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

Hendrik Dohme, Dennis Malcherczyk,
Kevin Leckey, Christine Müller

Nr. 24/2021

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

Hendrik Dohme[⋆], Dennis Malcherczyk[⋆], Kevin Leckey[⋆] and Christine Müller[⋆]

[⋆]Department of Statistics, TU Dortmund University, D-44227 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 modelling of crack width time series depending on temperature measurements in a bridge monitoring.

1 Introduction

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

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

for $t = 1, \dots, T$ where $(X_t)_{t=-q+2,-q+3,\dots,0,1,2,\dots,T}$ is a d-dimensional related process, $\theta_* \in \mathbb{R}^s$ is the model parameter, $g : \mathbb{R}^{s+p+q*d} \to \mathbb{R}$ the model function, and $(W_t)_{t=1,2,\dots,T}$ is a white noise process. We assume only that the error variables $W_1, \dots, W_T$ are independent and have a continuous distribution with median equal to zero. In particular, we do not assume variance homogeneity. The classical AR(p)-models are a special cases of these models.

In such time series, we can regard residuals given by

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

If θ is the true model parameter θ_* then $R_t(\theta_*) = W_t$ and $R_1(\theta_*), \dots, R_T(\theta_*)$ satisfy the following properties:

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

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 $r_1(\theta), \dots, r_T(\theta)$ of $R_1(\theta), \dots, R_T(\theta)$ 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 of 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 [1]. An additional problem for the application of the classical sign test for time series models is that the first condition in (2) 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 [2], see e.g. [3], pp. 78-86. It can be applied to the signs of the residuals $R_1(\theta_*), \dots, R_T(\theta_*)$ 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 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 (2) of a residual distribution with median equal to zero.

For testing model parameters, Leckey et al. [4] 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 K-subsets $\{t_1, \dots, t_K\} \subset \{1, \dots, T\}$ for which the corresponding residuals $R_{t_1}(\theta), \dots, R_{t_K}(\theta)$ have alternating signs. In a simplified version of the K-sign depth, only subsets of consecutive indices $t_1, t_1 + 1, \dots, t_1 + K - 1$ 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 demonstrate 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. [4] 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 [4, 5, 6, 7, 8], 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 [9], the Durbin-Watson test [10], the Ljung-Box test [11], Von-Neumann-Rank-Ratio test [12], and the Brook-Dechert-Schreinkamp test [13]. More details for these known independence tests are given in Section 2.2.

In Section 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 3.1 deals with the robustness of the tests with respect to innovation outliers and contaminations of the measurements. The behaviour for higher lags is investigated in Section 3.2 by considering second order autoregressive time series and seasonal autoregressive time series. Finally, Section 4 studies the behaviour of the tests in situations where the time series are corrupted by jumps and trends so that model deviations appear.

Moreover, we present in Section 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 one 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 selection. 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 tests 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 Section 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 are the properties given by (2) if θ_* is the true model parameter. In particular, the second assumption in (2) is satisfied if the residuals $R_n(\theta_*)$ have a continuous distribution with median equal to zero.

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 subsets $\{t_1, \ldots, t_K\} \subset \{1, \ldots, T\}$ with $\operatorname{sign}(r_{t_i}(\theta)) = -\operatorname{sign}(r_{t_{i+1}}(\theta))$ 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 second assumption of (2) 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

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

The **simplified K-sign depth** is defined as

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

Originally, the K-sign depth appeared in special situations of the simplicial regression depth introduced by Rousseeuw and Hubert [14] 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 Liu's simplicial depth [15, 16] based on simplexes for multivariate location data can be derived from Tukey's halfspace depth [17].

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 [5].

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

$$H_0 : \theta_* \text{ satisfies (2)},$$

a normalized version of the K-sign depth should be used, namely

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

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 (3). Then the **full K-depth test** rejects H_0 if

$$T_K(r_1(\theta_*),\ldots,r_T(\theta_*)) < q_{\alpha/2} \text{ or } T_K(r_1(\theta_*),\ldots,r_T(\theta_*)) > q_{1-\alpha/2} \tag{4}$$

and the **simplified K-depth test** rejects H_0 if T_K in (4) is replaced by T_K^S.

For small sample sizes T, the quantiles can be determined exactly by calculating the normalized depth in (3) 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 (3) 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 (2) as shown in [5]. 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 [18] and [6], 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 [19]. 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. [4] 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 cummulative sums.

Leckey et al. 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 test statistic of 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 [4, 5, 6, 7, 8] 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 [6, 7].

Here the behaviour 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 [20].

