Technische Universität Dortmund
Lehrstuhl Statistik in den Ingenieurwissenschaften
Gutachter: Prof. Dr. Christine Müller
Zweitgutachter: Prof. Dr. Roland Fried
Betreuer: M. Sc. Dennis Andreas Malcherczyk

Masterarbeit

Vergleich der Vorzeichentiefetests mit anderen Tests zur Überprüfung von Unabhängigkeitsannahmen in Zeitreihen

von

Hendrik Dohme

Dortmund, 20.01.2021

Inhaltsverzeichnis

1 Einleitung

Die Untersuchung und Spezifikation der Abhängigkeitsstrukturen in Zeitreihen sind wesentliche Bestandteile der Zeitreihenanalyse. Dabei besteht eine zentrale Voraussetzung für viele statistische Methoden darin, dass Zufallsvariablen $X_1, ..., X_N$ zufällig − also unabhängig und identisch verteilt − sind. So setzt z. B. die Methode der kleinsten Quadrate eine Unkorreliertheit der aus ihr resultierenden Fehler voraus. Und auch bei der Anpassung von autoregressiven bzw. Moving-Average-Modellen ist sicherzustellen, dass keine Reststruktur in den verbleibenden Residuen vorhanden ist. Außerdem ist es häufig von Interesse, ob die vorliegenden Beobachtungen zufällig ausgewählt wurden oder ob eine zeitliche Struktur, wie z. B. ein Trend, in der Zeitreihe vorhanden ist. Dabei existieren viele verschiedene Alternativen zur Zufälligkeit.

Aus diesem Grund stellt die Anwendung adäquater Unabhängigkeits- bzw. Zufälligkeitstests an Zeitreihen einen wichtigen Aspekt der Statistik dar. Dafür stehen diverse Testverfahren zur Verfügung, die verschiedene Kenngrößen in der Zeitreihe zur Entscheidungsfindung heranziehen. Dies führt dazu, dass die Verfahren in unterschiedlicher Weise auf verschiedene Alternativen zur Unabhängigkeit sowie auf andersartige Strukturen in der Zeitreihe reagieren. Beispielsweise kann sowohl auf Basis der empirischen Autokorrelationskoeffizienten auf Abhängigkeiten in der Zeitreihe geschlossen werden als auch anhand der Anzahl an Hoch- und Tiefpunkte. Liegt jedoch z. B. ein Trend in der Zeitreihe vor, so können die Autokorrelationskoeffizienten trotz fehlender Korrelation signifikant erhöht sein. Auf die Anzahl der Extrempunkte hat der Trend jedoch kaum einen Einfluss. Aus diesem Grund stellt das Verständnis der verwendeten Testverfahren und ihrer Statistiken einen wesentlichen Gesichtspunkt bei der Wahl eines geeigneten Testverfahrens für die jeweilige Situation dar.

Obwohl bereits einige Simulationsstudien zu den Trennschärfen unterschiedlicher Unabhängigkeits- und Zufälligkeitstests unter verschiedenen Szenarien existieren (u. a. Mateus und Caeiro, 2013; Harrison und McCabe, 1975; Islam und Toor, 2019; Gupta und Govindarajulu, 1980), werden in diesen Arbeiten oft nur Testverfahren mit ähnlichen Ansätzen in Bezug auf ihre Trennschärfe verglichen. Des Weiteren sind in vielen Arbeiten nur wenige Abweichungen von den Voraussetzungen der Testverfahren und Szenarien Bestandteil der Simulationen. Auch beschränkt sich die untersuchte Abhängigkeitsstruktur dabei in den meisten Fällen auf die von AR(1)-Prozessen.

Das zentrale Ziel dieser Arbeit besteht deshalb darin, das Verhalten parametrischer, nichtparametrischer sowie rangbasierter Unabhängigkeitstests in verschiedenen Szenarien zu untersuchen und zu vergleichen. Außerdem soll ihre Eignung in dem jeweiligen Szenario beurteilt

werden. Das Hauptaugenmerk liegt dabei auf den von Leckey et al. (2020) beschriebenen K-Vorzeichentiefetests. Kustosz et al. (2016a) konnten für diese Testverfahren im Kontext von explosiven Prozessen schon Überlegenheiten gegenüber dem F-Test sowie dem einfachen Vorzeichentest feststellen. Da über das Verhalten der K-Vorzeichentiefetests bei stationären Prozessen jedoch noch wenig bekannt ist, liegt der Fokus dieser Arbeit auf solchen Prozessen.

Im zweiten Kapitel dieser Arbeit werden die grundlegenden mathematischen Konzepte und Modelle, die des Weiteren von Interesse sind, definiert. Anschließend werden die verschiedenen zu untersuchenden statistischen Testverfahren vorgestellt und erläutert. Als parametrische Tests werden dabei der Durbin-Watson-Test sowie der Ljung-Box-Test betrachtet. Zur Auswahl nichtparametrischer Verfahren gehören der Runs-Test, der Turning-Point-Test, der Broock-Dechert-Schreinkman-Test und die K-Vorzeichentiefetests. Als Vertreter rangbasierter Tests wird der Von-Neumann-Ratio-Rang-Test betrachtet.

Das dritte Kapitel beinhaltet die statistische Auswertung. Dabei werden die verschiedenen Testverfahren auf simulierte Zeitreihen aus unterschiedlichen Prozessen sowie unter diversen Szenarien angewendet. Anschließend werden ihre Trennschärfen beschrieben, verglichen und erläutert. Bestandteil der Untersuchungen sind autoregressive Prozesse 1. und 2. Ordnung sowie saisonale Prozesse, Moving-Average-Prozesse 1. Ordnung und GARCH(1,1)-Prozesse. Als Szenarien werden Abweichungen von der Normalverteilung der Innovationen, innovative Ausreißer, Kontaminationen, wachsende Varianzen und Strukturen wie Trends, Sprünge sowie eine feste und eine wachsende Anzahl von Oszillationen in der Zeitreihe betrachtet. Außerdem werden weitere Untersuchungen zu den K-Vorzeichentiefetests durchgeführt. Dabei wird die Asymmetrie ihrer Trennschärfe thematisiert und vereinfachte Versionen dieser Testverfahren werden betrachtet und analysiert.

Im Fazit werden die wichtigsten gewonnen Erkenntnisse zusammengefasst und diskutiert. Dabei wird noch einmal ein Überblick über das Verhalten und die Eignung der einzelnen Testverfahren bei den verschiedenen zugrunde liegenden Prozessen und Szenarien gegeben. Im Ausblick werden mögliche Modifikationen der K-Vorzeichentiefetests für weitere Untersuchungen aufgezeigt, von denen verbesserte Trennschärfen bei den in dieser Arbeit betrachteten Alternativen zur Zufälligkeit in Zeitreihen zu erwarten wären.

