Self–consistency–based tests for bivariate distributions

Jochen Einbeck

Department of Mathematical Sciences

Durham University, UK

Simos Meintanis

Department of Economics, National and Kapodistrian University of Athens, Greece

Unit for Business Mathematics and Informatics, North–West University, South Africa

April 12, 2017

Abstract

A novel family of tests based on the self–consistency property is developed. Our developments can be motivated by the well known fact that a two–dimensional spherically symmetric distribution X is self–consistent w.r.t. to the circle $E||X||$, that is, each point on that circle is the expectation of all observations that project onto that point. This fact allows the use of the self–consistency property in order to test for spherical symmetry. We construct an appropriate test statistic based on empirical characteristic functions, which turns out to have an appealing closed–form representation. Critical values of the test statistics are obtained empirically. The nominal level attainment of the test is verified in simulation, and the test power under several alternatives is studied. A similar test based on the self–consistency property is then also developed for the question of whether a given straight line corresponds to a principal component. The extendibility of this concept to further test problems for multivariate distributions is briefly discussed.

Keywords: Self-consistency, empirical characteristic functions, spherical symmetry, principal curves, principal components

1 Introduction

Self–consistency is a fundamental but possibly under–rated concept in Statistics. In simple words, an object is called self–consistent if each point of the object is the mean of all data (or of the random vector) that project onto that point (Tarpey and Flury 1996). Examples for self–consistent structures include the cluster centers found by $k-$ means clustering (also referred to as 'principal points' (Flury 1990)), as well as principal curves and manifolds (Hastie and Stuetzle 1989).

To fix terms, let $X = (X_1, X_2)^T$ be a bivariate random vector. We assume that X is centered such that $E(X) = (0, 0)^T$ (otherwise we consider $X - E(X)$) and scaled such that

$$r \equiv E||X|| = 1$$

(otherwise we consider $r^{-1}X$). Define a curve $C : \lambda \mapsto C(\lambda)$; $\mathbb{R} \longrightarrow \mathbb{R}^2$ parametrized by $\lambda \in [a, b]$, where a and b may tend to minus and plus infinity, respectively. We assume C to be smooth (that is infinitely many times differentiable), unit-speed (that is, $||\nabla C(\lambda)|| = 1$ for all λ), and that the curve does not intersect itself except possibly at the start– and endpoint. In other words, we may or may not have $C(a) = C(b)$, but otherwise $\lambda_1 \neq \lambda_2$ will strictly imply $C(\lambda_1) \neq C(\lambda_2)$. A self–consistent curve (or a principal curve) $C(\cdot)$ fulfils

$$E(X|\lambda_C(X) = \lambda) = C(\lambda), \tag{1}$$

where the projection index λ_C of a point $x \in \mathbb{R}^2$ onto C is defined by

$$\lambda_C(x) = \sup\{\lambda : ||x - C(\lambda)|| = \inf_{\tau} ||x - C(\tau)||\}.$$

While principal curves can be colloquially described as "smooth curves passing through the middle of a data cloud" (in the sense of the self–consistency characteristic above), there do exist special cases of such curves which relate to other relevant statistical properties. Hastie and Stuetzle (1989) stated that,

> For any two–dimensional spherically symmetric distribution $X \in \mathbb{R}^2$,
>
> a circle with center at the origin and radius $E||X||$ is a principal curve, (2)

(that is, it is self–consistent) and

> If a straight line is self-consistent, then it is a principal component. (3)

These relationships open the intriguing possibility of using self–consistency as a device for statistical testing: If a circle with center at the origin and radius $E||X|| = 1$ is *not* self–consistent then X is *not* spherically symmetric. If a given line is *not* a principal component, then it is *not* self–consistent. We will, hence, use these relationships to devise tests for the questions of whether X is spherically symmetric, and whether a given vector v (centered at the origin) constitutes a principal component of X.

In either case, self–consistency just stands as a proxy for the item being tested. In the first case, this is a 'necessary' (but not sufficient) proxy and so the test is a 'necessary test' (using the terminology given in Henze, Hlávka, and Meintanis (2014)). In the second case, it is a sufficient (but not necessary) proxy. Further tests along these lines could be developed in principle, which will be touched upon briefly in the Discussion.

The restrictions $E(X) = (0, 0)^T$ and $E||X|| = 1$ are useful for our developments as they simplify notation and theoretical developments, and furthermore ensure the scale–independence of the developed test statistics. However, they do not form a restriction to the generality of our approach since, as explained above, each random vector can be standardized appro-

priately. There is an essential difference of this type of scaling as opposed to scaling with the inverse Σ^{-1} of the covariance matrix $\Sigma = \mathrm{Cov}(X)$. The inverse–covariance scaling would remove the differing variability in the directions of the two axes. This forms, at least in the case of spherical symmetry and possibly also in the PCA case, part of the characteristic that we want to test. Of course, the suggested standardization can also be applied for data sets: Suppose that $\{x_1, \ldots x_n\}$, where $x_i = (x_{i1}, x_{i2})^T$, $i = 1, \ldots, n$, is a random sample. Then one can mean–standardize the data by subtracting the vector $(\bar{x}_1, \bar{x}_2)^T$, where $\bar{x}_k = n^{-1} \sum_{i=1}^n x_{ik}$, and subsequently dividing all observations by $\hat{r} = n^{-1} \sum_{i=1}^n \sqrt{x_{i1}^2 + x_{i2}^2}$.

The divergence of the data from the self–consistency property needs to be captured by an adequate test statistic. This will be achieved through empirical characteristic functions, which are introduced in the next section, and the test statistic will be developed there in a general framework which encompasses both test problems mentioned above. Section 3 will then produce expressions of the test statistic, and discuss their empirical distribution, for the two cases separately. Section 4 will give simulations and a real data example, before we conclude this paper in Section 5.