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) [10] and the Ljung-Box test (LB test) [11] are used as representatives of parametric tests. The LB test utilizes the first `H=15` empirical autocorrelation coefficients $\hat{\rho}_h$ with $h \in \{1, ..., H\}$ and assumes that they are normally distributed. Under some general assumptions [**?**,]pp. 234 – 235]Gujarati.2009, the statistic of the DW test bases on the first autocorrelation coefficient $\hat{\rho}_1$ and its normality. While the LB test can be carried out by using the function `box.test` which is part of the basic `stats` package in R, the DW test is included in the `lmtest` package [22].

Furthermore, as non-parametric procedures, the runs test [3], the turning point test (TP test) [9], and the Broock-Dechert-Scheinkman test (BDS test) [13] 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 [23].

Moreover, a rank based test of independence, the Von-Neumann-Rank-Ratio test (VNRR-Test) [12], is discussed in this paper. This test can be performed with the `randtests` package [24].

3 Testing for independence

As a first step, the different independence tests were applied to stationary first order autoregressive time series. More precisely, the following model is used to generate data:

$$Y_t = \rho_1 Y_{t-1} + W_t, \quad |\rho_1| < 1, \quad W_t \sim N(0,1) \tag{5}$$

for $t \in \{2, \dots, T\}$ 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. 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.

In this context, the independence of the regarded time series is given when ρ_1 is equal to 0. 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 utilising the packages `lattice` [25], `viridis` [26], `magicaxis` [27] and `latex2exp` [28]. Note that a rejection rate of < 0.05 is associated with the colour 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 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 $\{t_1, \dots, t_K\} \subset \{2, \dots, T\}$ are considered and that subsets where

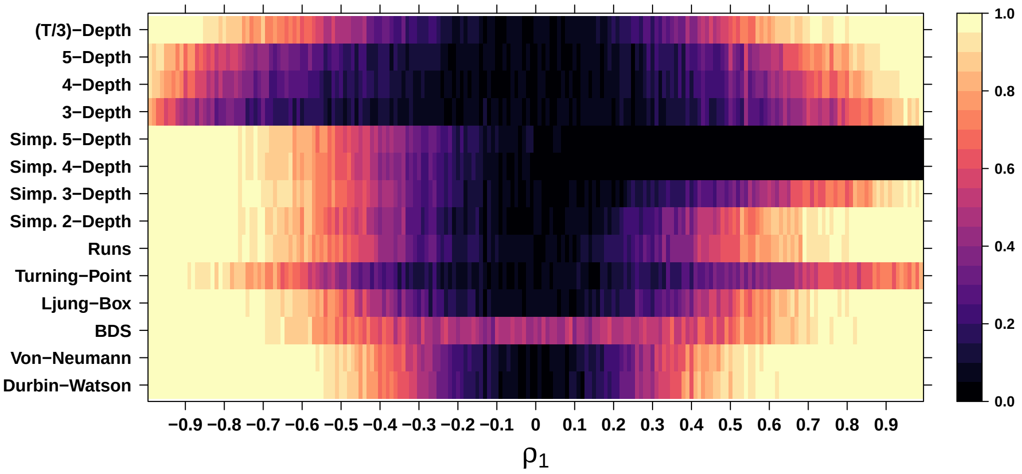

Figure 1: Simulated power of the different tests for stationary first order autoregressive time series with 50 observations.

$\{t_1, \ldots, t_K\}$ 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 $K = T/3$.

3.1 Robustness

In this section, the robustness of the different tests with respect to outliers is analysed.

3.1.1 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 [29]. In practice they often result from rare events that occur during the underlying process. Such a behaviour 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

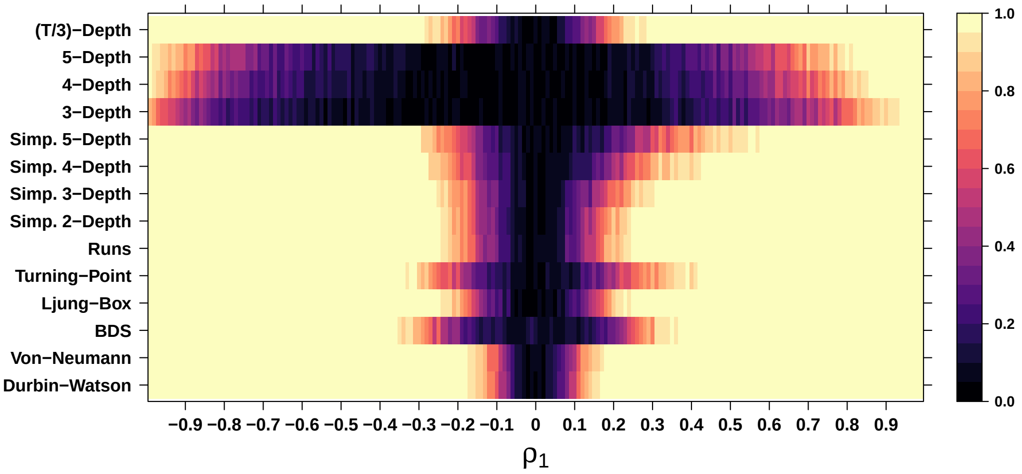

Figure 2: Simulated power of the different tests for stationary first order autoregressive time series with 500 observations.