2 Statistische Methoden

2.1 Definitionen

Stochastischer Prozess

Sei $(\Omega, \Sigma, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $I \neq \emptyset$ eine Indexmenge. Eine Folge von Zufallsvariablen $(X_t)_{t\in I}$ mit $X_t : \Omega \to \mathbb{R}$ wobei $I = \mathbb{Z}$ gilt und zugehöriger Wahrscheinlichkeitsdichtefunktionen f_{X_t} wird nach Brockwell und Davis (2006, S. 8) als zeitdiskreter, reellwertiger, stochastischer Prozess (oder hier kurz: stochastischer Prozess) bezeichnet. Eine Realisierung eines stochastischen Prozesses $(X_t(\omega),\, \omega \in \Omega)_{t\in I} = (x_t)_{t\in I}$ wird dabei als Pfad des Prozesses oder, im Folgenden vorwiegend, als Zeitreihe bezeichnet.

Ein Ziel der Zeitreihenanalyse besteht darin, auf Grundlage von beobachteten Zeitreihen Rückschlüsse auf den zugrunde liegenden Prozess zu ziehen. In der Praxis liegen in den meisten Fällen lediglich Ausschnitte einer Zeitreihe vor, die durch einen stochastischen Prozess erzeugt wurden. In dieser Arbeit werden deshalb weitestgehend Zeitreihen der Form $(x_1, ..., x_N)$ betrachtet, wobei $N \in \mathbb{N}$ gilt.

Erwartungswert, Varianz und Autokovarianz

Die Mittelwertfunktion eines stochastischen Prozesses $(X_t)_{t\in I}$ ist nach Brockwell und Davis (2010, S. 15) definiert als:

$$\mu_t = E[X_t] = \int_{-\infty}^{\infty} x f_{X_t}(x)\, dx,$$

falls dieses Integral existiert, wobei es sich bei f_{X_t} um die Wahrscheinlichkeitsdichtefunktion der Zufallsvariablen X_t des stochastischen Prozesses $(X_t)_{t\in I}$ handelt. Falls der Prozess ein endliches zweites Moment besitzt, das heißt, falls $E[X_t^2] < \infty$ für alle $t \in I$ gilt, so ist seine Varianz definiert als:

$$\sigma_t^2 = Var[X_t] = E[(X_t - \mu_t)^2]$$

und seine Autokovarianzfunktion entspricht:

$$\gamma_X(t, s) = Cov[X_t, X_s] = E[(X_t - \mu_t)(X_s - \mu_s)], \quad t, s \in I.$$

Liegt eine Zeitreihe $(x_1, ..., x_N)$ vor, so sind ihr empirischer Mittelwert und ihre empirische Autokovarianz zum Lag h nach Shumway und Stoffer (2017, S.27) definiert durch:

$$\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x_t,$$
$$\hat{\gamma}(h) = \frac{1}{N} \sum_{t=1}^{N-h} (x_{t+h} - \bar{x})(x_t - \bar{x}).$$

Stationarität

Ein stochastischer Prozess $(X_t)_{t \in I}$ heißt nach Brockwell und Davis (2006, S.12) schwach stationär (oder kurz: stationär), falls die folgenden Bedingungen erfüllt sind:

(i) $E[X_t^2] < \infty \ \forall \ t \in I$,

(ii) $E[X_t] = m \ \forall \ t \in I$, wobei $m \in \mathbb{R}$,

(iii) $\gamma_X(t, s) = \gamma_X(t + r, s + r) \ \forall \ s, t, r \in I$.

Die Autokovarianzfunktion stationärer, stochastischer Prozesse steht mit s und t lediglich durch ihre Differenz $|t - s| = h \in I$ in Zusammenhang, sodass sie häufig nur mit $\gamma_X(h)$ bezeichnet wird. Heuristisch gesehen bedeutet die schwache Stationarität eines Prozesses, dass sich seine statistischen Eigenschaften über die Zeit hinweg nicht verändern. Erst durch die Annahme der Stationarität kann geschlussfolgert werden, dass eine betrachtete Zeitreihe $(x_1, ..., x_N)$ charakteristisch für den gesamten Prozess ist.

Autokorrelationsfunktion

Handelt es sich bei $(X_t)_{t \in I}$ um einen (schwach) stationären Prozess, so kann seine Autokorrelationsfunktion zum Lag h nach Brockwell und Davis (2006, S. 16) definiert werden als:

$$\rho_X(h) = \frac{\gamma_X(t, t + h)}{\gamma_X(t, t)} = \frac{\gamma_X(h)}{\gamma_X(0)} \quad \forall \ t \in I.$$

Damit handelt es sich bei der Autokorrelationsfunktion um eine einheitenlose Normierung der Autokovarianzfunktion, die Werte zwischen -1 und 1 annehmen kann. Analog kann die empirische Autokorrelationsfunktion definiert werden als:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Korrelogramm

Als Korrelogramm wird nach Schlittgen (2001, S. 7) der Graph der Autokorrelationsfunktion einer Zeitreihe in Abhängigkeit vom Lag h bezeichnet. In der Zeitreihenanalyse wird es oft als Werkzeug für die Überprüfung der Unabhängigkeit einer Zeitreihe genutzt. Dabei können auf seiner Grundlage sowohl rein optische als auch statistisch fundierte Aussagen über die Korrelationsstrukturen einer Zeitreihe gefällt werden.

White-Noise-Prozess

Nach Schlittgen (2001, S. 9) wird eine Familie unkorrelierter, identisch verteilter Zufallsvariablen $(W_t)_{t\in I}$ als White-Noise-Prozess oder Weißes Rauschen bezeichnet, falls die folgenden Bedingungen erfüllt sind:

(i) $E[W_t] = 0 \;\; \forall\, t \in I$,

(ii) $Var[W_t] = \sigma_W^2 \;\; \forall\, t \in I$, wobei $\sigma_W^2 \in \mathbb{R}$.

In dieser Arbeit werden solche Prozesse mit $WN(0, \sigma_W^2)$ gekennzeichnet. Sind die Zufallsvariablen W_t sogar unabhängig, so wird der Prozess starkes Weißes Rauschen genannt und mit $iid(0, \sigma_W^2)$ bezeichnet. Insbesondere handelt es sich bei White-Noise-Prozessen aufgrund der oben eingeführten Eigenschaften um stationäre Prozesse.

Gaußscher Prozess

Der Gaußsche Prozess stellt nach Schlittgen (2001, S. 9) einen Spezialfall des Weißen Rauschens dar, bei dem die Zufallsvariablen $(W_t)_{t\in I}$ alle normalverteilt sind. Diese Art von Prozessen wird im Folgenden mit $GN(0, \sigma_G^2)$ bezeichnet.