2 A self–consistency–based test statistic

Recall the self–consistency property of principal curves (1). In order to develop a test statistic which can measure the degree of violation of the self–consistency property, we make use of a Fourier–type approach introduced by Bierens (1982): for real y and given a function $\omega(v)$ of a k–vector v, the equation $\mathbb{E}(y|v) = \omega(v)$ holds if and only if $\mathbb{E}[\{y - \omega(v)\}e^{it^T v}] = 0$, for all $t \in \mathbb{R}^k$. Writing the left–hand side of (1) short as $E(X|\lambda)$, we apply Bierens' result which

tells us

$$E(X|\lambda) = C(\lambda) \Longleftrightarrow E(X - C(\lambda))e^{it\lambda} = 0 \quad \forall t.$$

With $C(\lambda) = (C_1(\lambda), C_2(\lambda))^T$, the above implies

$$E \begin{pmatrix} (X_1 - C_1(\lambda))e^{it\lambda} \\ (X_2 - C_2(\lambda))e^{it\lambda} \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In order to measure deviation from the self–consistency property, define the empirical quantities

$$d_k(t) = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - C_k(\lambda_i))\, e^{it\lambda_i} \quad k = 1, 2, \tag{4}$$

where $\lambda_i \equiv \lambda_C(x_i)$ is the projection index for orthogonal projection of a data point $x_i = (x_{i1}, x_{i2})^T$ on $C(\cdot)$. Note that $d_k(t)$ can be considered as the empirical characteristic function of the random variable $X_k - C_k(\lambda_C(X_1, X_2))$, $k = 1, 2$. Then an omnibus test statistic is given by

$$\begin{aligned} D &= \int \left[||d_1(t)||^2 + ||d_2(t)||^2 \right] w(t)\, dt = \\ &= \int ||d_1(t)||^2 w(t)\, dt + \int ||d_2(t)||^2 w(t)\, dt \equiv D_1 + D_2 \end{aligned} \tag{5}$$

with some suitable weight function $w(\cdot)$, the choice of which we discuss further down the paper. Large values of this test statistic will indicate deviation from the self–consistency property.

The norm of the quantities (4) featuring in (5) expand as

$$\begin{aligned} ||d_k(t)||^2 &= \frac{1}{n^2} || \sum_{i=1}^{n} (x_{ik} - C_k(\lambda_i)) \cos t\lambda_i + i \sum_{i=1}^{n} (x_{ik} - C_k(\lambda_i)) \sin t\lambda_i ||^2 \\ &= \frac{1}{n^2} \left\{ \left[\sum_{i=1}^{n} (x_{ik} - C_k(\lambda_i)) \cos t\lambda_i \right]^2 + \left[\sum_{i=1}^{n} (x_{ik} - C_k(\lambda_i)) \sin t\lambda_i \right]^2 \right\} \end{aligned} \tag{6}$$

for $k = 1, 2$. For brevity of notation let now $\epsilon_{ik} = x_{ik} - C_k(\lambda_i)$ (which can, in either test scenario, be entirely determined from the data, see Section 3 for details). Then

$$\begin{aligned} D_k &= \int_t ||d_k(t)||^2 w(t)\, dt \\ &= \frac{1}{n^2} \int \left[\sum_{i=1}^{n} \epsilon_{ik} \cos t\lambda_i \right]^2 w(t)\, dt + \frac{1}{n^2} \int \left[\sum_{i=1}^{n} \epsilon_{ik} \sin t\lambda_i \right]^2 w(t)\, dt \\ &= \frac{1}{n^2} \int \sum_{i,j} \epsilon_{ik}\epsilon_{jk} \cos(t\lambda_i) \cos(t\lambda_j) w(t)\, dt + \frac{1}{n^2} \int \sum_{i,j} \epsilon_{ik}\epsilon_{jk} \sin(t\lambda_i) \sin(t\lambda_j) w(t)\, dt \\ &= \frac{1}{n^2} \sum_{i,j}^{n} \epsilon_{ik}\epsilon_{jk} \int \cos(t(\lambda_i - \lambda_j)) w(t)\, dt \end{aligned}$$

Hence,

$$\begin{aligned} D = D_1 + D_2 &= \frac{1}{n^2} \sum_{k=1}^{2} \sum_{i,j}^{n} \epsilon_{ik}\epsilon_{jk} \int \cos(t(\lambda_i - \lambda_j)) w(t)\, dt \\ &\equiv \frac{1}{n^2} \sum_{i,j}^{n} I_w(\lambda_i - \lambda_j)(\epsilon_{i1}\epsilon_{j1} + \epsilon_{i2}\epsilon_{j2}) \\ &= \frac{1}{n^2} \sum_{i,j}^{n} I_w(\lambda_i - \lambda_j) < \epsilon_i, \epsilon_j > \end{aligned} \tag{7}$$

where

$$I_w(\lambda) = \int \cos(t\lambda) w(t)\, dt$$

and $< \epsilon_i, \epsilon_j >= \epsilon_i^T \epsilon_j$ is the inner product between residuals $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})^T$ and $\epsilon_j = (\epsilon_{j1}, \epsilon_{j2})^T$.

Further simplification of quantity (7) requires the choice of a weight function, $w(t)$. The choice $w(t) = \exp(-at^2)$ is attractive in terms of computational convenience. Using this weight function, one finds

$$I_w(\lambda_i - \lambda_j) = \int \cos[t(\lambda_i - \lambda_j)] e^{-at^2}\, dt = \sqrt{\frac{\pi}{a}} \exp\left(-\frac{(\lambda_i - \lambda_j)^2}{4a} \right).$$

That is, using the index a to emphasize the dependence on the weight parameter, the statistic

D takes the form

$$D_a = \sqrt{\frac{\pi}{a}} \frac{1}{n^2} \sum_{i,j} \exp\left(-\frac{(\lambda_i - \lambda_j)^2}{4a}\right) < \epsilon_i, \epsilon_j > \tag{8}$$

It is clear from the shape of this statistic that a pair (x_i, x_j) of data points will only contribute relevant information to I_w if the corresponding projection indices λ_i and λ_j are 'close' (as measured by the weight function), where this contribution will be positive if their residuals point in the same direction and negative if they point in opposite direction. This makes intuitively sense: Spherical symmetry implies that the residual distribution on either side of the unity circle is balanced.