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 Section 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. See the supplementary material for $T = 500$.

3.1.2 Contaminations with additive outliers

Another type of outlier in the context of time series are 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

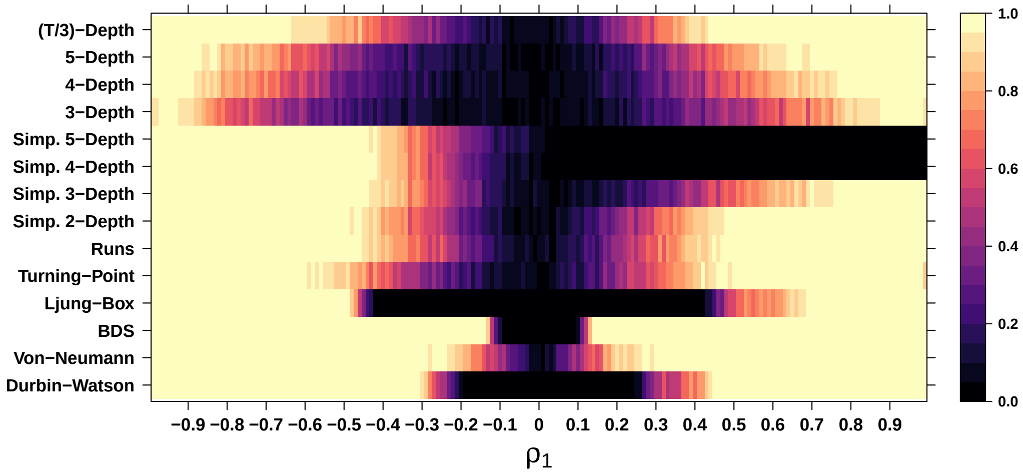

Figure 3: Simulated power of the different tests for stationary first order autoregressive time series with 50 observations and 3 innovation outliers.

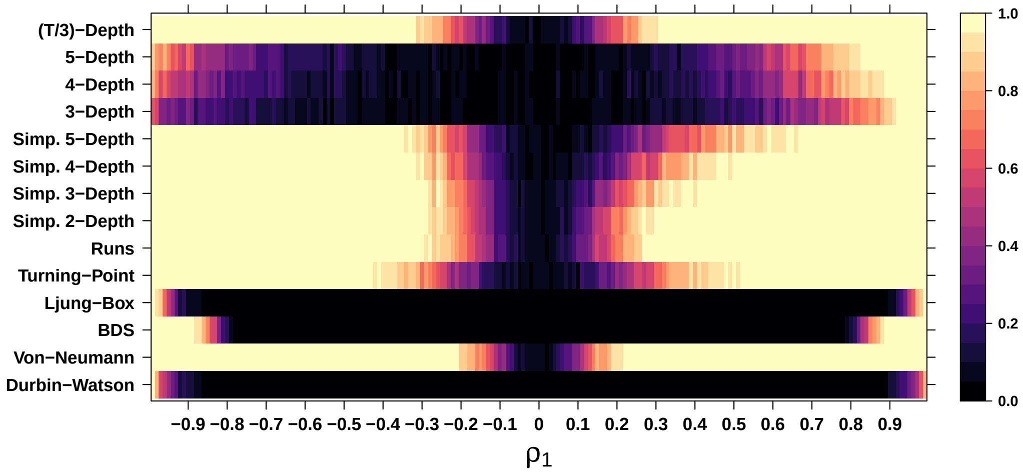

Figure 4: Simulated power of the different tests for stationary first order autoregressive time series with 500 observations and 25 additive outliers.

uncontaminated process $(Y_t)_{t=1,...,T}$ given in (5) to define a new process

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

where, as before, $\mathcal{I}$ denotes a set chosen uniformly at random among all subset of $\{1, ..., T\}$ 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 $\{-50, 50\}$. Moreover, the random variables $\mathcal{I}$ and $(V_t)_{t \in \mathbb{N}}$ are chosen independently of each other and independently of $(Y_t)_{t=1,...,T}$.

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.