2.2 Modelle

Autoregressive Prozesse (AR-Prozesse)

Ein stochastischer Prozess $(X_t)_{t\in\mathbb{Z}}$ wird nach Shumway und Stoffer (2017, S. 76) als autoregressiver Prozess der Ordnung p (AR(p)) bezeichnet, falls er von der Form:

$$x_t = \mu + \rho_1 x_{t-1} + ... + \rho_p x_{t-p} + w_t,\ t \in \mathbb{Z}$$

ist. Dabei sind $\mu, \rho_1, ..., \rho_p \in \mathbb{R}$ Konstanten mit $\rho_p \neq 0$ und w_t sind Realisationen eines White-Noise-Prozesses mit der Varianz $\sigma_W^2 > 0$, wobei μ als Mittelwert der Zeitreihe bezeichnet wird und $\rho_1, ..., \rho_p$ als AR-Koeffizienten. Für einen stationären, autoregressiven Prozess 1. Ordnung mit einem Mittelwert von 0 entspricht der Autokorrelationskoeffizient zum Lag h (s. Kap. 2.1) nach Shumway und Stoffer (2017, S.77) genau:

$$\rho(h) = \rho_1^h \quad h \geq 0.$$

Moving-Average-Prozesse (MA-Prozesse)

Ein stochastischer Prozess $(X_t)_{t\in\mathbb{Z}}$ wird nach Shumway und Stoffer (2017, S. 81) gleitender Durchschnitt-Prozess oder Moving-Average-Prozess der Ordnung q (MA(q)) genannt, wenn er die Form:

$$x_t = \mu + \nu_1 w_{t-1} + ... + \nu_q w_{t-q} + w_t,\ t \in \mathbb{Z}$$

hat. Bei $\mu, \nu_1, ..., \nu_q \in \mathbb{R}$ handelt es sich um Konstanten mit $\nu_q \neq 0$ und w_t sind Realisationen eines White-Noise-Prozesses mit der Varianz $\sigma_W^2 > 0$. In dem Zusammenhang heißen μ Mittelwert der Zeitreihe und $\nu_1, ..., \nu_q$ Moving-Average-Koeffizienten. Handelt es sich um einen Moving-Average-Prozess 1. Ordnung, der stationär ist und einen Mittelwert von 0 besitzt, so entspricht seine Autokorrelationsfunktion zum Lag h genau:

$$\rho(h) = \begin{cases} \nu_1/(1+\nu_1^2), & h = 1 \\ 0, & h > 1. \end{cases}$$

Insbesondere gilt für diese Funktion $|\rho(h)| \leq 0.5$, was sich durch einfachen Ableiten des obigen Terms nachvollziehen lässt.

Autoregressive Moving-Average Prozesse (ARMA-Prozesse)

Ein stochastischer Prozess $(X_t)_{t\in\mathbb{Z}}$ wird nach Shumway und Stoffer (2017, S. 83) als autoregressiver Moving-Average-Prozess der Ordnungen p und q (ARMA(p,q)) bezeichnet, wenn er stationär und von der Form

$$x_t = \mu + \rho_1 x_{t-1} + ... + \rho_p x_{t-p} + \nu_1 w_{t-1} + ... + \nu_q w_{t-q} + w_t,\ t \in \mathbb{Z}$$

ist. Dabei handelt es sich bei $\mu, \rho_1, ..., \rho_p, \nu_1, ..., \nu_q \in \mathbb{R}$ um Konstanten mit $\nu_q, \rho_p \neq 0$ und w_t sind Realisationen eines White-Noise-Prozesses. Die Parameter $\nu_1, ..., \nu_q$ werden als MA-Koeffizienten, $\rho_1, ..., \rho_p$ als AR-Koeffizienten und μ als Mittelwert der Zeitreihe bezeichnet.

Generalized-Autoregressive-Conditional-Heteroscedasticity-(GARCH-)Modell

Bei einem Generalized-Autoregressive-Conditional-Heteroscedasticity-(GARCH-)Modell handelt es sich um eine von Bollerslev (1986) erstmals beschriebene Verallgemeinerung des von Engle (1982) eingeführten Autoregressive-Conditional-Heteroscedasticity-(ARCH-)Modells.

Im Gegensatz zu herkömmlichen Zeitreihenmodellen, die eine konstante Varianz der Fehlerterme voraussetzen, unterstellt dieses Modell, dass die bedingte Varianz der Fehler zum Zeitpunkt t als eine Funktion der quadrierten vorangegangenen Fehler variiert. Eine Motivation für dieses Modell stellen Finanzzeitreihen dar, in denen typischerweise sogenannte Volatilitätscluster zu beobachten sind. Das bedeutet, dass sie sowohl Phasen enthalten, in denen die Zeitreihenwerte eine hohe Volatilität aufweisen, als auch andere, in denen sie eine vergleichsweise sehr geringe Varianz besitzen. In seiner einfachsten Form ist das ARCH(p)-Modell nach Verbeek (2012, S. 298) definiert als

$$x_t = \epsilon_t, \quad \epsilon_t = \sigma_t w_t, \quad w_t \sim WN(0,1)$$
$$\sigma_t^2 = E\left[\epsilon_t^2 |\, \mathbb{I}_{t-1}\right] = \omega + \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_p \epsilon_{t-p}^2.$$

Dabei handelt es sich bei $\mathbb{I}_{t-1}$ um ein Informationsset, das typischerweise ϵ_{t-1} und seine gesamte Historie enthält. Eine notwendige Restriktion ist dabei, dass $\omega \geq 0$ und $\alpha_i \geq 0$ für $i \in \{1, ..., p\}$ gelten, um sicherzustellen, dass σ_t nichtnegativ ist. Im Fall $p = 1$ sagt das ARCH(p)-Modell aus, dass auf eine betragsmäßig große Innovation zum Zeitpunkt (t - 1) tendenziell wieder eine betragsmäßig große Innovation folgt. Die Spezifikation des Modells impliziert also, dass die Terme ϵ_t und ϵ_{t-1} korreliert sind.

Eine nützliche und in der empirischen Anwendung stark verbreitete Verallgemeinerung des ARCH(p)-Modells stellt das generalisierte ARCH(p)-Modell oder GARCH(p,q)-Modell nach Bollerslev (1986) dar. Dabei wird die bedingte Varianz der Fehler modelliert durch:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2.$$

In der empirischen Anwendung schneidet das GARCH(1,1)-Modell häufig sehr gut ab. Die bedingte Varianz kann in diesem Fall geschrieben werden als

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Um die Nichtnegativität von σ_t^2 zu gewährleisten, müssen hier die Parameter ω, α_1 sowie β_1 alle nichtnegativ sein.

