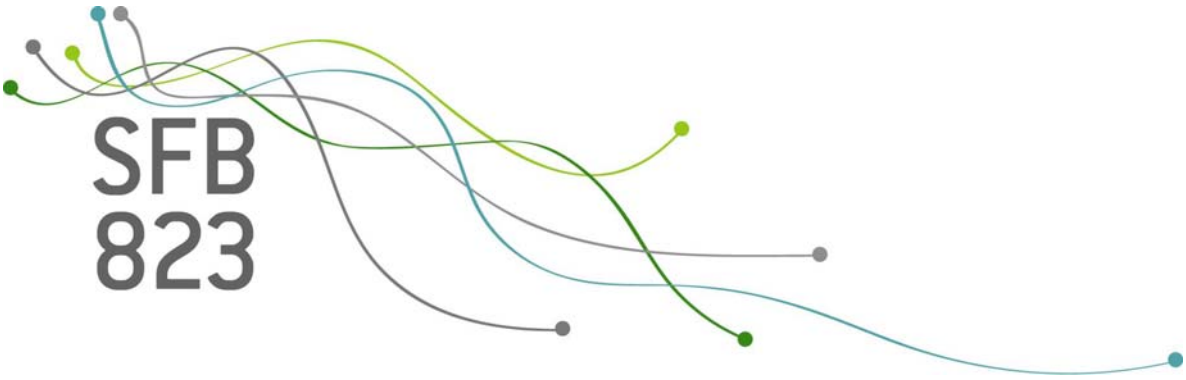
SFB 823

Discussion Paper

# Generalized sign tests based on sign depth

Kevin Leckey, Dennis Malcherczyk,
Christine H. Müller

Nr. 37/2018



# Generalized sign tests based on sign depth

Kevin Leckey, Dennis Malcherczyk,
and Christine H. Müller *

December 14, 2018

**Abstract**

We introduce generalized sign tests based on $K$-sign depth, shortly denoted by $K$-depth. These so-called $K$-depth tests are motivated by simplicial regression depth. Since they depend only on the signs of the residuals, these test statistics are easy to comprehend and outlier robust. We show that the $K$-depth test with $K = 2$ is equivalent to the classical sign test so that $K$-depth tests with $K > 2$ are generalizations of the classical sign test. Since the $K$-depth test with $K = 2$ is equivalent to the classical sign test, it has the same drawbacks as the classical sign test. However, the generalized sign tests with $K > 2$ are much more powerful. We show this by deriving their behavior at observations with few sign changes. Thereby we also prove an upper bound for the $K$-depth which is attained by observations with alternating signs of residuals. Furthermore, we prove the consistency of the $K$-depth. Finally, we demonstrate the good power of the $K$-depth tests for relevance testing, quadratic regression, and tests for explosive AR(2) and nonlinear AR(1) regression.

*Keywords:* Simplicial regression depth, $K$-sign depth, $K$-depth test, sign test, relevance test, quadratic regression, nonlinear AR(1) regression, AR(2) regression.

# 1 Introduction

We consider stochastic models where a parameter $\theta \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$, is unknown and where residuals $R_1(\theta), \ldots, R_N(\theta)$ of $N$ observations in $\mathbb{R}$ are independent and identically distributed with

$$P_\theta(R_n(\theta) > 0) = \frac{1}{2} = P_\theta(R_n(\theta) < 0). \tag{1}$$

*Department of Statistics, TU Dortmund University, D-44227 Dortmund, Germany, kevin.leckey@tu-dortmund.de, dennis.malcherczyk@tu-dortmund.de, cmueller@statistik.tu-dortmund.de

Examples of such models are linear and nonlinear regression models with additive errors $E_n$ where the observations are of the form $Y_n = g(x_n, \theta) + E_n$ with $x_n \in \mathbb{R}^q$ so that the residuals are $R_n(\theta) = Y_n - g(x_n, \theta)$. Generalized linear and nonlinear models are further examples if the link function can be expressed by the median of the observations $Y_n$, i.e. if $\text{med}(Y_n) = g(x_n, \theta)$. More examples are given by stochastic processes with i.i.d. increments as AR(p) processes given by $Y_n = g(Y_{n-1}, \ldots, Y_{n-p}, \theta) + E_n$.

In models given by (1), the classical sign test can be used for testing hypotheses $H_0 : \theta = \theta^0$ and for deriving confidence sets. The classical sign tests counts the positive (or negative) residuals and uses the fact that the number of positive (negative) residuals has a binomial distribution with parameter $\frac{1}{2}$. In particular, it does not reject the null hypothesis $H_0 : \theta = \theta^0$ if half of the residuals $R_n(\theta^0)$ are positive and half of them are negative. However, this can happen also for parameters far away from $\theta^0$, typically in situations where the first half of residuals are negative (positive) and the second half of residuals are positive (negative). Hence the power of the classical sign test for such alternatives is very bad.

The bad power of the classical sign test can be seen, for example, in the simulations studies of Kustosz et al. (2016a) and Kustosz et al. (2016b) where the classical sign test was compared to tests based on simplicial regression depth for linear and nonlinear regression and autoregression with two unknown regression parameters. Simplicial regression depth is a modification of the regression depth introduced by Rousseeuw and Hubert (1999) to generalize the depth notion to regression. Originally, the half space depth of Tukey (1975) was used to get a generalization of the median for multivariate data. Liu (1988, 1990) extended this to simplicial depth. Afterwards many depth notions were introduced, see e.g. Zuo and Serfling (2000); Mosler (2002); Mizera (2002); Mizera and Müller (2004); López-Pintado and Romo (2007, 2009); Agostinelli and Romanazzi (2011); Denecke and Müller (2011); Lok and Lee (2011); Paindaveine and van Bever (2013); Claeskens et al. (2014); López-Pintado et al. (2014); Paindaveine and Van Bever (2018); Nagy and Ferraty (2018); Wang (2019). Regression depth and simplicial regression depth are two of these depth notions and the relation between them is the same as between Liu's simplicial depth and the half space depth. Both simplicial depth notions are given by the relative number of subsets with $p + 1$ observations where the half space depth or the regression depth, respectively, of a $p$-dimensional parameter vector is greater than zero.

Simplicial depth has the advantage that it is a U-statistics although it is often a degenerated U-statistic so that more effort is necessary to get the asymptotic distribution, see Dümbgen (1992); Müller (2005); Wellmann et al. (2009); Wellmann and Müller (2010a,b); Kustosz et al. (2016a). Moreover, for its calculation, Rousseeuw and Hubert (1999) and Müller (2005) noted that the regression depth of a $p$-dimensional parameter vector within $p + 1$ observations is greater than zero if and only if the residuals have alternating signs. Sufficient conditions and a proof for this property are given by Kustosz et al. (2016b). One of the sufficient conditions is that the observations are given by a natural order as this is the case for time series. Moreover, the proof of the asymptotic distribution of the simplicial regression depth for $p = 2$ given by Kustosz et al. (2016a) is not restricted to AR(1) regression since it uses only the alternating signs of $p + 1 = 3$ residuals. In particular, the derived asymptotic distribution can be used as soon as there is a natural

ordering of the observations and the median of the residuals is zero. This leads to the idea to define simplicial depth not via regression depth but via alternating signs of residuals.

We call this depth notion $K$-sign depth or shortly $K$-depth where $K$ stands for the number of residuals used in the simplicial depth. It is not necessary any more to choose $K = p + 1$ if the unknown parameter vector is $p$-dimensional. Tests based on this depth notion are called $K$-depth tests. We show in this paper that the $K$-depth test with $K = 2$ is equivalent to the classical sign test so that $K$-depth tests with $K > 2$ are indeed generalizations of the classical sign test. Moreover, we demonstrate that $K$-depth tests with $K > 2$ are much more powerful than the classical sign test. In particular, they do not have the drawback of bad power as soon as half of the residuals are positive and the others are negative as is the case for the classical sign test. Since they are based only on signs of residuals, these tests are outlier robust.

In Section 2, we introduce the $K$-depth and the $K$-depth tests, show the consistency of $K$-depth, and provide the equivalence of the $K$-depth test with $K = 2$ and the classical sign test. A comparison between the $K$-depths and $K$-depth tests for different values of $K$ is given in Section 3 by considering their behavior in situations where only few sign changes appear in the residuals. This is done by p-values of the tests and by the test statistics themselves. Moreover, we derive the maximum possible value of the test statistics which is always attained for alternating signs of residuals. Section 4 demonstrates the good power of the $K$-depth tests for $K = 3$ and $K = 4$ via simulations for relevance testing, quadratic regression, AR(2)- and nonlinear AR(1)-models. Finally, a discussion of the results and an outlook is given in Section 5. All proofs are given in the appendix.

# 2 $K$-sign depth and $K$-depth tests

## 2.1 $K$-depth

If $r_1 := r_1(\theta), \ldots, r_N := r_N(\theta)$ are realized residuals for the parameter $\theta$ then the $K$-sign depth or shortly $K$-depth $d_K(r_1, \ldots, r_N)$ of these residuals is the relative number of subsets with $K$ elements with alternating signs of the residuals, i.e.

$$\begin{aligned} &d_K(r_1, \ldots, r_N) \\ &:= \frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left( \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \right), \end{aligned} \quad (2)$$

if $K \ge 2$. Thereby $\mathbb{1}\{\ldots\}$ denotes the indicator function, i.e. the function which equals 1 if the condition in the brackets is satisfied and 0 otherwise. The definition depends strongly on the order of the residuals so that an order of the residuals must be specified in advance.

Since the definition in (2) only makes sense for $K \ge 2$, we define for $K = 1$

$$d_1(r_1, \ldots, r_N) := \frac{2}{N} \sum_{n=1}^{N} \mathbb{1}\left\{r_n < 0\right\},$$

on which the test statistic of the classical sign test is based.

## 2.2 Consistency of $K$-depth

If $r_1 := r_1(\theta), \ldots, r_N := r_N(\theta)$ are realizations of independent random variables $R_1 := R_1(\theta), \ldots, R_N := R_N(\theta)$ satisfying (1) then the expectation of the $K$-depth is given by

$$E_\theta\big(d_K(R_1(\theta), \ldots, R_N(\theta))\big) = \frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left( \left(\frac{1}{2}\right)^K + \left(\frac{1}{2}\right)^K \right) = \left(\frac{1}{2}\right)^{K-1}. \quad (3)$$

To see that the $K$-depth converges to this expected value, we need the following representation of $K$ alternating signs.