3.2 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 behaviours 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 equations are satisfied:

$$\rho_1 + \rho_2 < 1, \;\; \rho_1 - \rho_2 < 1, \;\; -1 < \rho_2 < 1.$$

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 colour 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 analysed. The most simple versions of

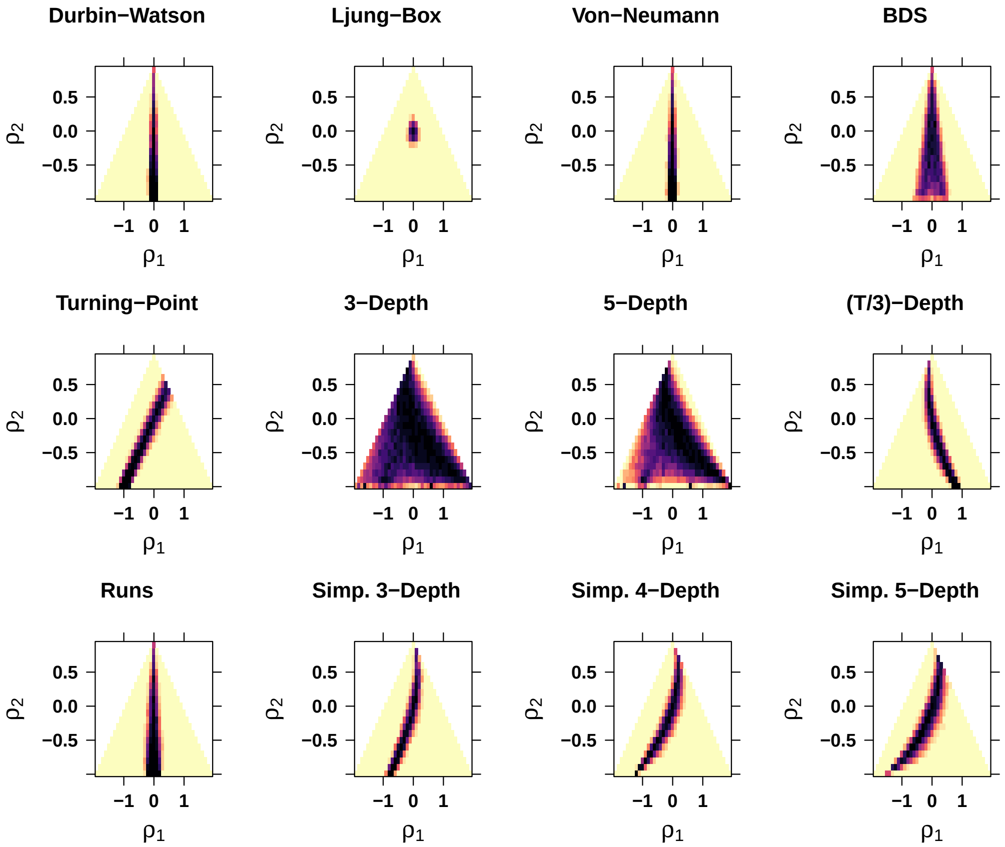

Figure 5: Simulated power of the different tests for stationary second order autoregressive time series with 500 observations.

these processes are stationary, seasonal 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 $t \in \{S+1, \ldots, T\}$ 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.

