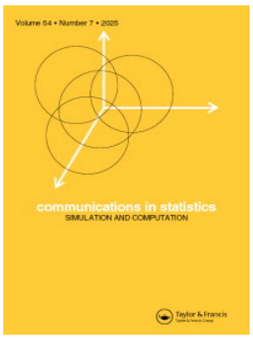

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: www.tandfonline.com/journals/lssp20

K-depth tests for testing simultaneously independence and other model assumptions in time series

Hendrik Dohme, Dennis Malchercyzk, Kevin Leckey & Christine H. Müller

To cite this article: Hendrik Dohme, Dennis Malchercyzk, Kevin Leckey & Christine H. Müller (23 Oct 2024): *K*-depth tests for testing simultaneously independence and other model assumptions in time series, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2024.2413905

To link to this article: https://doi.org/10.1080/03610918.2024.2413905

Published online: 23 Oct 2024.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2024.2413905

K-depth tests for testing simultaneously independence and other model assumptions in time series

Hendrik Dohme, Dennis Malchercyzk, Kevin Leckey, and Christine H. Müller

Department of Statistics, TU Dortmund University, Dortmund, Germany

ABSTRACT

We consider the recently developed *K*-depth tests for testing simultaneously independence and other model assumptions for univariate time series with a potentially related *d*-dimensional process of explanatory variables. Since these tests are based only on signs of residuals, they are easy to comprehend. They can be used in a full version and in a simplified version. While former investigations already showed that the full version is appropriate for testing model assumptions, we concentrate here on either testing the independence assumption on its own or on simultaneously testing independence and model assumptions with both types of tests. In an extensive simulation study, we compare these tests with several known independence test such as the runs test, the Durbin-Watson test, and the Von-Neumann-Rank-Ratio test. Finally, we demonstrate how the *K*-depth tests can be used for improved modeling of crack width time series depending on temperature measurements in a bridge monitoring.

ARTICLE HISTORY
Received 25 May 2022
Accepted 2 October 2024

KEYWORDS
K-sign depth; Independence; Tests; Time series; Model selection

1. Introduction

We consider here univariate time series $(Y_t)_{t=-p+1,-p+2,\ldots,0,1,2,\ldots,T}$ given by

$$Y_t = g(\theta, Y_{t-p}, \ldots, Y_{t-2}, Y_{t-1}, X_{t-q+1}, \ldots, X_{t-1}, X_t) + E_t \tag{1}$$

for $t = 1, \ldots, T$ where $(X_t)_{t=-q+2,-q+3,\ldots,0,1,2,\ldots,T}$ is a d-dimensional related process, $\theta \in \mathbb{R}^s$ is the model parameter, $g : \mathbb{R}^{s+p+q\cdot d} \to \mathbb{R}$ the model function, and $(E_t)_{t=1,2,\ldots,T}$ is an error process. We assume only that the error variables $E_1, \ldots, E_T$ have a continuous distribution. The classical AR(p) models are special cases of these models.

Given the parameter θ_*, maybe estimated by a previous time series, we want to test

$$H_0\colon \text{ Model } (1) \text{ holds with } \theta{=}\theta_* \text{ and error variables } (\mathrm{E_t})_{\mathrm{t}\ =\ 1,2,\ldots,\mathrm{T}} \text{ such that } E_1, \ldots, E_T \text{ are independent and their median is equal to zero.} \tag{2}$$

This means that we want to test that the time series satisfies a specific model assumption where the error process fulfills quite general conditions besides the independence assumption. In particular, we do not assume variance homogeneity.

CONTACT Christine H. Müller cmueller@statistik.tu-dortmund.de Department of Statistics, TU Dortmund University, Dortmund, Germany.

In such time series, we can regard residuals given by

$$R_t(\theta) := Y_t - g(\theta, Y_{t-p}, ..., Y_{t-2}, Y_{t-1}, X_{t-q+1}, ..., X_{t-1}, X_t).$$

If θ_* is the true model parameter then $R_t(\theta_*) = E_t$ and the test problem (2) is equivalent with:

$$\begin{aligned} &1.\ \ R_1(\theta_*), ..., R_T(\theta_*) \text{ are independent,} \\ &2.\ \ P(R_t(\theta_*) > 0) = \frac{1}{2} = P(R_t(\theta_*) < 0) \text{ for } t = 1, ..., T. \end{aligned} \tag{3}$$

A very simple test for such models is the classical sign tests which uses the fact that, under the true model with true model parameter, the signs of the residuals are Bernoulli distributed with a probability of $\frac{1}{2}$ for a positive sign. Hence, e.g. the number of positive signs can be used as a test statistic, yielding a symmetric binomial distribution under the true model. If this test statistic is too small or too large for realizations $r_1(\theta_*), ..., r_T(\theta_*)$ of $R_1(\theta_*), ..., R_T(\theta_*)$ then the postulated parameter θ_* cannot be the true parameter or the model at all is not correct. However for most models, this simple test has the drawback that deviations $\theta \neq \theta_*$ from the true parameter θ_* exist where the power of the test at θ is very bad since the expected value for a positive residual $R_T(\theta_*) > 0$ remains close to $\frac{1}{2}$. The only exception is the univariate location problem where the sign test is a quite powerful nonparametric test (Lehmann 1994), see also (Kitani and Murakami 2022). An additional problem for the application of the classical sign test for time series models is that the first condition in (3) of independent residuals is not checked with this test.

A similarly simple test for independence in time series is the runs test of Wald and Wolfowitz (1940), see, e.g. Gibbons and Chakraborti (2003, 78–86), or Cui et al. (2018) for a recent modification. It can be applied to the signs of the residuals $R_1(\theta_*), ..., R_T(\theta_*)$ and counts the number of runs. A run is a sequence of equal signs. Note that if N_R is the number of runs and N_S is the number of sign changes then $N_R = N_S + 1$. A low number of runs indicates positive correlation and a high number negative correlation. However, the runs test is not constructed for testing the second condition in (3) of a residual distribution with median equal to zero.

For testing model parameters, Leckey et al. (2023) proposed the so-called K-sign depth tests or shortly K-depth tests. Since these tests are only based on signs of residuals, they are nearly as simple as the sign test and the runs test. These tests are based on the K-sign depth. The full K-sign depth is the relative number of ordered K-subsets $\{t_1, ..., t_K\} \subset \{1, ..., T\}$ with $t_1 < ... < t_K$ for which the corresponding residuals $R_{t_1}(\theta_*), ..., R_{t_K}(\theta_*)$ have alternating signs. In a simplified version of the K-sign depth, only subsets of consecutive indices $t_1, t_1 + 1, ..., t_1 + K - 1$ are used. Leckey et al. showed in particular that the full K-depth test based on the full version of K-sign depth is equivalent to the classical sign test for $K = 2$ and demonstrated theoretically and by simulations that the full K-depth tests are much more powerful than the classical sign test for $K > 2$. They also mentioned that the simplified 2-depth test based on the simplified version of 2-sign depth is similar to the runs test but did not study this case any further.

In particular, a low K-sign depth indicates a bad fit of the model and/or positive correlation while a high K-sign depth may indicate negative correlation. Since Leckey et al. (2023) used the full K-depth tests only for testing model parameters, they used the full K-depth tests in a one-sided version where a null hypothesis is rejected if the K-sign depths is too small. Here, we study the K-depth tests in a two-sided version and compare the full K-depth tests with the simplified K-depth tests. Especially, we are interested in the efficiency of these tests to detect deviations from the independence assumption.

Section 2.1 provides a detailed discussion of K-depth tests and further references which showed the efficiency of K-depth tests for testing model parameters. Since this efficiency was already studied in several other publications (Falkenau 2016; Kustosz, Leucht, and Müller 2016; Kustosz, Müller, and Wendler 2016; Horn 2021b; Malcherczyk 2022; Horn and Müller 2023; Leckey et al. 2023), we concentrate on either testing the independence assumption on its own or on simultaneously testing independence and model assumptions in the subsequent sections. To this end, we compare the K-depth tests

with known tests for independence such as the runs test, the turning point test (Kendall 1973), the Durbin-Watson test (Verbeek 2012), the Ljung-Box test (Ljung and Box 1978), Von-Neumann-Rank-Ratio test (Bartels 1982), and the Brook-Dechert-Schreinkamp test (Broock et al. 1996). More details for these known independence tests are given in Sec. 2.2

In Sec. 3, the simulated power of the various tests are given for testing the null hypothesis $H_0 : \rho = 0$ where ρ is the autocorrelation coefficient of an AR(1) model so that $\rho = 0$ is equivalent with the independence assumption. Section 3.2 deals with the robustness of the tests with respect to innovation outliers and contaminations of the measurements. The behavior for higher lags is investigated in Sec. 3.3 by considering second order autoregressive time series and seasonal autoregressive time series. Finally, Sec. 4 studies the behavior of the tests in situations where the time series are corrupted by jumps and trends so that model deviations appear.

Moreover, we present in Sec. 5 an application to crack data in a bridge monitoring. In this monitoring, the width of a crack and the temperature below and above the bridge are observed over 1 year. Since the crack width varies with the temperature, it is difficult to find an adequate model for these crack data. Section 5 shows how the full 3-depth test leads to a reasonable model.

The conclusion of the simulation studies and the application is that the simplified K-depth tests and the full K-depth test with $K = T/3$ can compete with the classical independence tests in terms of power when only testing the independence assumption. All K-depth tests are very outlier robust and are able to detect model deviations. Hence they can also be used for model testing including the independence assumption of the errors. However, the Ljung-Box test is the best test for detecting model deviation but, similar to the Durbin-Watson test, struggles when outliers occur since both tests base on the outlier sensitive autocorrelation coefficient. The runs test behaves often similarly to the simplified K-depth tests and the full K-depth test with $K = T/3$ but the K-depth tests are superior in the case of seasonal autoregressive time series. A more detailed discussion is given in Sec. 6.

