-Leibler counterparts. However, the choice of divergence does not affect the results as much as the

divergence estimation technique. The tests using the numerical estimators or the decomposed estimators lead to similar and more stable results, whereas the ones relying on the natural estimators $\hat{D}_H$ and $\hat{D}_{KL}$ perform quite differently, cf. Tables 5 and 9. For example in the setting of different scales the latter detect departures from the null hypothesis more often if the variance of the second sample exceeds the one of the first, but reject rarely compared to other methods in the opposite case. This behaviour could be caused by the asymmetry of the estimation procedure discussed on page 6. Overall, the decomposed and numerical estimators lead to higher rejection rates in most of the cases we study.
All in all, permutation tests using divergence estimators detect discrepancies between distributions less often than the Wilcoxon, Kolmogorov-Smirnov and 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 numerical divergence estimators and the decomposed estimators 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, like in the scale and the skewness setting as well as for the comparison of Gaussian and t-distributed data. The two Kanamori tests show even better results than the permutation tests as long as the exponential density ratio model is correct. However, if the model is inadequate, they do not hold the nominal significance level and lead to worse results than the permutation tests, cf. Table 4, 7 and 8.

Since the two best permutation tests using $\hat{D}_H^N$ and $\hat{D}_H^D$ lead to quite similar results, they are evaluated in terms of runtime. We apply the methods to equally large samples of varying size $n = m = 50, 100, 200 \ldots, 1000$ and determine the mean computation time over 200 replications for each sample size. All tests are conducted using 1000 permutations and data stemming from the standard Gaussian distribution in both samples, respectively. We also checked the runtime in the case of different Gaussian distributions and got essentially the same results. According to the results given in Table 10, the runtime of the test based on $\hat{D}_H^N$ are always smaller and increase notably slower in the sample size than the runtime for the decomposed estimator $\hat{D}_H^D$. Since both methods led to comparable rejection rates in our simulations, we recommend the numerical divergence estimator for applications.

3.3 Application to real data

We want to get an impression of the performance of our tests on real data and thus consider so called ion mobility spectrometry (IMS) measurements, which are carried out 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. Since the spectrograms are generated one-by-one in real-time with a high frequency, the amount of data grows very quickly. In order to represent the given information in a compressed form and thereby minimise the amount of storage, they are typically analysed with regard to major peaks. In this way, the position and shape of the detected peaks is stored instead of the corresponding measurements. In an effort to automate and speed-up the computations, D'Addario et al. propose to model the

spectrograms in each of the two dimensions independently by finite mixtures of inverse Gaussian probability density functions, cf. [9]. The parameters within this model are estimated using a version of the EM algorithm. We make use of homogeneity tests to evaluate this modelling.
From several minutes of IMS measurement we obtain 500 one-dimensional spectrograms by conditioning on a certain value for one of the two dimensions for each spectrogram and focussing on the other dimension, cf. [14]. For every spectrogram two equally sized data sets containing $m = n = 500$ observations are investigated. Hereby, one of the data sets is generated from the corresponding real spectrogram observed, while the other is sampled from the fitted mixture model. We apply the permutation test based on the numerical divergence estimator $\hat{D}_H^N$ as well as the Anderson-Darling test both at a significance level of five percent to each of these 500 sample pairs.
In general, the results for both tests suggest that the inverse Gaussian models fit the spectrograms quite well. They reject the null hypothesis of equal distribution for only 62 and 51 of the 500 spectrograms, respectively. For 91 spectrograms they come to different test decisions. We focus on two of these 91 situations in the following by looking at kernel density estimates associated with the data and the corresponding mixture model, see Figure 1.

Most of the 91 spectrograms, where the tests come to different decisions, are unimodal or almost unimodal like spectrogram A. Among these, there are both cases where the Anderson-Darling test rejects the null hypothesis of equal distributions while the divergence test does not and vice versa. Presumable, most of them are false rejections of one or the other test. However, for all of the few multimodal situations similar to spectrogram B, the Anderson-Darling test does not reject $H_0 : P = Q$ in contrast to the divergence test. Since the discrepancies between the densities in spectrogram B look notedly larger than in spectrogram A, the test based on $\hat{D}_H^N$ is preferable to the common 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.