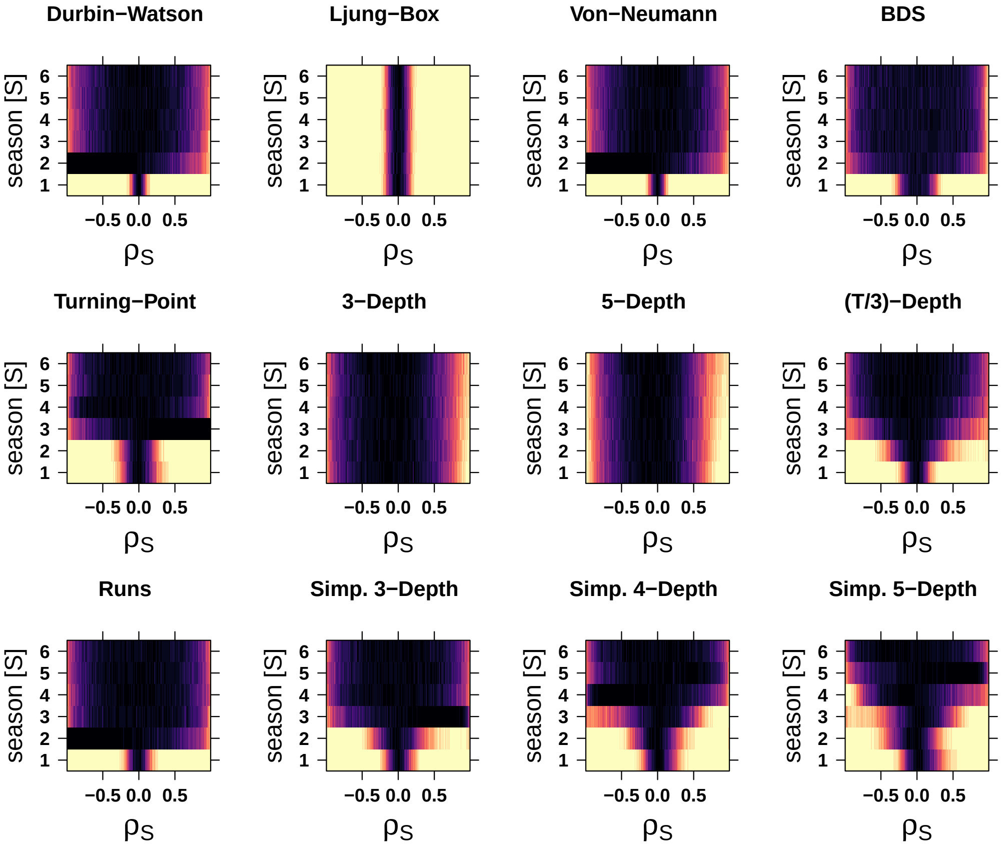

Figure 6: Simulated power of the different tests for stationary seasonal autoregressive time series with 500 observations.

4 Detection of model deviations

While the previous section was solely focused on detecting independence, we will now consider testing independence in situations where deviations from the model assumptions might also be present.

As a first example, we consider first order autoregressive models which might have a visible jump in their average behaviour. To this end, we consider the AR(1) process $(Y_t)_{t=1,\dots,T}$ given in (5) 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

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

in which the autoregression coefficient ρ_1 from $(Y_t)_{t=1,\ldots,T}$ is the model parameter and the jump height h can be considered as a nuisance parameter. In this model, the jump does not influence subsequent observations just like in the case of contaminated time series (see 3.1.2).

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. 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.

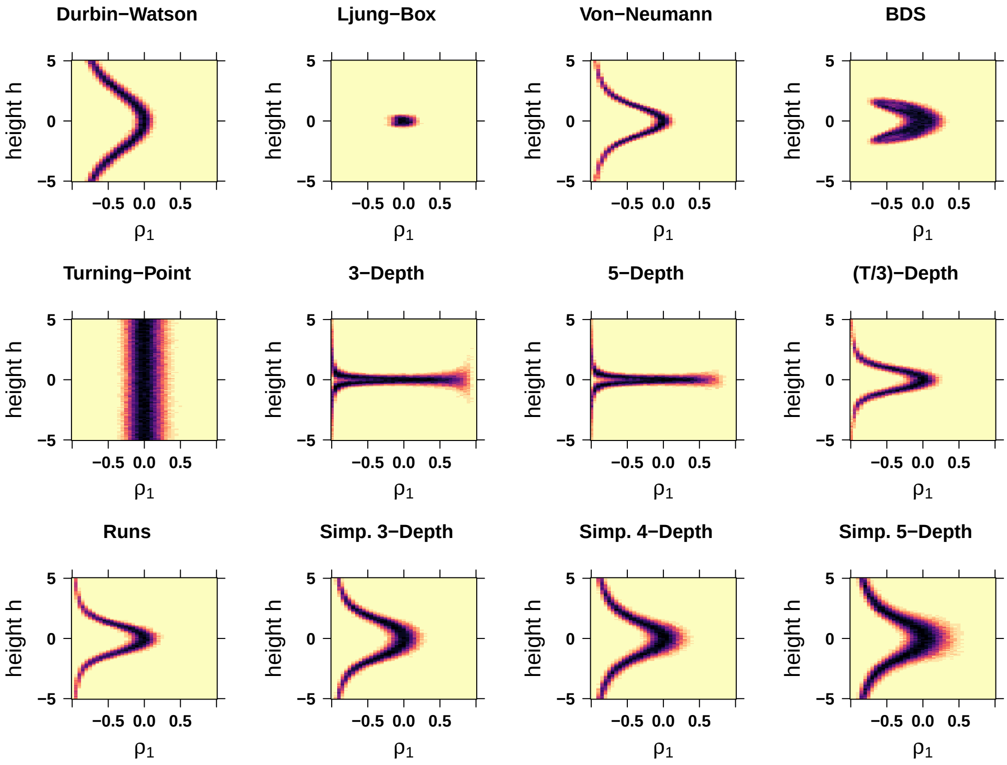

Figure 7: Simulated power of the different tests for stationary first order autoregressive time series with 500 observations and a jump after the 250th observation.

Another deviation from the model assumptions that was considered in this paper 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 time series, they were simulated according to the formula

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

where $(Y_t)_{t=1,\ldots,T}$ is the process given in (5), the autoregression coefficient ρ_1 from $(Y_t)_{t=1,\ldots,T}$ is the model parameter, and the drift parameter δ is the nuisance parameter. 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 second. For more details see [30]. 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 [30] 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 minutes 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 a.m., for the days of one 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 minutes 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 eight explanatory variables are considered: 1. Time, the day, 2. TBr, the current temperature below the bridge, 3. TSun, the current temperature above the bridge, 4. TBr4h, the temperature 4 hours ago below the bridge, 5. TSun4h, the temperature 4 hours ago above the bridge, 6. TBrM, the mean temperature of the previous 7 days below the bridge, 7. TSunM, the mean temperature of the previous 7 days above the bridge, 8. WN2(-1), the crack