2. Statistical tests for independence

2.1. K-depth tests

A K-depth test is based on the K-sign depth which is a measure of fit of a given model. Let $(y_t)_{t=-p+1,-p+2,\ldots,0,1,2,\ldots,T}$ be the realization of the time series $(Y_t)_{t=-p+1,-p+2,\ldots,0,1,2,\ldots,T}$ satisfying (1) and $r_1(\theta), \ldots, r_T(\theta)$ the realizations of the residuals $R_1(\theta), \ldots, R_T(\theta)$ for a given model parameter θ. The only assumptions for the null hypothesis (2) are the properties given by (3) if θ_* is the true model parameter.

Then, the full K-sign depth of a model with model parameter θ in the realized time series $(y_t)_{t=-p+1,-p+2,\ldots,0,1,2,\ldots,T}$ is the relative number of all subsets with K residuals so that the residuals have alternating signs, i.e. it is the relative number of ordered subsets $\{t_1, \ldots, t_K\} \subset \{1, \ldots, T\}$ with $t_1 < \ldots < t_K$ satisfying $\operatorname{sign}(r_{t_i}(\theta)) = -\operatorname{sign}(r_{t_{i+1}}(\theta))$ for $i = 1, \ldots, K-1$. Here, sign denotes the sign-function, i.e. $\operatorname{sign}(z) = 1$ if $z > 0$, $\operatorname{sign}(z) = -1$ for $z < 0$, and $\operatorname{sign}(0) = 0$. The assumption of a continuous distribution of the errors means that we can assume without loss of generality that all signs are nonzero. Hence, the **full K-sign depth** can be given formally as

$$d_K(r_1(\theta), \ldots, r_T(\theta)) := \frac{1}{\binom{T}{K}} \sum_{1 \le t_1 < \ldots < t_K \le T} \left(\prod_{k=1}^{K} \mathbb{1}\left\{ r_{t_k}(\theta)(-1)^k > 0 \right\} + \prod_{k=1}^{K} \mathbb{1}\left\{ r_{t_k}(\theta)(-1)^k < 0 \right\} \right).$$

The **simplified *K*-sign depth** is only based on subsets with K consecutive residuals so that it is defined as

$$d_K^S(r_1(\theta),...,r_T(\theta)) := \frac{1}{T-K+1}\sum_{t=1}^{T-K+1}\left(\prod_{k=1}^{K}\mathbb{1}\left\{r_{t+k-1}(\theta)(-1)^k > 0\right\} + \prod_{k=1}^{K}\mathbb{1}\left\{r_{t+k-1}(\theta)(-1)^k < 0\right\}\right).$$

Originally, the K-sign depth appeared in special situations of the simplicial regression depth introduced by Rousseeuw and Hubert (1999) who proposed regression depth and simplicial regression depth as a measure of fit of a regression model. The name simplicial regression depth originated from the fact that it is derived from the regression depth in the same way as Liu's simplicial depth (Liu 1988, 1990) based on simplices for multivariate location data can be derived from Tukey's halfspace depth (Tukey 1975).

While location depth measures the depth of a location parameter in the data set, regression depth measures the depth of the regression function in the data set. However, the notion of regression depth and simplicial regression depth is quite complicated. In particular, simplicial regression depth becomes more manageable when it is equivalent to K-sign depth where sufficient conditions for this equivalence are given in Kustosz, Müller, and Wendler (2016).

If the model with model parameter θ_* is the correct model then the K-sign depth should be high. A small K-sign depth indicates either a wrong model parameter or that the model is not correct at all. This works as a model check quite well as long as the independence of the residuals is ensured which is the case for regression models with independent observations. However, in time series, too many alternating signs of residuals may indicate a negative correlation between the residuals and thus a violation of the independence assumption.

For calculating critical values for testing the null hypothesis of the form

$$H_0 : \theta_* \quad \text{satisfies} \quad (3), \tag{4}$$

normalized versions of the K-sign depths should be used, namely

$$\begin{aligned} T_K(r_1(\theta),...,r_T(\theta)) &= T\left(d_K(r_1(\theta),...,r_T(\theta)) - \frac{1}{2^{K-1}}\right), \\ T_K^S(r_1(\theta),...,r_T(\theta)) &= \sqrt{T-K+1}\ \frac{d_K^S(r_1(\theta),...,r_T(\theta)) - \frac{1}{2^{K-1}}}{\sqrt{\frac{1}{2^{K-1}}\left(3 - \frac{K}{2^{K-2}} - \frac{3}{2^{K-1}}\right)}}, \end{aligned} \tag{5}$$

for the full K-sign depth d_K and for the simplified K-sign depth d_K^S, respectively.

Let q_α be the α-quantile of the distribution of the normalized version of the K-sign depth given in (5). Then, for testing (2) or (4), respectively, we propose and study the following two-sided tests here: the **full *K*-depth test** rejects H_0 if

$$T_K(r_1(\theta_*),...,r_T(\theta_*)) < q_{\alpha/2} \ \text{ or } \ T_K(r_1(\theta_*),...,r_T(\theta_*)) > q_{1-\alpha/2} \tag{6}$$

and the **simplified *K*-depth test** rejects H_0 if T_K in (6) is replaced by T_K^S.

For small sample sizes T, the quantiles can be determined exactly by calculating the normalized depth in (5) for all 2^T combinations of positive and negative signs. However, if T is too large, one can use the fact that the normalized depth in (5) converges to an asymptotic distribution.

The advantage of the simplified K-sign depth is that its asymptotic distribution can be easily derived under the assumptions (3) as shown in Kustosz, Müller, and Wendler (2016). The asymptotic distribution is the normal distribution so that the symmetric quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are

the best choice of quantiles. The asymptotic distribution of the full K-sign depth is more complicated. For $K = 2$ and $K = 3$, the asymptotic distribution was derived in Kustosz and Müller (2014) and Kustosz, Müller, and Wendler (2016), respectively. In these papers, the asymptotic distribution was derived for simplicial regression depth in special autoregressive models but the proofs base only on the signs of the residuals so that they hold for 2-sign depth and 3-sign depth. For general $K \geq 2$, the asymptotic distribution of the K-sign depth is derived in Malcherczyk, Leckey, and Müller (2021). It is an asymmetric distribution given by an integrated transformed Brownian motion. Hence, the symmetric quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ could be replaced by quantiles which minimized $q_{\alpha_2} - q_{\alpha_1}$ with $\alpha_2 - \alpha_1 = 1 - \alpha$. However, since asymmetric quantiles did not provide relevant visible improvements in the simulation studies, we use here only the symmetric quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$.

Another advantage of the simplified K-sign depth is that it can be calculated in linear time with growing sample size T while a naive algorithm for the full K-sign depth has complexity of $\binom{T}{K}$. However, Leckey et al. (2023) and Malcherczyk (2022) provide a more efficient algorithm for the full K-sign depth called *block implementation*. This algorithm manages to compute the full K-sign depth with a linear time complexity in T for any fixed $K \geq 2$ by rearranging the sum over K-tuples as well as precomputing the resulting cumulative sums.

Leckey et al. (2023) also show that the full 2-depth test is equivalent to the classical sign test. They also mention that the simplified 2-depth test is closely related to the runs test. The only major difference is that the runs test considers the number of runs conditioned on the number of positive signs while the simplified 2-depth test considers the number of runs/sign changes without conditioning.

Since the K-depth tests are only based on signs of residuals, they are robust against outliers, heavy tailed distribution and heteroscedasticity. The simulations in Leckey et al. (2023); Kustosz, Leucht, and Müller (2016); Kustosz, Müller, and Wendler (2016); Falkenau (2016); Horn (2021b); Malcherczyk (2022) and Horn and Müller (2023) additionally indicate a high power of the one-sided K-depth tests for $K \geq 3$ in the case of testing hypotheses on the model parameter θ in linear, nonlinear and multiple regression as well as in linear and nonlinear autoregressive models and thus are much better than the classical sign test. In particular, the power of the full K-depth tests reaches the power of classical parametric tests as t- and F-tests while the simplified K-depth tests are a little bit less powerful (Falkenau 2016; Kustosz, Müller, and Wendler 2016). While the full 2-depth test is usually not a consistent test, Leckey, Jakubzik, and Müller (2023) can show that the full K-depth test is consistent for quite general testing problems if already $K = 3$ holds.

Here, the behavior of the two-sided K-depth tests for testing simultaneously the independence of the residuals and the model is of interest. The K-depth tests were carried out by using the `GSignTest` package (Horn 2021a).

2.2. Other reference tests

As a benchmark in terms of testing independence for stationary time series, several other tests are considered in this paper. The Durbin-Watson test (DW test) (Verbeek 2012) and the Ljung-Box test (LB test) (Ljung and Box 1978) are used as representatives of parametric tests. The LB test utilizes the first H empirical autocorrelation coefficients $\hat{\rho}_h$ with $h \in \{1, ..., H\}$ and assumes that they are normally distributed. Under some general assumptions (Gujarati and Porter 2009, 234–235), the statistic of the DW test bases on the estimator of the first autocorrelation coefficient $\hat{\rho}_1$ and its normality. While the LB test with $H = 15$ was carried out by using the function `box.test` which is part of the basic `stats` package in R, the DW test was obtained from the `lmtest` package (Zeileis and Hothorn 2002).

Furthermore, as non-parametric procedures, the runs test (Gibbons and Chakraborti 2003), the turning point test (TP test) (Kendall 1973), and the Broock-Dechert-Scheinkman test (BDS test) (Broock et al. 1996) are considered. Those procedures are mainly based on the sequential scheme of observations in a time series. The BDS test is an exception here, because its statistic utilizes concrete distances between observations. The BDS test and the runs test are included in the `tseries` package, the TP test is implemented in the `spgs` package (Hart and Martínez 2019).