The precise shape of the projections λ_i and the residuals ϵ_{ik}, $i = 1, \ldots, n$, $k = 1, 2$, depend on the geometry of the specific test problem. Explicit expressions for these quantities will be provided for both considered test scenarios in Section 3. For now, it is sufficient to emphasize that these quantities are entirely known from the data.

The choice for the weight function $w(\cdot)$ is usually based upon computational considerations. In fact if $w(\cdot)$ integrates to one (even after some scaling) and satisfies $w(-t) = w(t)$ then the function $I_w(\cdot)$ figuring below eqn. (7) may be interpreted as the characteristic function of a symmetric around zero random variable λ having density $w(\cdot)$. Hence $w(\cdot)$ can be chosen as the density of any such distribution. Clearly then the choice e^{-at^2} corresponds to a zero–mean normal density. Other convenient examples for $w(\cdot)$ are the Laplace density, the density of mixtures of normals, symmetric stable or Laplace distributions, and combinations thereof. In fact, one might wonder whether there is a weight function which is optimal in some sense. The problem is still open but based on earlier finite–sample results it appears that the issue of the choice of w is similar to the corresponding problem of choosing a kernel and a bandwidth in nonparametric estimation: Most weight functions (kernels) render similar behavior of the test statistic, but there is some sensitivity with respect to the "bandwidth"

parameter $a > 0$ figuring in (8). This is a highly technical problem that has been tackled only under the restrictive scenario of testing goodness–of–fit for a given parametric distribution, and even then a good choice of a depends on the direction away from the null hypothesis; see Tenreiro (2009). Thus in our context the approach to the weight function is in some sense pragmatic: We use the Gaussian weight function which has become something of a standard, and investigate the behavior of the criterion over a grid of values of the weight parameter a; see Sections 3 and 4.

3 Specific test scenarios

3.1 Testing spherical symmetry

In what follows, the general concept introduced in Section 2 is used to devise a test for the hypothesis H_0: 'The distribution of the bivariate random vector X is spherically symmetric' against the alternative that this is not the case. Inspired by property (2), consider a curve C parametrized by $\lambda \in \mathbb{R}$ which proceeds along the unit circle, that is, one has

$$||C(\lambda)|| = E||X|| = 1 \text{ for all } \lambda.$$

We use the convention to parametrize the curve by the angle $\lambda \in [0, 2\pi)$ between $C(\lambda)$ and the abscissa (that is, allowing to set $b = 2\pi$ would close the circle at the point $(1, 0)^T$). Each data point x_i will then orthogonally project onto some point on the circle, and the resulting projection index $\lambda_i = \lambda_C(x_i)$ can be interpreted as an angle. Hence, due to the unitary property of the circle, it follows

$$\begin{aligned} C_1(\lambda_i) &= \cos \lambda_i \,, \\ C_2(\lambda_i) &= \sin \lambda_i \,. \end{aligned}$$

Using further that $\tan\lambda_i = x_{i2}/x_{i1}$, one can compute the projection of x_i onto the circle as

$$\begin{aligned}
\begin{pmatrix} C_1(\lambda_i) \\ C_2(\lambda_i) \end{pmatrix} &= \begin{pmatrix} \cos\lambda_i \\ \sin\lambda_i \end{pmatrix} = \begin{pmatrix} \pm\frac{1}{\sqrt{1+\tan^2\lambda_i}} \\ \pm\frac{\tan\lambda_i}{\sqrt{1+\tan^2\lambda_i}} \end{pmatrix} = \pm\frac{1}{\sqrt{1+\tan^2\lambda_i}} \begin{pmatrix} 1 \\ \tan\lambda_i \end{pmatrix} \\
&= \pm\frac{1}{\sqrt{1+\left(\frac{x_{i2}}{x_{i1}}\right)^2}} \begin{pmatrix} 1 \\ \frac{x_{i2}}{x_{i1}} \end{pmatrix} = \frac{x_{i1}}{\sqrt{x_{i1}^2+x_{i2}^2}} \begin{pmatrix} 1 \\ \frac{x_{i2}}{x_{i1}} \end{pmatrix} \\
&= \frac{1}{\sqrt{x_{i1}^2+x_{i2}^2}} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = \frac{1}{||x_i||} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}, \qquad (9)
\end{aligned}$$

where the sign '+' applies for $\lambda_i \in [0, \pi/2) \cup [3\pi/2, 2\pi)$ and '$-$' otherwise. The result (9) is quite obvious; these are just the original data standardized by their norm. Then, it follows also that

$$\epsilon_{ik} = x_{ik} - C_k(\lambda_i) = x_{ik} - \frac{1}{||x_i||}x_{ik} = x_{ik}\left(1 - \frac{1}{||x_i||}\right). \qquad (10)$$

[For a graphical illustration of residuals (green) and projections (red), see Figure 1 (left) for a spherically symmetric distribution (bivariate normal with zero covariance), and Figure 1 (right) for a non–symmetric data set (the Old Faithful Geyser data; for further details on these data see Section 4).] Hence,

$$\begin{aligned}
< \epsilon_i, \epsilon_j > &= \left(1 - \frac{1}{||x_i||}\right)\left(1 - \frac{1}{||x_j||}\right)(x_{i1}x_{j1} + x_{i2}x_{j2}) \\
&\equiv \left(1 - \frac{1}{||x_i||}\right)\left(1 - \frac{1}{||x_j||}\right) < x_i, x_j > .
\end{aligned}$$

So, from (8),

$$D_a^S = \sqrt{\frac{\pi}{a}}\frac{1}{n^2}\sum_{i,j} < x_i, x_j > \left(1 - \frac{1}{||x_i||}\right)\left(1 - \frac{1}{||x_j||}\right)\exp\left(-\frac{(\lambda_i - \lambda_j)^2}{4a}\right)$$

where the hyperscript S stands for the *Symmetry* test. Computation of the squared differences $(\lambda_i - \lambda_j)^2$ for use in this equation needs some care: Note that $\lambda_i = \arctan(x_{i2}/x_{i1}) + c_i$, where $c_i = 0$ [if $x_1 \geq 0, x_2 \geq 0$)], $c_i = \pi$ [if $x_1 < 0$] or $c_i = 2\pi$ [if $x_1 \geq 0, x_2 < 0$], respectively, and