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

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

width of the day before (AR(1) component). Let p denote the number of selected explanatory variables, i.e. $p \in \{0,\ldots,8\}$ here, $r_t^i(\theta)$ be the t'th residual at θ of the i'th time interval, $t = 1,\ldots,364$, $i = 1,\ldots,96$. For all 2^8 possible selections of the eight variables the following model selection criteria are used:

1) AIC: $\sum_{i=1}^{96}\left(\ln\left(\frac{1}{364-p}\sum_{t=1}^{364}(r_t^i(\widehat{\theta}_i)^2\right)+2p\right)$ where $\widehat{\theta}_i$ is the least squares estimator in the i'th time series,

2) trim. AIC: sum of the AICs of 10% trimmed sum of squared residuals for i'th time series where $\widehat{\theta}_i$ 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 $d_3(r_1(\widehat{\theta}_i),\ldots,r_T(\widehat{\theta}_i))$ where $\widehat{\theta}_i$ is the MM-estimator of θ 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 $H_0 : \widehat{\theta}_i$ satisfies (2) at the i'th time series where $\widehat{\theta}_i$ 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 no reduction of the variables while the criteria based on the full 3-depth lead to the smallest models, namely a model with variables Time 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.

Figure 8 provides the boxplots of the 96 p-values of some independence tests at the 96 time series when they are used for the residuals of the full

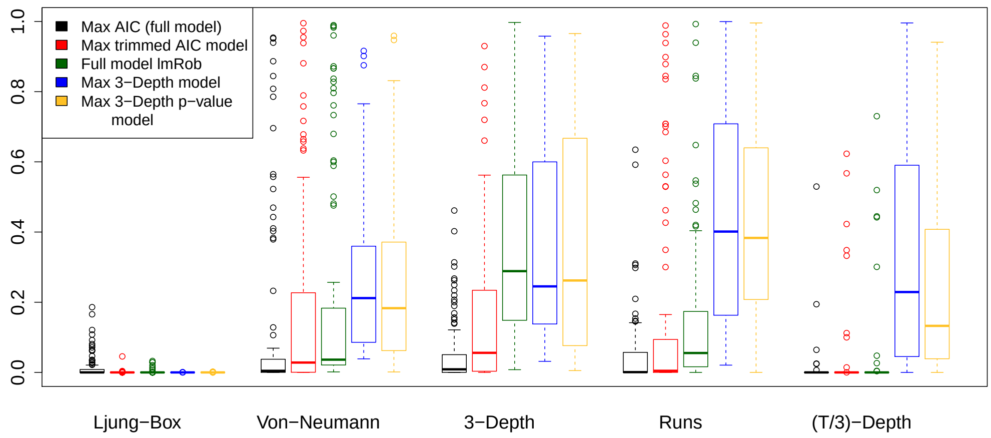

Figure 8: Boxplots of the p-values of the independence tests for different optimal models.

model with least squares estimates (which is also the selected model by maximizing the AIC criterion), the full model with the robust MM-estimate of `lmRob`, the selected model by maximizing the 10% trimmed AIC, and the selected model by maximizing the 3-depth and the p-value of the one-sided 3-depth tests, respectively. One can see that almost all tests reject the independence of the residuals if the model is chosen with the classical AIC criterion. If the full model is used with the robust MM-estimate or the model is chosen by maximizing the trimmed AIC-critrion, then there are some time series where the independence of the residuals is not rejected. However, much more time series with no independence rejection exist if the small models are used which were found by using the 3-depth criteria. This holds for the Von-Neumann-Rank-Ratio test, the runs test, the full 3-depth test, and the full $(T/3)$-depth test. Note that the boxplots concern the p-values of the two-sided K-depth tests while the fourth model selection uses the one-sided 3-depth test. Furthermore, note that the Ljung-Box test is the only one that rejects independence in all models. The full $(T/3)$-depth test is similarly strict in all cases except for models found by the 3-depth criteria for which the independence assumption is not rejected in most times series.

In order to see that the model chosen according to the 3-depth criterion fits the data well, Figures 9 and 10 contain the crack widths and their predictions according to this 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 time series according to the two-sided full 3-depth test. Note that even the data with the smallest p-value is still fitted fairly well.