**Lemma 2.1.**
*If $E_{n_1}, ..., E_{n_K}$ are random variables with $P(E_{n_i} \neq 0) = 1$ for $i = 1, ..., K$ and $K \in \mathbb{N}\setminus\{1\}$ then we have*

$$\begin{aligned} &\prod_{k=1}^{K} \mathbb{1}\{E_{n_k}(-1)^k > 0\} + \prod_{k=1}^{K} \mathbb{1}\{E_{n_k}(-1)^k < 0\} - \left(\frac{1}{2}\right)^{K-1} \\ =&\frac{1}{2^{K-1}} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{1 \le i(1) < \ldots < i(2L) \le K} (-1)^{i(1)+\ldots+i(2L)} \prod_{j=1}^{2L} \Phi(E_{n_{i(j)}}) \ P\text{-almost surely}, \end{aligned} \quad (4)$$

*where $\Phi(x) := \mathbb{1}\{x > 0\} - \mathbb{1}\{x < 0\}$.*

In particular, Formula (4) yields for $K = 2, 3, 4$

$$\begin{aligned} K = 2: \ & -\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}), \quad (5) \\ K = 3: \ & -\frac{1}{4}\left(\Phi(E_{n_1})\Phi(E_{n_2}) - \Phi(E_{n_1})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_3})\right), \\ K = 4: \ & \frac{1}{8}\Big( \prod_{i=1}^{4} \Phi(E_{n_i}) - \Phi(E_{n_1})\Phi(E_{n_2}) + \Phi(E_{n_1})\Phi(E_{n_3}) \\ & \quad - \Phi(E_{n_1})\Phi(E_{n_4}) - \Phi(E_{n_2})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_4}) - \Phi(E_{n_3})\Phi(E_{n_4})\Big). \end{aligned}$$

**Theorem 2.2.**
*If $R_1(\theta), \ldots, R_N(\theta)$ are satisfying (1) then*

$$d_K(R_1(\theta), \ldots, R_N(\theta)) \longrightarrow \left(\frac{1}{2}\right)^{K-1}$$

*$P_\theta$-almost surely for $N \to \infty$ for all $K \in \mathbb{N}$.*

## 2.3 $K$-depth tests

Up to now the asymptotic distribution of

$$T_K(\theta) := N\left(d_K(R_1(\theta),\dots,R_N(\theta)) - \left(\frac{1}{2}\right)^{K-1}\right) \tag{6}$$

is only known for $K = 2$ and $K = 3$. It is the same as for the simplicial depth for autoregressive models derived in Kustosz and Müller (2014) and Kustosz et al. (2016a) since only the signs of the residuals are used in these derivations. However, if $N$ is not too large, the finite sample distribution for any $K$ can be easily simulated since the determination of the $K$-depth with an underlying C++ algorithm is fairly fast.

If a $K$-depth is used with $K \geq 2$ then a hypothesis of the form $H_0 : \theta \in \Theta^0$ shall be rejected if the $K$-depth $d_K(r_1(\theta),\dots,r_N(\theta))$ of $\theta$ or $T_K(\theta)$ is too small for all $\theta \in \Theta^0$. Hence, if $q_\alpha$ is the $\alpha$-quantile of the finite sample or asymptotic distribution of $T_K(\theta)$ under $\theta$ then the $K$-depth test for $H_0 : \theta \in \Theta^0$ is given by

$$\text{reject } H_0 : \theta \in \Theta^0 \text{ if } \sup_{\theta\in\Theta^0} T_K(\theta) < q_\alpha. \tag{7}$$

That this is indeed an (asymptotic) $\alpha$-level test can be seen as in Müller (2005) for simplicial regression tests.

## 2.4 2-depth test and the classical sign test

At first, we show that the 2-depth test, i.e. the $K$-depth test with $K = 2$, and the classical sign test for $H_0 : \theta \in \Theta^0$ are equivalent. We define the classical sign test in its asymptotic form here. I.e. the classical sign test is given by

$$\text{reject } H_0 : \theta \in \Theta^0 \text{ if } \inf_{\theta\in\Theta^0} T_{\text{sign}}(\theta)^2 > \chi^2_{1,1-\alpha}, \tag{8}$$

where $T_{\text{sign}}(\theta) := \dfrac{1}{\sqrt{N}}\displaystyle\sum_{n=1}^{N}\left(\frac{\mathbb{1}\{R_n(\theta) < 0\} - \frac{1}{2}}{\frac{1}{2}}\right) = \sqrt{N}\,(d_1(R_1(\theta),\dots,R_N(\theta)) - 1)$ and $\chi^2_{1,\alpha}$ is the $\alpha$-quantile of the $\chi^2_1$ distribution. Hence the test statistic of the classical sign test depends only on the number $N_-$ of negative residuals. Equivalently, it can be defined via the number $N_+$ of positive residuals. Thereby $T_{\text{sign}}(\theta)^2$ is minimized if $N_- = N_+ = \frac{N}{2}$.

The test statistic (6) for the 2-depth test also only depends on the number $N_+$ of positive residuals since the 2-depth satisfies almost surely

$$d_2(R_1(\theta),\dots,R_N(\theta)) = \frac{1}{\binom{N}{2}}\,N_+\,(N - N_+).$$

In particular the 2-depth and thus the corresponding test statistic is maximized at $N_+ = \frac{N}{2}$ leading to

$$d_2(R_1(\theta),\dots,R_N(\theta)) = \frac{1}{\binom{N}{2}}\,\frac{N}{2}\,\frac{N}{2} = \frac{1}{2}\,\frac{N}{N-1}.$$

If the finite sample distribution of $T_{\text{sign}}(\theta)^2$ and $T_K(\theta)$ with $K = 2$ is used then the corresponding tests are equivalent since both test statistics depend only on $N_+$ or $N_-$, respectively. However, if both tests are used in their asymptotic versions then they are only asymptotically equivalent.

To see this, note at first that the asymptotic distribution of the test statistic $T_K(\theta)$ with $K = 2$ was derived by Müller (2005). It is the distribution of a random variable $\frac{1}{2}(1 - X)$ where $X$ has a $\chi^2_1$ distribution. In particular, $q_\alpha$ in (7) is then the $\alpha$-quantile of the distribution of this random variable and it satisfies $q_\alpha = \frac{1}{2}(1 - \chi^2_{1,1-\alpha})$.

The following lemma provides the relationship between the two test statistics. Thereby, the representation of alternating signs given by Lemma 2.1 is used.

**Lemma 2.3.**

$$T_2(\theta) = \frac{N}{2(N-1)} - \frac{N}{2(N-1)} T_{sign}(\theta)^2.$$

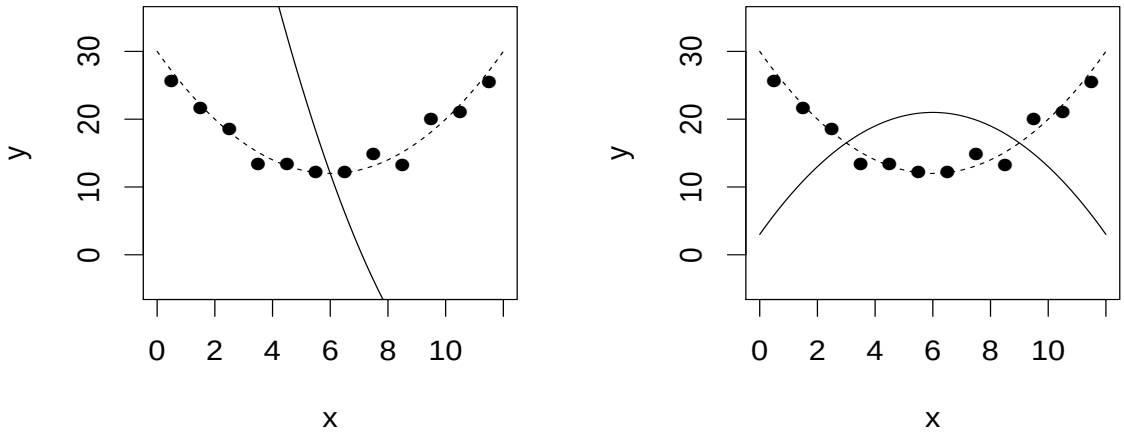



Figure 1: 12 observations generated by $Y_n = g(x_n, \theta^0) + E_n$ with $g(x, \theta^0) = 30 - 6\,x + 0.5\,x^2$ (dashed line, $\theta^0 = (30, -6, 0.5)^\top$) and $E_n \sim \mathcal{N}(0, 1.5^2)$. Left hand side: nonfit with one sign change of $\theta^1 = (120, -24, 1)^\top$ providing the function $g(x, \theta^1) = 120 - 24x + x^2$ (solid line). Right hand side: nonfit with two sign changes of $\theta^2 = (3, 6, -0.5)^\top$ providing the function $g(x, \theta^2) = 3 + 6\,x - 0.5\,x^2$ (solid line).

Since the 2-depth test as well as the sign test depend only on the number $N_+$ of positive residuals, the ordering of the positive and negative residuals is not relevant. In particular maximum 2-depth $d_2$ and minimum deviation of $d_1$ from 1 is also achieved if e.g. the residuals of the first half are negative and the residuals of the second half are positive. But this is a typical situation of a nonfit. Thereby, according to Rousseeuw and Hubert (1999), a parameter $\theta$ is called a nonfit if there is another parameter $\tilde{\theta}$ so that $|r_n(\tilde{\theta})| < |r_n(\theta)|$ for all $n = 1, \ldots, N$. Figure 1 shows two cases of nonfit for the quadratic regression. On the

left hand side, the first 6 residuals are negative while the last 6 residuals are positive, i.e. there is only one sign change of the residuals. On the right hand side, the first 3 residuals are positive followed by 6 negative residuals and the last 3 residuals are positive again so that here are two sign changes. In both cases, 6 residuals are positive and 6 residuals are negative so that maximum 2-depth and minimum deviation of $d_1$ from 1 is achieved. However, for $K$-depth with $K \geq 3$ this drawback does not appear which is shown in the next section.

# 3 Comparison of $K$-depths and $K$-depth tests

## 3.1 $K$-depth for alternating signs

At first we study the behavior of $K$-depth for alternating signs of residuals. The residuals $r_1, \ldots, r_N$ have alternating signs if $\operatorname{sign}(r_n) = -\operatorname{sign}(r_{n+1})$ for $n = 1, \ldots, N-1$ is satisfied. Alternating signs is the best situation for a good fit and $K$-depth attains its maximum value in this situation. Therefore it is of interest what exactly this maximum value is. This is given by the following lemma. As usual, we use the convention $\binom{n}{k} = 0$ for $n < k$.

**Theorem 3.1.** *Suppose the residuals $r_1, \ldots, r_N$ have alternating signs. Then the following holds for all $K \leq N$:*

*(a) If $N + K$ is odd then*

$$d_K(r_1, \ldots, r_N) = \frac{2}{\binom{N}{K}} \binom{(N+K-1)/2}{K}$$

*(b) If $N + K$ is even then*

$$d_K(r_1, \ldots, r_N) = \frac{1}{\binom{N}{K}} \left( \binom{(N+K)/2}{K} + \binom{(N+K-2)/2}{K} \right)$$

Theorem 3.1 provides that the $K$-depth of alternating signs converges for $N \to \infty$ to the expected values of the $K$-depth. This holds also if the residuals are alternating in blocks of size $M$, i.e. if $\operatorname{sign}(r_n) = \operatorname{sign}(r_1)$ for $n = 2LM + m$ and $\operatorname{sign}(r_n) = -\operatorname{sign}(r_1)$ for $n = (2L+1)M + m$ for $L = 0, 1, 2, \ldots$ and $m = 1, \ldots, M$.