Multiple Regression

Unter einer multiplen Regression versteht man nach Bender et al. (2002) eine Verallgemeinerung des einfachen linearen Regressionsmodells auf ein Modell mit mehreren erklärenden Variablen. Seine allgemeine Form entspricht:

$$y_i = \theta_0 + \theta_1 x_{i1} + ... + \theta_m x_{im} + \epsilon_i$$

für $i \in \{1, ..., N\}$ oder in Matrixschreibweise:

$$y = X\theta + \epsilon.$$

Dabei handelt es sich bei $X = (x_1^T, ..., x_m^T) \in \mathbb{R}^{m \times N}$ um die Designmatrix der erklärenden Variablen oder Regressoren $x_j = (x_{1j}, ..., x_{Nj})$ für $j \in \{1, ..., m\}$, deren Einfluss auf den Vektor der sogenannten Zielvariablen $y = (y_1, ..., y_N)^T \in \mathbb{R}^N$ mithilfe des Regressionsmodells beschrieben werden soll. Der Vektor $\epsilon = (\epsilon_1, ..., \epsilon_N)^T \in \mathbb{R}^N$ entspricht dabei dem Fehlervektor, dessen Einträge Realisierungen unkorrelierter, identisch verteilter Zufallsvariablen $E_1, ..., E_N$ mit Erwartungswerten $E[E_i] = 0$ und konstanter Varianz $Var[E_i] = \sigma_e^2 > 0 \ \forall i \in \{1, ..., N\}$ darstellen. Bei $\theta = (\theta_0, \theta_1, ..., \theta_m)^T \in \mathbb{R}^m$ handelt es sich um den sogenannten Regressionskoeffizienten, der den Einfluss der erklärenden Variablen auf die Zielvariable quantifiziert. Dabei wird θ_0 als Regressionskonstante oder Interzept bezeichnet und die Parameter $\theta_1, ..., \theta_m$ heißen Steigungsparameter. Die Analyse dieses Regressionskoeffizienten ist bei der multiplen Regression von zentralem Interesse. Seine Schätzung kann beispielsweise durch die Methode der kleinsten Quadrate (KQ-Schätzung) erfolgen.

Nachdem ein multiples Regressionsmodell angepasst worden ist, stellt die Analyse der Residuen $r_i(\theta) := y_i$ - $\tilde{x_i}\theta$ einen wichtigen Schritt zur Überprüfung der Annahmen der Schätzmethode des Regressionskoeffizienten dar. Dabei handelt es sich bei $\tilde{x}_i$ mit $i \in \{1, ..., N\}$ um die Zeilen der Matrix X. So wird z. B. bei der KQ-Schätzung vorausgesetzt, dass die Residuen unkorreliert sind und einen Erwartungswert von 0 sowie eine homogene Varianz aufweisen. Diese Annahmen können mit numerischen und mit grafischen Methoden überprüft werden.

2.3 Durbin-Watson-Test (DW-Test)

Beim Durbin-Watson-Test handelt es sich um einen von Durbin und Watson (1950) eingeführten parametrischen Test auf das Vorliegen einer Autokorrelation zum Lag 1 in einer Zeitreihe $(x_1, ..., x_N)$. Besondere Anwendung findet dieser Test im Kontext von multiplen linearen Regressionen, um die Unabhängigkeitsannahmen von den aus einer KQ-Schätzung resultierenden Residuen, die im Folgenden als $e = (e_1, ..., e_N)$ bezeichnet werden, zu überprüfen. Sie sind in einem solchen Fall mit den Bezeichungen aus Kapitel 2.2 definiert als:

$$e = \left[I - X\left(X^T X\right)^{-1} X^T\right] y,$$

wobei I die Einheitsmatrix entsprechender Dimension bezeichnet. Eine Anwendung des DW-Tests macht dabei lediglich Sinn, wenn die Schätzung einen zeitlich geordneten Prozess betrifft. Für die Residuen wird dann der Zusammenhang

$$e_t = \rho_1 e_{t-1} + w_t, \quad \text{mit} \quad |\rho_1| \leq 1, \quad w_t \sim GN(0, \sigma_G^2)$$

für $t \in \{2, ..., N\}$ angenommen und die Nullhypothese $\rho_1 = 0$ wird gegen die Alternative $\rho_1 \neq 0$ getestet. Eine Ablehnung der Nullhypothese würde nach Durbin und Watson (1950) dafür sprechen, dass die Voraussetzung der KQ-Schätzungen, nach denen aufeinanderfolgende Fehler unabhängig voneinander sind, verletzt ist. Als Konsequenz kann nach dem Gauß-Markov-Theorem beispielsweise nicht mehr garantiert werden, dass der berechnete KQ-Schätzer die geringste Varianz unter der Menge der erwartungstreuen Schätzer hat. Aufgrund der Linearität des Erwartungswertes hat die Kovarianz der Residuen jedoch keinen Einfluss auf die Erwartungstreue des Schätzers.

Voraussetzungen für die Anwendung des Durbin-Watson-Tests sind nach Verbeek (2012, S. 102), dass das Regresssionsmodell einen Interzept beinhaltet und die Regressoren als deterministisch betrachtet werden. Damit darf die Regressormatrix beispielsweise keine verzögerten, abhängigen Variablen beinhalten.

Die Teststatistik des Durbin-Watson-Tests ist definiert als:

$$T_{DW} = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2}.$$

Dabei ist $T_{DW} \geq 0$, da sowohl im Zähler als auch im Nenner quadratische Terme summiert werden. Durch das Ausmultiplizieren obiger Formel und indem ausgenutzt wird, dass für hinreichend großes N approximativ $\sum_{t=2}^{N} e_t \approx \sum_{t=2}^{N} e_{t-1}$ gilt, ergibt sich die alternative Darstellung:

$$T_{DW} = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} = \frac{\sum_{t=2}^{N}(e_t^2 - 2e_{t-1}e_t + e_{t-1}^2)}{\sum_{t=1}^{N} e_t^2} \approx 2\left(1 - \frac{\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=1}^{N} e_t^2}\right).$$

Da die Residuen eines KQ-Modells definitionsgemäß einen Erwartungswert von 0 aufweisen, entspricht der Ausdruck $\sum_{t=2}^{N} e_t e_{t-1} / \sum_{t=1}^{N} e_t^2$ dem empirischen Korrelationskoeffizienten $\hat{\rho}_1$.

Mit der Voraussetzung, dass $|\rho_1| \leq 1$ erfüllt ist, sollte dies auch für den empirischen Korrelationskoeffizienten gelten, sodass für die Teststatistik insgesamt $0 \leq T_{DW} \leq 4$ gilt. Der Extremfall $T_{DW} \approx 0$ tritt dabei genau dann auf, wenn eine perfekte positive Korrelation zwischen den Residuen besteht, also $\rho_1 = 1$ ist. Im Fall, dass eine perfekte negative Korrelation vorliegt, also $\rho_1 =$ -1 gilt, nimmt die Teststatistik ihren maximalen Wert von ungefähr 4 an. Unter der Nullhypothese beträgt der Wert der Teststatistik 2 und deutet somit auf ein Fehlen von Autokorrelationen 1. Ordnung in der betrachteten Stichprobe hin.