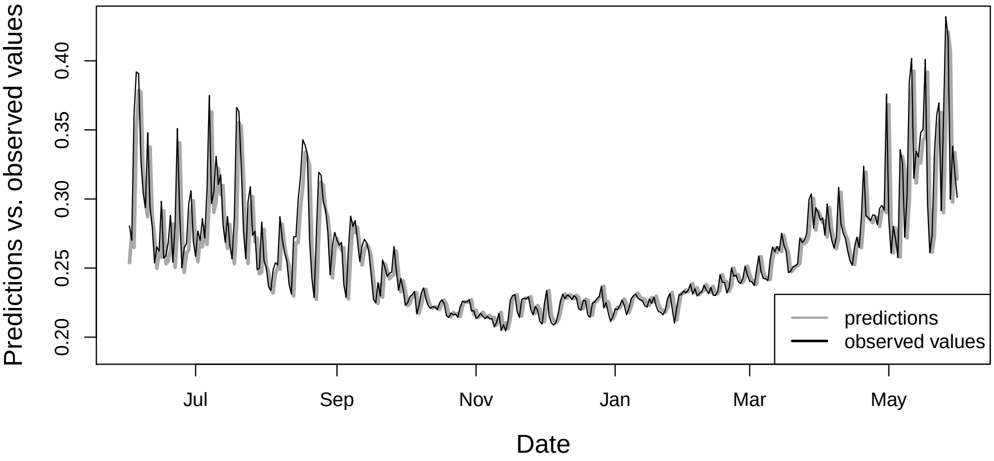

Figure 9: Predicted and observed values of the time series with the maximal p-value of the 3-depth test in its optimal model found by maximizing the full 3-sign depth.

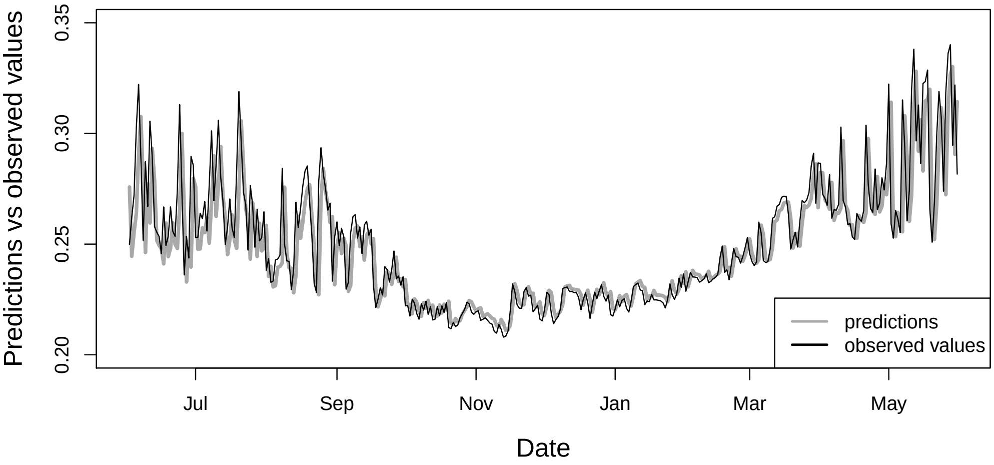

Figure 10: Predicted and observed values of the time series with the minimal p-value of the 3-depth test in its optimal model found by maximizing the full 3-sign depth.

Supplementary material

Further simulation results, the data set, and the R-code can be found under `https://www.statistik.tu-dortmund.de/2273.html`.

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 test 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 behaviour 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.

In an application with data from a bridge monitoring, we demonstrated how the K-depth tests can be used to improve the modelling of time series which depend on a related d-dimensional co-process. In this context, classical model selection methods fail while two methods based on the 3-sign depth provided reasonable models with few variables. However, more investigations are necessary to confirm the suitability of the K-sign depth and K-depth tests for model selection. In particular, we did not treat the situation where the model parameter must be estimated and what happens when the model parameter is estimated in a wrong model.

Acknowledgements

The authors gratefully acknowledge support from the Collaborative Research Center "Statistical Modeling of Nonlinear Dynamic Processes" (SFB 823, B4, B5) of the German Research Foundation (DFG).

References

[1] E. L. Lehmann, Testing Statistical Hypothesis, 2nd Edition, Chapman & Hall, New York, 1994.

[2] A. Wald, J. Wolfowitz, On a test whether two samples are from the same population, The Annals of Mathematical Statistics 11 (2) (1940) 147–162.
URL `http://www.jstor.org/stable/2235872`

[3] J. Gibbons, S. Chakraborti, Nonparametric statistical inference, Statistics, textbooks and monographs, Marcel Dekker Incorporated, 2003.
URL `https://books.google.pt/books?id=dPhtioXwI9cC`

[4] K. Leckey, D. Malcherczyk, C. H. Müller, Powerful generalized sign tests based on sign depth, SFB - discussion paper No. 823 (2020).
URL `https://eldorado.tu-dortmund.de/handle/2003/39099`

[5] C. P. Kustosz, C. H. Müller, M. Wendler, Simplified simplicial depth for regression and autoregressive growth processes, Journal of Statistical Planning and Inference 173 (2016) 125–146.