**Corollary 3.1.** *If the residuals $r_1, \ldots, r_N$ are alternating in blocks of size $M$, then*

$$\lim_{N \to \infty} d_K(r_1, \ldots, r_N) = \left(\frac{1}{2}\right)^{K-1}.$$

*Remark* 3.2.
a) Since the maximum possible $K$-depth is converging to the expected value of the $K$-depth according to Corollary 3.1 and the minimum possible depth is always zero, the distribution of the $K$-depth cannot be symmetric around its expectation.

b) If the residuals are alternating in blocks then this is an indicator of a good fit of the overall model to the data although the residuals are not independent. This happens in particular if the data contain some vibration behavior which is difficult to filter out. Hence it is desirable to not reject the model when the residuals are alternating in blocks. Fortunately, the $K$-depth converges to the maximum possible value in these situations which is a necessary condition for the test not to reject. Note that a more careful asymptotic study along the lines of Corollary 3.1 reveals that, for $r_1, \ldots, r_N$ as in the corollary, $N(d_K(r_1, \ldots, r_N) - (1/2)^{K-1})$ converges to the maximum possible value $(K-1)K/2^K$ obtained for residuals with alternating signs. Hence, the $K$-depth test does not reject the model if the sample size is sufficiently large. This is not the case for a simplified $K$-depth which uses only subsequent residuals. The simplified $K$-depth can be defined as in Kustosz et al. (2016b) for $K \geq 2$ by

$$\begin{aligned} & d_K^S(r_1, \ldots, r_N) \\ & := \frac{1}{N-K+1} \sum_{n=1}^{N-K+1} \left( \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n+k-1} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n+k-1} < 0\right\} \right). \end{aligned}$$

Although the asymptotic distribution of the simplified $K$-depth is known for each $K \geq 2$ according to Kustosz et al. (2016b) and it is faster to compute, it has the drawback that a test based on it rejects models if the data are contaminated by some vibration noise. Moreover, since the simplified $K$-depth only considers $N-K+1$ subsets instead of $\binom{N}{K}$, tests based on it are less powerful than tests based on the full $K$-depth. This can be clearly seen from the examples in Kustosz et al. (2016a) and Falkenau (2016) for AR(1)-models.

## 3.2 Comparison via the test statistics

In Section 2.4, it was shown that maximum 2-depth $d_2$ and minimum deviation of $d_1$ from 1 is achieved if e.g. the residuals of the first half are negative and the residuals of the second half are positive. This is is a typical situation of a nonfit with only one sign change in the residuals. However, for $K$-depth with $K \geq 3$ the following lemma is easy to see.

**Lemma 3.3.**
*a) If there is only one sign change in $r_1, \ldots, r_N$ then $d_K(r_1, \ldots, r_N) = 0$ for all $K \geq 3$.*
*b) If there are $L$ sign changes in $r_1, \ldots, r_N$ then $d_K(r_1, \ldots, r_N) = 0$ for all $K \geq L+2$.*

Note that a $K$-depth of zero is the smallest possible value of the $K$-depth. Hence this will lead always to a rejection of the hypotheses $H_0 : \theta = \theta^0$ by the $K$-depth test if the sample size is high enough that a rejection at level $\alpha$ is possible.

Usually a nonfit of a $p$-dimensional parameter is expressed by at most $p-1$ sign changes. Hence a $K$-depth test with $K = p+1$ will protect against bad power at nonfits. However, $K = p+1$ is not necessary for consistency of the $K$-depth tests at alternatives. Usually already 3-depth tests and 4-depth tests are consistent tests with good power at nonfits.

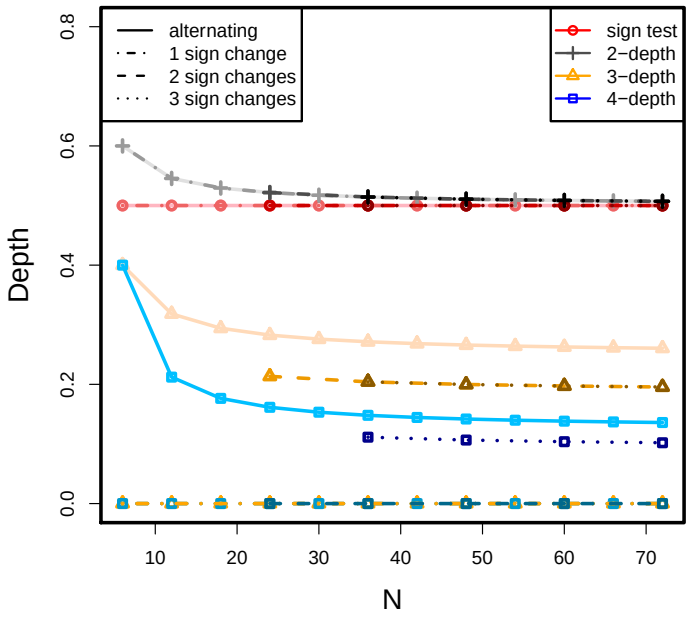



Figure 2: Comparison of 2-depth, 3-depth, 4-depth and the test statistic of the sign test for alternating signs of residuals and residuals with one, two or three sign changes for different samples sizes $N$. (For interpretation of this Figure, the reader is referred to the web version of this article.)

Figure 2 shows a comparison of 2-depth, 3-depth, 4-depth and the test statistic of the sign test for alternating signs of residuals and residuals with one, two or three sign changes for different samples sizes $N$. In the case of one sign change, the sign change happens after $\frac{N}{2}$ positive residuals. In the case of two sign changes, the first sign change is after $\frac{N}{4}$ positive residuals and the second sign change after $\frac{N}{2}$ negative residuals so that the last $\frac{N}{4}$ residuals are positive again. In the case of three sign changes, the sign changes are after $\frac{N}{4}$, $\frac{N}{2}$, and $\frac{3N}{4}$ residuals. Hence in all cases, $\frac{N}{2}$ residuals are positive and $\frac{N}{2}$ residuals are negative. In order to differentiate the various cases for the number of sign changes, different line types and several color brightnesses are used in Figure 2. E.g. the 4-depth is represented by different blue color levels.

At first note that the case of alternating signs always leads to highest depth which was derived in Section 3.1. However, Figure 2 demonstrates clearly that there is no difference between the 2-depth and sign test statistics for one, two, and three sign changes and alternating signs. This is opposite to 3-depth and 4-depth where the depth for alternating signs is always above the depth of few sign changes. Although 3-depth is not zero for two and three sign changes, it is constantly smaller than the 3-depth at alternating signs. The 3-depth for two and three sign changes is the same here and is equal to $\binom{N}{3}^{-1} \frac{N}{4} \frac{N}{2} \frac{N}{4} = \frac{3}{16} \frac{N^2}{(N-1)(N-2)}$. This converges to $\frac{3}{16}$ for $N \to \infty$ while the 3-depth for alternating signs converges to $\frac{1}{4}$ according to Corollary 3.1 so that the asymptotic difference is $\frac{1}{16}$. Hence for $N$ large enough, 3-depth for two and three sign changes will reject $H_0 : \theta = \theta^0$

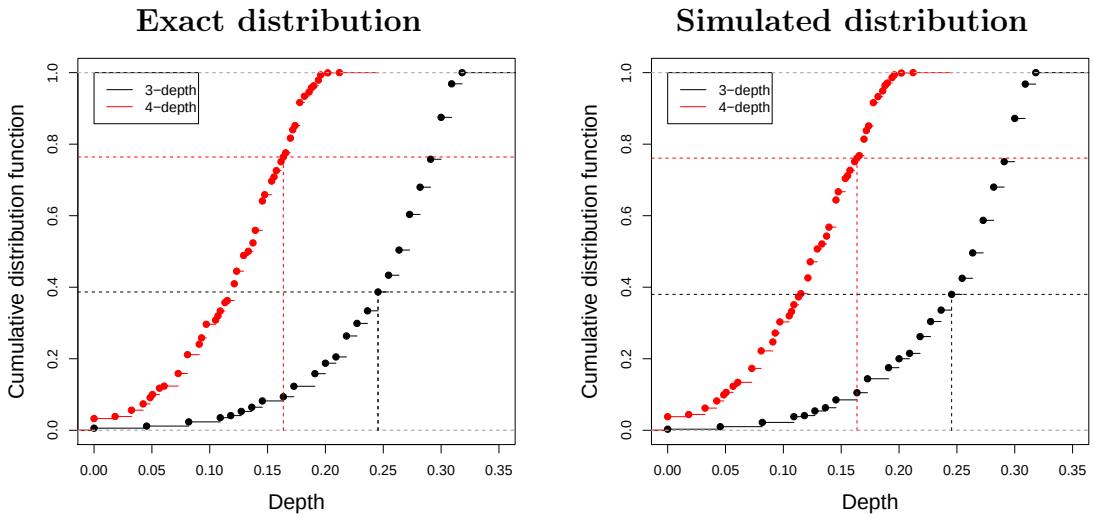



Figure 3: The exact (left) and the simulated (right) cumulative distribution function of 3-depth and 4-depth for $N = 12$ and the depths and the p-values for $N = 12$ observations with three sign changes (dashed lines).

since $T_3(\theta^0)$ converges in distribution (Kustosz et al., 2016a, Theorem 1) and therefore $d_3$ shrinks to $\frac{1}{4}$ under $H_0$. While the 4-depth for two sign changes is still zero, it becomes $\binom{N}{4}^{-1}\left(\frac{N}{4}\right)^4 = \frac{3}{32}\,\frac{N^3}{(N-1)(N-2)(N-3)}$ for three sign changes. This converges to $\frac{3}{32}$ for $N \to \infty$ while the 4-depth for alternating signs converges to $\frac{1}{8}$ according to Corollary 3.1. Here the asymptotic difference between depth for alternating signs and depths for three sign changes is $\frac{1}{32}$ which is smaller than the difference for 3-depth. If the shrinkage of the distribution of $d_4$ to $\frac{1}{8}$ is the same as for $d_3$ to $\frac{1}{4}$, then this would mean that the 3-depth test is more powerful for this case of three sign changes.

## 3.3 Comparison via p-values

Another indicator that the 3-depth test seems to be more powerful than the 4-depth test can be found in Figure 3. The left hand side of Figure 3 shows the exact cumulative distribution function (cdf) for 3-depth and 4-depth where the cdf was determined by calculating the depth of the $2^{12} = 4096$ possible residual vectors. Moreover, the dashed lines in Figure 3 show the 3-depth and the 4-depth and the corresponding p-values for the case of three sign changes in $N = 12$ observations, i.e. for $(1, 1, 1, -1, -1, -1, 1, 1, 1, -1, -1, -1)^\top$. The 3-depth of this residual vector is 0.2454545 which leads to a p-value of 0.3867188 while the 4-depth of this vector is 0.1636364. This leads to a p-value of 0.7641602 which is much higher than the p-value of the 3-depth test. The right hand side of Figure 3 provides a simulated cumulative distribution function where 1 000 times $N = 12$ observations were simulated. Each observation is chosen uniformly at random from $\{-1, 1\}$. Figure 3 shows that the exact and the simulated cdf are very similar. In particular, the simulated cdf yields similar p-values (0.38 and 0.761) for $(1, 1, 1, -1, -1, -1, 1, 1, 1, -1, -1, -1)^\top$ in the 3-depth test and 4-depth test.