Moreover, a rank based test of independence, the Von-Neumann-Rank-Ratio test (VNRR-Test) (Bartels 1982), is discussed in this paper. This test can be performed with the `randtests` package (Caeiro and Mateus 2014).

3. Behavior at deviations from the independence assumption

At first, we only regard deviations from the independence assumption, i.e. deviations from the first assumption in (3), and therefore test

$$\begin{aligned} &H_0 : R_1(\theta_*), ..., R_T(\theta_*) \text{ are independent,} \\ &\text{against} \\ &H_1 : R_1(\theta_*), ..., R_T(\theta_*) \text{ are not independent.} \end{aligned} \tag{7}$$

Here, we can assume without loss of generality that $g(\theta_*, Y_{t-p}, ..., Y_{t-1}, X_{t-q+1}, ..., X_t) = 0$ for each $t = 1, ..., T$ is satisfied for the true parameter θ_* so that $R_t(\theta_*) = Y_t = E_t$ (if $g(\theta_*, Y_{t-p}, ..., Y_{t-1}, X_{t-q+1}, ..., X_t) \neq 0$ then it can be subtracted from Y_t to get a new time series). This avoids to specify g and θ_* in the following simulations.

3.1. AR(1) time series

At first, we consider the special situation that E_t and thus Y_t is a stationary first order autoregressive time series so that

$$Y_t = \rho_1 Y_{t-1} + W_t, \quad |\rho_1| < 1, \quad W_t \sim N(0, 1), \tag{8}$$

holds for $t \in \{2, ..., T\}$ where T is the sample size, $(W_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. standard normally distributed random variables, Y_1 is the starting value and ρ_1 the first order autocorrelation coefficient. Then, the testing problem (7) is equivalent to testing $H_0 : \rho_1 = 0$ against $H_1 : \rho_1 \neq 0$.

The simulation of these time series was carried out by the function `arima.sim` of the `stats` package. Also note that additionally a burn-in period of $T/2$ observations was simulated at the beginning of the time series and eliminated afterwards in order to remove the effect of a starting value. For the purpose of judging the powers of the different tests, alternatives with values of ρ_1 on a grid from $-0.99 - 0.99$ with a fineness of 0.01 were tested at an α-level of 0.05. For each of the grid points, 100 repetitions of testing were carried out and the powers of the tests was determined by their relative rejection rates. These estimated powers were then displayed graphically by utilizing the packages `lattice` (Sarkar 2008), `viridis` (Garnier 2018), `magicaxis` (Robotham 2019) and `latex2exp` (Meschiari 2015). Note that a rejection rate of < 0.05 is associated with the color black in order to assess whether the α-levels of the tests are met. Due to the relatively small number of repetitions, it is necessary to keep in mind that minor transgressions of this value are no clear indication for tests not reaching the significance level under the null hypothesis.

The results of a simulation for samples of $T = 50$ and $T = 500$ observations are displayed in Figures 1 and 2. It can be seen that the full K-depth test with $K = T/3$ and the simplified K-depth tests with $K = 2, 3$ behave similarly to the runs test, the Von-Neumann-Rank-Ratio test, the Ljung-Box test and the Durbin-Watson test for small ($T = 50$) and large time series ($T = 500$). The bad power of the simplified K-depth tests with $K = 4, 5$ for positive ρ_1 and $T = 50$ is

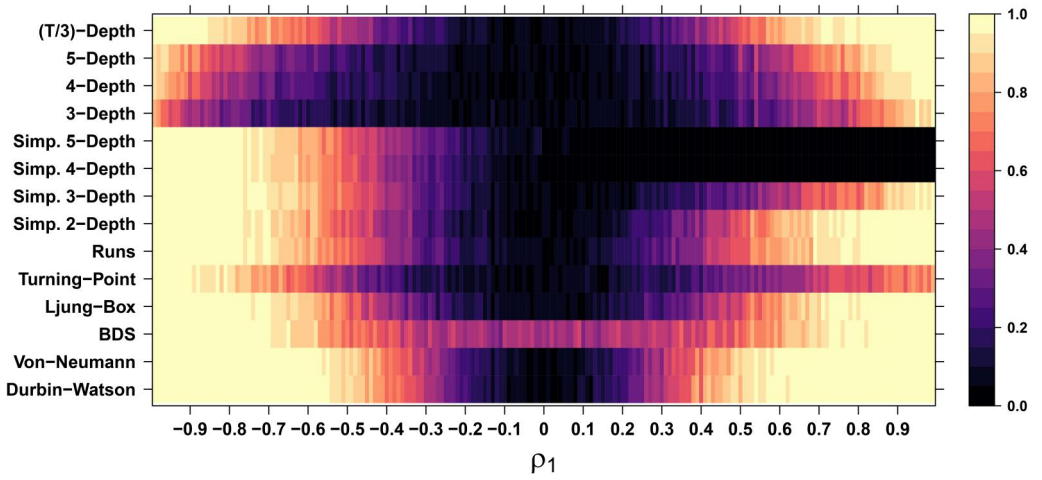

Figure 1. Simulated power of the different tests for stationary first order autoregressive time series with standard normally distributed innovations and 50 observations.

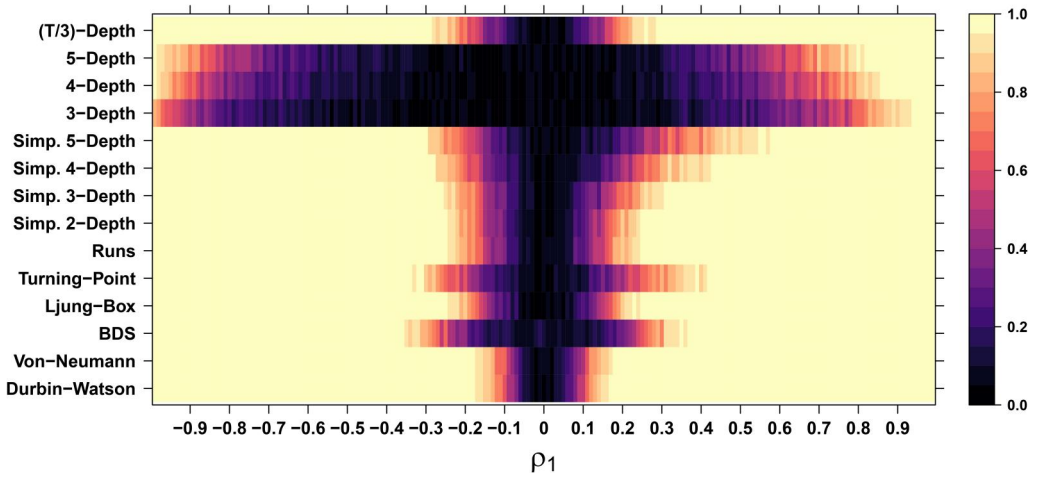

Figure 2. Simulated power of the different tests for stationary first order autoregressive time series with standard normally distributed innovations and 500 observations.

caused by the fact that the probability of the simplified depth d_K^S attaining the smallest possible value of zero is greater than $\alpha/2$ for $T = 50$. This effect disappears for $T = 500$ so that then these tests are also quite powerful for larger sample sizes. The Broock-Dechert-Scheinkman test does not keep the level for the small sample size of $T = 50$. Moreover, the power of the full K-depth test with $K = 3, 4, 5$ is not much worse than the power of the other tests for $T = 50$ but does not improve significantly with the larger sample size of $T = 500$. The reason is that all subsets $\{t_1, ..., t_K\} \subset \{2, ..., T\}$ are considered and that subsets where $\{t_1, ..., t_K\}$ are spread over the whole time series do not contribute to the detection of the dependence structure. This effect becomes less important with growing K and disappears if K is chosen depending on the sample size, for example as $K = T/3$.

3.2. Robustness for AR(1) time series

In this section, the robustness of the different independence tests with respect to outliers is analyzed for AR(1) time series.

3.2.1. Innovation outliers

First, situations are regarded in which the simulated time series contain obvious outliers that influence subsequent values of the time series. These kinds of outliers are referred to as innovation outliers or random shocks (Fox 1972). In practice, they often result from rare events that occur during the underlying process. Such a behavior can be simulated as follows:

$$Y_t = \rho_1 Y_{t-1} + \mathbb{1}\{t \in \mathcal{I}\} \cdot V_t + W_t, \quad |\rho_1| < 1, \quad W_t \sim N(0,1),$$

where $\mathcal{I}$ denotes a set chosen uniformly at random among all subset of $\{1, \ldots, T\}$ with size $\lceil T \cdot 0.05 \rceil$ and $(V_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. random variables drawn uniformly at random from the set $\{-50, 50\}$. All random variables $\mathcal{I}$, $(V_t)_{t\in\mathbb{N}}$, and $(W_t)_{t\in\mathbb{N}}$ are chosen independently of each other.

The power values of the different tests are evaluated in the same fashion as in Sec. 3 and the results for $T = 50$ observations are displayed in Figure 3. Here, we see that the Von-Neumann-Rank-Ratio test provides the best power followed by the runs tests, simplified K-depth tests with $K = 2, 3$, full K-depth test with $K = T/3$, turning point test, and the Broock-Dechert-Scheinkman test. The power of the full K-depth tests, in particular for $K = 5$, are only slightly worse. The simplified K-depth tests with $K = 4, 5$, the Ljung-Box test, and the Durbin-Watson test are worse. However, the power of simplified K-depth tests with $K = 4, 5$ becomes much better for larger samples sizes while again the power of the full K-depth tests does not improve with growing sample size. The Supplementary Material contains the simulations for $T = 500$ as well as simulations for processes with Cauchy distributed innovations which show similar behavior.