4 Conclusions

This paper deals with the estimation of f-divergences and the testing of homogeneity of two samples using such quantities. Since divergences are density based distance-like measures between distributions, they are capable of detecting any departure between the corresponding distributions and are not restricted to discrepancies in location or scale. Working in the one-dimensional setting in the case of continuous distributions, we propose a new nonparametric divergence estimation technique involving kernel density estimation and numerical integration and compare it to several standard estimators by Monte Carlo experiments. The new estimator shows a stable performance and leads to quite good results especially for the unbounded Kullback-Leibler divergence.
In addition, we make use of the permutation technique to tackle the two-sample homogeneity problem based on arbitrary divergence estimators. Just like the new estimator, the method does not require any assumptions on the underlying distributions and is

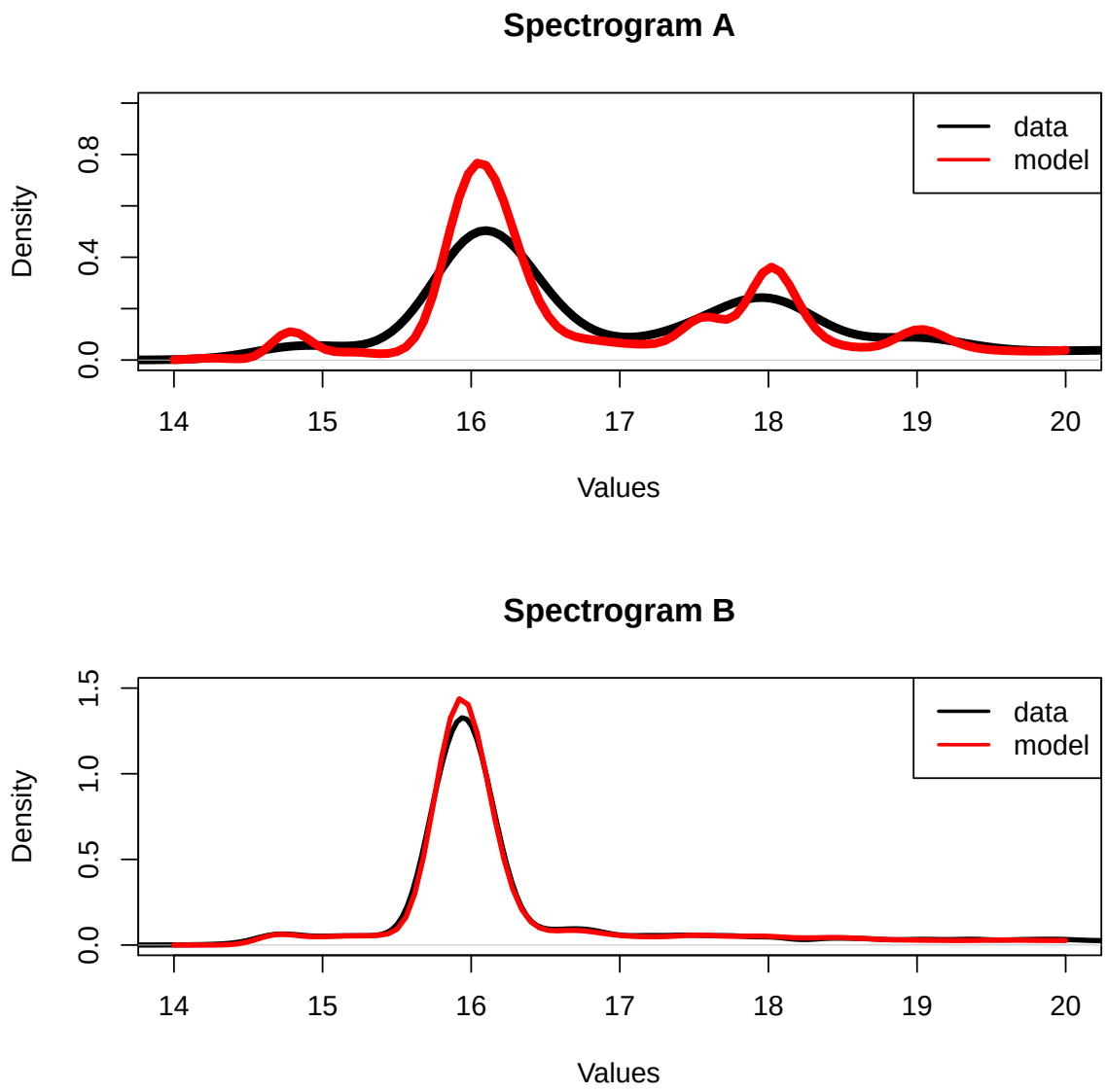

Figure 1: Kernel density estimation based and data and fitted model for two spectrograms.