Die exakte Verteilung der Teststatistik T_{DW} ist komplex und hängt nicht nur vom Stichprobenumfang N und der Anzahl der Regressoren K ab, sondern auch von den Werten, die sie annehmen. Aus diesem Grund beschäftigten sich viele wissenschaftliche Studien mit der Herleitung von oberen und unteren Schranken d_L und d_U der kritischen Werte von T_{DW} (also: $d_L < d_{crit} < d_U$) in Abhängigkeit von K und N und damit, die Trennschärfe des Tests zu beurteilen (Savin und White, 1977; Durbin und Watson, 1971; L'Esperance und Taylor, 1975; Krämer, 1985; Farebrother, 1980). Obwohl z. B. Imhof (1961) oder Farebrother (1980) bereits Methoden zur Bestimmung von d_L und d_U vorgestellt haben, werden aufgrund des großen Rechenaufwandes, den sie benötigen, häufig Tabellen für die Schranken herangezogen. Durbin und Watson (1971) und Savin und White (1977) erkannten dabei zumindest in Stichproben mit $N < 15$ die Überlegenheit der von Farebrother (1980) beschriebenen Methode in Bezug auf die Rechenkomplexität gegenüber der von Imhof (1961). So tabellierten Savin und White (1977) entsprechende Werte von d_L und d_U bei Stichprobengrößen von $6 - 200$ Beobachtungen und bis zu 6 Regressoren, wobei sie auf Grundlage ihrer Berechnungen empfehlen, für Stichprobengrößen mit $N > 80$ die von Imhof entwickelte Methode zu verwenden. Für größere Stichproben erachteten Savin und White (1977) den erforderlichen Rechenaufwand als zu groß und eine weitere Tabellierung als impraktikabel. Weiterhin bemerkten sie, dass die berechneten Werte d_L und d_R den Quantilen der exakten Verteilung bis auf die dritte Nachkommastelle entsprechen.

Unter der Nullhypothese, dass $\rho_1 = 0$ ist, gilt damit für die Schranken zum Niveau $\alpha = 0.05$:

$$P(T_{DW} < d_L) \leq P(T_{DW} < d_{crit}) = 0.05 \leq P(T_{DW} < d_U).$$

Für die einseitige Alternative $\rho_1 > 0$ gibt es damit nach Verbeek (2012, S. 103) drei mögliche Szenarien, eine Testentscheidung zu treffen:

(i) $T_{DW} < d_L$: Damit gilt auf jeden Fall auch $T_{DW} < d_{crit}$ und die Nullhypothese wird abgelehnt.

(ii) $T_{DW} > d_U$: Damit gilt auf jeden Fall auch $T_{DW} > d_{crit}$ und die Nullhypothese kann nicht abgelehnt werden.

(iii) $d_L \leq T_{DW} \leq d_U$: In diesem Fall kann keine Aussage getroffen werden, da nicht sicher ist, ob T_{DW} den Wert von d_{crit} unterschreitet oder nicht.

Um die Alternative $\rho_1 < 0$ zu testen, wird die Teststatistik (4 - T_{DW}) herangezogen, sodass obige Entscheidungsregeln analog verwendet werden können. Im Fall einer beidseitigen Alternative müssen beide erwähnten Teststatistiken betrachtet werden und zwei Entscheidungen zum

Niveau $\alpha/2$ müssen nach obigem Prinzip gefällt werden. Erst wenn in einem der beiden Fällen eine Ablehnung der Nullhypothese stattfindet, kann die beidseitige Nullhypothese abgelehnt werden.

Die Existenz eines Wertebereichs, in dem keine Testentscheidung getroffen werden kann, stellt dabei einen wesentlichen Nachteil des Durbin-Watson-Tests dar. Allerdings gilt zu beachten, dass die Größe dieses Bereichs mit zunehmender Stichprobenanzahl abnimmt, sodass die Wahrscheinlichkeit, dass die Teststatistik einen Wert in diesem Bereich annimmt, zunehmend kleiner wird.

In `R` kann der Durbin-Watson-Test mit der Funktion `dwtest` aus dem Paket `lmtest` von Zeileis und Hothorn (2002) durchgeführt werden. Als Eingabe akzeptiert diese Funktion entweder ein `lm`-Objekt oder eine Regressionsformel. Im Fall, dass keine Regressoren vorhanden sind, müssen die Beobachtungen in Abhängigkeit ihrer Indizes modelliert werden. Die Art der Alternative kann über den Parameter `alternative` eingestellt werden. Für die Berechnung des p-Wertes kann mit Hilfe des Parameters `exact` entweder der von Farebrother (1980) beschriebene Algorithmus (`exact=TRUE`) oder eine Normalapproximation mit den Momenten der DW-Statistik (`exact=FALSE`) gewählt werden. Allerdings geben Zeileis und Hothorn (2002) zu bedenken, dass der obige, exakte Algorithmus für große Stichproben zu rechenintensiv ist, sodass in solchen Fällen automatisch eine Normalapproximation vorgenommen wird.

2.4 Box-Pierce- & Ljung-Box-Test (LB-Test)

Beim Box-Pierce-Test handelt es sich um einen von Box und Pierce (1970) entwickelten, parametrischen Test auf das Vorhandensein von Autokorrelationen bis zu einem vorher festgelegten Lag $H \in \mathbb{N}$ in einer Zeitreihe $(x_1, .., x_N)$. Dazu werden die empirischen Autokorrelationen $\hat{\rho}(h)$ mit $h \in \{1, ..., H\}$ herangezogen (s. Kap. 2.1). Typischerweise findet der Box-Pierce-Test Anwendung, um die Residuen nach der Anpassung eines ARMA(p,q)-Prozesses (s. Kap. 2.2) an eine Zeitreihe auf ihre Unabhängigkeit zu überprüfen. So würde eine Ablehnung der Nullhypothese dafür sprechen, dass das vorgeschlagene Modell nicht die gesamte Struktur des Prozesses erfasst und somit auf eine schlechte Anpassung hindeuten. Tests, die auf diese Weise die Anpassungsgüte eines geschätzten Modells überprüfen, werden auch als Portmanteau-Tests bezeichnet.

Die Teststatistik zur Überprüfung der Nullhypothese, dass sich kein $\hat{\rho}(h)$ signifikant von 0 unterscheidet, ist gegeben durch:

$$T_{BP} = N \sum_{h=1}^{H} \hat{\rho}_h^2.$$

Dabei wird gegen die Alternativhypothese getestet, dass es mindestens ein $h \in \{1, ..., H\}$ gibt, für das dies nicht der Fall ist.