3.2.2. Contaminations with additive outliers

Another type of outliers in the context of time series is given by so-called additive outliers or contaminations, which have no impact on subsequent observations. They typically arise from measurement errors and are no part of the underlying process. Here, the contaminated time series are simulated by using the uncontaminated process $(Y_t)_{t=1,\ldots,T}$ given in (8) to define a new process

$$\tilde{Y}_t = Y_t + \mathbb{1}\{t \in \mathcal{I}\} \cdot V_t, \quad t = 1, \ldots, T,$$

where, as before, $\mathcal{I}$ denotes a set chosen uniformly at random among all subset of $\{1, \ldots, T\}$ with size $\lceil T \cdot 0.05 \rceil$ and $(V_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. random variables drawn uniformly at random

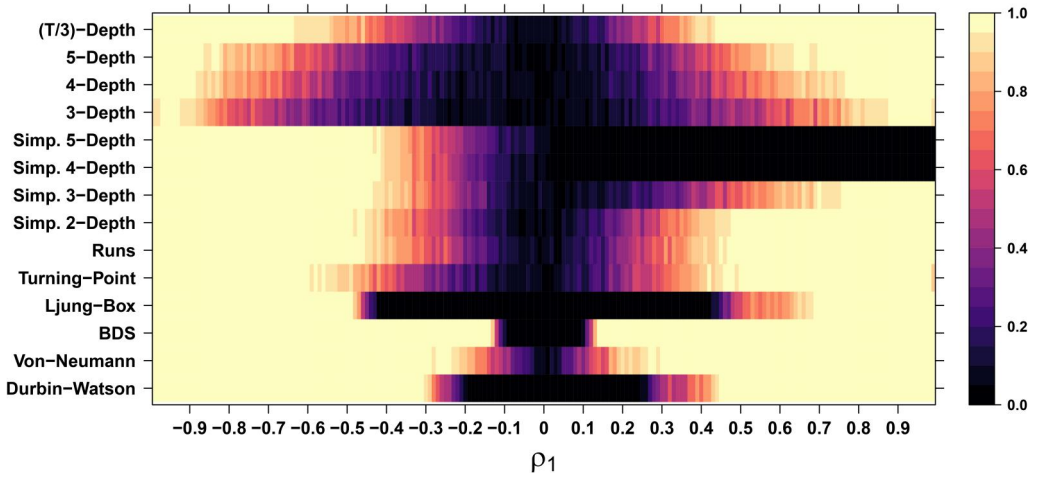

Figure 3. Simulated power of the different tests for stationary first order autoregressive time series with 50 observations and 3 innovation outliers.

from the set $\{-50, 50\}$. Moreover, the random variables $\mathcal{I}$ and $(V_t)_{t\in\mathbb{N}}$ are chosen independently of each other and independently of $(Y_t)_{t=1,\ldots,T}$. In the simulation study, the tests are applied to $(\tilde{Y}_t)_{t=1,\ldots,T}$.

The results of this simulation for $T = 500$ observations are displayed in Figure 4. Here, the power of the Ljung-Box test, the Broock-Dechert-Scheinkman test and the Durbin-Watson test is much worse than the power of the other tests. A similar result was obtained for the smaller sample size of $T = 50$, see the Supplementary Material.

3.3. Dependencies to higher lags

Hereinafter, the powers of the different independence tests were applied to time series that have dependencies to higher lags. As a first step, the test behaviors in stationary, second order autoregressive time series have been investigated. Samples of this time series are simulated according to the formula

$$Y_t = \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + W_t, \quad W_t \sim N(0,1),$$

where ρ_2 is the second order autocorrelation coefficient and $(W_t)_{t\in\mathbb{N}}$ is a sequence of independent standard normally distributed random variables as in the previous models. Such processes are independent if $\rho_1 = \rho_2 = 0$ holds and they are stationary if and only if the following inequalities are satisfied:

$$-1 < \rho_2 < 1, \ \ |\rho_1| < 1 - \rho_2.$$

When the values of the two autoregressive coefficients are shown as a surface, the three equations define the so-called stationarity triangle, which is also visible in the upcoming figures.

The power of the tests was evaluated as in the previous sections on a grid with a fineness of 0.1 for both parameters and the results for time series with $T = 500$ observations are displayed in Figure 5. The color scale of the rejection rates is not displayed but can be found in the previous figures. Here, the Ljung-Box test is clearly the best test while all other tests show power problems in some subsets of the considered (ρ_1, ρ_2). The largest subset with these power problems appear for the full K-depth tests with $K = 3, 4, 5$.

Furthermore, the powers of the independence tests in the context of seasonal autoregressive time series were analyzed. The most simple versions of these processes are stationary, seasonal

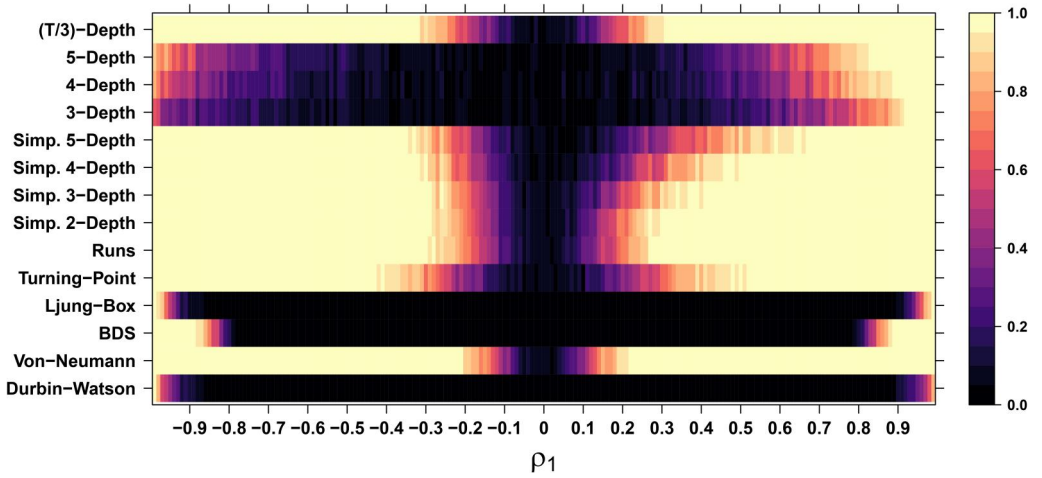

Figure 4. Simulated power of the different tests for stationary first order autoregressive time series with 500 observations and 25 additive outliers.

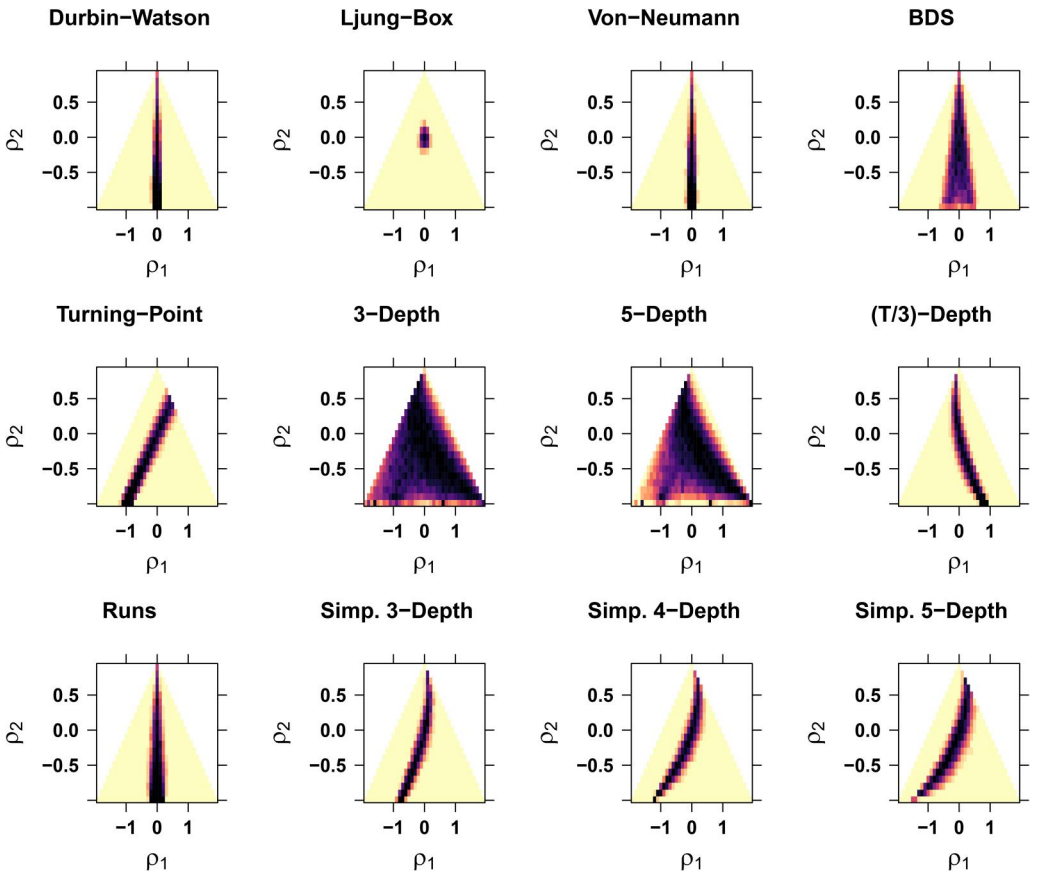

Figure 5. Simulated power of the different tests for stationary second order autoregressive time series with 500 observations.

first order autoregressive time series to the parameter $S \in \mathbb{N}$. In this kind of processes, the value of an observation depends on the observation which lies S time units in the past. This relationship can be simulated according to the formula

$$Y_t = \rho_S Y_{t-S} + W_t, \quad |\rho_S| < 1, \quad W_t \sim N(0,1),$$

for $t \in \{S+1, ..., T\}$ and where ρ_S is the autocorrelation coefficient of order S.

