Powerful generalized sign tests based on sign depth
Supplementary results

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

November 23, 2019

1 Further Simulation Results

We complement here the simulation results for the quadratic regression in Subsection 1.1. Additionally, we provide simulation results for two other models with three unknown parameters: for the explosive nonlinear AR(1) model in Subsection 1.2 and for a linear AR(2) model in Subsection 1.3.

All simulations were done with 100 repetitions since there were no visible different when 500 repetitions were used. The tests were performed at level $\alpha = 0.05$ which mean that they reject the null hypotheses if the p-value is smaller than $\alpha = 0.05$. For getting the p-values for the 3-depth test and the 4-depth test for $N \geq 24$, 3-depth and 4-depth were simulated 10 000 times. For $N = 12$, the exact distribution was used.

1.1 Quadratic regression

In the quadratic regression model given by

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

we consider the problem of testing the null hypothesis $H_0 : \theta = (1, 0, 1)^\top$.

Figures 1 and 2 show the simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for $N = 12$ and $x_1 = -5.5$, $x_2 = -4.5$, ... $x_6 = -0.5$, $x_7 = 0.5$, ... $x_{12} = 5.5$ where E_n has a standard normal distribution. For each simulation, a 41×41 grid of alternatives was used. The parameter of the null hypothesis is given by the intersection of the two dotted lines.

*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

In Figure 1, where the the component θ_2 was fixed to 1, the power of the 3-depth test is slightly worse than the power of the F test and better than the power of the 4-depth test. However, the power of the 3-depth tests is much worse in Figure 2, where the component θ_1 was fixed to 0 in the upper part and the component θ_0 was fixed to 1 in the lower part. In both cases, the 3-depth test is even worse than the sign test while only the 4-depth test is slightly worse than the F test. The 3-depth test possesses as the sign test an unbounded area of power less or equal $\alpha = 0.05$. This is due to the fact that in these cases often two sign changes appear so that the 3-depth test cannot reject the null hypothesis because of the small sample size. In particular, 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 shows 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 the power of the 3-depth test becomes much better for $N = 96$. For this sample size, $x_n = -6 + (2n-1)/16$ for $n = 1, \ldots, 96$ was used as design points. Again, a 41×41 grid of alternatives was used for each simulation. Figures 3, 4, and 5 show the results for the cases where at first component θ_2 was fixed to 1 (Figure 3), then component θ_1 was fixed to 0 (Figure 4), and at last component θ_0 was fixed to 1 (Figure 5). Each of the three figures provides the results for errors with standard normal distribution in the upper part and the results for errors with standard Cauchy distribution in the lower part.

For the normal distribution, the area of small power of the 3-depth test is now bounded. Only in the case where θ_1 is fixed to zero, this area is much larger than the area of small power of the F test. But in the two other cases, the 3-depth test behaves similar to the F test. The 4-depth test behaves in all three cases similar to the F test. The reason that the 3-depth test behaves now more similar to the 4-depth test is given by Figure 7. In particular the maximum depth for two sign changes is $\frac{32^3 \cdot 6}{96 \cdot 95 \cdot 94} = 0.229$ providing a p-value of 0.014 which is smaller than the significance level $\alpha = 0.05$.

If the errors follows 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 are only slightly changed. Although Figure 4 may indicate that the area of small power of the 3-depth test and the F test is unbounded, this is not correct. This is only caused by the zoom. If the plot area of Figures 1 and 2 is used then it is clear that the area of small power of both tests is bounded as for the normal distribution. This shows Figure 6 which also provides that the area of small power is much larger for the F test than for the 3-depth test. Hence the 3-depth test and the 4-depth test are much more robust against outliers than the F test.

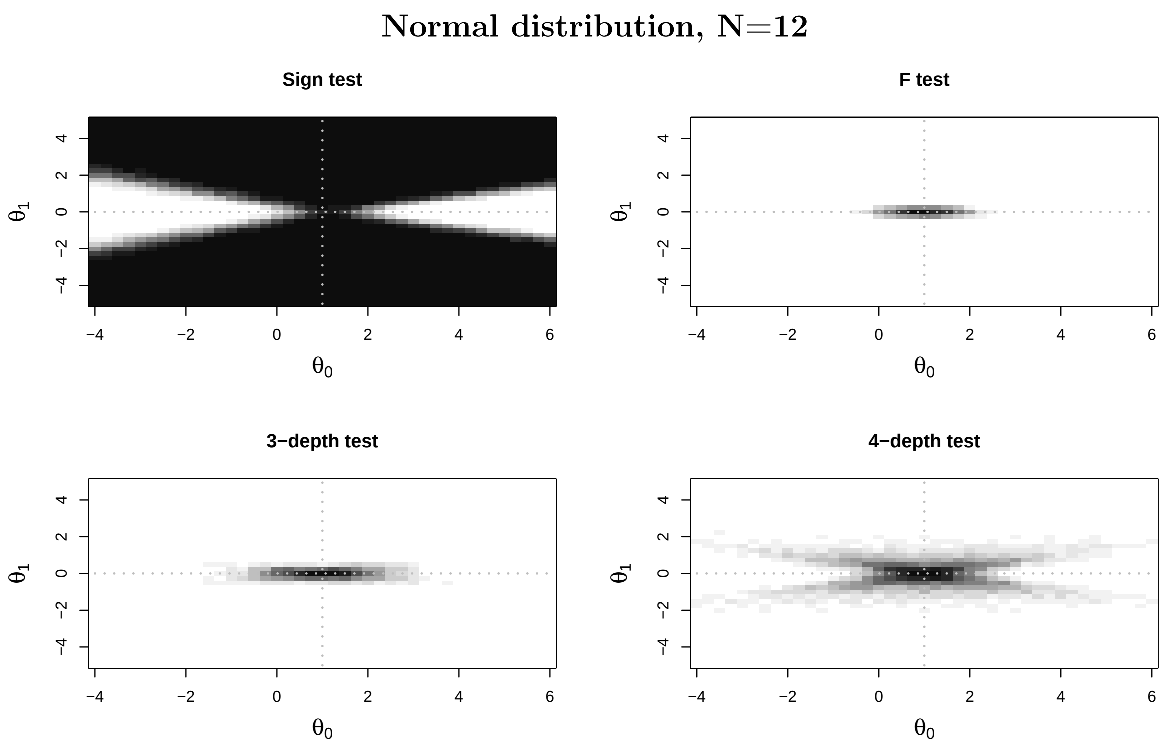

Figure 1: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for normally distributed errors for sample size $N = 12$, where component θ_2 is fixed to 1 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