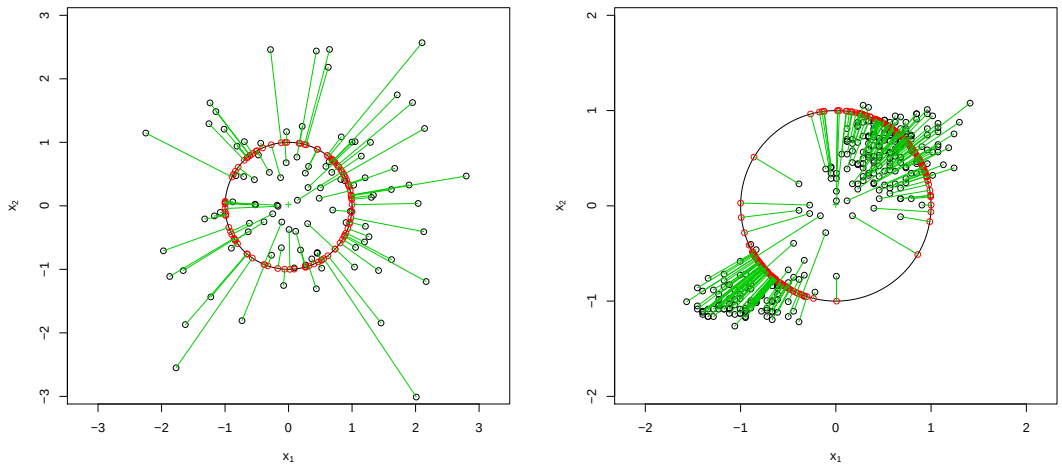


Figure 1: Residuals (green) and projected points (red) for a bivariate normal data set (left) and the Old Faithful Geyser data set (right).

that we are operating on a circle with maximum parametrization 2π; that is any distance $|\lambda_i - \lambda_j| > \pi$ needs to be mapped to $2\pi - |\lambda_i - \lambda_j|$.

In order to use this test statistic for practical purposes, one needs insight into the null distribution of D_a^S. It turns out (not shown) that this depends on the sample size, in the sense that $D_a^S \propto 1/n$. Therefore, we use

$$T_a^S := nD_a^S$$

as test statistic for the spherical symmetry test. Figure 2 shows the distribution of T_a^S (using $a = 0.5$) for 100000 samples from a bivariate normal distribution $N[(0,0)^T, I_2)]$ of size $n = 50$, $n = 100$, $n = 200$, and $n = 400$, where I_2 stands for the identity matrix of dimension two. One finds that the distribution of $T_{0.5}^S$ takes the same location and shape irrespective of the value of n, and is skewed to the right.

Table 1 shows that the quantiles are consistent across different sample sizes. Of course, these quantiles still depend on a. A repetition of the above experiments using $a = 0.1, 0.2, \ldots, 1$

Table 1: Empirical quantiles of the statistic $T_{0.5}^S$ under H_0, for sample size n.

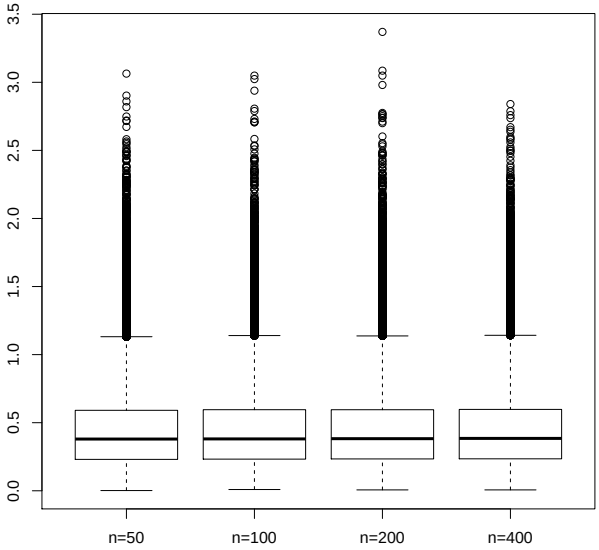
quantile	50%	90%	95%	97.5%	99%	99.9%
n						
50	0.38	0.85	1.04	1.22	1.46	2.06
100	0.38	0.85	1.04	1.23	1.45	1.99
200	0.38	0.85	1.03	1.22	1.45	2.07
400	0.38	0.85	1.03	1.21	1.43	1.98
median	0.38	0.85	1.04	1.22	1.45	2.03

gives the 95% quantiles (that is, the critical values for the 5% level of significance) as reported in Table 2. We again observe very clearly the independence of the quantiles from the sample size. While it cannot be inferred from this table whether certain values of a are preferred to other ones, the choice $a = 1/2$, with associated critical value of 1.04 at the 5% level of significance, is convenient for general use. It should be noted that these quantiles are also independent of the location and scale of the random vector, due to the normalization to expectation equal to 0 and expected norm equal to 1. So, as a general reference, the values provided in the row labelled **median** in Tables 1 and 2, which give the respective column medians (taken before rounding), can be used as critical values for this test, for bivariate data of arbitrary size, location and scale. A word of caution does apply to data which are strongly non–normal. We elaborate on this point in the Discussion.

Table 2: Empirical 95% quantiles of the statistic T_a^S under H_0, for weight parameter a and sample size n.

a	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
n										
50	2.32	1.62	1.32	1.15	1.04	0.96	0.90	0.85	0.82	0.79
100	2.30	1.62	1.33	1.15	1.04	0.96	0.90	0.85	0.82	0.78
200	2.30	1.62	1.32	1.15	1.03	0.95	0.90	0.85	0.81	0.78
400	2.28	1.61	1.31	1.14	1.03	0.95	0.89	0.85	0.81	0.78
median	2.30	1.62	1.32	1.15	1.04	0.96	0.90	0.85	0.81	0.78

Figure 2: Distribution of test statistic $T_{0.5}^S$, with $n = 50, 100, 200$ and 400 (from left to right).