therefore widely applicable. It is compared to a semiparametric asymptotic divergence based test procedure as well as to standard parametric and nonparametric tests. Our simulation experiments suggest that the traditional nonparametric procedures like the Kolmogorov-Smirnov and the Anderson-Darling test outperform divergence based tests if the distributions differ in location only. However, for scale alternatives, skewness alternatives and for the comparison of different distribution classes the divergence test yield better results. The asymptotic test proposed by Kanamori et al. ([13]) performs better than the permutation tests as long as its density ratio model is adequate, but breaks down otherwise. We also apply the test based on the new divergence estimator to real data from bioinformatics and get results at least comparable to the Anderson-Darling

test. On the basis of our findings, we recommend to use the permutation test based on the stable numerical estimator of the Hellinger divergence for testing the two-sample homogeneity problem in cases with no a priori information of discrepancies in location only. Also, a combination of this method and the Anderson-Darling test via a Bonferoni correction or the like might be a reasonable option.

5 Appendix

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), 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 ($\mathrm{Kan_{KL}}$), Kanamori test based on the Hellinger divergence ($\mathrm{Kan_{H}}$), parametric test for two exponential distributions (Exp), naive kernel density ratio estimator (KD), uLSIF density ratio estimator (uLSIF) and moment matching density ratio estimator (MM). Permutation tests based on a divergence estimator are denoted by the estimator's label. All rejection rates are given in percent and the null hypothesis is always $H_0 : P = Q$.

Table 1: Empirical mean square errors for estimators of the Kullback-Leibler divergence in situations (i) to (iv), cf. page 7.

	KD		uLSIF		MM		
	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}_{KL}$	$\hat{D}^D_{KL}$	$\hat{D}^N_{KL}$
(i)	0.0014	0.0019	0.0027	0.0041	0.0004	0.0004	0.0015
(ii)	8.2009	17.8172	91.3088	7.1588	10.9003	1.1715	3.9507
(iii)	0.9449	0.9276	0.3728	0.4041	0.0611	0.0610	0.1801
(iv)	77.2846	79.2853	9.6543	10.0414	0.8303	0.8276	0.6996

Table 2: Empirical mean square errors for estimators of the Hellinger divergence in situations (i) to (iv), cf. page 7, multiplied by 10^4.

	KD		uLSIF		MM		
	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}_H$	$\hat{D}^D_H$	$\hat{D}^N_H$
(i)	0.20	0.24	0.37	0.56	0.07	0.07	0.22
(ii)	685.34	15.79	7699.41	161.73	466.80	11.05	18.03
(iii)	3.05	3.02	6.31	7.56	3.35	3.34	3.53
(iv)	18.13	11.78	27.89	25.79	10.05	8.88	11.99

Table 3: Rejection rates of the Kanamori test under H_0 for different sample sizes. Considered distributions: a) standard Gaussian, b) and c) exponential with mean 1. Considered 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 4: Rejection rates under H_0 for $m = n = 300$.

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

Table 5: Rejection rates 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 11.