Normal distribution, N=12

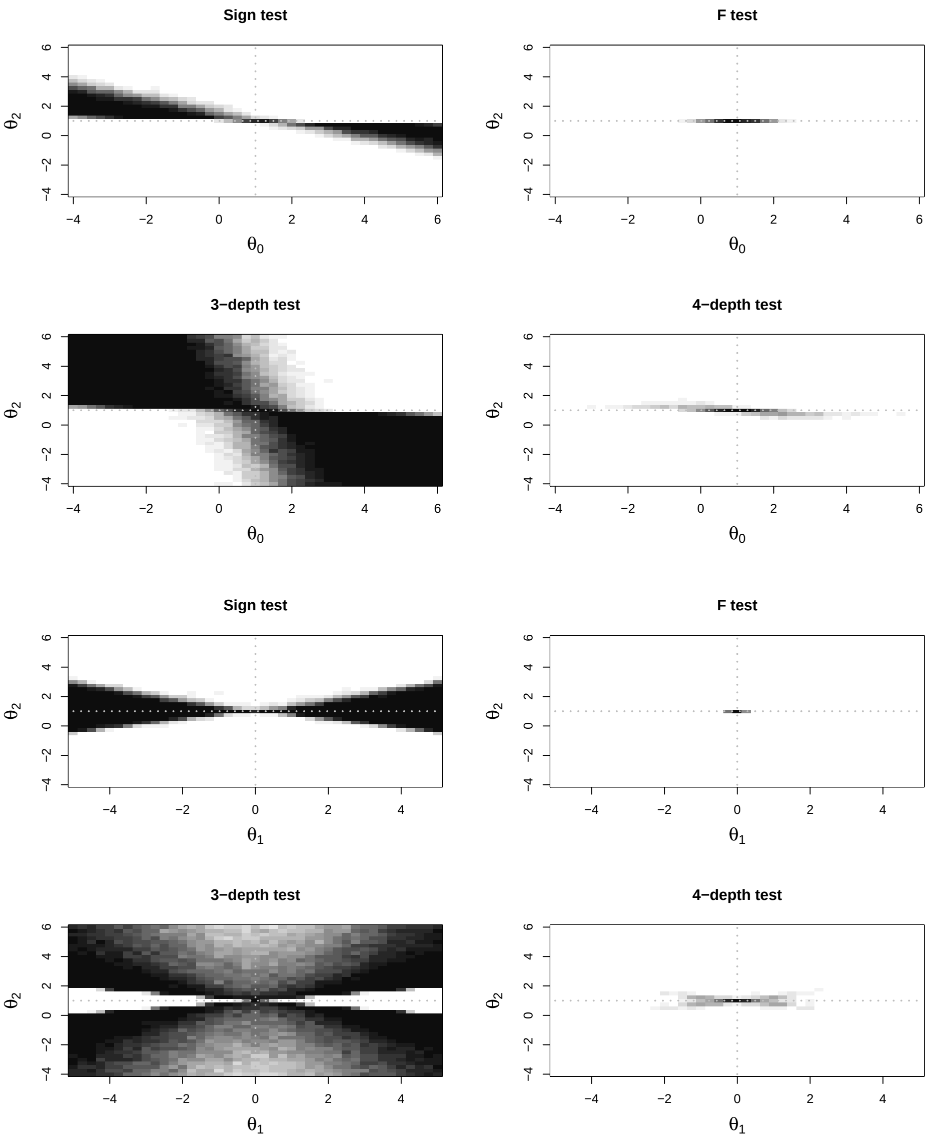

Figure 2: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for normally distributed errors for sample size $N = 12$, where the component θ_1 is fixed to 0 in the upper part and the the component θ_0 is fixed to 1 in the lower part (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

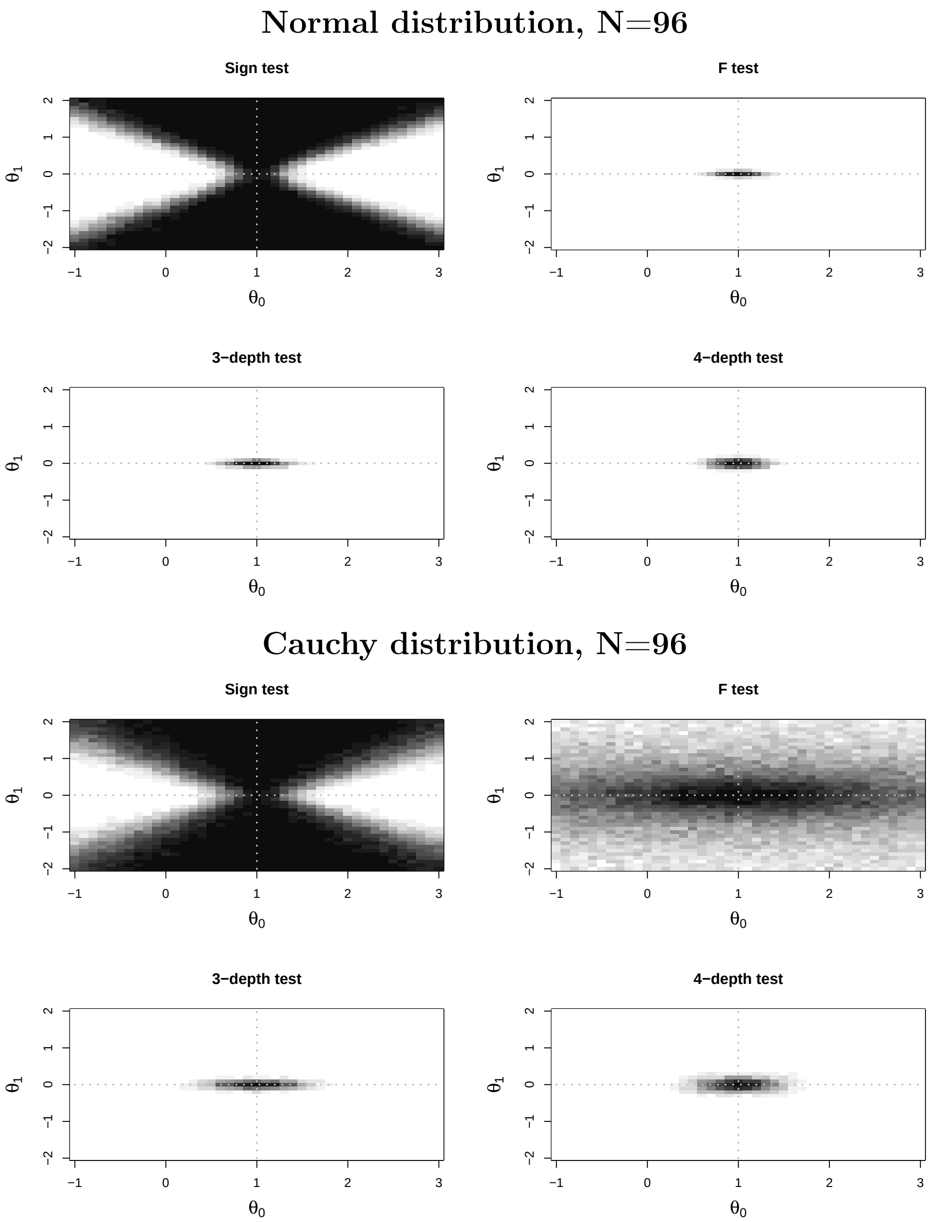

Figure 3: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for errors with normal distribution (upper part) and with Cauchy distribution (lower part) for sample size $N = 96$, where the component θ_2 is fixed to 1 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