3.2 Testing for principal components

Let us consider again a bivariate random vector with $E(X) = (0,0)^T$ and $E||X|| = 1$, and data $(x_{i1}, x_{i2})^T, i = 1, \ldots, n$ generated as a random sample from X. In this second scenario we are given a vector v through the origin, and would like to test the hypothesis H_0: 'the vector v is a principal component', against the alternative that this is not the case. Equivalently, one could say 'H_0: the vector $C(\lambda) = \lambda v$, $\lambda \in \mathbb{R}$ is a principal component line'.

The geometry of this scenario is different. Firstly, the parametrization is now allowed to move freely on the real line, that is, we have (in principle) $a \longrightarrow -\infty$ and $b \longrightarrow \infty$. However, due to the scaling $E||X|| = 1$, most observations will still orthogonally project into a range of values of λ which is unlikely to increase beyond, say, $[-3, 3]$, unless one or both of the involved random variables are highly skewed.

The projection index is now given by $\lambda_i \equiv \lambda_C(x_i) = v^T x_i$, and so

$$\begin{aligned} I_w(\lambda_i - \lambda_j) &= \sqrt{\frac{\pi}{a}} \exp\left(-\frac{[v^T x_i - v^T x_j]^2}{4a}\right) \\ &= \sqrt{\frac{\pi}{a}} \exp\left(-\frac{(x_i - x_j)^T v v^T (x_i - x_j)}{4a}\right) \end{aligned}$$

To compute the residuals ϵ_i, note that the projected point $x_i^{'}$ on the line $C(\lambda) = \lambda v$ is given by

$$x_i^{'} = v v^T x_i$$

and so the i-th residual vector is obtained as

$$\epsilon_i = x_i - x_i^{'} = x_i - v v^T x_i = (I_2 - v v^T) x_i.$$

Hence,

$$\begin{aligned} < \epsilon_i, \epsilon_j > &= < (I_2 - vv^T)x_i, (I_2 - vv^T)x_j > \\ &= x_i^T [I_2 - vv^T]^2 x_j \\ &= x_i^T [I_2 - vv^T] x_j \end{aligned}$$

Hence, from (8),

$$D_a^P = \sqrt{\frac{\pi}{a}} \frac{1}{n^2} \sum_{i,j} x_i^T \left[I_2 - vv^T \right] x_j \exp\left(\frac{(x_i - x_j)^T vv^T (x_i - x_j)}{4a} \right) \tag{11}$$

where the hyperscript P now stands for the *Principal Component* test. Figure 3 illustrates the empirical distributions of

$$T_a^P := nD_a^P$$

for $a = 0.5$ and $n = 50$, $n = 100$, $n = 200$, and $n = 400$, and Table 3 gives again the tabulated empirical quantiles, based on 100000 simulation runs. In Table 4 we also provide a list of quantiles across different values of a. As for the symmetry test, we find that the empirical quantiles of T_a^P are very stable across different sample sizes. The last row of each table gives again the column medians which can be used as critical values for general reference.

4 Examples and simulations

We begin with a real data example which gives some insight how the proposed tests work in practice. We then proceed with detailed simulation studies in order to assess the nominal level attainment and power, thereby focussing on the spherical symmetry test.

Figure 3: Distribution of test statistic $T_{0.5}^P$, with $n = 50$, 100, 200 and 400 (from left to right).

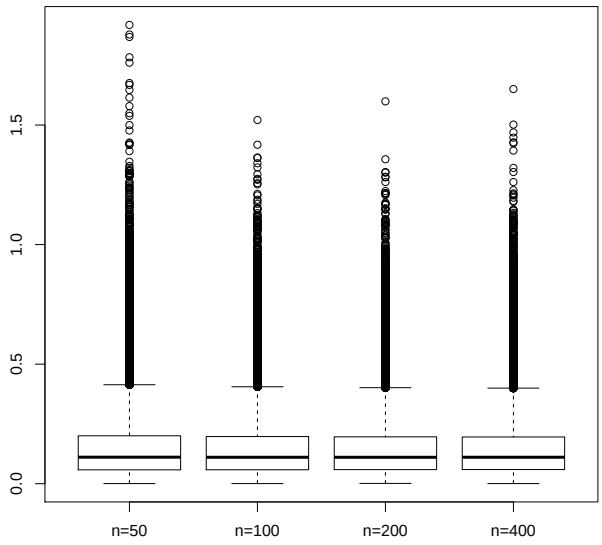


Table 3: Empirical quantiles of the statistic $T_{0.5}^P$ under H_0, for sample size n.

quantile	50%	90%	95%	97.5%	99%	99.9%
n						
50	0.11	0.32	0.42	0.52	0.66	1.05
100	0.11	0.31	0.41	0.50	0.63	0.94
200	0.11	0.31	0.40	0.49	0.62	0.91
400	0.11	0.31	0.40	0.49	0.62	0.94
median	0.11	0.31	0.40	0.50	0.62	0.94

Table 4: Empirical 95% quantiles of the statistic T_a^P under H_0, for weight parameter a and sample size n.

a	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
n										
50	1.22	0.80	0.61	0.49	0.42	0.36	0.32	0.25	0.23	0.21
100	1.18	0.79	0.61	0.49	0.41	0.36	0.31	0.25	0.23	0.23
200	1.14	0.77	0.60	0.49	0.40	0.36	0.32	0.25	0.23	0.22
400	1.13	0.76	0.58	0.48	0.40	0.35	0.30	0.27	0.24	0.22
median	1.16	0.78	0.60	0.49	0.40	0.36	0.31	0.25	0.23	0.22

4.1 Real data examples

We consider the Old Faithful Geyser data for this example. The data were presented in Härdle (1991) and are available as data set `faithful` in the statistical programming environment R (R Core Team 2016). They consist of $n = 272$ observations ($\texttt{waiting}_i$, $\texttt{eruption}_i$) on the waiting time between two consecutive eruptions (in minutes) and the subsequent eruption time (in minutes). A scatterplot of the mean–centered and unit-norm scaled data is provided in Figure 1 (right). Applying now the test for spherical symmetry onto this data set, one finds $T_{0.5}^S = 14.87$ and $T_1^S = 11.51$. From Tables 1 and 2 it is clear that this means strong rejection of the null hypothesis 'Spherical Symmetry' at any reasonable significance level — of course, this is in line with what we would expect for this data set.