The powers of the tests for $S \in \{1, ..., 6\}$ is displayed in Figure 6 for time series with $T = 500$ observations. In this scenario, the Ljung-Box test is again the best. The Durbin-Watson test, Von-Neumann-Rank-Ratio test, Broock-Dechert-Scheinkman test, and the runs test struggle with lags $S \geq 2$ and the turning point test with lags $S \geq 3$. The simplified K-depth test and the full K-depth test with $K = T/3$ can deal with higher lags while the performance of the full K-depth tests with $K \in \{3, 5\}$ is nearly equally bad for all lags S. However, the power increases with increasing K for all K-depth tests.

4. Behavior at deviations from the independence assumption and the model assumption

While the previous section was solely focused on detecting deviations from the independence assumption, we will now consider the test power in situations where deviations from the model assumptions given by the model function g and the model parameter θ_* might also be present. As mentioned in the

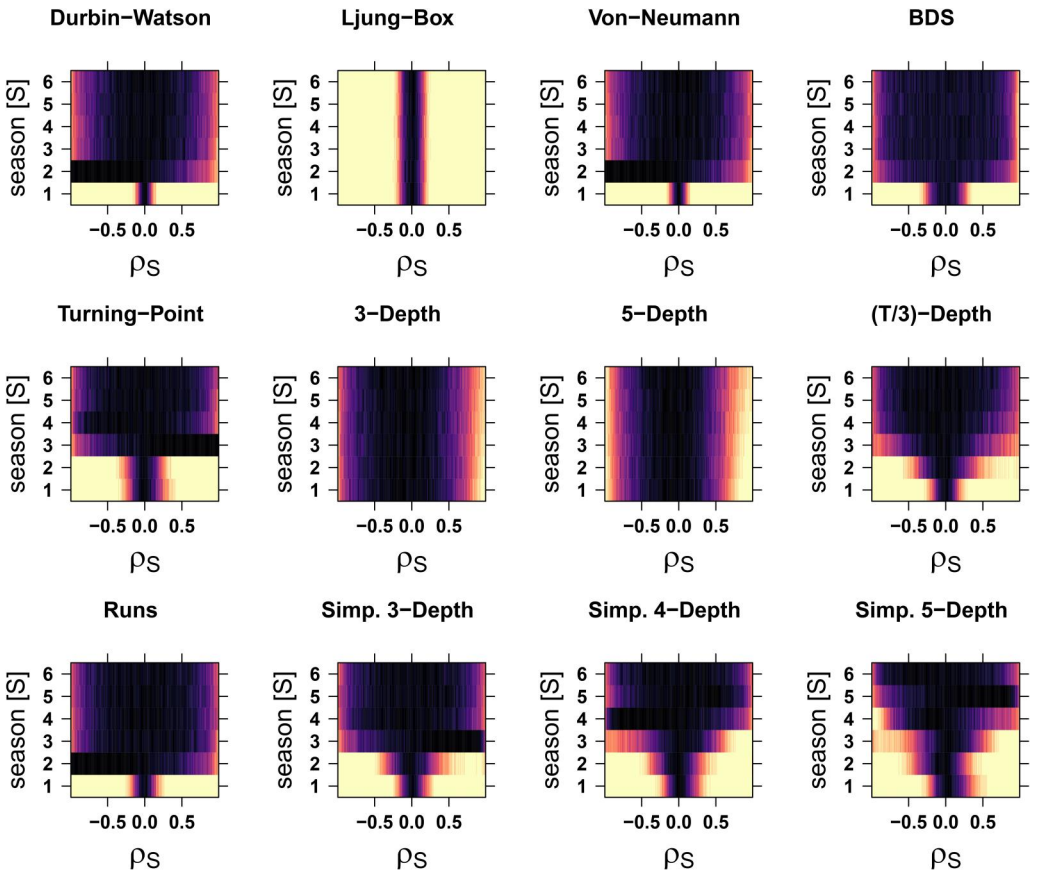

Figure 6. Simulated power of the different tests for stationary seasonal autoregressive time series with 500 observations.

beginning of Sec. 3, we assume without loss of generality $g(\theta_*, Y_{t-p}, ..., Y_{t-1}, X_{t-q+1}, ..., X_t) = 0$ for each $t = 1, ..., T$ so that the specification of g and θ_* is avoided in the following simulations.

As a first example, we consider first order autoregressive processes which might have a visible jump in their average behavior. To this end, we simulate the AR(1) process $(Y_t)_{t=1,...,T}$ given in (8) and shift the first and second half of the observations by $-h/2$ and $h/2$, respectively. Thus, the new process is formally defined as

$$\tilde{Y}_t = Y_t - \mathbb{1}\{t \leq T/2\} \cdot \frac{h}{2} + \mathbb{1}\{t > T/2\} \cdot \frac{h}{2}, \quad t = 1, ..., T,$$

and the tests are applied to $(\tilde{Y}_t)_{t=1,...,T}$. Here, the autoregression coefficient ρ_1 from $(Y_t)_{t=1,...,T}$ with $\rho_1 \neq 0$ provides the deviation from the independence assumption in the first assumption in (3) while the jump height h with $h \neq 0$ yields the deviation from the model with parameter θ_* so that the second assumption in (3) is not satisfied. Note that it does not matter whether θ_* includes a component $h_* = 0$ or not. If h is not included in θ, then $h \neq 0$ means a model deviation. Otherwise it means a deviation from θ_*. In this process $(\tilde{Y}_t)_{t=1,...,T}$, the jump does not influence subsequent observations just like in the case of contaminated time series (see 3.2.2). In particular, the independence assumption of the residuals are not violated for $\rho_1 = 0$ and any $h \neq 0$.

The power of the tests was then evaluated for parameters h from -5 to 5 with a fineness of 0.1 in combination with values of ρ_1 from -0.95 to 0.95 with a fineness of 0.05. The corresponding results are displayed in Figure 7. Again, the Ljung-Box test shows the best power.

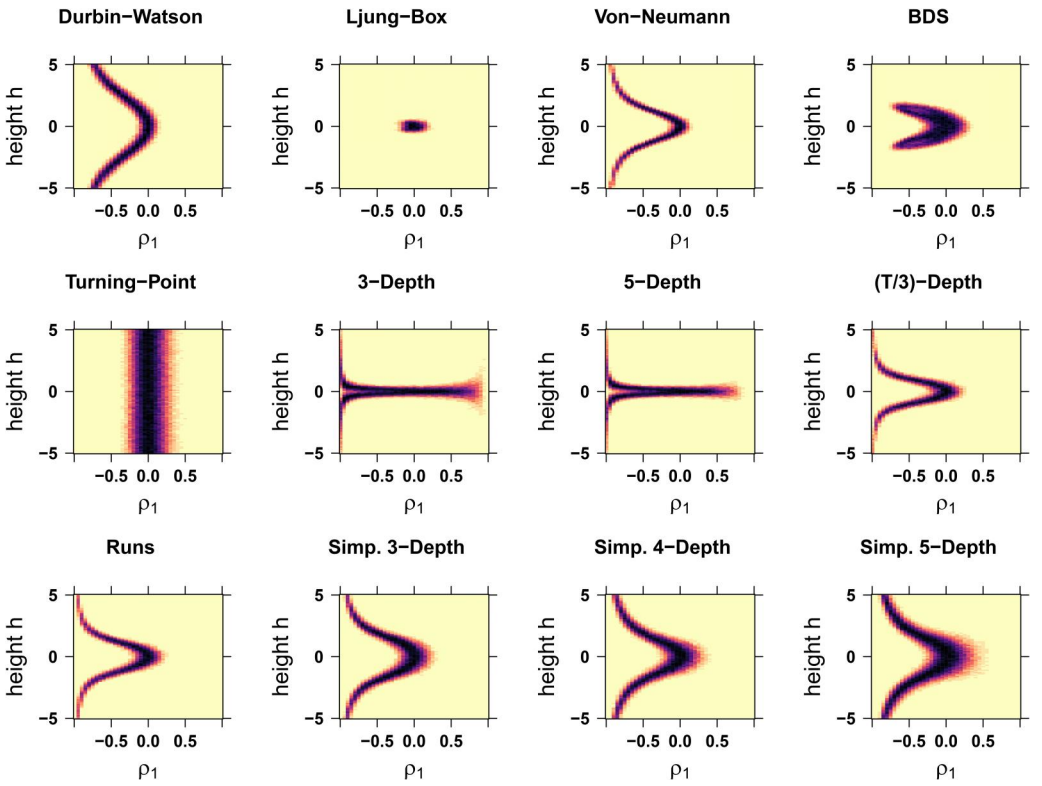

Figure 7. Simulated power of the different tests for stationary first order autoregressive time series with 500 observations and a jump after the 250th observation.

However all other tests except the turning point test are able to detect the existence of a jump. This holds also for all full K-depth tests.

Another deviation from the model assumptions, that was considered in this study, concerns the presence of a trend or drift in the examined time series. For this purpose, stationary first order autoregressive time series were simulated just like in the case of a jump and a drift of the intensity δ was included post hoc. In order to ensure a theoretical median of 0 in the whole time series, they were simulated according to the formula

$$\tilde{Y}_t = Y_t + \delta \cdot (t - T/2)/T, \quad t = 1, ..., T,$$

where $(Y_t)_{t=1,...,T}$ is the process given in (8). Again, the autoregression coefficient ρ_1 from $(Y_t)_{t=1,...,T}$ with $\rho_1 \neq 0$ provides the deviation from the independence assumption in the first assumption in (3) while the drift parameter δ with $\delta \neq 0$ yields the deviation from the model with parameter θ_* so that the second assumption in (3) is not satisfied. By varying the parameter δ from -0.01 to 0.01, very similar results were obtained as for the model deviation with a jump. See the Supplementary Material.