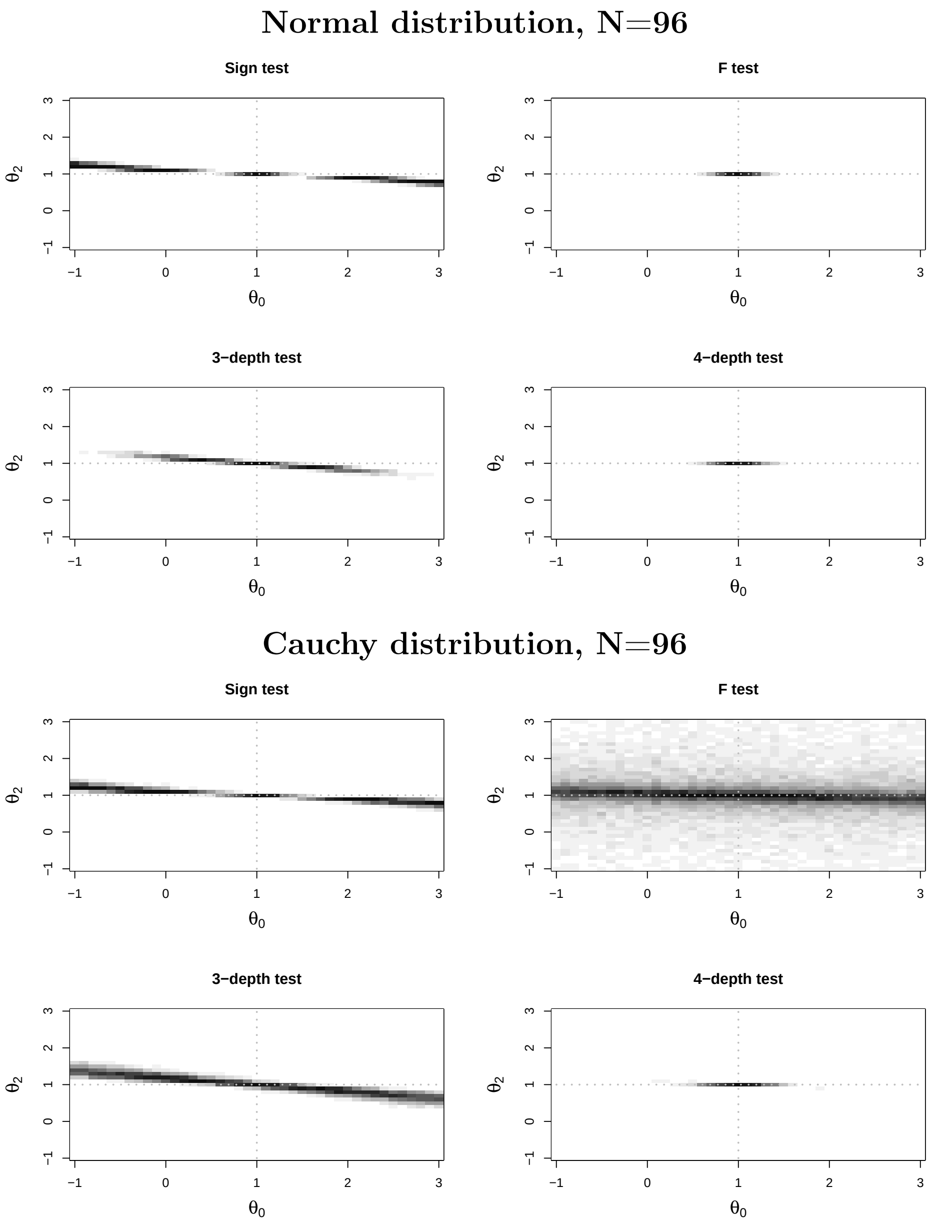

Figure 4: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for errors with normal distribution (upper part) and with Cauchy distribution (lower part) for sample size $N = 96$, where the component θ_1 is fixed to 0 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