Next, we consider the test for the principal component property. Specifically, we consider five different null hypotheses v:

(i) the actual first principal component line, which is given by the vector $v = (0.9971, 0.0755)^T$,

Table 5: Value of T_a^P for null hypotheses (i) to (v), and two different values of a.

H_0	(i)	(ii)	(iii)	(iv)	(v)
a					
0.5	0.029	1.738	6.266	0.030	0.057
1	0.007	1.040	2.228	0.007	0.035

(ii) the abscissa, represented by $v = (1, 0)^T$,

(iii) the ordinate, defined by $v = (0, 1)^T$,

(iv) the regression of `eruption` on `weighting`, given by $v = (0.9972, 0.0754)^T$,

(v) the regression `weighting` on `eruption`, given by $v = (0.9957, 0.0928)^T$.

All five lines are depicted in Figure 4. It is clear that we would expect hypothesis (i) to be accepted and hypotheses (ii) and (iii) to be rejected. It is also clear that, conceptually, the regression line (iv) needs to be 'flatter' than the principal component line, and the regression line (v) needs to be steeper. In this case, line (iv) almost coincides with line (i). It is, hence, interesting to test whether a statistically significant difference between the regression lines and the principal component line is observed.

Table 5 gives the values of the test statistic T_a^P with $a = 0.5$ and $a = 1$, respectively. By comparison of Table 5 with Table 4 one sees that the test behaves plausibly. As expected, the test fails to reject hypotheses (i) and (iv), and correctly (and strongly) rejects hypotheses (ii) and (iii). The actually interesting case is hypothesis (v), where Figure 4 suggests that the decision will not be obvious. Focusing on, say, $a = 0.5$, we see via Table 4 that $0.057 < 0.40$ and hence the null hypothesis is not rejected. Hence, in terms of its orientation, this regression

line can not be significantly distinguished from the principal component line.

For completeness, it should be noted that we have not scaled the two variables of this data set to unity variance (as sometimes done in PCA) before carrying out the two test procedures of the two data sets. Of course, one could do this, which would need to happen *before* scaling the data to unity norm. (The scaling to unity norm would move the two variances away from the value 1 again, but this does not matter — important is just that the two values would be equal). Finally, it is worth saying that all results, for both tests, would stay exactly the same if the two variables were considered in reverse order, that is (`eruption`, `waiting`) rather than (`waiting`, `eruption`).

4.2 Simulations for the spherical symmetry test

We consider three simulation scenarios for bivariate distributions. In the first scenario, we generate bivariate data sets of sample size $n = \text{Poisson}(200)$ from a mixture distribution $m \times N(0, 5uI_2) + (1 - m) \times N(4, uI_2)$, where $u \sim U(0, 1)$ (of course, for all observations generated *within* each data set, u is constant). The mixture proportion m is varied between $m = 0, 0.1, 0.2, \ldots, 0.9, 1$, and the weight parameter a is varied from $a = 0.1$ to $a = 1$ in steps of 0.1. We create 5000 data sets for each value of m and a, which are considered as the design variables of the experiment. Clearly, the cases $m = 0$ and $m = 1$ correspond to spherical symmetry, so that in this case the proportion of rejection should correspond to the nominal significance level, which we set to be $\alpha = 0.05$. For each simulated data set, we compute the test statistic T_a^S and then calculate the proportion of times that this value exceed the critical value 1.04 motivated earlier.

The resulting powers are given in Table 6. We find that the powers are correctly sitting at about 5% in the first and last row of the table, as expected. For other values of m, there is

Figure 4: Graphical illustration of the lines corresponding to hypotheses (i)-(v). In the legend, `w` stands for variable `waiting`, and `e` for `eruption`. [Note that line (iv) is in fact almost exactly overlaying line (i), and has been moved upwards in vertical direction by the value 0.001 to make it optically visible.]