5. Application to bridge monitoring

In a bridge monitoring running from June 2016 to October 2018, the width of eight cracks and the temperature above and below a bridge in Bochum (Germany) were monitored every 2 s. For more details, see Abbas et al. (2019). The attempt to derive a reasonable model for these crack data with classical model selection methods was not successful. In particular, the crack width depends strongly on the temperature and on the traffic. Moreover, there are anomalous crack

sequences. The attempt to filter out these anomalous sequences as described in Abbas et al. (2019) did not help in modelling the crack width. Hence, the idea was to smooth the time series by calculating the median in time intervals of 15 min and considering the 96 time intervals of a day separately. This leads to 96 time series where each time series consists of the median crack width and median temperature of a specific time interval, say 7:00 to 7:15 am, for the days of 1 year, namely from June 2016 to May 2017. Considering the 96 time intervals of a day separately should reduce the influence of the traffic. The smoothing with the median over 15 min should reduce the influence of the anomalous sequences. However, this does not work completely so that some outliers remain as contamination. This is probably the reason that even in this reduced setup, classical model selection methods still fail.

Therefore, for modelling the crack width called WN2, the following seven explanatory variables are considered: 1. TBr, the current temperature below the bridge, 2. TSun, the current temperature above the bridge, 3. TBr4h, the temperature 4 h ago below the bridge, 4. TSun4h, the temperature 4 h ago above the bridge, 5. TBrM, the mean temperature of the previous 7 days below the bridge, 6. TSunM, the mean temperature of the previous 7 days above the bridge, 7. WN2(-1), the crack width of the day before (AR(1) component). To test some models based on these variables, the data set of 363 days is divided in a training data set of 303 days and a test data set of 60 days.

In the training step, relevant variables out of the seven variables are selected by four model selection criteria: Let p denote the number of selected explanatory variables, i.e. $p \in \{0, \ldots, 7\}$ here, $r_t^i(\theta)$ be the t'th residual at θ of the i'th time interval, $t = 1, \ldots, 303$, $i = 1, \ldots, 96$. For all 2^7 possible selections of the seven variables, the following model selection criteria are used:

1. AIC: $\sum_{i=1}^{96}\left(\ln\left(\frac{1}{303-p}\sum_{t=1}^{303}(r_t^i(\hat{\theta}_i)^2\right) + 2p\right)$ where $\hat{\theta}_i$ is the least squares estimator in the i'th time series of the training data set of,
2. trim. AIC: sum of the AICs of 10% trimmed sum of squared residuals for i'th time series of the training data set where $\hat{\theta}_i$ is the 10% least trimmed squares estimator calculated with `lqs` of the R package `MASS`,
3. 3-Depth: mean of the full 3-sign depths $d_3(r_1^i(\hat{\theta}_i), \ldots, r_{303}^i(\hat{\theta}_i))$ where $\hat{\theta}_i$ is the MM-estimator of θ in the i'th times series of the training data set calculated with `lmRob` of the R package `robustbase`.
4. 3-Depth p-values: mean of the p-values of the one-sided full 3-depth tests for $H_0 : \hat{\theta}_i$ satisfies (3) at the i'th time series of the training data set where $\hat{\theta}_i$ is the MM-estimator of θ calculated with `lmRob` of the R package `robustbase`.

In Table 1, values of 1 indicate those variables which are selected by maximizing the four model selection criteria. We see that maximizing the classical AIC criterion leads to a reduction of only one variable while the criteria based on the full 3-depth lead to the smallest models, namely a model with variables TBrM and WN2(-1) for the criterion based directly on the 3-sign depth and a model with the variables TBrM, TSunM, and WN2(-1) for the criterion based on the p-values of the one-sided 3-depth test.

For the testing step, let $\hat{\theta}_i$ be the estimated parameter based on the training data in the selected model using the estimation method of the corresponding selection method. Now, this estimated

Table 1. Table of the selected variables by the different model selection criteria (1 $\hat{=}$ selected, 0 $\hat{=}$ not selected).

Model	TBr	TSun	TBr4	TSun4	TBrM	TSunM	WN2(-1)
AIC	1	1	1	1	0	1	1
trim. AIC	1	1	0	1	0	0	1
3-Depth	0	0	0	0	1	0	1
3-Depth p-value	0	0	0	0	1	1	1

parameter is used as candidate model parameter, i.e. $\theta_{i*} = \hat{\theta}_i$, for the i'th time series. Additionally, the robust MM-estimate of `lmRob` in the full model with seven variables applied to the training data set is considered, leading to a fifth model. Then the hypotheses "$H_0^i : \theta_{i*}$ satisfies (3)" are tested for the five models at the 96 time series of the test data set using the residuals $r_t^i(\theta_{i*})$ for $t = 304, \ldots, 363$, $i = 1, \ldots, 96$. This leads to 96 p-values for each of the five models for each considered test.

Figure 8 provides the boxplots of the 96 p-values of five independence tests at the 96 time series. The results are very different although the sample size of $T = 60$ of the 96 test time series is not very high. The Ljung-Box test rejects almost all models at test level of 0.1 while the Von-Neumann-Rank-Ratio test and the runs test provide quite high p-values at almost all models. Only the two two-sided full depth tests yield more different results for the five different models. In particular, the two-sided full 3-depth test rejects almost all models obtained by the AIC selection method while the majority of p-values is above 0.2 for the two depth based models. In general, the depth based tests lead to much higher p-values at the depth based models than at the other models, which is not the case for the Von-Neumann-Rank-Ratio test and the runs test. This may be explained by the similarities of the selection and the test methods. However, the depth based model selection criteria are based on the 3-depth itself or on the p-values of the one-sided 3-depth test while the depth tests are two-sided depth tests and one of them is a $(T/3)$-depth test with $T/3 = 20$.

It might be surprising that the depth tests produce much smaller p-values at the full model than at the small models selected with a depth criterion although `lmRob` was used as estimator in all these models. However, the `lmRob` estimator is not an estimator which maximizes the depth. As a robust estimator, it only produces rather high depth. The determination of an estimator, which maximizes the depth, is too complicated so that the `lmRob` estimator was used for simplifying the calculations. Hence, in the training phase with the depth criteria, the full model was sorted out because of too small depth. Therefore, the full model also provides smaller p-values than the depth trained models in the testing phase.

In order to see that the model trained with the 3-depth criterion provides good predictions for the test data, Figures 9 and 10 contain the observed crack widths, the fitted values at the training data, and their predictions according to the trained model for two different time series (i.e. two different times of the day). These two time series correspond to the largest p-value (Figure 9) and smallest p-value (Figure 10) among all 96 time series according to the two-sided full 3-depth test applied to the test data. Note that even the data with the smallest p-value still provide fairly good predictions.

Figure 11 shows the same time series as in Figure 9, but now trained with the AIC criterion. Here the fits at the training data are not as good as with the 3-depth criterion, and this is carried over to the predictions.

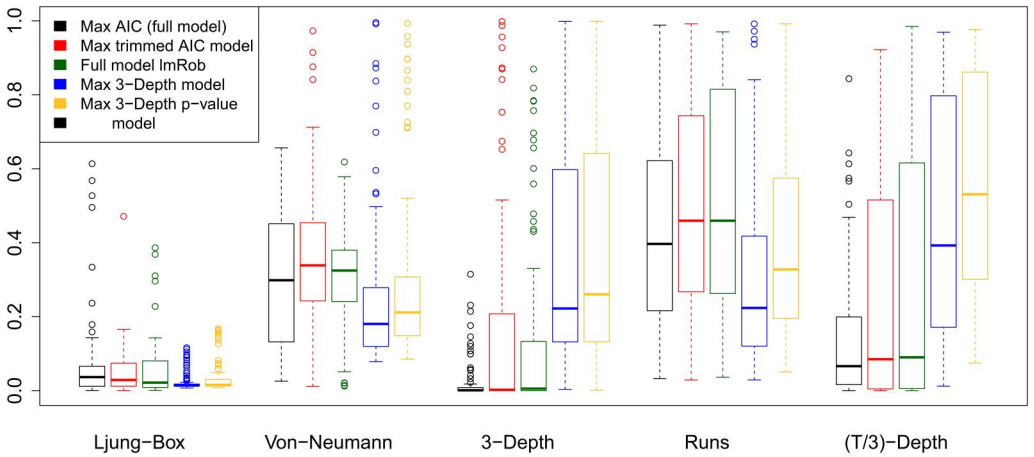

Figure 8. Boxplots of the p-values of the independence tests for different optimal models.