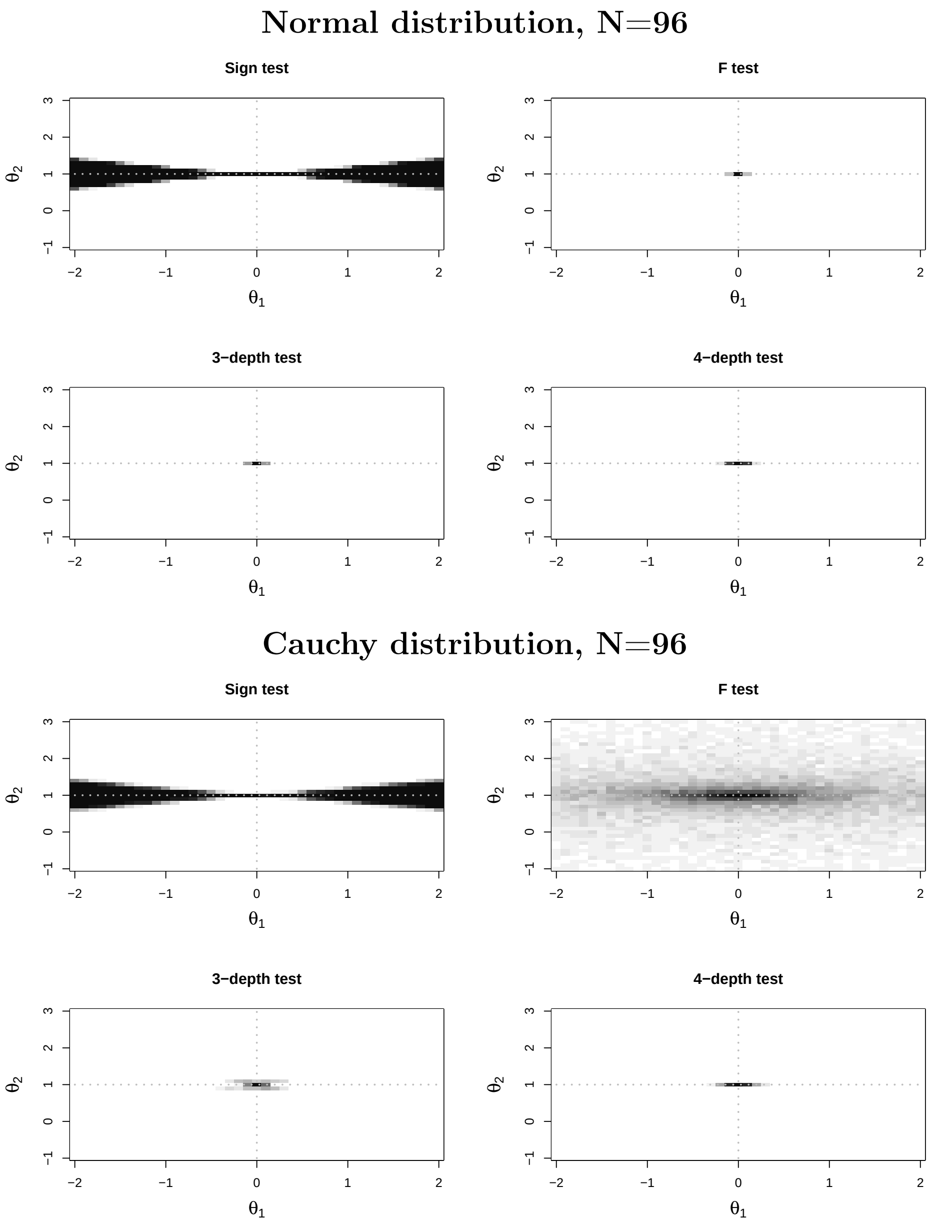

Figure 5: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for errors with normal distribution (upper part) and with Cauchy distribution (lower part) for sample size $N = 96$, where the component θ_0 is fixed to 1 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

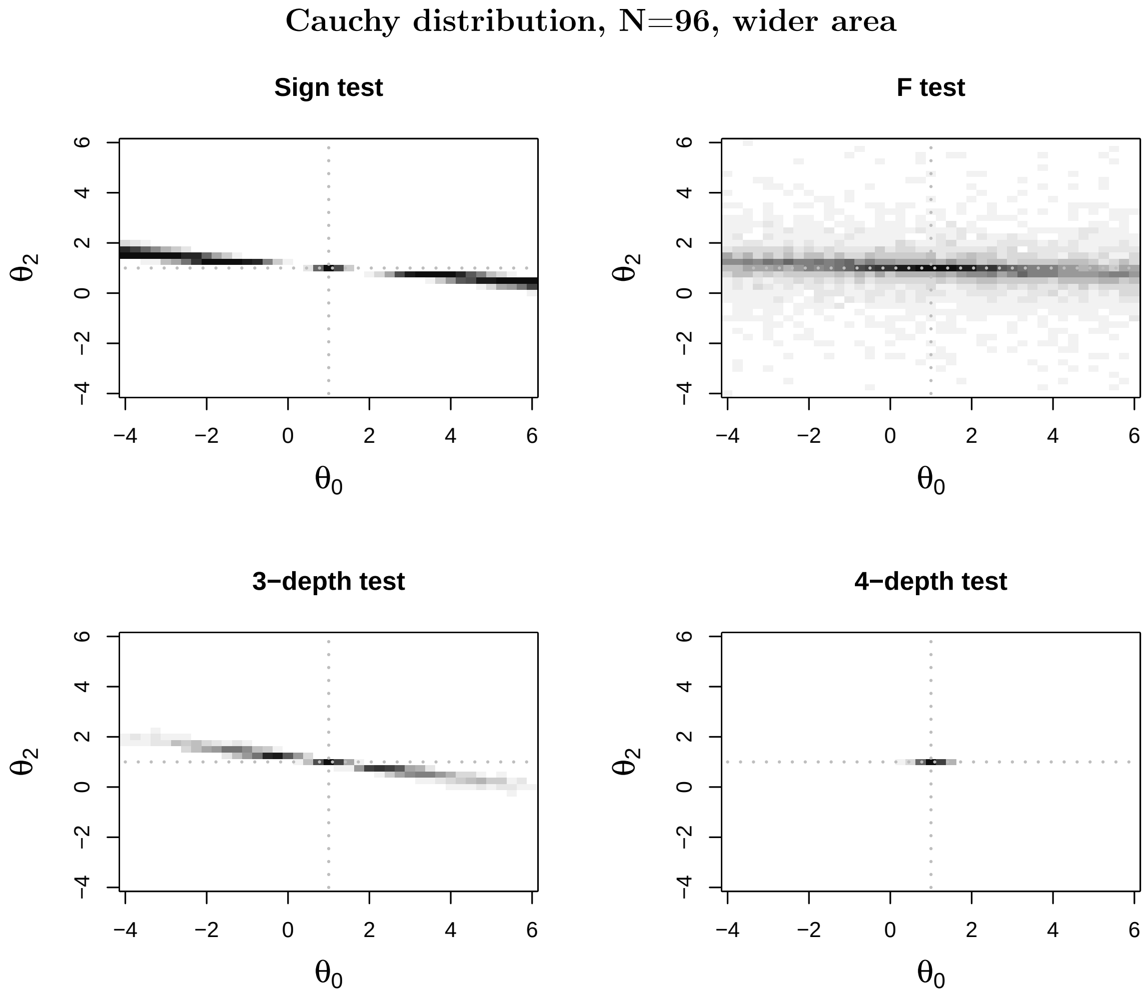

Figure 6: Simulated power of the sign test, the F test, the 3-depth test, and the 4-depth test for errors with Cauchy distribution for sample size $N = 96$ in a wider area of alternatives, where the component θ_1 is fixed to 0 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

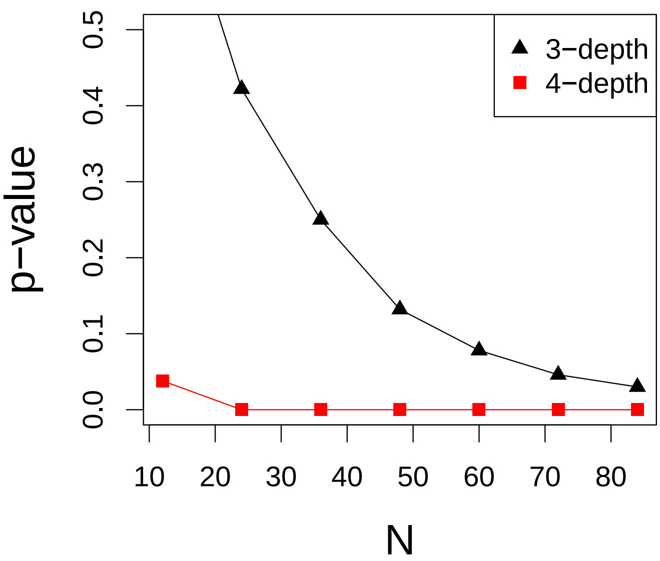

Figure 7: Simulated p-values of 3-depth tests and 4-depth tests for two sign changes using three blocks of equal size at different sample sizes N.