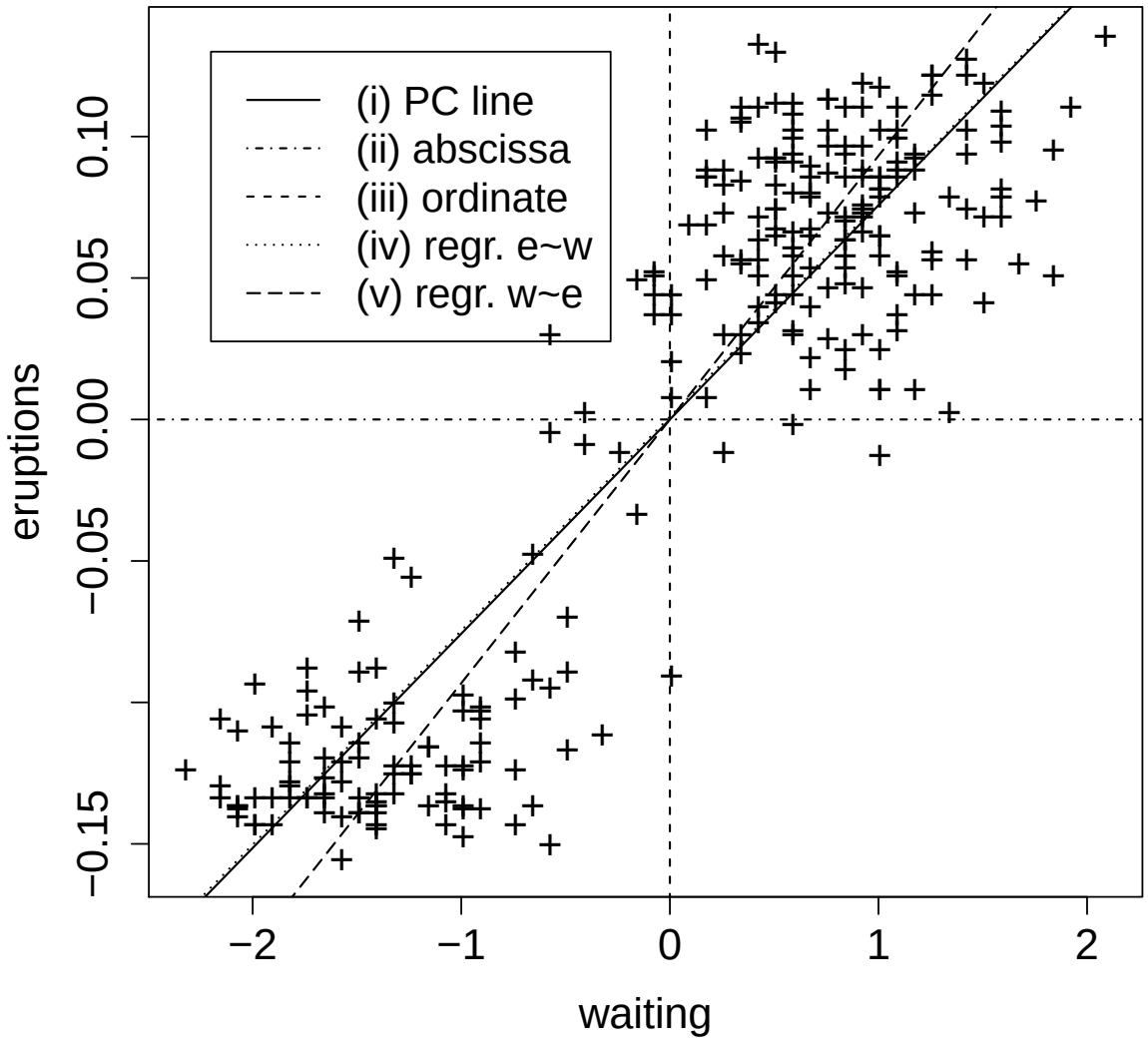


Table 6: Rejection proportions of the spherical symmetry test, for the first simulation experiment.

a	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
m										
0	0.048	0.049	0.049	0.049	0.049	0.048	0.047	0.047	0.046	0.049
0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.4	0.995	0.996	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
0.5	0.625	0.646	0.651	0.653	0.654	0.656	0.657	0.659	0.659	0.662
0.6	0.632	0.638	0.638	0.635	0.631	0.630	0.627	0.622	0.614	0.614
0.7	0.949	0.944	0.939	0.934	0.931	0.928	0.924	0.821	0.916	0.914
0.8	0.998	0.997	0.996	0.995	0.993	0.992	0.991	0.989	0.987	0.987
0.9	0.985	0.978	0.972	0.967	0.963	0.958	0.954	0.950	0.946	0.943
1	0.053	0.049	0.049	0.050	0.051	0.050	0.050	0.051	0.049	0.051

deviation from spherical symmetry, and this is correctly picked up by the test, though the powers depend on the degree of asymmetry of the mixture problem.

As a second example, we consider the alternative $H_1^{(2)}$ from Henze, Hlávka, and Meintanis (2014), namely

$X = (Z_1, Z_2)^T$ has independent components with $Z_1 \sim N(0,1)$ and $Z_2 \sim \text{Exp}(1)$.

We produce 1000 data sets of size $n = 100$ and $n = 200$ respectively, and obtain the powers as given in Table 7, for the weight parameters $a = 0.2$, $a = 0.5$, and $a = 0.8$, respectively. We also provide comparison to the powers attained via the Koltchinksi-Li test (KL, Koltchinskii and Li 1998), the local empirical likelihood (LEL, Einmahl and Gantner 2012), and a Cramér–

Table 7: Powers of second simulation experiment.

Test	T_a^S			KL	LEL	CM			
a	0.2	0.5	0.8			0.05	0.2	0.5	0.8
$n = 100$	0.97	0.97	0.97	0.90	0.89	0.96	0.91	0.70	0.60
$n = 200$	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	0.99

von Mises– type test statistic (CM, Henze, Hlávka, and Meintanis 2014). We refer to the latter manuscript for a detailed description of all tests.

Notably, the CM test is based on the empirical characteristic function as well, and hence it also needs a weight function. In concordance with the techniques presented in here, we consider for this purpose the weight $w(t) = \exp(-at^2)$. Since the CM test varies more strongly with the weight parameter a than the self–consistency based test, we provide for the CM test additionally the result for $a = 0.05$. Results are provided in Table 7. It is clear that the proposed self–consistency test demonstrates superior behavior to all other tests: In the case $n = 100$, the powers are uniformly higher, and in the case $n = 200$, it shares a power of 1 with all tests except the KL test. As indicated earlier, it is evident that the performance of the self–consistency test is not sensitive to the choice of the weight parameter at all. It is further noted at this occasion that we have only considered values of a in the range $0 < a \leq 1$ in all our experiment, since, considering the impact of the weight parameter in (8), this setting seems to be at the correct scale given the range of projection parameters. Setting a larger than 1 or closer to 0 will still lead to essentially unchanged result as long as no numerical problems occur (if a tends to 0 it will cause a singularity due to its position in the denominator, and if it is excessively large it will annihilate all information).

Finally, following the setup of a simulation experiment in Henze, Hlávka, and Meintanis (2014), we investigate the test power against two–dimensional normal distributions with correlation $\rho \in \{0, 0.2, 0.4, 0.6\}$ (that is, $\rho = 0$ corresponds to H_0). We use 10000 data sets of size $n = 200$. We see from the upper part of Table 8 that the nominal level is well attained and that powers increase successively with increasing ρ (that is, with decreasing spherical symmetry), up to almost the value 1 for $\rho = 0.6$.

In order to illustrate the limitations of our approach, we repeat this analysis using a bivariate Laplace distribution, using the same variance matrices as used for the bivariate normal distribution. We see from the lower part of Table 8 that the powers are still good and build up in the correct direction, but that the nominal level is now poorly aligned. So, it appears to be the case that the critical values reported in Tables 1 and 2 are only correct for Gaussian(–like) base distributions. See the Discussion for some further comments on this matter.