Box und Pierce (1970) konnten in ihrer wissenschaftlichen Arbeit zeigen, dass die Teststatistik asymptotisch chiquadratverteilt mit H Freiheitsgeraden ist. Falls es sich um eine Anwendung auf die Residuen eines ARMA(p,q)-Prozesses handelt, so weist die asymptotische Verteilung lediglich (H - p - q) Freiheitsgerade auf. Um diese Asymptotik zu begründen, nutzten Box und Pierce (1970) die von Anderson (1942) gewonnene Erkenntnis, dass die empirischen Autokorrelationskoeffizienten asymptotisch normalverteilt sind und einen Erwartungswert von 0 und eine Varianz von $1/\sqrt{N}$ aufweisen. Anderson (1942) konnte dieses Resultat unter der Voraussetzung beweisen, dass es sich bei $(X_1, ..., X_N)$ um unabhängige identisch-verteilte Zufallsvariablen handelt. Als Summe quadrierter, standardnormalverteilter Zufallsvariablen ist die asymptotische Verteilung von T_{BP} offensichtlich, wobei es sich bei N um einen Skalierungsfaktor handelt.

Allerdings gaben Davies et al. (1977) zu bedenken, dass der Box-Pierce-Test bei kleineren Stichproben eine schlechte Approximation an die Chiquadratverteilung liefert. Insbesondere neigt der Test in Situationen, in denen N im Verhältnis zu H klein ist, dazu, konservativ auszufallen, sodass die Nullhypothese deutlich zu selten verworfen wird. Davies et al. (1977) konnten diese Probleme auf eine deutliche Unterschreitung des Erwartungswertes und der Varianz der Teststatistik von den entsprechenden Werten der asymptotischen Verteilung zurückführen. Eine Ursache dafür sahen sie in Abweichungen von der Normalität der empirischen Autokorrelationskoeffizienten. Dies führte vor allem bei praktischen Anwendungen, in denen meist $H > 15$ zur Erfassung diverser Abhängigkeitsstrukturen gewählt wird, zu Problemen.

Aus diesem Grund schlugen Ljung und Box (1978) die Verwendung einer bereits von Box und Pierce (1970) erwähnten, modifizierten Version der Box-Pierce-Statistik vor, die in Folge dessen

als Ljung-Box-Statistik bekannt wurde. Sie ist definiert als:

$$T_{LB} = N(N+2)\sum_{h=1}^{H}\frac{1}{N-h}\hat{\rho}_h^2.$$

In Simulationsstudien konnten sie zeigen, dass diese Teststatistik, trotz der von Davies et al. (1977) auch für diese Statistik gezeigten Abweichung der Momente von denen der asymptotischen Verteilung, eine beträchtlich bessere Approximation liefert als die herkömmliche Box-Pierce-Statistik. Zumindest für große Stichproben bestätigten Davies und Newbold (1979) die verbesserte Anpassung der Ljung-Box-Statistik in einer Simulationsstudie.

Die kritischen Werte beider Teststatistiken können aus den Quantilen der entsprechenden Chiquadratverteilung abgeleitet werden. Die Nullhypothesen können also zum Niveau α abgelehnt werden, falls

$$T_{BP} \text{ bzw. } T_{LB} > \mathcal{X}_{H-p-q,\,(1-\alpha)},$$

wobei $\mathcal{X}_{H-p-q,\,(1-\alpha)}$ das (1 - α)-Quantil der Chiquadratverteilung mit (H - p - q) Freiheitsgraden darstellt. In dieser Arbeit wird lediglich der Ljung-Box-Test aufgrund seiner überlegenen asymptotischen Eigenschaften und der ansonsten großen Ähnlichkeiten der Teststatistiken betrachtet. Als Anzahl der zu betrachtenden Lags wurde in dieser Arbeit $H = 15$ gewählt.

In `R` können der Box-Pierce- und der Ljung-Box-Test durch die Funktion `box.test` aus dem `stats`-Paket, das zur Basisversion von `R` gehört, angewendet werden. Dabei kann der auszuführende Test über die Parametereinstellung `type` („Box-Pierce“ bzw. „Ljung-Box“) spezifiziert werden. Die Anzahl der zu betrachteten Lags H kann durch den Parameter `lag` geregelt werden. Im Fall, dass die Anpassung eines ARMA(p,q)-Prozesses überprüft werden soll, kann die Anzahl der zu korrigierenden Freiheitsgrade durch den Parameter `df` eingestellt werden.

2.5 Runs-Test

Bei dem Runs-Test handelt es sich um ein von Wald und Wolfowitz (1940) entwickeltes, nichtparametrisches, statistisches Testverfahren zur Überprüfung der Zufälligkeit (hier: Annahme einer unabhängigen, identischen Verteilung der Zufallsvariablen $(X_1, ..., X_N)$) einer Zeitreihe $(x_1, ..., x_N)$. Im Gegensatz zu vielen anderen Verfahren werden dabei nicht die Größen der einzelnen Beobachtungen für die Überprüfung der Unabhängigkeit herangezogen, sondern vielmehr ihr sequenzielles Schema (Madansky, 1988, S.106 ff.).

Der Runs-Test eignet sich für jegliche Beobachtungen, die sich in dichotome Merkmale transformieren lassen. Wird z. B. eine Zeitreihe mit quantitativen Ausprägungen betrachtet, kann eine Dichotomisierung durch die Zentrierung der Beobachtungen mit einem Zentrierungspunkt (z. B. dem Median oder dem arithmetischen Mittel) erreicht werden.

Beim herkömmlichen Runs-Test nach Wald und Wolfowitz (1940) wird zunächst eine Zentrierung der Beobachtungen $(x_1, ..., x_N)$ durch den empirischen Median $x_{\hat{Med}}$ vorgenommen. Eine Voraussetzung dafür ist, dass die Zufallsvariablen des zugrunde liegenden Prozesses denselben theoretischen Median besitzen, das heißt, es muss einen Wert x_{Med} geben, sodass für jedes X_n mit $n \in \{1, ..., N\}$ gilt:

$$P(X_n \leq x_{Med}) = \frac{1}{2} = P(X_n \geq x_{Med}).$$

Eine hinreichende Bedingung dafür wäre beispielsweise, dass die Zufallsvariablen eine identische Verteilung und damit dieselbe Wahrscheinlichkeitsdichtefunktion f_{X_n} besitzen. Abhängig davon, ob eine Beobachtung größer oder kleiner als $x_{\hat{med}}$ ist, wird sie im Folgenden dichotom durch 1 oder 0 codiert.