1.2 Nonlinear AR(1)-model

Motivated by crack growth analysis, Kustosz et al. (2016) 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 θ_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. This 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 $P(E_n > 0) = P(E_n < 0) = \frac{1}{2}$. 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, the 3-depth test, the 4-depth test, and the t-test for $N = 96$ is given by the Figures 8, 9, and 10. Thereby a 81×71 grid was used for the presentation of alternatives in θ_0 and θ_1, a 81×101 grid for the presentation of alternatives in θ_0 and θ_2, and a 65×74 grid of for the presentation of alternatives in θ_1 and θ_2. As in Section 1.1, the parameters of the null hypothesis are given by the intersections of the two dotted lines.

Figure 8 shows the behaviour when θ_2 is fixed to 1.002, Figure 9 when θ_1 is fixed to 0.005, and Figure 10 when θ_0 is fixed to 0.01. All three figures show, that the classical sign test and the t-test have a much worse power than the 3-depth test and the 4-test test. In particular, they have an unbounded area of power below 0.05 which is not the case for the depth tests. The results for fixed θ_0 are very similar to results for the classical sign test and the 3-depth test provided in Kustosz et al. (2016) with $\theta_0 = 0$. Here, 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.

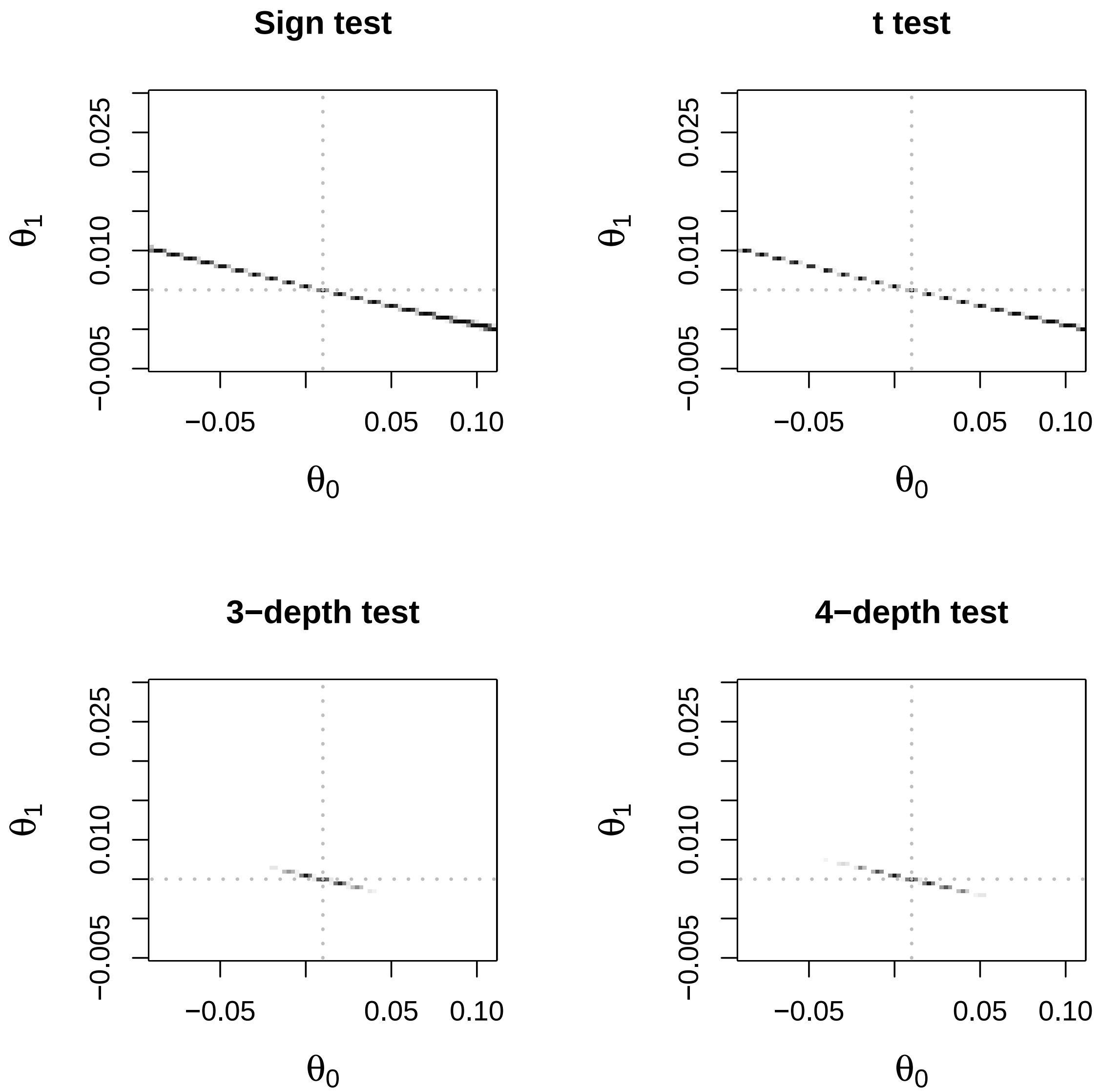

Figure 8: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the nonlinear AR(1)-model where θ_2 is fixed to 1.002 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