5 Discussion

In this work, we have combined several different statistical concepts which do not appear immediately related in order to develop an entirely novel class of statistical tests for bivariate distributions. The developed family of tests statistics make use of empirical characteristic functions to measure the deviation from the self–consistency property. Different test problems can then be addressed by considering different properties of self–consistency for bivariate distributions. Specifically, in this paper, we have considered two special cases: Firstly, using that for centred and spherically symmetric distributions, each circle with radius equal to the mean norm is self–consistent, we have constructed a test for spherical symmetry. Secondly, using the property that, if a straight line is self–consistent then it is a principal component,

Table 8: Powers under the third simulation scenario.

Bivariate Normal

a	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ρ										
0	0.048	0.047	0.047	0.047	0.048	0.048	0.048	0.049	0.048	0.049
0.2	0.265	0.255	0.246	0.238	0.226	0.218	0.208	0.202	0.189	0.184
0.4	0.862	0.856	0.844	0.830	0.815	0.802	0.787	0.770	0.743	0.729
0.6	0.999	0.999	0.998	0.998	0.998	0.997	0.997	0.996	0.994	0.994

Bivariate Laplace

a	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	10
ρ										
0	0.665	0.612	0.588	0.574	0.563	0.558	0.554	0.551	0.548	0.542
0.2	0.782	0.744	0.724	0.709	0.697	0.688	0.680	0.673	0.667	0.659
0.4	0.945	0.932	0.925	0.917	0.909	0.902	0.895	0.890	0.883	0.876
0.6	0.998	0.996	0.996	0.994	0.993	0.992	0.991	0.991	0.989	0.988

we devise a test for the null hypothesis that a given line is a principal component. It has been noted already in the introduction that, in either case, the self–consistency property just serves as a (necessary and sufficient, respectively) proxy for the question being tested. A further 'necessary' test could be constructed based on the statement (Hastie and Stuetzle 1989)

$$\begin{aligned} &\text{For any two–dimensional ellipsoidal distribution } X \in \mathbb{R}^2, \\ &\text{the principal components are principal curves.} \end{aligned} \tag{12}$$

Hence, one could construct a test for H_0: *X follows an ellipsoidal distribution* by checking the first two principal components for self–consistency. If only one principal component is considered for this purpose, then the test statistic for this test would in fact take identical shape as the one produced based on (11). Actually, since one of the two tests is based on a necessary proxy and the other one on a sufficient one, the interplay of the two principal–component based tests is interesting: If one *knows* that X is ellipsoidal and the test discussed in Subsection 3.2 rejects the null hypothesis, then it is clear with certainty that the given line was not a principal component. If one *knows* that a given line is a principal component and H_0 (as formulated just after (12)) is rejected, then the distribution of X is proven to be not ellipsoidal. If one does not have certainty about either property, then rejection of the null hypothesis (of either test) could strictly mean that the distribution of X is either not ellipsoidal, or the given line was not a principal component. To avoid such potential ambiguities, it would be desirable to develop self–consistency based criteria which relate more directly to statistical properties of interest.

In the cases considered, the produced test statistics had a simple analytic expression, and its null distribution was independent on location and scale of the random vector, as well as the sample size. The distribution did depend on the specific value of the weight parameter a,

though this is not an obstacle in practice since the critical values for given a can be tabulated, and the choice of a is down to the data analyst. The *performance* of the test procedure (in terms of power and nominal level attainment) did *not* depend on the weight parameter, with excellent powers achieved under all simulation scenarios. We should, however, add that all critical values reported in this manuscript were obtained from distributions X of bivariate normal character. For non–normal base distributions, it appears from our limited simulation studies that, in order to attain the target significance level, the critical values as provided in Tables 1, 2, 3, and 4 would need to be multiplied by a 'small' constant γ which is largely independent of n and a but is characteristic for the test problem and the underlying distribution. For instance, in case of the bivariate Laplace distribution, γ takes a value of about 3 for the symmetry test and about 1.7 for the principal component test. What remains true, for any distribution of X, is that large values of the test statistic corresponds to more evidence against the respective null hypothesis. Further research which gives insight into the theoretical distribution of the self-consistency based test statistics would be desirable.

It is finally noted that the computation of the test statistics is extremely quick and so attractive from a computational point of view. Also, even though the statistic T_a^S contains denominators which could – in principle – cause computational instabilities when taking the value 0, we did not observe this to happen in practice, even in tens of thousands of simulation runs.

The test ideas presented still have considerable scope for extension beyond the scope of bivariate distributions, and beyond the use of self–consistent *curves*. For instance, one could be interested in testing whether a multidimensional clustering routine has actually identified the *principal points* (Flury 1990). These are cluster centres which constitute the average of all points which are closest to those centres. For tests based on such principal (self–consistent)

points, test statistic (8) cannot be used anymore and therefore needs to be developed from scratch.

Acknowledgements

The authors wish to thank two referees for their constructive comments.

References

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*, 105–134.

Einmahl, J. H. J. and M. Gantner (2012). Testing for bivariate spherical symmetry. *TEST 21*(1), 54–73.

Flury, B. D. (1990). Principal points. *Biometrika 77*, 33–41.

Härdle, W. (1991). *Smoothing Techniques. With Implementation in S.* New York: Springer Verlag.

Hastie, T. and W. Stuetzle (1989). Principal curves. *J. Amer. Statist. Assoc. 84*, 502–516.

Henze, N., Z. Hlávka, and S. Meintanis (2014). Testing for spherical symmetry via the empirical characteristic function. *Statistics 48*(6), 1282–1296.

Koltchinskii, V. and L. Li (1998). Testing for spherical symmetry of a multivariate distribution. *Journal of Multivariate Analysis 65*(2), 228–244.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tarpey, T. and B. Flury (1996). Self-consistency: A fundamental concept in statistics. *Statistical Science 11*, 229–243.

Tenreiro, C. (2009). On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Computational Statistics and Data Analysis 53*(4), 1038–1053.