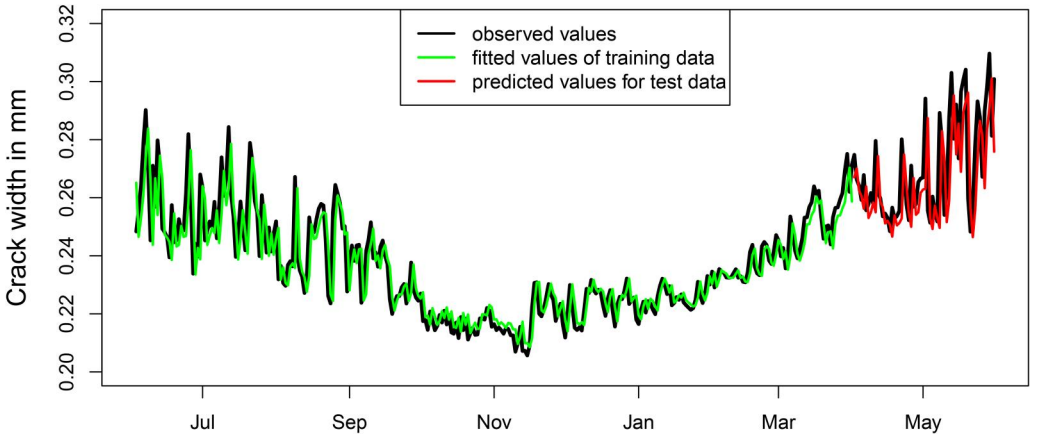

Figure 9. Observed values, fitted values at the training data set using model selection *via* 3-sign depth and estimation *via* lmRob, and predicted values *via* the obtained lmRob estimate for the time interval 7:30 to 7:45 am which leads to the maximal p-value of the 3-depth test for the test data.

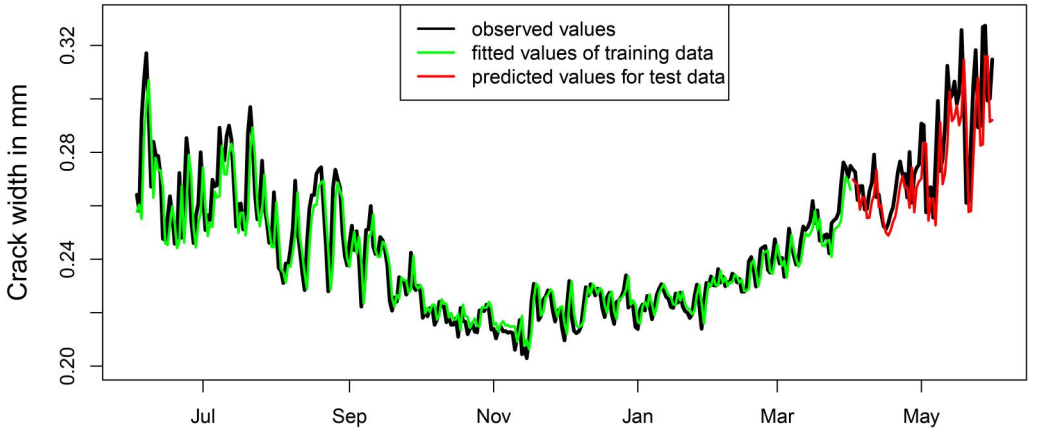

Figure 10. Observed values, fitted values at the training data set using model selection *via* 3-sign depth and estimation *via* lmRob, and predicted values *via* the obtained lmRob estimate for the time interval 11:30 to 11:45 am which leads to the minimal p-value of the 3-depth test for the test data.

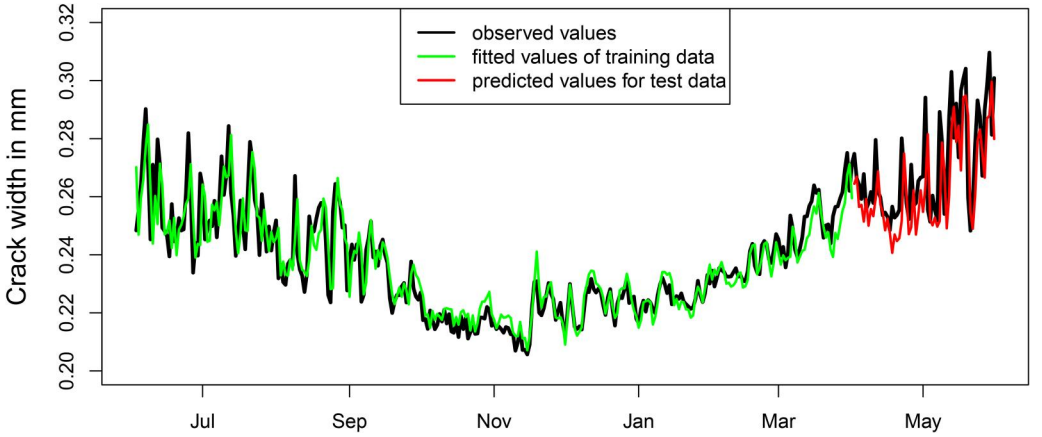

Figure 11. Observed values, fitted values at the training data set using model selection *via* AIC and least squares estimation, and predicted values *via* the obtained least squares estimate for the time interval 7:30 to 7:45 am

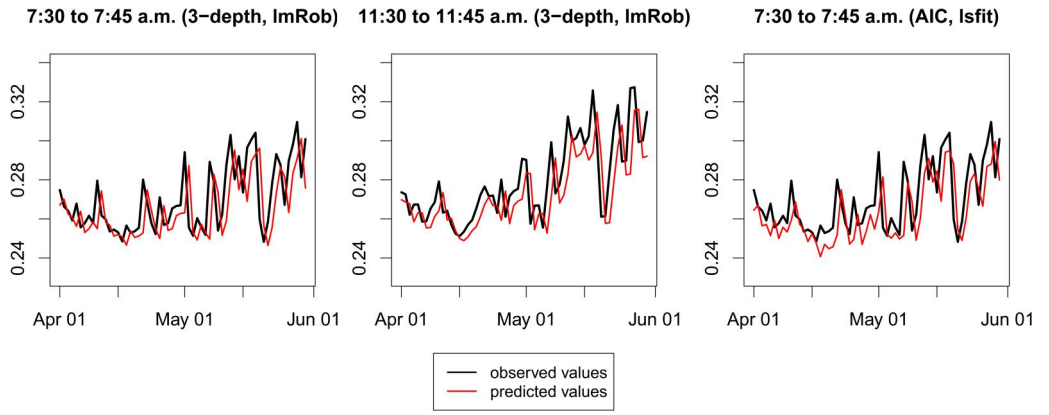

Figure 12. Predicted and observed values of the time series of the time intervals 7:30 to 7:45 am and 11:30 to 11:45 am where the predictions are given after training with the 3-depth criterion (on the left and in the Middle) and the AIC criterion (on the right).

Table 2. Table of the corresponding estimates obtained in the models selected with the 3-depth criterion and the AIC criterion for the time intervals / quarters of an hour 7:30 (to 7:45 am) and 11:30 (to 11:45 am).

Model Quarter	Intercept	TBr	TSun	TBr4	TSun4	TBrM	TSunM	WN2(-1)
AIC 7:30	0.0540	−0.0043	0.0022	0.0023	0.0005		−0.0005	0.7805
3-Depth 7:30	0.0253					−0.0000		0.8911
3-Depth 11:30	0.0276					0.0000		0.8801

Figure 12 compares the predictions for the three cases of Figures 9, 10, and 11 with more details. Here it is visible that all predictions are shifted from the observations by 1 day. This is caused by the high value of the estimate for the AR(1) component WN2(-1) as shown in Table 2. Although the estimates in Table 2 and the predictions given by Figure 12 look rather similar, the p-values of the two-sided 3-depth test are very different, namely 0.9983 for the 3-depth selection criterion and time interval 7:30 to 7:45 am (best time interval for the two-sided 3-depth test), 0.0036 for the 3-depth selection criterion and time interval 11:30 to 11:45 am (worst time interval for the two-sided 3-depth test), and 0.0084 for the AIC selection criterion and time interval 7:30 to 7:45 am. This also explains why the 3-depth test rejects most of the models obtained by the AIC criterion as can be seen in Figure 8.

On the other hand, the very small p-values visible in Figure 8 for the Ljung-Box test at almost all models and in particular at the models obtained with the depth selection criteria may be explained by the shift between observed and predicted values caused by a high value of the estimated AR(1) component WN2(-1). This means that all model selection procedures do not lead to an optimal model. Although the crack width depends strongly on the temperature, it seems that the regarded temperature variables are not enough to explain this temperature dependence completely. Hence there is room for improvement. Nevertheless, the models found by the 3-depth selection criterion are not that bad as indicated with the Ljung-Box test. In contrast to the Ljung-Box test, the Von-Neumann-Rank-Ratio test and the runs test have problems to detect any problems with the five derived models. The depth based test are between these extreme cases. They indicate problems with some models as those given by the AIC criterion, but show that models given by depth based model selection criteria are not that bad.

6. Discussion

The *K*-depth tests can be used to test simultaneously the independence of residuals of a given model and whether these residuals are distributed with a median equal to zero. They can be used

in a full version where they are based on the full K-sign depth which is the relative number of all subsets with K residuals showing alternating signs. In the simplified version, they are based on the simplified K-sign depth which only uses subsets of subsequent residuals. The performance of these tests when only testing the median property in models that always yield independent residuals was already investigated in former studies. We therefore concentrated here on studying the behavior of these tests in the context of independence testing and simultaneous testing of independence and model deviations. We compared them with classical independence tests in a simulation study. It turned out that in particular the simplified K-depth tests can compete with these classical tests. The full K-depth tests show a quite good power for moderate sample sizes but the power does not increase for larger sample sizes. This is explained by the fact that when considering the relative number of K-tuples with alternating signs, the overwhelming majority of these K-tuples concern positions in the residual vector that are far apart and therefore do not contribute to the detection of (local) dependence structures. This can be avoided by choosing the hyperparameter K in dependence of the sample size T, a reason why we also considered full $(T/3)$-depth tests. Indeed, these tests are a good alternative to the classical tests. Only the Ljung-Box test is superior in some situations but has massive problems with outliers while the K-depth tests basing only on signs of residuals are outlier robust. Often, the simplified K-depth tests and the full $(T/3)$-depth tests behave similarly to the runs test but they are superior to the runs test in the case of seasonal autoregressive time series.