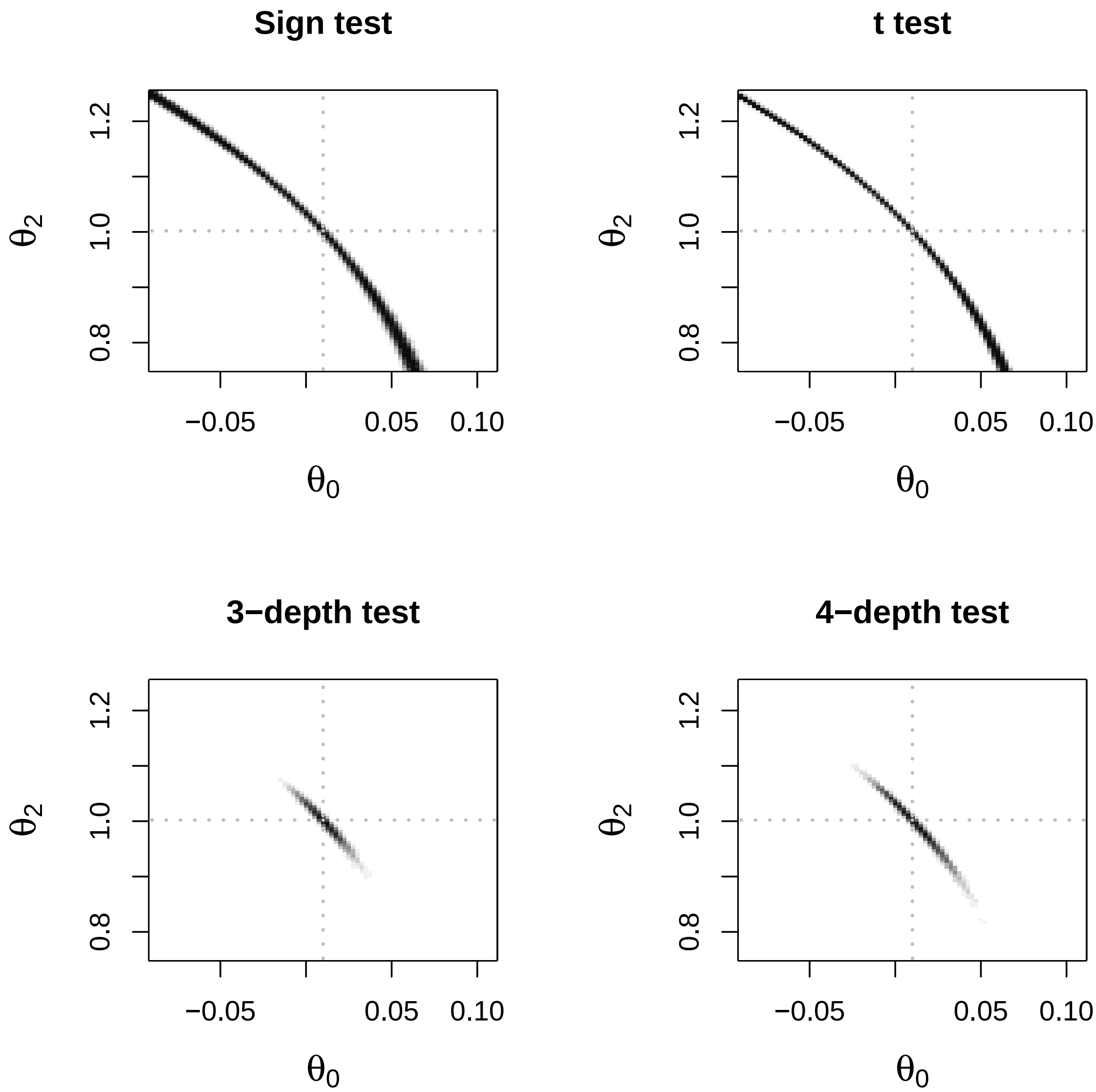

Figure 9: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the nonlinear AR(1)-model where θ_1 is fixed to 0.005 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

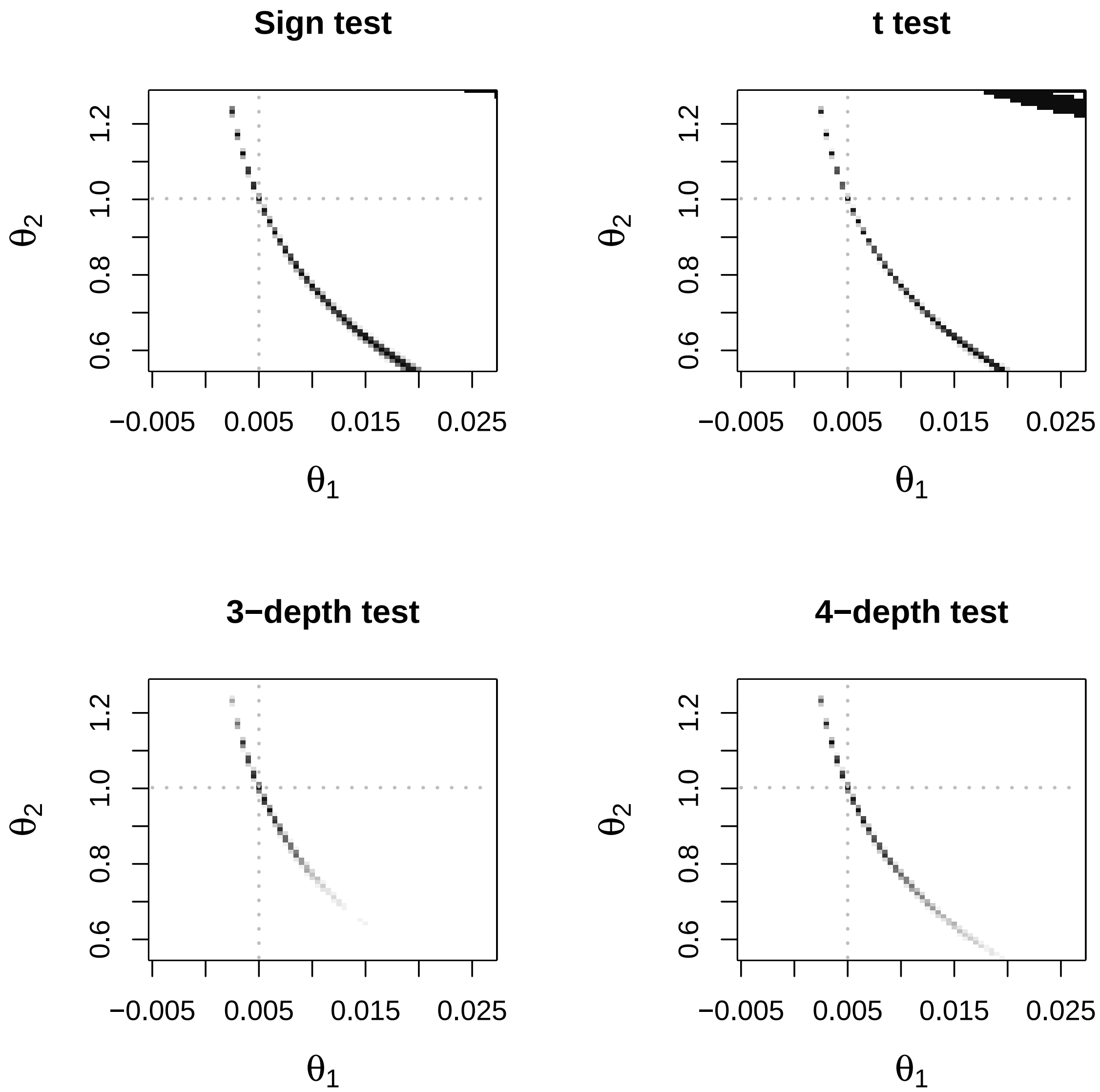

Figure 10: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the nonlinear AR(1)-model where θ_0 is fixed to 0.01 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

1.3 AR(2)-model

Here we consider the autoregressive model given by

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

with $Y_{-1} = Y_0 = 5$. The aim is to test $H_0 : \theta = (0.2, 0.8, 0.21)^\top$. In particular, we have an explosive process without stationarity under the null hypothesis. 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 α-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 a t-test for testing $H_0 : \theta = (0.2, 0.8, 0.21)^\top$ for $N = 96$ with $\alpha = 0.05$ are given by the Figures 11, 12, and 13. Thereby a 41×41 grid of alternatives was used. As in Section 1.1, the parameters of the null hypothesis are given by the intersections of the two dotted lines.