Dazu wird die geordnete Sequenz $(\tilde{x_1}, ..., \tilde{x_N})$ auf der Grundlage von den Beobachtungen $(x_1, ..., x_N)$ mit $\tilde{x_i} = x_i$ - $x_{\hat{Med}}$ für $i \in \{1, ..., N\}$ berechnet. Im Anschluss werden alle $\tilde{x_i}$, die den Wert 0 annehmen, entfernt und basierend auf $(\tilde{x_1}, ..., \tilde{x_M})$ mit $M \leq N$ wird eine neue Sequenz $(u_1, ..., u_M)$ gebildet, wobei die u_j für $j \in \{1, ..., M\}$ definiert werden als:

$$u_j = \begin{cases} 1, & \text{falls } \tilde{x_j} > 0 \\ 0, & \text{falls } \tilde{x_j} < 0. \end{cases}$$

Bedingt auf die Anzahl der Einsen in dieser Sequenz, die im Folgenden mit n_1 bezeichnet wird, gibt es $\binom{M}{n_1}$ mögliche Anordnungen, die unter der Unabhängigkeitsannahme der Sequenz alle mit derselben Wahrscheinlichkeit auftreten können. Die Anzahl der Nullen in der Sequenz wird im Folgenden mit $n_2 = M$ - n_1 bezeichnet.

Das Maß, das zur Überprüfung der Nullhypothese der Zufälligkeit der Beobachtungen herangezogen wird, ist die Anzahl der sogenannten „Runs“ R. In einem Run werden dabei alle benachbarte Beobachtung mit derselben Ausprägung zusammengefasst und die einzelnen Runs werden durch Runs der jeweils anderen Ausprägung separiert. Auf diese Weise kann die Sequenz $(u_1, ..., u_M)$ in Runs unterteilt werden und zu jeder möglichen Sequenzen lässt sich eine Anzahl

von Runs R berechnen. Formal kann die Anzahl der Runs dieser Sequenz dann ermittelt werden durch:

$$R = 1 + \sum_{j=2}^{M} \mathbb{1}\{u_{j-1} \neq u_j\}.$$

Hier gilt es zu beachten, dass die obige Summe der Anzahl der Vorzeichenwechsel in der Sequenz $(u_1, ..., u_M)$ entspricht. Beispielsweise kommen in der Sequenz

$$\underbrace{00}_{1.}\,\underbrace{111}_{2.}\,\underbrace{0}_{3.}\,\underbrace{11}_{4.}\,\underbrace{000}_{5.}\,\underbrace{1111}_{6.}\,\underbrace{0}_{7.}\,\underbrace{1}_{8.}$$

genau 8 Runs vor, wobei jeweils 4 Runs der Einsen und 4 Runs der Nullen vorhanden sind. Bei Gültigkeit der Nullhypothese würde man nicht zu viele Runs erwarten, da dies auf ein alternierendes Verhalten und z. B. negative Korrelationen hindeuten würde. Andererseits sind aber auch nicht zu wenige Runs zu erwarten, wie es im Fall von positiven Korrelationen der Fall wäre. Insgesamt umfasst die Alternativhypothese ein breites Spektrum von Abhängigkeitsstrukturen, die eine Abweichungen von der Zufälligkeit der Zeitreihe darstellen.

Die Stichprobenverteilung von R kann nach Wang (2003, S. 4 f.) durch einfache kombinatorische Überlegungen ermittelt werden. Dabei gilt, dass im Fall, wenn die beobachtete Anzahl von Runs, die im Folgenden mit r bezeichnet wird, gerade ist, genau $r/2$ Runs von Einsen und $r/2$ Runs von Nullen in der Sequenz vorkommen müssen. Weiter kann die Sequenz $(u_1, ..., u_M)$ entweder mit einer Eins oder einer Null beginnen. Beginnt sie mit einer Eins, so ist es für eine gerade Anzahl von Runs notwendig, dass der letzte Eintrag der Sequenz eine Null ist. Um nun eine Anzahl von $r/2$ Runs der Einsen zu erhalten, müssen die n_1 Einsen jeweils durch Blöcke von Nullen in genau $r/2$ Gruppen unterteilt werden. Die Anzahl der Möglichkeiten, die ($r/2$ - 1) Separatoren auf die (n_1 - 1) Räume zwischen den Einsen zu verteilen, entspricht dabei genau $\binom{n_1-1}{r/2-1}$. Da nun noch berücksichtigt werden muss, dass die einzelnen Runs der Einsen jeweils nicht nur durch alleinstehende Nullen separiert werden können, sondern umgekehrt die Einsen als Separatoren zwischen Runs der Nullen betrachtet werden können, gibt es hier $\binom{n_2-1}{r/2-1}$ Möglichkeiten, die Einsen auf die (n_2 - 1) Zwischenräume zwischen den Nullen zu verteilen. Jede Anordnung der $r/2$ Runs von Einsen mit den $r/2$ Runs der Nullen kann dabei miteinander kombiniert werden. Die Anzahl der Möglichkeiten r Runs zu erhalten, die mit einer 1 beginnen, entspricht demnach $\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}$.

Beginnt die Sequenz anderseits mit einer 0, so endet sie mit einer 1 und die Anzahl der Möglichkeiten r Runs zu erhalten, entspricht mit äquivalenten Überlegungen wieder $\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}$. Insgesamt ergibt sich also für die Wahrscheinlichkeit, eine gerade Anzahl von Runs r in einer Sequenz der Länge $M = n_1 + n_2$ mit n_1 Einsen vorzufinden genau:

$$P\left(R = r \mid \sum_{i=1}^{M} u_i = n_1\right) = \frac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{M}{n_1}}.$$

Im Fall, dass r ungerade ist, können ähnliche Überlegungen angestellt werden. Beginnt die betrachtete Sequenz mit einer 1, so muss sie diesmal auch mit einer 1 enden.

Die Anzahl der Runs von Einsen entspricht demnach dann genau $(r+1)/2$ wobei die restlichen $(r$ - $1)/2$ Runs von Nullen stammen. Die Anzahl der Möglichkeiten, die Nullblöcke zwischen den Runs der Einsen zu verteilen, entspricht dann genau $\binom{n_1-1}{(r-1)/2}$. Andererseits gibt es $\binom{n_2-1}{(r-3)/2}$ Möglichkeiten, die Nullblöcke durch die Einserblöcke zu separieren, sodass es in diesem Fall insgesamt $\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}$ mögliche Sequenzen mit r Runs gibt. Beginnt und endet die Sequenz mit einer 0, so gibt es mit äquivalenten Überlegungen genau $\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}$ mögliche Sequenzen.