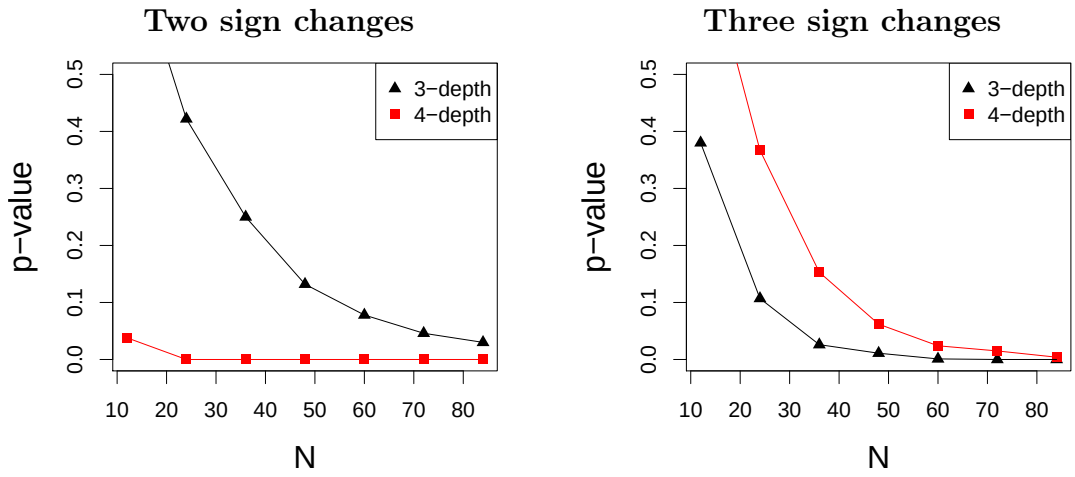



Figure 4: Simulated p-values of 3-depth tests and 4-depth tests for two sign changes (left) and for three sign changes (right) at different sample sizes $N$.

Simulated distributions based on 1 000 repetitions are also used in Figure 4. On the left hand side of this figure the simulated p-values of the 3-depth test and the 4-depth are given for the case of two sign changes. Here the situation is considered that the first third of the data has positive signs, the second third of the data has negative signs, and the last third of the data has positive signs. This leads to the highest possible 3-depth in the case of two sign changes. The right hand side provides the situation of three sign changes where the first quarter of signs is positive, the second negative, the third positive and the last negative. Here, the p-value of the sign test is always 1. The left hand side of Figure 4 shows that the 4-depth test leads to smaller simulated p-values than the 3-depth test in the situation of two sign changes. This is due to the fact that the 4-depth for two sign changes is always zero according to Lemma 3.3. However, the p-values of 3-depth tests are strictly falling and are already very close to the p-value of the 4-depth tests for $N = 96$. Moreover, the p-values of the 3-depth tests are smaller than those of the 4-depth tests if there are three sign changes as is demonstrated on the right hand side of Figure 4. Hence the 3-depth tests, which are easier to calculate, seem to have a very good power at least for large sample sizes. This is supported by the following applications.

# 4 Applications

The high power of 3-depth tests in the case of two unknown parameters was already shown for explosive AR(1) models, namely in Kustosz et al. (2016a) for linear AR(1)-models given by $Y_n = \theta_0 + \theta_1 Y_{n-1} + E_n$ and in Kustosz et al. (2016b) for nonlinear AR(1)-models given by $Y_n = Y_{n-1} + \theta_1 Y_{n-1}^{\theta_2} + E_n$, see also Falkenau (2016). In particular these results showed for normally distributed errors $E_n$ that 3-depth tests possess similar high power compared to classical tests based on least squares. Here we will provide more applications. At first we show the high power of 3-depth tests for testing of relevant difference in two samples where the unknown parameter vector is again two dimensional.

This is a special situation where the classical sign test is ill-suited since it cannot reject in some of the alternatives. Afterwards, we consider models with three unknown parameters. One model is a quadratic regression and the other models are an explosive AR(2)-model with intercept and a nonlinear AR(1)-model with intercept.

Since for relevance testing only the 3-depth test was studied, we used the quantiles of the asymptotic distribution there. For the other applications, we used the exact distribution for the small sample size $N = 12$ and a simulated distribution for the large sample size $N = 96$ since both 3-depth and 4-depth tests were considered and no asymptotic distribution for the 4-depth test is known up to now. Thereby, the simulated distribution of the both depth tests were obtained with 10 000 repetitions.

In all applications with a three dimensional unknown parameter vector, we used 100 repetitions for each considered alternative. In some cases we used also 500 repetitions. However, since a difference was not visible for us, we speeded up the computation by using only 100 repetitions.

## 4.1 Relevance tests based on 3-depth

Here we consider two samples given by

$$\begin{aligned} Y_n &= \mu_1 + E_n, \quad \text{for} \quad n = 1, \dots, M, \\ Y_n &= \mu_2 + E_n, \quad \text{for} \quad n = M+1, \dots, N, \end{aligned}$$

so that $\theta = (\mu_1, \mu_2)^\top$ is the unknown parameter vector. The ordering of the observations within the two samples is given by the ordering how the observations appeared within the two samples. The hypotheses for relevance testing are given by

$$H_0 : |\mu_1 - \mu_2| \leq \delta, \quad H_1 : |\mu_1 - \mu_2| > \delta.$$

Often we have $M \approx \frac{N}{2}$. Then a sign test will never reject the null hypothesis since we can set $\theta^1 = \left(\frac{\mu_1^0+\mu_2^0}{2}, \frac{\mu_1^0+\mu_2^0}{2}\right)^\top$ for any true $\theta^0 = (\mu_1^0, \mu_2^0)^\top$ so that approximately half of the residuals are positive. Hence a sign test makes no sense here.

If the errors are normally distributed then two classical two-sample t-tests for level $\frac{\alpha}{2}$ can be used as follows

$$\begin{aligned} &\text{reject } H_0 \text{ if} \\ &H_0^1 : \mu_1 - \mu_2 \leq \delta \text{ is rejected by the corresponding one-sided t-test} \\ &\text{or} \\ &H_0^2 : \mu_1 - \mu_2 \geq -\delta \text{ is rejected by the corresponding one-sided t-test.} \end{aligned} \tag{9}$$

Since two t-tests are used, the Bonferroni correction of $\frac{\alpha}{2}$ is necessary. If the variance $\sigma^2$ of the normal distribution is known or a lower bound $\sigma^2$ of it is known, then a more powerful $\alpha$-level test for

$$H_0^\sigma : |\mu_1 - \mu_2| \leq \delta\sigma, \quad H_1 : |\mu_1 - \mu_2| > \delta\sigma$$

is given by

$$\text{reject } H_0^\sigma \text{ if } T > c_\alpha \tag{10}$$

where $T$ is the classical test statistic of the two-sample t-test and $c_\alpha$ satisfies

$$\alpha = 1 - F(c_\alpha) + F(-c_\alpha).$$

Thereby $F$ is the cumulative distribution function of the t-distribution with $N-2$ degrees of freedom and noncentrality parameter $\sqrt{\frac{M(N-M)}{N}}\,\delta$. The deviation is the same as for equivalence tests given by Wellek (2010), pp. 119.

The $K$-depth relevance test is of the form given by (7) with $\Theta^0 = \{(\mu_1, \mu_2)^\top \in \mathbb{R}^2;\ |\mu_1 - \mu_2| \le \delta\}$. Hence the $K$-depth must be calculated for $\{(\mu_1, \mu_2)^\top \in \mathbb{R}^2;\ |\mu_1 - \mu_2| \le \delta\}$. This can be done on a grid given by classical confidence intervals for $\mu_1$ and $\mu_2$. Let $[c_l^i, c_u^i]$ be a $(1 - \frac{\alpha}{L})$-confidence intervals for $\mu_i$, $i = 1, 2$, and set

$$C^i := \{c_l^i, c_l^i + \epsilon, c_l^i + 2\epsilon, \ldots, c_u^i - \epsilon, c_u^i\} \quad \text{with} \quad \epsilon = \frac{c_u^i - c_l^i}{J}$$

for some $L > 0,\ J \ge 10$, where $L$ determines how large the grid is and $J$ determines how fine the grid is. Then we have

$$\{(\mu_1, \mu_2)^\top \in \mathbb{R}^2;\ |\mu_1 - \mu_2| \le \delta\} \approx \{(\mu_1, \mu_2)^\top \in C^1 \times C^2;\ |\mu_1 - \mu_2| \le \delta\} =: C.$$

Hence $d_K$ and $T_K$ must be calculated only for $(\mu_1, \mu_2)^\top \in C$. Thereby it is advantageous to start the iteration in the middle of the intervals $C^i$ and to stop as soon as a parameter $(\mu_1, \mu_2)^\top$ is found so that the null hypothesis is not rejected. In the simulations below, $L = 2$ and $J = 50$ was used. Moreover for the Cauchy distribution, the median instead of the arithmetic mean was used in the confidence intervals. This choice of the middle of the confidence interval should be used as soon as the distribution is unknown.

Figure 5 provides the power function of the t-test given by (10) with $\delta = 1 = \sigma^2$ for $20 = M = N - M$ and errors with standard normal distribution. Thereby a $41 \times 41$ grid with step length 0.1 was used and the number of repetitions per grid point was 100. The border of the nonrelevance set $\Theta^0$ is given by the dotted white lines. It is clear that this is an infinite band where the behavior parallel to the dashed line given by $\mu_1 = -\mu_2$ is the same. Hence to obtain a power comparison, we can restrict ourselves to the set $\{(\mu, -\mu)^\top;\ \mu \in [-a, a]\}$ for $a \in \{2, 8\}$ where the nonrelevance set is given by $\{(\mu, -\mu)^\top;\ \mu \in \left[-\frac{1}{2}, \frac{1}{2}\right]\}$.

The power comparison on this diagonal for normal and Cauchy distribution for $20 = M = N - M$ for the two t-tests and the 3-depth test is shown in Figure 6. Thereby $\delta = \sigma = 1$, the standard normal distribution and the standard Cauchy distribution are used as errors and the power is simulated 4000 times for the normal distribution and 1000 times for the Cauchy distribution for the t-tests. The smaller simulation number for the Cauchy distribution was used since no improvement was visible with a higher number. Since the calculation of the 3-depth test needs more time, the power of the 3-depth test was simulated only 100 times for both distributions. Every power was simulated in steps

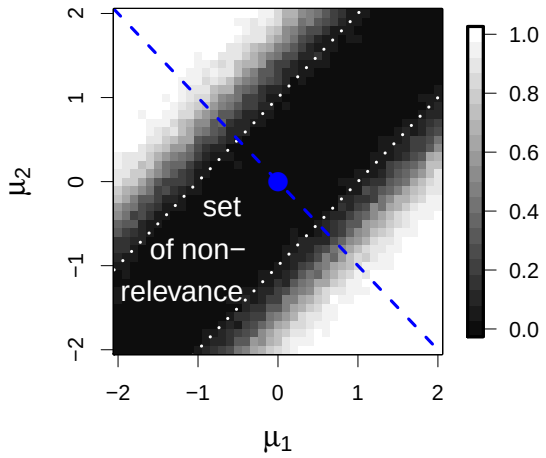



Figure 5: Simulated power of the t-test given by (10) for normally distributed errors for sample sizes $M = N - M = 20$, $\delta = 1 = \sigma^2$, and $\alpha = 0.05$.