Figure 11 shows the results for the situation where θ_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 θ_1 is fixed to 0.8, the value of the null hypothesis, see Figure 12. A different behavior appears when θ_0 is fixed 0.2, the value of the null hypothesis. This behaviour is given by Figure 13. Here all four methods behave similarly and are struggling with an identifiability problem. This identifiability problem disappears for larger values of θ_1 and θ_2. However, then no difference between the methods is visible anymore.

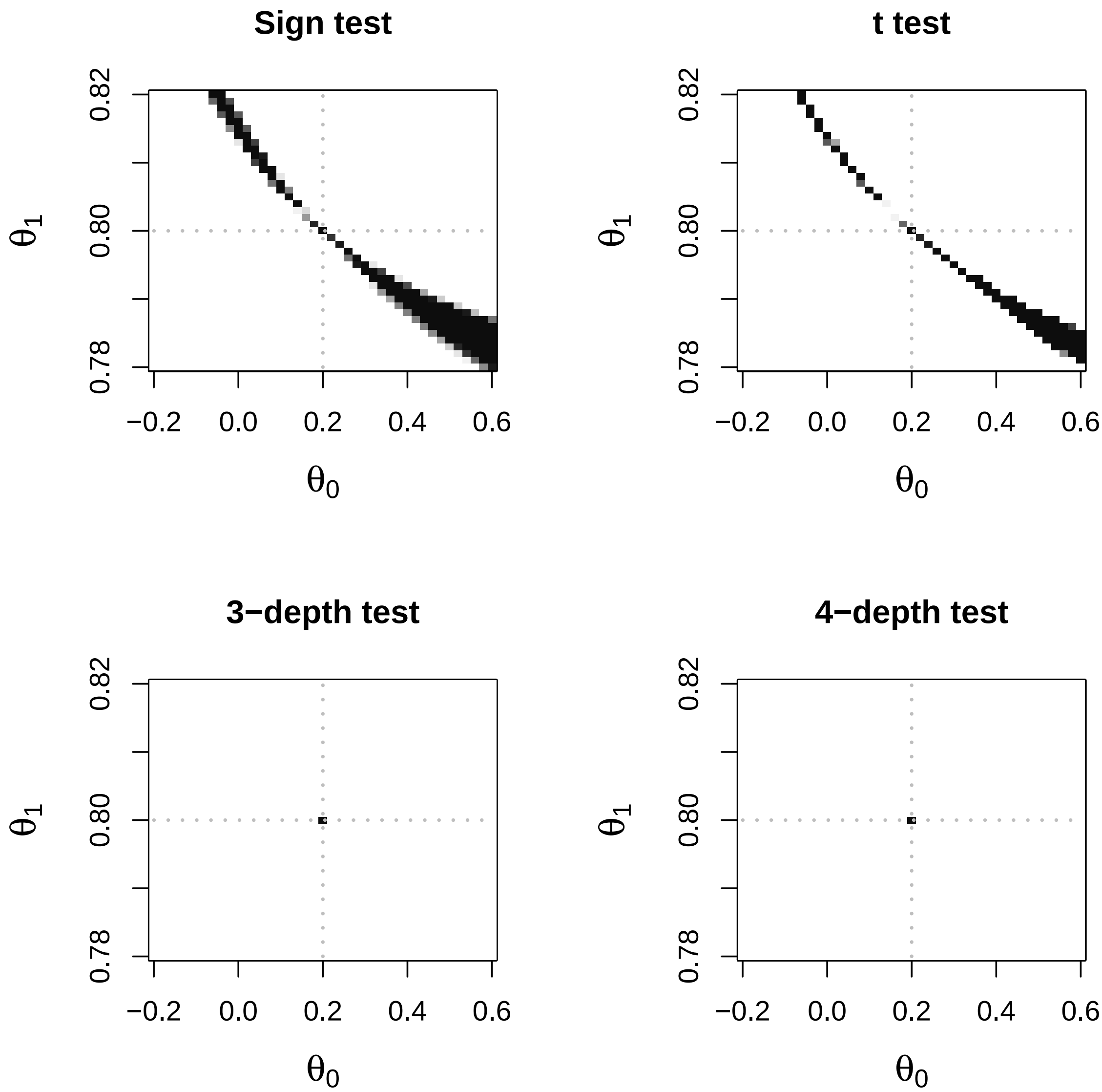

Figure 11: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the AR(2)-model where θ_2 is fixed to 0.21 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