We also studied the behavior of the tests in GARCH processes and the results can be found in the Supplementary Material. However, none of the tests except for the Broock-Dechert-Scheinkman test were able to detect any deviations from the independence. Since all of the nonparametric procedures do not use the actual value of points but only signs or ranks of the residuals and thus cannot detect changes in variance, this observation is comprehensible.

In an application with data from a bridge monitoring, we demonstrated how the K-depth tests can be used to improve the modeling of time series which may depend on a related d-dimensional co-process. In the application, it was not clear whether all $d = 7$ co-processes are relevant. To avoid simultaneous model selection, estimation, and testing, we divided the data set in a training data set and a test data set. The training data is used to select an appropriate model including an estimated model parameter. We considered four model selection methods including two new methods based on 3-depth. Then five independence tests including two depth based tests were applied to the selected models using the test data. In this context, classical independence tests tended to either reject all models or to produce high p-values for all models. Only the depth based tests provided different results for the different models. In particular, the depth based methods led to reasonable models with few variables. Nevertheless, it seems that some improvements by using other explanatory processes may be possible.

In the training step, the 3-sign depth was only used to find two additional models besides the models selected by AIC related criteria so that the different tests studied in this paper could be applied to different models. A by-product is that selection methods based on depth could be reasonable alternatives to classical model selection methods. However, the suitability of the K-sign depth and K-depth tests for model selection is beyond this paper and more investigations are necessary. To avoid the splitting of a data set in a training and a test data set, further research should also consider the situation of simultaneous estimation and testing and what happens when the model parameter is estimated in a wrong model.

Moreover, it might be possible to extend the presented approach to multivariate time series by using multivariate sign-type tests as studied, e.g. in Dyckerhoff, Ley, and Paindaveine (2015) and Dovoedo and Chakraborti (2016). In particular, multivariate runs and sign changes based on the multivariate spatial sign of Möttönen and Oja (1995) could be used as in Paindaveine (2009) for counting the K-tuples with $K - 1$ sign changes. This would lead to a multivariate K-sign depth.

Acknowledgements

The authors thank the reviewers for their comments and suggestions which improved the paper. They also gratefully acknowledge support from the Collaborative Research Center "Statistical Modeling of Nonlinear Dynamic Processes" (SFB 823, B4, B5) of the German Research Foundation (DFG).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Supplementary material

Further simulation results, the data set, and the R-code can be found under https://stabi.statistik.tu-dortmund.de/mitarbeiterinnen/mueller/publikationen/.

Funding

This article was supported by Deutsche Forschungsgemeinschaft.

References

Abbas, S., Fried, R., Heinrich, J., et al. 2019. Detection of anomalous sequences in crack data of a bridge monitoring. In *Applications in statistical computing - from music data analysis to industrial quality improvement*, ed. K. Ickstadt, H. Trautmann, and G. Szepannek. New York: Springer; p. 251–69.

Bartels, R. 1982. The rank version of von Neumann's ratio test for randomness. *Journal of the American Statistical Association* 77 (377):40–6. doi: 10.1080/01621459.1982.10477764.

Broock, W. A., J. A. Scheinkman, W. D. Dechert, and B. LeBaron. 1996. A test for independence based on the correlation dimension. *Econometric Reviews* 15 (3):197–235. doi: 10.1080/07474939608800353.

Caeiro, F., and A. Mateus. 2014. randtests: Testing randomness in r. R package version 1.0; Available from: https://CRAN.R-project.org/package=randtests.

Cui, J., C. Li, K. Yang, and W. Xu. 2018. Adaptive positive and negative runs test. *Journal of Statistical Computation and Simulation* 88 (7):1314–35. doi: 10.1080/00949655.2018.1430802.

Dovoedo, Y. H., and S. Chakraborti. 2016. On the robustness to symmetry of some nonparametric multivariate one-sample sign-type tests. *Journal of Statistical Computation and Simulation* 86 (10):1936–53. doi: 10.1080/00949655.2015.1092540.

Dyckerhoff, R., C. Ley, and D. Paindaveine. 2015. Depth-based runs tests for bivariate central symmetry. *Annals of the Institute of Statistical Mathematics* 67 (5):917–41. doi: 10.1007/s10463-014-0480-y.

Falkenau, C. P. 2016. Depth based estimators and tests for autoregressive processes with application on crack growth and oil prices. Dissertation, TU Dortmund. doi: 10.17877/DE290R-17269.

Fox, A. 1972. Outliers in time series. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34 (3): 350–63. doi: 10.1111/j.2517-6161.1972.tb00912.x.

Garnier, S. 2018. viridis: Default color maps from 'matplotlib'; R package version 0.5.1. Available from: https://CRAN.R-project.org/package=viridis.

Gibbons, J., and S. Chakraborti. 2003. Nonparametric statistical inference. Marcel Dekker Incorporated; Statistics, Textbooks and Monographs. Available from: https://books.google.pt/books?id=dPhtioXwI9cC.

Gujarati, D. N., and D. C. Porter. 2009. Basic econometrics, 5th ed. Boston, MA: *McGraw-Hill Irwin; The McGraw-Hill Series Economics*.

Hart, A., and S. Martínez. 2019. spgs: Statistical patterns in genomic sequences. R package version 1.0-3. https://CRAN.R-project.org/package=spgs

Horn, M. 2021a. GSignTest: Robust tests for regression-parameters via sign depth. R package version 1.0.7. https://github.com/melaniehorn/GSignTest

Horn, M. 2021b. Sign depth for parameter tests in multiple regression. Dissertation TU Dortmund. doi: 10.17877/DE290R-22355.

Horn, M., and C. H. Müller. 2023. Sign depth tests in multiple regression. *Journal of Statistical Computation and Simulation* 93 (7):1169–91. doi: 10.1080/00949655.2022.2130922.

Kendall, M. G. 1973. *Time Series*. London: Griffin.

Kitani, M., and H. Murakami. 2022. One-sample location test based on the sign and wilcoxon signed-rank tests. *Journal of Statistical Computation and Simulation* 92 (3):610–22. doi: 10.1080/00949655.2021.1968399.
Kustosz, C. P., A. Leucht, and C. H. Müller. 2016. Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis* 37 (6):763–84. doi: 10.1111/jtsa.12186.
Kustosz, C. P., and C. H. Müller. 2014. Analysis of crack growth with robust, distribution-free estimators and tests for non-stationary autoregressive processes. *Statistical Papers* 55 (1):125–40. doi: 10.1007/s00362-012-0479-5.
Kustosz, C. P., C. H. Müller, and M. Wendler. 2016. Simplified simplicial depth for regression and autoregressive growth processes. *Journal of Statistical Planning and Inference* 173:125–46. doi: 10.1016/j.jspi.2016.01.005.
Leckey, K., M. Jakubzik, and C. H. Müller. 2023. On the consistency of K-sign depth tests. *Econometrics and Statistics*. doi: 10.1016/j.ecosta.2023.10.002.
Leckey, K., D. Malcherczyk, M. Horn, and C. H. Müller. 2023. Simple powerful robust tests based on sign depth. *Statistical Papers* 64 (3):857–82. doi: 10.1007/s00362-022-01337-5.
Lehmann, E. L. 1994. *Testing statistical hypothesis*, 2nd ed. New York: Chapman & Hall.
Liu, R. Y. 1988. On a notion of simplicial depth. *Proceedings of the National Academy of Sciences of the United States of America* 85 (6):1732–4. doi: 10.1073/pnas.85.6.1732.
Liu, R. Y. 1990. On a notion of data depth based on random simplices. *The Annals of Statistics* 18 (1):405–14. doi: 10.1214/aos/1176347507.
Ljung, G. M., and G. E. P. Box. 1978. On a measure of lack of fit in time series models. *Biometrika* 65 (2):297–303. doi: 10.1093/biomet/65.2.297.
Malcherczyk, D. 2022. K-sign depth: Asymptotic distribution, efficient computation and applications. Dissertation TU Dortmund. doi: 10.17877/DE290R-22644.
Malcherczyk, D., K. Leckey, and C. H. Müller. 2021. K-sign depth: From asymptotics to efficient implementation. *Journal of Statistical Planning and Inference* 215:344–55. https://www.sciencedirect.com/science/article/pii/S0378375821000458. doi: 10.1016/j.jspi.2021.04.006.
Meschiari, S. 2015. latex2exp: Use latex expressions in plots. R package version 0.4.0. https://CRAN.R-project.org/package=latex2exp.
Möttönen, J., and H. Oja. 1995. Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics* 5 (2):201–13. doi: 10.1080/10485259508832643.
Paindaveine, D. 2009. On multivariate runs tests for randomness. *Journal of the American Statistical Association* 104 (488):1525–38. doi: 10.1198/jasa.2009.tm09047.
Robotham, A. 2019. magicaxis: Pretty scientific plotting with minor-tick and log minor-tick support. R package version 2.0.10. https://CRAN.R-project.org/package=magicaxis.
Rousseeuw, P. J., and M. Hubert. 1999. Regression depth. *Journal of the American Statistical Association* 94 (446):388–402. doi: 10.1080/01621459.1999.10474129.
Sarkar, D. 2008. *Lattice: Multivariate data visualization with r*. New York: Springer. ISBN 9780-387-75968-5; http://lmdvr.r-forge.r-project.org
Tukey, J. W. 1975. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians* 2:523–31.
Verbeek, M. 2012. *A guide to modern econometrics*. 4th ed. Chichester, West Sussex: Wiley.
Wald, A., and J. Wolfowitz. 1940. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics* 11 (2):147–62. http://www.jstor.org/stable/2235872. doi: 10.1214/aoms/1177731909.
Zeileis, A., and T. Hothorn. 2002. Diagnostic checking in regression relationships. *R News* 2 (3):7–10. https://CRAN.R-project.org/doc/Rnews/