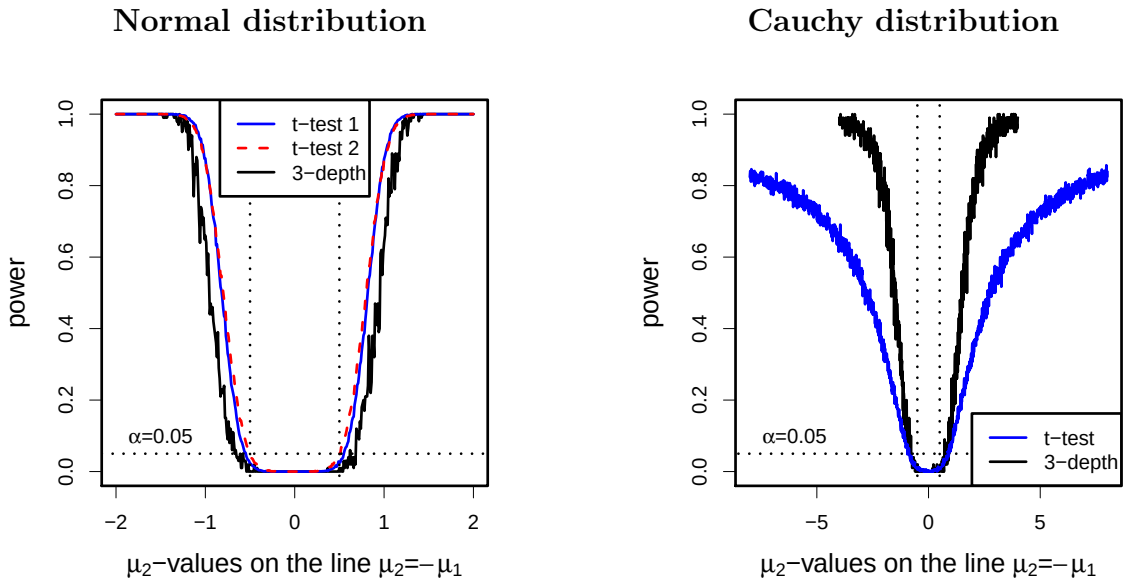



Figure 6: Simulated power of the t-test given by (9) (t-test 1 and t-test, respectively), the t-test given by (10) (t-test 2), and the 3-depth test on the diagonal $\{(\mu, -\mu)^\top;\ \mu \in [-2, 2]\}$ for normally distributed errors (left) and on the diagonal $\{(\mu, -\mu)^\top;\ \mu \in [-8, 8]\}$ for Cauchy distributed errors (right) for sample sizes $M = N - M = 20$. (For a colored version of this figure, the reader is referred to the web version of the article.)

of 0.01 along the diagonal $\{(\mu, -\mu)^\top;\ \mu \in [-a, a]\}$ for $a \in \{2, 8\}$. The left hand side of Figure 6 concerns the normal distribution. Here the power of the 3-depth test is slightly worse than that of the t-tests and the t-test given by (10) (t-test 2) shows the best power.

Since both t-tests are very similar, only the t-test given by (9) is used in the simulation with Cauchy distributed errors shown on the right hand side. Here the power of the 3-depth test is much better than that of the t-test. More results of power comparison of these three tests can be found in Malcherczyk (2018).

## 4.2 Quadratic regression

In the quadratic regression model given by

$$Y_n = \theta_0 + \theta_1\, x_n + \theta_2\, x_n^2 + E_n, \;\; n = 1, \ldots, N, \;\; \theta = (\theta_0, \theta_1, \theta_2)^\top,$$

we consider the problem of testing the null hypothesis $H_0 : \theta = (1, 0, 1)^\top$ with a test with level $\alpha = 0.05$ and samples sizes $N = 12$ and $N = 96$. For each simulation, a $41 \times 41$ grid of alternatives and 100 repetitions for each alternative were used. For $N = 12$, we used $x_1 = -5.5$, $x_2 = -4.5$, $\ldots$ $x_6 = -0.5$, $x_7 = 0.5$, $\ldots$ $x_{12} = 5.5$ as explanatory variables and the exact distribution for the 3-depth and the 4-depth was used to obtain the p-values. For $N = 96$, the explanatory variables were chosen as $x_1 = -5.9375$, $x_2 = -5.875$, $\ldots$ $x_{48} = -0.0625$, $x_{49} = 0.0625$, $\ldots$ $x_{96} = 5.9375$.

Figure 7 shows the simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for the case where $E_n$ has a standard normal distribution and the component $\theta_1$ is fixed to 0. The parameter $\theta_0$ and $\theta_2$ of the null hypothesis is given by the intersection of the two dotted lines. The results for $N = 12$ are shown in the upper part of this figure. Here, the 3-depth test is even worse than the sign test while only the 4-depth test is slightly worse than the F test. Similar to the sign test, the 3-depth test possesses an unbounded area of power less or equal to $\alpha = 0.05$. This is due to the fact that parameter choices in this area often lead to exactly two sign changes which cannot be rejected by the 3-depth test because of the small sample size. More precisely, the maximum 3-depth for two sign changes is $\frac{4^3 \cdot 6}{12 \cdot 11 \cdot 10} = 0.291$ providing a p-value of 0.758 so that a rejection of the null hypothesis is not possible. In this case, the numbers of positive and negative signs are not equal so that the sign test often has a better power for two sign changes than the 3-depth test. Since the 4-depth is zero for two sign changes, the power of the 4-depth test is similar to the F test. However, as indicated by Figure 4, the power of the 3-depth test becomes much better for $N = 96$ which shows the lower part of Figure 7. In particular the maximum 3-depth for two sign changes is now $\frac{32^3 \cdot 6}{96 \cdot 95 \cdot 94} = 0.229$ providing a p-value of 0.014 which is smaller than the significance level $\alpha = 0.05$. A similar result was obtained when the component $\theta_0$ was fixed to 1. However, when the component $\theta_2$ was fixed to 1, then, even for $N = 12$, the power of the 3-depth test is slightly worse than the power of the F test and even better than the power of the 4-depth test. The reason is that two sign changes does not appear in this case. These results are given in the supplementary material.

Figure 8 shows what happens when the normal distribution for the errors is replaced by the Cauchy distribution. Here the component $\theta_2$ is fixed to 1, but similar results were obtained when the other two components were fixed (see the supplementary material). If the errors follow a Cauchy distribution, then the power of the F test becomes very bad while the power functions of the sign test, the 3-depth test and the 4-depth test change

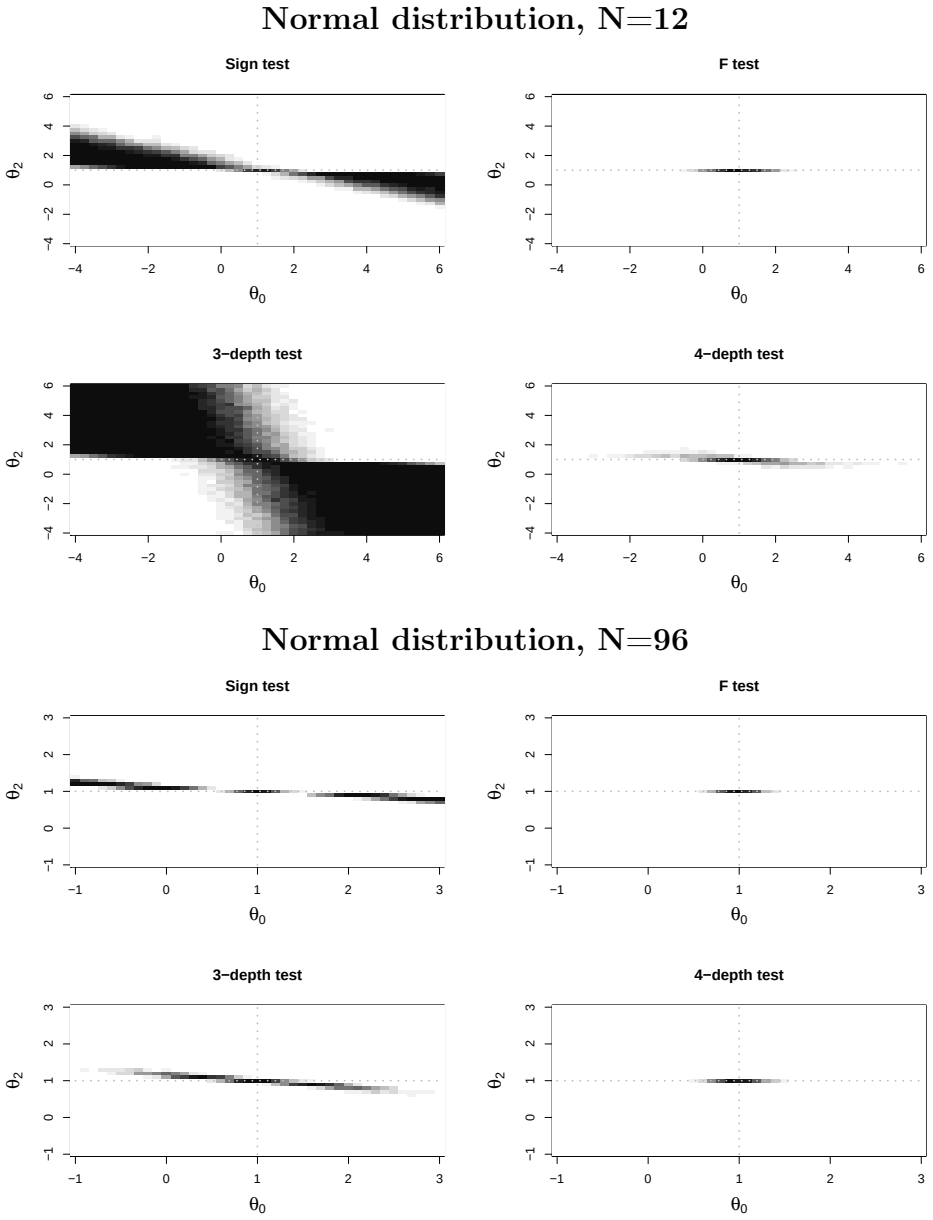



Figure 7: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for normally distributed errors for sample size $N = 12$ (upper part) and $N = 96$ (lower part) where the component $\theta_1$ is fixed to 0 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

only slightly. Hence the 3-depth test and the 4-depth test are much more robust against outliers than the F test.