	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	34.6	66.8	37.8
F	5	5	7	7	20	20	100	79	100	80	97	73	92.2	61.4	92.0	59.0	84.6	60.2
Wil	68	68	73	69	80	80	5	6	5	4	4	5	63.2	35.8	67.6	33.2	75.4	44.4
KS	58	52	53	54	67	69	36	11	36	13	29	10	72.8	36.4	77.6	36.2	84.8	42.8
AD	70	63	68	66	77	78	75	23	72	23	57	20	85.6	50.4	89.4	45.0	89.0	55.8
$\hat{D}_{KL}$	46	40	42	40	43	49	33	68	26	67	12	54	41.6	64.2	41.0	60.6	40.2	58.2
$\hat{D}_H$	46	41	43	43	49	52	40	65	35	65	22	56	47.4	63.4	45.4	59.4	47.2	60.0
$\hat{D}_{KL}^D$	50	45	46	45	48	54	94	52	93	51	75	34	89.2	54.2	90.0	48.4	85.4	42.8
$\hat{D}_H^D$	50	45	47	46	52	56	94	52	94	51	84	37	89.0	54.4	90.0	46.4	88.8	45.6
$\hat{D}_{KL}^N$	50	44	47	45	46	50	94	51	94	53	83	40	90.6	53.4	90.6	48.0	87.0	44.2
$\hat{D}_H^N$	51	45	48	48	54	55	94	50	94	51	85	41	90.8	54.2	92.0	48.2	91.8	49.4

Table 6: 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 simulaneously, cf. page 11.

	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
$\hat{D}_{KL}$	40	45	41	44	37	37	59	75	50	70	20	48	84	90	81	85	60	75
$\hat{D}_H^N$	42	47	44	47	45	46	88	69	82	60	62	44	98	87	96	81	91	72

Table 7: Rejection rates for testing the equality of the standard Gaussian and a skewed Gaussian distribution with skewness parameter λ for $m = n = 300$.

λ	t	F	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_{KL}$	$\hat{D}_H^N$
-5	6.8	7.8	13.6	35.0	51.8	7.4	7.0	75.4	96.4
-3	6.6	7.2	11.0	21.2	23.8	6.2	5.8	51.4	66.4
-1	7.4	5.0	7.0	5.6	7.0	6.4	6.4	6.8	6.2
0	6.0	5.2	5.6	4.2	5.4	4.4	4.4	5.2	5.0
1	4.8	4.2	4.4	4.4	5.8	6.0	6.0	5.4	5.6
3	3.0	6.4	9.8	19.0	20.4	5.0	5.0	45.2	63.4
5	3.8	7.2	13.8	31.8	46.8	5.4	5.2	75.0	95.0

Table 8: Rejection rates for testing the equality of the standard Gaussian and a standardised t-distribution for varying degrees of freedom and $m = n = 300$.

d.o.f.	Wil	KS	AD	Kan_{KL}	Kan_{H}	$\hat{D}_H$	$\hat{D}_H^N$
3	4.8	74.2	90.8	24.0	23.2	97.8	99.4
4	5.0	25.2	35.8	9.8	9.6	59.6	72.6
5	5.6	14.0	16.4	7.6	7.0	34.2	44.4
10	4.8	5.2	6.8	6.6	6.4	9.0	11.0

Table 9: Rejection rates for testing the equality of two exponential distributions with parameters $\lambda_P = 1$ and varying λ_Q for $m = n = 50$.

λ_Q	Exp	Wil	KS	AD	$\hat{D}_{KL}$	$\hat{D}_H$	$\hat{D}^D_{KL}$	$\hat{D}^D_H$	$\hat{D}^N_{KL}$	$\hat{D}^N_H$
0.7	98.8	96.0	93.8	97.4	83.0	95.4	60.0	85.8	81.2	88.0
0.8	79.4	67.2	56.4	71.6	42.6	61.6	23.0	41.2	36.2	41.6
0.9	26.2	22.4	15.2	22.6	17.4	20.6	9.6	12.6	11.8	12.8
1	6.6	5.4	6.2	6.2	6.6	4.8	6.0	4.4	5.2	3.8
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	26.8
1.3	88.4	77.4	68.4	80.0	3.4	12.6	31.4	54.6	49.6	57.8

Table 10: Runtimes of the permutation tests using the estimators $\hat{D}^D_H$ and $\hat{D}^N_H$ on samples from the standard Gaussian distribution in seconds.

n	50	100	150	200	250	300	350	400	450	500
$\hat{D}^D_H$	16.9	23.9	31.4	39.1	46.6	54.2	62.4	70.3	77.8	85.6
$\hat{D}^N_H$	13.6	13.8	14.1	14.4	14.8	15.2	15.6	16.1	16.7	17.2

References

[1] Ahrens J *et al.*.'Sensitivity of the IceCube detector to astrophysical sources of high energy moun neutrinos', Astropart Phys, 507:532, 20, 2004.