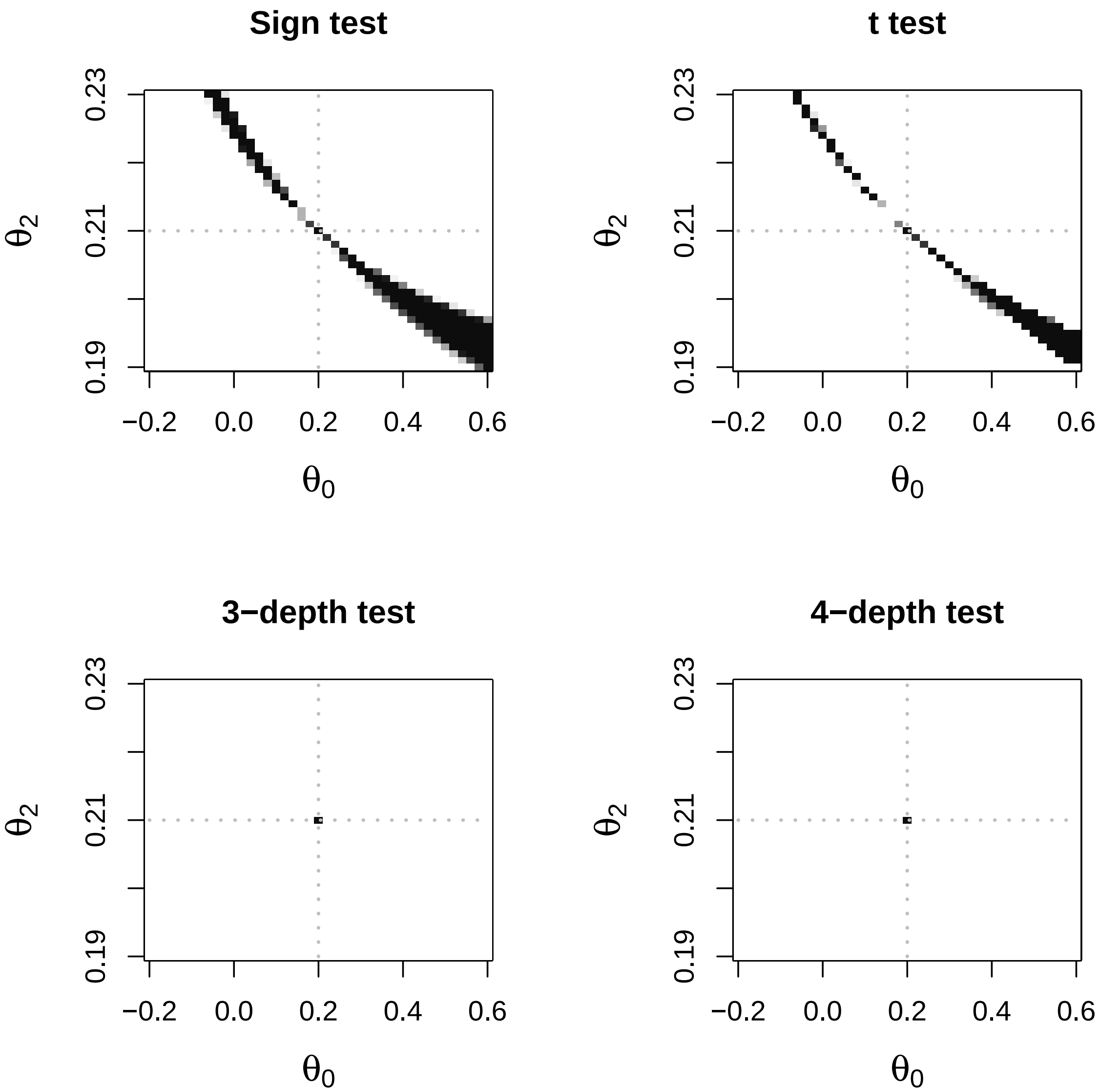

Figure 12: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the AR(2)-model where θ_1 is fixed to 0.8 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

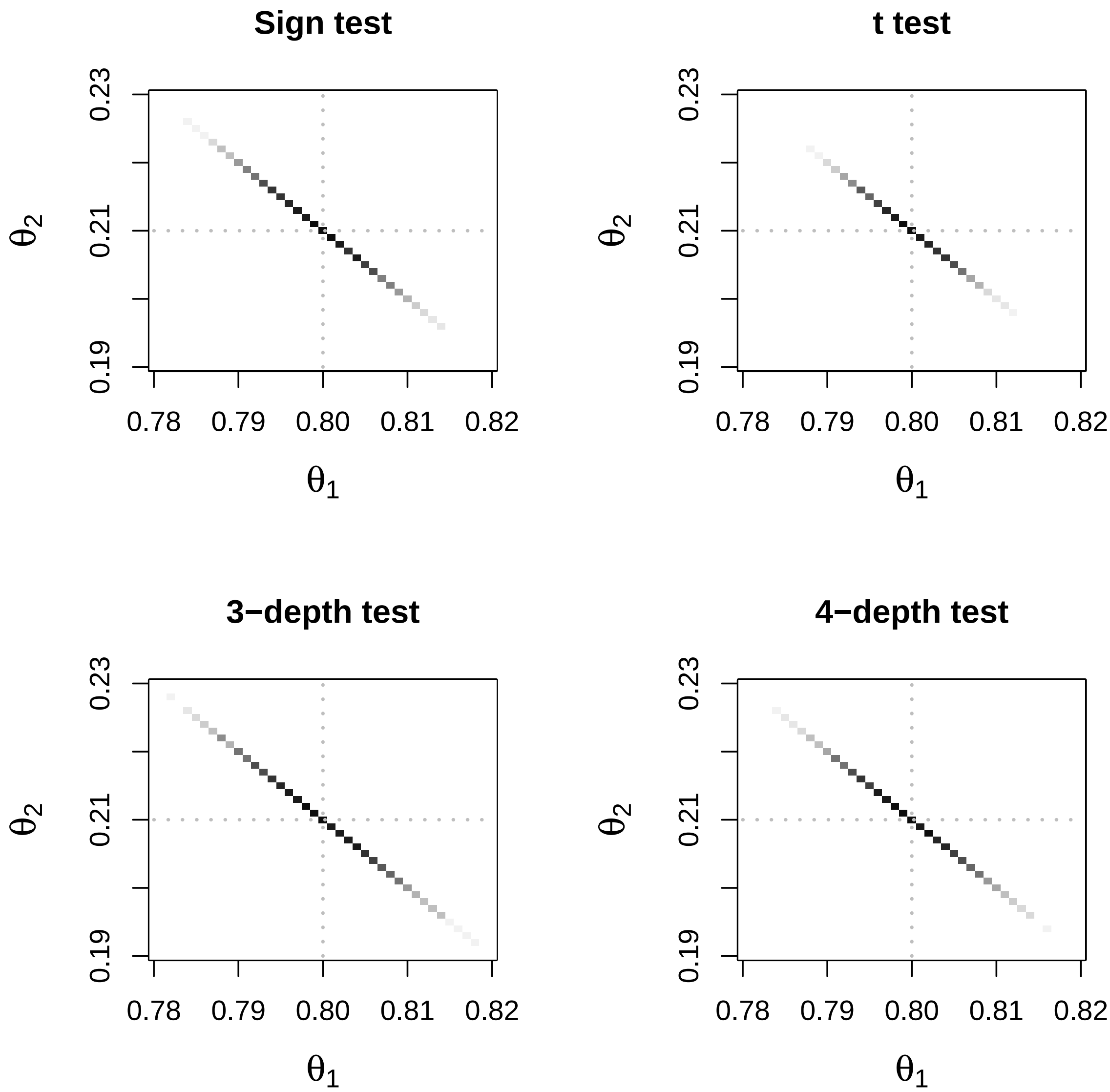

Figure 13: Simulated power of the sign test, 3-depth test, 4-depth test and t test for the AR(2)-model where θ_0 is fixed to 0.2 (20 gray levels were used, where black corresponds to $[0, 0.05]$ and white to $(0.95, 1]$).

References

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

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

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. (2016). Simplified simplicial depth for regression and autoregressive growth processes. *Journal of Statistical Planning and Inference*, 173:125–146.

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