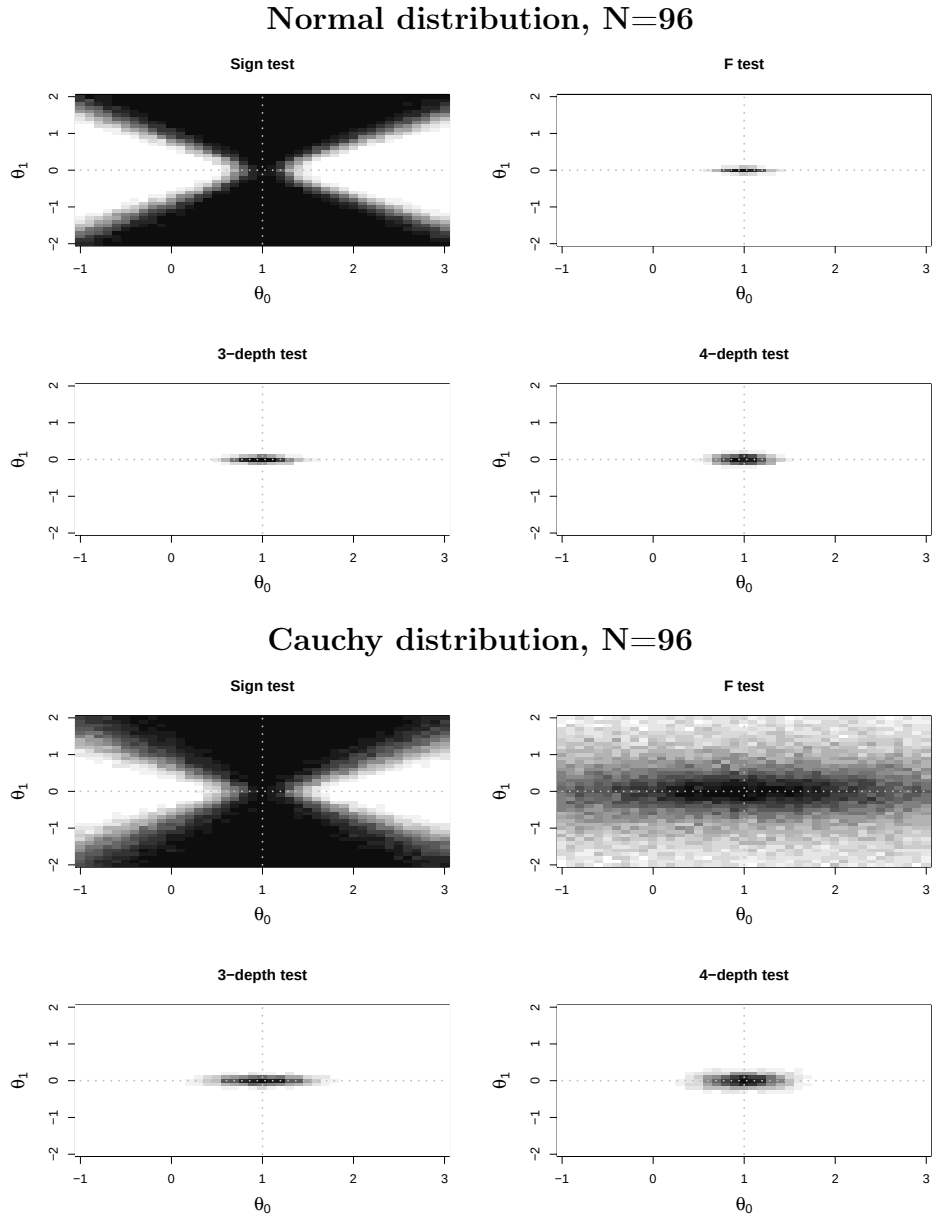



Figure 8: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for errors with normal distribution (upper part) and with Cauchy distribution (lower part) for sample size $N = 96$, where the component $\theta_2$ is fixed to 1 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

## 4.3 AR(2)-model

Here we consider the autoregressive model given by

$$Y_n = \theta_0 + \theta_1 Y_{n-1} + \theta_2 Y_{n-2} + E_n,\ n = 1, \ldots, N,\ \ \theta = (\theta_0, \theta_1, \theta_2)^\top,$$

with $Y_{-1} = Y_0 = 5$. The aim is to test $H_0 : \theta = (0.2, 0.8, 0.21)^\top$ with $\alpha = 0.05$. In particular, we have an explosive process without stationarity under the null hypothesis.

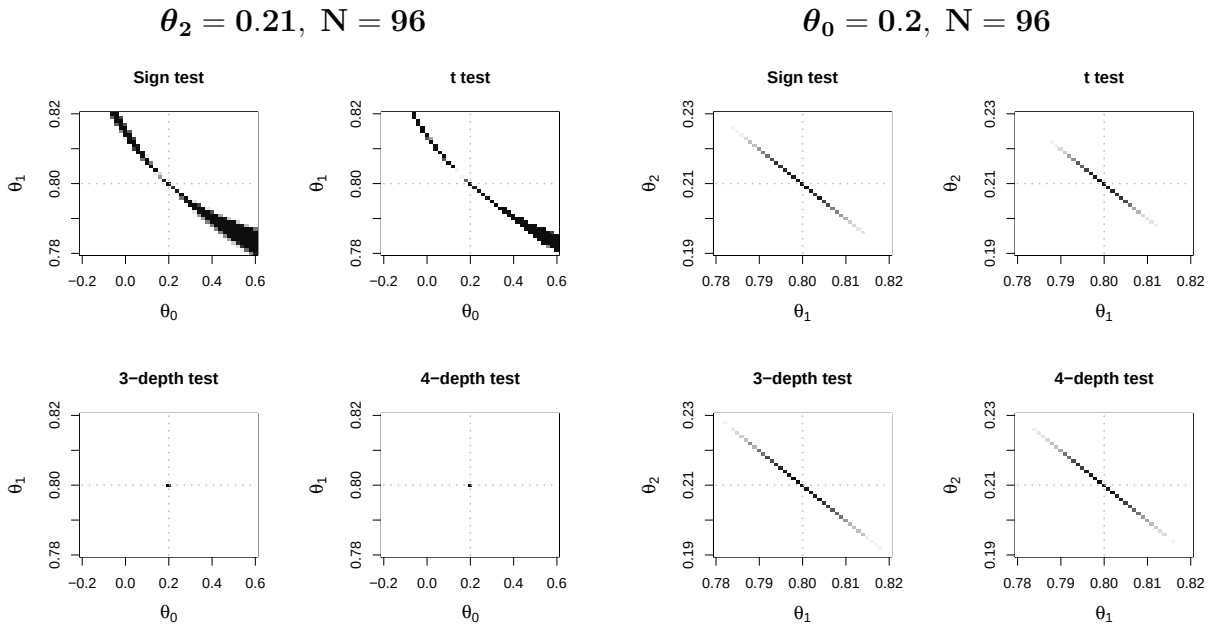



Figure 9: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the AR(2)-model where $\theta_2$ is fixed to 0.21 on the left-hand side and $\theta_0$ is fixed to 0.2 on the right hand side (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

Classical methods are not working for this situation. However, the sign tests based on the residuals $R_n(\theta) = Y_n - \theta_0 - \theta_1 Y_{n-1} - \theta_2 Y_{n-2}$ can be used and for them only the assumption $P(E_n > 0) = P(E_n < 0) = \frac{1}{2}$ is needed. To compare the sign tests with a more classical test, we test with the t-test whether the residuals have mean zero. This t-test is an $\alpha$-level test for $H_0$ under the assumption of normally distributed errors although it might be not very powerful. To give this t-test a chance in the simulations, we used a normal distribution with mean 0 and standard deviation 0.01 for the distribution of the errors $E_n$. A comparison of the sign test, the 3-depth test, the 4-depth test, and this t-test for testing $H_0 : \theta = (0.2, 0.8, 0.21)^\top$ with $N = 96$ is given by the Figure 9. Thereby a $41 \times 41$ grid of alternatives and 100 simulations for each alternative were used. As in Section 4.2, the parameters of the null hypothesis are given by the intersections of the two dotted lines.

The left-hand side of Figure 9 shows the results for the situation where $\theta_2$ is fixed to the value of the null hypothesis, i.e. $\theta_2 = 0.21$. Here the classical sign test and the t-test have a problem since they have an unbounded area with very bad power (the black area). The opposite is true for the 3-depth test and the 4-depth test. The power of the two is only in a small area around the null hypothesis below $\alpha = 0.05$. The same result was obtained when $\theta_1$ is fixed to 0.8, the value of the null hypothesis, see the supplementary material. A different behavior appears when $\theta_0$ is fixed 0.2, the value of the null hypothesis. This behavior is given on the right-hand side of Figure 9. Here all four methods behave similarly and are struggling with an identifiability problem. This identifiability problem disappears for larger values of $\theta_1$ and $\theta_2$. However, then no difference between the methods is visible

anymore.

## 4.4 Nonlinear AR(1)-model

Motivated by crack growth analysis, Kustosz et al. (2016b) and Falkenau (2016) considered already an explosive nonlinear AR(1)-model but without intercept since their method could be used only for a two-dimensional unknown parameter. However, the Euler-Maruyama approximation (Iacus, 2008) applied to the stochastic differential equation given by the deterministic Paris-Erdogan equation (Pook, 2000) for crack growth leads, in its general form, to the following nonlinear autoregressive model with intercept $\theta_0$:

$$Y_n = \theta_0 + Y_{n-1} + \theta_1 Y_{n-1}^{\theta_2} + E_n,\ n = 1, \ldots, N,\ \ \theta = (\theta_0, \theta_1, \theta_2)^\top,$$

see also Kustosz and Müller (2014).