[2] Ali SM and Silvey SD: 'A general class of coefficients of divergence of one distribution from another', J R Stat Soc (B), 131:142, 28, 1966.

[3] Alin A and Kurt S. 'Ordinary and penalized minimum power-divergence estimators in two-way contingency tables', Computat Stat, 455:468, 23, 2008.

[4] A. Azzalini. 'A class of distributions which includes the normal ones', Scand J Stat, 171:178, 12, 1985.

[5] Basu A, Harris IR, Hjort NL and Jones MC .'Robust and efficient estimation by minimising a density power divergence', Biometrika, 549:559, 85, 1998.

[6] Beran R. 'Minimum Hellinger distance estimates for parametric models', Ann Stat, 445:463, 3, 1977.

[7] Bischl B, Lang M and Mersmann O. 'BatchExperiments: Statistical experiments on batch computing clusters', R package version 1.0-968, `http://CRAN.R-project.org/package=BatchExperiments/`, 2013.

[8] Cardot H, Prchal L and Sarda P. 'No effect and lack-of-fit permutation tests for functional regression', Comput Stat, 371:390, 22, 2007.

[9] D'Addario M, Kopczynski D, Baumbach J I and Rahmann S. 'A modular computational framework for automated peak extraction from ion mobility spectra', BMC Bioinform, 25:36, 15, 2014.

[10] Fisher RA, *The design of experiments*, Oliver and Boyd, Edinburgh, UK, 1935.

[11] Green PJ and Silverman BW. *Nonparametric regression and generalized linear models: a roughness penalty approach*, CRC Monogr Stat Appl Probab (Book 58), Chapman and Hall, New York, 1994.

[12] Govindarajulu Z. *Nonparametric inference*, World Scientific Pub Co , 2007.

[13] Kanamori T, Suzuki T and Sugiyama M. 'F-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models', IEEE Transact Inf Theor, 708:720, 58, 2012.

[14] Kopczynski D, Baumbach J I and Rahmann S. 'Peak modeling for ion mobility spectrometry measurements', Proc. of the 20th Eur. Signal Process. Conf. (EUSIPCO 2012), 1801:1805, 2012.

[15] Lee ET, Desu MM and Gehan EA. 'A monte carlo study of the power of some two-sample tests', Biometrika, 425:432, 62, 1975.

[16] Lee S, Na O. 'Test for parameter change based on the estimator minimizing density-based divergence measures', Ann Inst Stat Mat, 553:573, 57, 2005.

[17] Nelder JA and Mead R. 'A simple algorithm for function minimization', Computer J, 308:313, 7, 1965.

[18] Qin J. 'Inferences for case control and semiparametric two-sample density ratio models', Biometrika, 619:630, 58, 1998.

[19] R Development Core Team. 'R: A language and environment for statistical computing. R Foundation for Statistical Computing', Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`, 2013.

[20] Seghouane AK and Amari SI. 'The AIC criterion and symmetrizing the Kullback-Leibler divergence', IEEE Transact Neural Netw, 97:104, 18, 2007.

[21] Sheather SJ and Jones MC. 'A reliable data-based bandwidth selection method for kernel density estimation', J R Stat Soc (B), 683:690, 53, 1991.

[22] Sohn S, Jung BC and Jhun M. 'Permutation tests using least distance estimator in the multivariate regression model', Comput Stat, 191:201, 27, 2012.

[23] Sugiyama M, Kanamori T, Suzuki T, Hido S, Sese J, Takeuchi I and Wei L. 'A density-ratio framework for statistical data processing', IPSJ Transact Comput Vis Appl, 183:208, 1, 2009.

[24] Turlach BA. *Bandwidth selection in kernel density estimation: A review*, Université catholique de Louvain, 1993.

[25] Zeileis A, Hothorn T. 'A toolbox of permutation tests for structural change', Stat Pap, 931:954, 54, 2013.

[26] Zhu Y, Wu J and Lu X. 'Minimum Hellinger distance estimation for a two-sample semiparametric cure rate model with censored survival data', Comput Stat, 2495:2518, 28, 2013.