[6] C. P. Kustosz, A. Leucht, C. H. Müller, Tests based on simplicial depth for AR(1) models with explosion, Journal of Time Series Analysis 37 (2016) 763–784.
URL `http://dx.doi.org/10.1111/jtsa.12186`

[7] C. P. Falkenau, Depth based estimators and tests for autoregressive processes with application on crack growth and oil prices, Dissertation, TU Dortmund, 2016.
URL `http://dx.doi.org/10.17877/DE290R-17269`

[8] M. Horn, C. H. Müller, Tests based on sign depth for multiple regression, SFB Discussion Paper 20 (07) (2020).
URL `https://www.statistik.tu-dortmund.de/2630.html`

[9] M. G. Kendall, Time-series, Griffin, London, 1973.

[10] M. Verbeek, A guide to modern econometrics, 4th Edition, Wiley, Chichester, West Sussex, 2012.

[11] G. M. Ljung, G. E. P. Box, On a measure of lack of fit in time series models, Biometrika 65 (2) (1978) 297–303.

[12] R. Bartels, The Rank Version of von Neumann's Ratio Test for Randomness, Journal of the American Statistical Association 77 (377) (1982) 40–46.

[13] W. A. Broock, J. A. Scheinkman, W. D. Dechert, B. LeBaron, A test for independence based on the correlation dimension, Econometric Reviews 15 (3) (1996) 197–235.

[14] P. J. Rousseeuw, M. Hubert, Regression depth, Journal of the American Statistical Association 94 (446) (1999) 388–402.

[15] R. Y. Liu, On a notion of simplicial depth, Proceedings of the National Academy of Sciences of the United States of America 85 (1988) 1732–1734.

[16] R. Y. Liu, On a notion of data depth based on random simplices, The Annals of Statistics 18 (1990) 405–414.

[17] J. W. Tukey, Mathematics and the picturing of data, Proceedings of the International Congress of Mathematicians 2 (1975) 523–531.

[18] C. P. Kustosz, C. H. Müller, Analysis of crack growth with robust, distribution-free estimators and tests for non-stationary autoregressive processes, Statistical Papers 55 (1) (2014) 125–140.
URL `http://dx.doi.org/10.1007/s00362-012-0479-5`

[19] D. Malcherczyk, K. Leckey, C. H. Müller, K-sign depth: From asymptotics to efficient implementation, Journal of Statistical Planning and Inference 215 (2021) 344–355.
URL `https://www.sciencedirect.com/science/article/pii/S0378375821000458`

[20] M. Horn, GSignTest: Robust Tests for Regression-Parameters via Sign Depth, R package version 1.0.7 (2021).
URL `https://github.com/melaniehorn/GSignTest`

[21] D. N. Gujarati, D. C. Porter, Basic econometrics, 5th Edition, The McGraw-Hill series Economics, McGraw-Hill Irwin, Boston, Mass., 2009.

[22] A. Zeileis, T. Hothorn, Diagnostic checking in regression relationships, R News 2 (3) (2002) 7–10.
URL `https://CRAN.R-project.org/doc/Rnews/`

[23] A. Hart, S. Martínez, spgs: Statistical Patterns in Genomic Sequences, R package version 1.0-3 (2019).
URL `https://CRAN.R-project.org/package=spgs`

[24] F. Caeiro, A. Mateus, randtests: Testing randomness in R, R package version 1.0 (2014).
URL `https://CRAN.R-project.org/package=randtests`

[25] D. Sarkar, Lattice: Multivariate Data Visualization with R, Springer, New York, 2008, iSBN 978-0-387-75968-5.
URL `http://lmdvr.r-forge.r-project.org`

[26] S. Garnier, viridis: Default Color Maps from 'matplotlib', R package version 0.5.1 (2018).
URL `https://CRAN.R-project.org/package=viridis`

[27] A. Robotham, magicaxis: Pretty Scientific Plotting with Minor-Tick and Log Minor-Tick Support, R package version 2.0.10 (2019).
URL `https://CRAN.R-project.org/package=magicaxis`

[28] S. Meschiari, latex2exp: Use LaTeX Expressions in Plots, R package version 0.4.0 (2015).
URL `https://CRAN.R-project.org/package=latex2exp`

[29] A. Fox, Outliers in time series, Journal of the Royal Statistical Society. Series B 34 (3) (1972) 350–363.

[30] S. Abbas, R. Fried, J. Heinrich, M. Horn, M. Jakubzik, J. Kohlenbach, R. Maurer, A. Michels, C. Müller, Detection of anomalous sequences in crack data of a bridge monitoring, in: K. Ickstadt, H. Trautmann, G. Szepannek, N. Bauer, K. Lübke, M. Vichi (Eds.), Applications in Statistical Computing - From Music Data Analysis to Industrial Quality Improvement, Springer, New York, 2019, pp. 251–269.