Here we test $H_0 : \theta = (0.01, 0.005, 1.002)^\top$ with $\alpha = 0.05$ and set $Y_0 = 15$ which may be interpreted as an initial crack length. Again, as in Section 4.3, the process is nonstationary so that classical methods cannot be applied. However, the sign tests based on the residuals $R_n(\theta) = Y_n - \theta_0 - Y_{n-1} - \theta_1 Y_{n-1}^{\theta_2}$can be applied and they need only the assumption $P(E_n > 0) = P(E_n < 0) = \frac{1}{2}$. As in Section 4.3, we compare the sign tests with a t-test applied to the residuals. Therefore, we used a normal distribution for the errors $E_n$ with mean 0 and standard deviation 0.01 in the simulations. A comparison of the sign test,

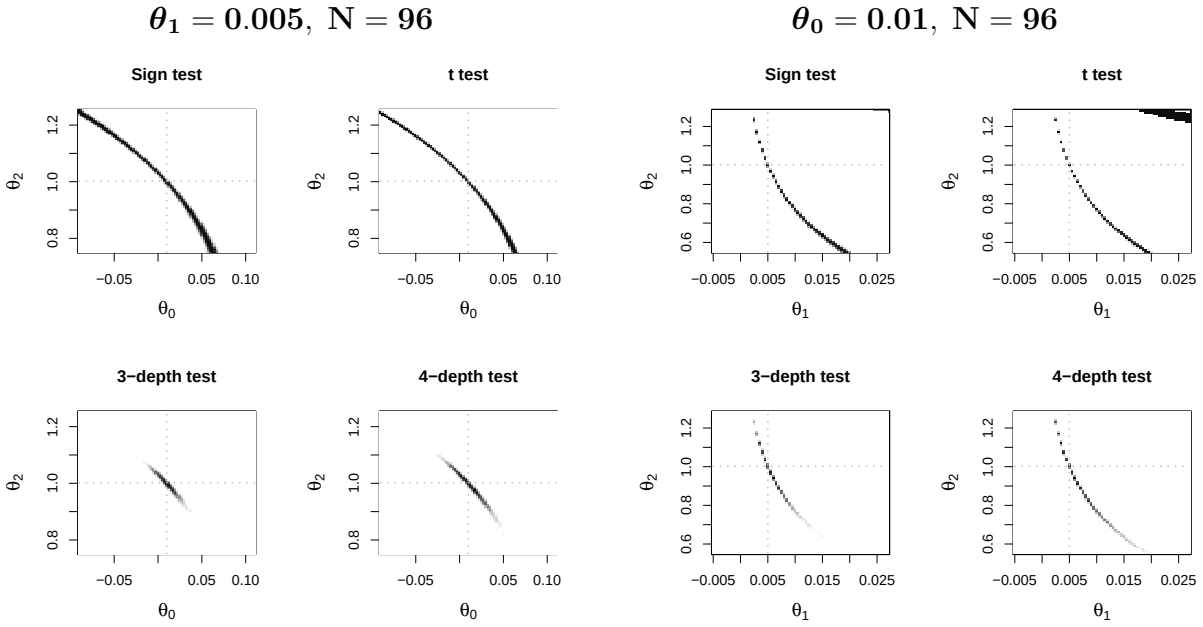



Figure 10: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the AR(2)-model where $\theta_1$ is fixed to 0.005 on the left-hand side and $\theta_0$ is fixed to 0.01 on the right-hand side (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

the 3-depth test, the 4-depth test, and the t-test for $N = 96$ is given by Figure 10. Again, 100 simulations for each alternative were used.

The left-hand side of Figure 10 shows the behavior when $\theta_1$ was fixed to 0.005 and a $81 \times 101$ grid of alternatives in $\theta_0$ and $\theta_2$ was used. Here the classical sign test and the t-test have bad power in an infinite area (the black area). This is not the case for the 3-depth test and the 4-depth test where the 3-depth test shows a better power. A similar result was obtained when $\theta_2$ was fixed to 1.002, see the supplementary material.

In opposite to the results in Section 4.3, the classical sign test and the t-test have a much worse power than the 3-depth test and the 4-test test when $\theta_0$ is fixed to 0.01. In particular, they have an unbounded area of power below 0.05 which is not the case for the depth tests. This is shown on the right-hand side of Figure 10 where a $65 \times 74$ grid of alternatives in $\theta_1$ and $\theta_2$ was used. This result is very similar to results for the classical sign test and the 3-depth tests provided in Kustosz et al. (2016b) with $\theta_0 = 0$. Here we see again that the 3-depth test has a better power than the 4-depth tests. The right upper corner of the results for the t-test shows also a specific problem of the t-test: because of the explosion of the process, it can happen that the test statistic of the t-test gets numerical problems.

# 5 Discussion and outlook

We introduced $K$-sign depth, shortly denoted by $K$-depth, and proposed $K$-depth tests based on the $K$-depth. We show that the $K$-depth test with $K = 2$ is equivalent to the classical sign depth applied to residuals. After a comparison via maximum depth, depth at few signs changes, and p-values at few sign changes in the residuals, we provided a comparison in four simulation studies including relevance testing, quadratic regression, explosive AR(2)- and nonlinear AR(1)-models. The example of relevance testing demonstrates how quite general problems can be treated with the generalized sign tests, even in situations where the classical sign test cannot be applied. The other three applications demonstrate the behavior for a three-dimensional unknown parameter vector where the classical sign tests can be applied. As expected from the theoretical considerations concerning the behavior at few sign changes, the classical sign test yields bad results for nonfits leading to at least one sign change in the residuals. We observed similar problems for the 3-depth test if two sign changes can appear and the sample size is small. However, these problems disappear for larger sample sizes. For large samples it seems like there is no advantage in choosing the 4-depth instead of the 3-depth. However, it is not clear whether this is always the case so that further research is necessary.

For small samples sizes, we used the exact distribution of the $K$-depth. For larger sample sizes, the distribution was determined via simulation since the asymptotic distribution of $K$-depth is only known for $K = 2$ and $K = 3$ up to now. Note that we are currently working on limit laws for $K > 3$. Our research indicates that an implementation of the $K$-depth is possible which is faster than the $O(N^K)$ implementation we have used here. Thereby, Lemma 2.1 plays an important role.

Another drawback in the usage of $K$-depth tests is that a natural ordering of the residuals is necessary. Various results regarding the ordering of residuals are given in Horn and Müller (2019).

# Acknowledgments

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

# Appendix

**Proof of Lemma 2.1**
In order to simplify the notation, we assume $(n_1, \dots, n_K) = (1, \dots, K)$. Note for $x \neq 0$

$$\mathbb{1}\{x > 0\} = \frac{1}{2}\left(\Phi(x) + 1\right), \qquad \mathbb{1}\{x < 0\} = \frac{1}{2}\left(-\Phi(x) + 1\right).$$

It is straight forward to check $\prod_{i=1}^{K}(a_i + 1) = \sum_{\ell=1}^{K} \sum_{1 \le i(1) < \ldots < i(\ell) \le K} \prod_{j=1}^{\ell} a_{i(j)} + 1$ for arbitrary $a_1, \dots, a_K$. Hence this implies almost surely

$$\begin{aligned}\prod_{k=1}^{K} \mathbb{1}\{E_k(-1)^k > 0\} &= \frac{1}{2^K} \prod_{k=1}^{K} \left((-1)^k \Phi(E_k) + 1\right) \\ &= \frac{1}{2^K} \left( \sum_{\ell=1}^{K} \sum_{1 \le i(1) < \ldots < i(\ell) \le K} (-1)^{i(1)+\cdots+i(\ell)} \prod_{j=1}^{\ell} \Phi(E_{i(j)}) + 1 \right)\end{aligned}$$

Similarly

$$\begin{aligned}\prod_{k=1}^{K} \mathbb{1}\{E_k(-1)^k < 0\} &= \frac{1}{2^K} \left( \sum_{\ell=1}^{K} \sum_{1 \le i(1) < \ldots < i(\ell) \le K} (-1)^{i(1)+\cdots+i(\ell)+\ell} \prod_{j=1}^{\ell} \Phi(E_{i(j)}) + 1 \right) \\ &= \frac{1}{2^K} \left( \sum_{\substack{\ell=1,\ldots,K \\ \ell \text{ even}}} \sum_{1 \le i(1) < \ldots < i(\ell) \le K} (-1)^{i(1)+\cdots+i(\ell)} \prod_{j=1}^{\ell} \Phi(E_{i(j)}) + 1 \right) \\ &\quad - \frac{1}{2^K} \sum_{\substack{\ell=1,\ldots,K \\ \ell \text{ odd}}} \sum_{1 \le i(1) < \ldots < i(\ell) \le K} (-1)^{i(1)+\cdots+i(\ell)} \prod_{j=1}^{\ell} \Phi(E_{i(j)})\end{aligned}$$

Therefore

$$\prod_{k=1}^{K} \mathbb{1}\{E_k(-1)^k > 0\} + \prod_{k=1}^{K} \mathbb{1}\{E_k(-1)^k < 0\}$$

$$=\frac{1}{2^{K-1}}\left(\sum_{\substack{\ell=1,\dots,K\\ \ell \text{ even}}}\sum_{1\le i(1)<\dots<i(\ell)\le K}(-1)^{i(1)+\dots+i(\ell)}\prod_{j=1}^{\ell}\Phi(E_{i(j)})+1\right)$$

$$=\left(\frac{1}{2}\right)^{K-1}+\frac{1}{2^{K-1}}\sum_{L=1}^{\lfloor\frac{K}{2}\rfloor}\sum_{1\le i(1)<\dots<i(2L)\le K}(-1)^{i(1)+\dots+i(2L)}\prod_{j=1}^{2L}\Phi(E_{i(j)})$$

and the assertion follows.

**Proof of Theorem 2.2**
Set $R_n = R_n(\theta)$. According to Lemma 2.1, it holds

$$d_K(R_1,\dots,R_N)-\left(\frac{1}{2}\right)^{K-1}$$

$$=\frac{1}{\binom{N}{K}}\sum_{1\le n_1<n_2<\dots<n_K\le N}\frac{1}{2^{K-1}}\sum_{L=1}^{\lfloor\frac{K}{2}\rfloor}\sum_{1\le i(1)<\dots<i(2L)\le K}(-1)^{i(1)+\dots+i(2L)}\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})$$

with $\Phi(x) := \mathbb{1}\{x>0\}-\mathbb{1}\{x<0\}$. Set

$$v := \sum_{L=1}^{\lfloor\frac{K}{2}\rfloor}\sum_{1\le i(1)<\dots<i(2L)\le K}1$$

for the number of summands in the representation of $K$ alternating signs given by Lemma 2.1. This number depends only on $K$ and not on $N$. First of all, we show that each of these $v$ summands is converging in probability to zero.

To this end, let $L=1,\dots,\lfloor\frac{K}{2}\rfloor$ and $1\le i(1)<\dots<i(2L)\le K$ be arbitrary. We consider the summand multiplied by the factor $2^{K-1}$.

Because $\mathrm{E}_\theta(\Phi(R_n))=0$ and $R_1,\dots,R_N$ are independent, we get at once for this summand

$$\mathrm{E}_\theta\left(\frac{1}{\binom{N}{K}}\sum_{1\le n_1<n_2<\dots<n_K\le N}\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})\right)=0.$$

Moreover, $\Phi(R_n)^2=1$ almost surely implies

$$\mathrm{E}_\theta\left(\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})\prod_{j=1}^{2L}\Phi(R_{\tilde{n}_{i(j)}})\right)$$
$$=\begin{cases}1, & \text{if } n_{i(j)}=\tilde{n}_{i(j)} \text{ for } j=1,\dots,2L,\\ 0, & \text{else.}\end{cases}$$

Then

$$\mathrm{var}_\theta\left(\frac{1}{\binom{N}{K}}\sum_{1\le n_1<\dots<n_K\le N}(-1)^{i(1)+\dots+i(2L)}\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})\right)$$

$$
\begin{aligned}
&= \mathrm{E}_\theta\left(\left(\frac{1}{\binom{N}{K}}\sum_{1\le n_1<\ldots<n_K\le N}(-1)^{i(1)+\ldots+i(2L)}\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})\right)^2\right)\\
&= \frac{1}{\binom{N}{K}^2}\sum_{1\le n_1<\ldots<n_K\le N}\ \sum_{1\le \tilde{n}_1<\ldots<\tilde{n}_K\le N}\mathrm{E}_\theta\left(\prod_{j=1}^{2L}\Phi(R_{n_{i(j)}})\prod_{j=1}^{2L}\Phi(R_{\tilde{n}_{i(j)}})\right)\\
&= \frac{1}{\binom{N}{K}^2}\sum_{\substack{1\le n_1<\ldots<n_K\le N,\ 1\le \tilde{n}_1<\ldots<\tilde{n}_K\le N\\ n_{i(j)}=\tilde{n}_{i(j)}\text{ for }j=1,\ldots,2L}}1\\
&\le \frac{1}{\binom{N}{K}^2}\sum_{1\le n_1<\ldots<n_{2L}\le N}\ \sum_{n_{2L+1},\ldots,n_K\in\{1,\ldots,N\}}\ \sum_{\tilde{n}_{2L+1},\ldots,\tilde{n}_K\in\{1,\ldots,N\}}1\\
&= \frac{\binom{N}{2L}N^{K-2L}N^{K-2L}}{\binom{N}{K}^2}\le\frac{(K!)^2}{(2L)!}\frac{N^{2L+2K-4L}}{(N-(K+1))^{2K}}=\frac{(K!)^2}{(2L)!}\ \frac{1}{N^{2L}}\ \frac{1}{\left(1-\frac{K+1}{N}\right)^{2K}}\longrightarrow 0
\end{aligned}
$$

for $N\to\infty$ so that Chebyshev inequality provides the convergence in probability to zero. Furthermore, the convergence in probability is sufficiently quick of order $O(N^{-2L})$ so that the Borel-Cantelli lemma implies the convergence to zero almost surely.

**Proof of Lemma 2.3**
Lemma 2.1 yields for $K=2$ using (5)

$$
\begin{aligned}
T_2(\theta) &= N\left(d_2(R_1,\ldots,R_N)-\frac{1}{2}\right)\\
&= \frac{N}{\binom{N}{2}}\left(\sum_{1\le n_1<n_2\le N}\left(\mathbb{1}\{R_{n_1}>0,R_{n_2}<0\}+\mathbb{1}\{R_{n_1}<0,R_{n_2}>0\}-\frac{1}{2}\right)\right)\\
&= \frac{N}{\binom{N}{2}}\sum_{1\le n_1<n_2\le N}\left(-\frac{1}{2}\Phi(R_{n_1})\Phi(R_{n_2})\right)=\frac{N}{4\binom{N}{2}}\sum_{1\le n_1\ne n_2\le N}(-\Phi(R_{n_1})\Phi(R_{n_2}))\\
&= -\frac{N}{2N(N-1)}\left(\sum_{n_1=1}^N\sum_{n_2=1}^N\Phi(R_{n_1})\Phi(R_{n_2})-\sum_{n=1}^N\Phi(R_n)^2\right)\\
&= -\frac{N}{2(N-1)}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^N\Phi(R_n)\right)^2+\frac{1}{2(N-1)}\sum_{n=1}^N\Phi(R_n)^2\\
&= -\frac{N}{2(N-1)}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^N\Phi(R_n)\right)^2+\frac{N}{2(N-1)}=\frac{N}{2(N-1)}-\frac{N}{2(N-1)}T_{\mathrm{sign}}(\theta)^2.
\end{aligned}
$$

**Proof of Theorem 3.1**
Set

$$
d_K^0(r_1,\ldots,r_N):=\binom{N}{K}d_K(r_1,\ldots,r_N)
$$

for $K\ge 2$ to simplify the notation and assume without loss of generality $r_1>0$ and $r_2<0$. Then note at first the following recursion for $K\ge 3$

$d_K^0(r_1,\ldots,r_N)$

$$
\begin{aligned}
&= \sum_{1\le n_1<n_2<\ldots<n_K\le N} \left( \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \right) \\
&\overset{(r_1>0,\, r_2<0)}{=} \sum_{2\le n_2<n_3<\ldots<n_K\le N} \mathbb{1}\{r_1 > 0\} \prod_{k=2}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \\
&\quad + \sum_{2<n_2<n_3<\ldots<n_K\le N} \mathbb{1}\{r_2 < 0\} \prod_{k=2}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} \\
&\quad + \sum_{2<n_1<n_2<\ldots<n_K\le N} \left( \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \right) \\
&\overset{(r_2\le 0)}{=} \sum_{2\le n_2<n_3<\ldots<n_K\le N} \left( \prod_{k=2}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} + \prod_{k=2}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} \right) \\
&\quad + \quad d_K^0(r_3, \ldots, r_N).
\end{aligned}
$$

Hence it holds

$$
d_K^0(r_1, \ldots, r_N) = d_{K-1}^0(r_2, \ldots, r_N) + d_K^0(r_3, \ldots, r_N). \tag{11}
$$

For arbitrary integers $N, K$ with $2 \le K \le N$ let

$$
f(N, K) := \binom{N}{K} d_K(r_1, \ldots, r_N).
$$

First note that

$$
f(K, K) = 1 \quad \text{and} \quad f(K+1, K) = 2 \quad \text{for every } K \ge 2. \tag{12}
$$

Moreover, it is not hard to check that

$$
f(N, 2) = \begin{cases} \left(\frac{N}{2}\right)^2, & \text{if } N \text{ is even,} \\ \frac{(N-1)(N+1)}{4}, & \text{if } N \text{ is odd,} \end{cases} \tag{13}
$$

since a pair $r_i, r_j$ has alternating signs if and only if either $i$ is even and $j$ is odd or vice versa. Since there are exactly $\lfloor N/2 \rfloor$ even and $\lceil N/2 \rceil$ odd indices to choose from, one obtains (13).

Furthermore, note that (11) implies

$$
f(N, K) = f(N-1, K-1) + f(N-2, K). \tag{14}
$$

Since (12), (13) and (14) already uniquely determine $f$, it only remains to find a function satisfying these equations. To this end, recall that the binomial coefficients satisfy

$$
\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \quad \text{for all } 1 \le k < n. \tag{15}
$$

With this property it is not hard to check that the following definition yields a function that satisfies (12), (13) and (14):

$$f(N,K) = \begin{cases} 2\binom{(N+K-1)/2}{K}, & \text{if } N+K \text{ is odd,} \\ \binom{(N+K)/2}{K} + \binom{(N+K-2)/2}{K}, & \text{if } N+K \text{ is even.} \end{cases}$$

The details of checking (12), (13) and (14) by induction are left to the reader. Finally, the assertion follows since $d_K(r_1,\ldots,r_N) = f(N,K)/\binom{N}{K}$ by definition.

**Proof of Corollary 3.1**
If $N = 2LM$ with $L = 2,3,\ldots$ and the residuals $r_1,\ldots,r_N$ are alternating in blocks of size $M$ then Theorem 3.1 provides for odd $K$

$$\begin{aligned} \lim_{N\to\infty} d_K(r_1,\ldots,r_N) &= \lim_{L\to\infty} \frac{1}{\binom{2LM}{K}}\, M^K\, 2 \binom{(2L+K-1)/2}{K} \\ &= \lim_{L\to\infty} \frac{2\,M^K\,(L+\frac{K-1}{2})\,(L+\frac{K-1}{2}-1)\,\ldots\,(L+\frac{K-1}{2}-K+1)}{2LM\,(2LM-1)\,\ldots\,(2LM-K+1)} \\ &= \lim_{L\to\infty} \frac{2\,M^K\,(L+\frac{K-1}{2})\,(L+\frac{K-1}{2}-1)\,\ldots\,(L+\frac{K-1}{2}-K+1)}{(2M)^K\,L\,(L-\frac{1}{2M})\,\ldots\,(L-\frac{K-1}{2M})} = \left(\frac{1}{2}\right)^{K-1}. \end{aligned}$$

The result follows similarly for even $K$.

# Supplementary material

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

# References


Agostinelli, C. and Romanazzi, M. (2011). Local depth. *Journal of Statistical Planning and Inference*, 141:817–830.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109:411–423.

Denecke, L. and Müller, C. H. (2011). Robust estimators and tests for copulas based on likelihood depth. *Computational Statistics and Data Analysis*, 55:2724–2738.

Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statistics and Probability Letters*, 14:119–128.

Falkenau, C. P. (2016). *Depth Based Estimators and Tests for Autoregressive Processes with Application on Crack Growth and Oil Prices.* Dissertation, TU Dortmund.

Horn, M. and Müller, C. H. (2019). Tests based on sign depth for multiple regression. *In preparation.*

Iacus, S. (2008). *Simulation and Inference for Stochastic Differential Equations*. Springer, New York.

Kustosz, C. P., Leucht, A., and Müller, C. H. (2016a). Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis*, 37:763–784.

Kustosz, C. P. and Müller, C. H. (2014). Analysis of crack growth with robust, distribution-free estimators and tests for non-stationary autoregressive processes. *Statistical Papers*, 55(1):125–140.

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

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

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

Lok, W. S. and Lee, S. M. S. (2011). A new statistical depth function with application to multimodal data. *Journal of Nonparametric Statistics*, 23:617–631.

López-Pintado, S. and Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

López-Pintado, S., Sun, Y., Lin, J. K., and Genton, M. G. (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8(3):321–338.

Malcherczyk, D. A. (2018). Vergleich von Zwei-Stichproben-Relevanz-Tests basierend auf t-Tests und Datentiefen. *Bachelor Thesis, TU Dortmund*, https://www.statistik.tu-dortmund.de/1255.html.

Mizera, I. (2002). On depth and deep points: A calculus. *The Annals of Statistics*, 30(6):1681–1736.

Mizera, I. and Müller, C. H. (2004). Location-scale depth (with discussion). *Journal of the American Statistical Association*, 99:949–966.

Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach.* Lecture Notes in Statistics, 165, Springer, New York.

Müller, C. H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis*, 95(1):153–181.

Nagy, S. and Ferraty, F. (2018). Data depth for measurable noisy random functions. *Journal of Multivariate Analysis*, doi.org/10.1016/j.jmva.2018.11.003.

Paindaveine, D. and van Bever, G. (2013). From depth to local depth: A focus on centrality. *Journal of the American Statistical Association*, 108:1105–1119.

Paindaveine, D. and Van Bever, G. (2018). Halfspace depths for scatter, concentration and shape matrices. *The Annals of Statistics*, 46(6B):3276–3307.

Pook, L. (2000). *Linear Elastic Fracture Mechanics for Engineers: Theory and Application*. WIT Press, Southampton.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94(446):388–402.

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

Wang, J. (2019). Asymptotics of generalized depth-based spread processes and applications. *Journal of Multivariate Analysis*, 169:363–380.

Wellek, S. (2010). *Testing Statistical Hypothesis of Equivalence and Noninferiority*. 2. Edition, Chapman and Hall/CRC, Heidelberg.

Wellmann, R., Harmand, P., and Müller, C. H. (2009). Distribution-free tests for polynomial regression based on simplicial depth. *Journal of Multivariate Analysis*, 100(4):622 – 635.

Wellmann, R. and Müller, C. H. (2010a). Depth notions for orthogonal regression. *Journal of Multivariate Analysis*, 101(10):2358 – 2371.

Wellmann, R. and Müller, C. H. (2010b). Tests for multiple regression based on simplicial depth. *Journal of Multivariate Analysis*, 101(4):824 – 838.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28:461–482.