Insgesamt ergibt sich also bei einer ungeraden Anzahl von Runs für die Wahrscheinlichkeit, dass bei n_1 Einsen genau r Runs auftreten:

$$P\left(R = r \mid \sum_{i=1}^{M} U_i = n_1\right) = \frac{\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}} + \binom{n_1-1}{\frac{r-3}{2}}\binom{n_2-1}{\frac{r-1}{2}}}{\binom{M}{n_1}}.$$

Im Fall, dass sowohl n_1 als auch n_2 groß sind, kann eine Approximation von R durch eine Normalverteilung mit einem Mittelwert von $\mu_R = 2n_1n_2/M + 1$ und mit einer Varianz von $\sigma_R^2 = 2n_1n_2(2n_1n_2$ - $M)/M^2(M$ - $1)$ vorgenommen werden (Wald und Wolfowitz, 1940). Dies folgt mit der Konvergenz der Binomialverteilung gegen die Normalverteilung, wobei Q definiert ist als $(2 \cdot \min\{n_1, n_2\} + 1)$ und sich der entsprechende bedingte Erwartungswert und die Varianz von R über die Summendarstellung berechnen lassen durch:

$$E\left[R \mid \sum_{i=1}^{M} u_i = n_1\right] = 1 + \sum_{r=2}^{Q} r \cdot P\left(R = r \mid \sum_{i=1}^{M} u_i = n_1\right)$$

$$Var\left[R \mid \sum_{i=1}^{M} u_i = n_1\right] = 1 + \sum_{r=2}^{Q}\left(r - E\left[R \mid \sum_{i=1}^{M} u_i = n_1\right]\right)^2 \cdot P\left(R = r \mid \sum_{i=1}^{M} u_i = n_1\right),$$

falls $n_1 \neq n_2$ und sonst:

$$E\left[R \mid \sum_{i=1}^{M} u_i = n_1\right] = \sum_{r=2}^{M} r \cdot P\left(R = r \mid \sum_{i=1}^{M} u_i = n_1\right)$$

$$Var\left[R \mid \sum_{i=1}^{M} u_i = n_1\right] = \sum_{r=2}^{M}\left(r - E\left[R \mid \sum_{i=1}^{M} u_i = n_1\right]\right)^2 \cdot P\left(R = r \mid \sum_{i=1}^{M} u_i = n_1\right).$$

Dabei wird für den Fall, dass $n_1 \neq n_2$ gilt, der entsprechende Summand für $r = 1$ aus der Summe gezogen, da der dazugehörige Fall implizieren würde, dass entweder $n_1 = 0$ oder $n_2 = 0$ gilt. Die Wahrscheinlichkeit, dann nur einen Run zu erhalten, entspricht 1.

Die Nullhypothese der Zufälligkeit der Beobachtungen kann somit zu einem Signifikanzniveau von α abgelehnt werden, falls gilt:

$$\frac{|R - \mu_R|}{\sigma_R} \geq z_{1-\alpha/2}.$$

Dabei handelt es sich bei $z_{1-\alpha/2}$ um das (1 - $\alpha/2$)-Quantil der Standardnormalverteilung, für das aufgrund der Symmetrie der Normalverteilung gilt:

$$P(Z \geq z_{1-\alpha/2}) = \frac{\alpha}{2} = P(Z \leq -z_{1-\alpha/2}),$$

wobei Z eine standardnormalverteile Zufallsvariable ist.

In `R` kann der Runs-Test mit der Funktion `runs.test` aus dem Paket `tseries` von Trapletti und Hornik (2019) für eine dichotome Zeitreihe durchgeführt werden. Als Alternativhypothesen können dabei einseitige Hypothesen (`less` oder `greater`) oder die zweiseitige Hypothese (`two.sided`) über die Parametereinstellung `alternative` eingestellt werden.

2.6 Turning-Point-Test (TP-Test)

Der Turning-Point-Test stellt einen Test auf die Zufälligkeit (also die Annahme einer unabhängigen identischen Verteilung) von Zufallsvariablen $(X_1, ..., X_N)$ mit kontinuierlichen Ausprägungen dar und wurde erstmals von Bienaymé (1873) beschrieben. Kendall (1973) schlägt den Test dabei als eine einfach zu berechnende Alternative zu statistischen Tests vor, die auf den Autokovarianzen einer Zeitreihe beruhen (s. Kap. 2.3 u. 2.4).

Eine Voraussetzung für die Anwendung des Testes ist, dass zwei beliebige Beobachtungen der Zufallsvariablen x_i und x_j P-fast sicher paarweise verschieden sind, das heißt, es muss gelten: $P(X_i \neq X_j) = 1$ für alle $i \neq j \in \{1, ..., N\}$. Diese Voraussetzung ist beispielsweise erfüllt, wenn die Zufallsvariablen $(X_1, ..., X_N)$ eine stetige Wahrscheinlichkeitsverteilung besitzen.

Die wesentliche Idee des Turning-Point-Tests besteht darin, die Anzahl der Maxima und Minima T in einer Zeitreihe $(x_1, ..., x_N)$, die im Kontext dieses Tests als Turning-Points bezeichnet werden, als Teststatistik heranzuziehen. Ein Turning-Point liegt dabei zum Zeitpunkt $i \in \{2, ..., N \text{ - } 1\}$ genau dann vor, wenn entweder $x_{i-1} < x_i$ und $x_{i+1} < x_i$ oder aber $x_{i-1} > x_i$ und $x_{i+1} > x_i$ gilt (Brockwell und Davis, 2010, S. 36 f.). Unter der Nullhypothese, dass es sich bei $(x_1, ..., x_N)$ um eine durch unabhängige und identisch verteilte und kontinuierliche Zufallsvariablen erzeugte Zeitreihe handelt, kommt jede der 6 möglichen Anordnungen von 3 aufeinander folgenden Beobachtungen mit der gleichen Wahrscheinlichkeit vor, wobei 4 dieser Anordnungen einem Turning-Point entsprechen (s. Abb. 2.1).

Die Tatsache, dass jede mögliche Anordnung gleich wahrscheinlich ist, kann durch die folgenden Überlegungen nachvollzogen werden: π_1, π_2 und π_3 seien dafür beliebige Permutationen von $\{1, 2, 3\}$. Handelt es sich bei (X_1, X_2, X_3) um unabhängige und identisch verteilte Zufallsvariablen, dann folgt, dass $(X_{\pi_1}, X_{\pi_2}, X_{\pi_3})$ ebenfalls unabhängig und identisch verteilt sind und damit insbesondere dieselbe Verteilung besitzen. Das bedeutet, dass für ein beliebiges Ereignis $A \in \Sigma$ die Beziehung:

$$P((X_1, X_2, X_3) \in A) = P((X_{\pi_1}, X_{\pi_2}, X_{\pi_3}) \in A)$$

gelten muss. Wird A als eine beliebige Ordnungsstatistik definiert (z. B. die eines Minimums oder Maximums), so folgt mit der Voraussetzung, dass 2 Zufallsvariablen nicht denselben Wert annehmen können, dass alle Permutationen einer unabhängigen und identisch verteilten Sequenz von Zufallsvariablen gleich wahrscheinlich sind. Konkret beträgt diese Wahrscheinlichkeit in dem hier betrachteten Fall 1/3!.

Mit obigen Überlegungen zum Auftreten eines Turning-Points und der Definition der Zufallsvariablen T_i für $1 < i < N$ als:

$$T_i = \begin{cases} 1 & \text{, falls zum Zeitpunkt } i \text{ ein Turning-Point vorliegt} \\ 0 & \text{, sonst,} \end{cases}$$

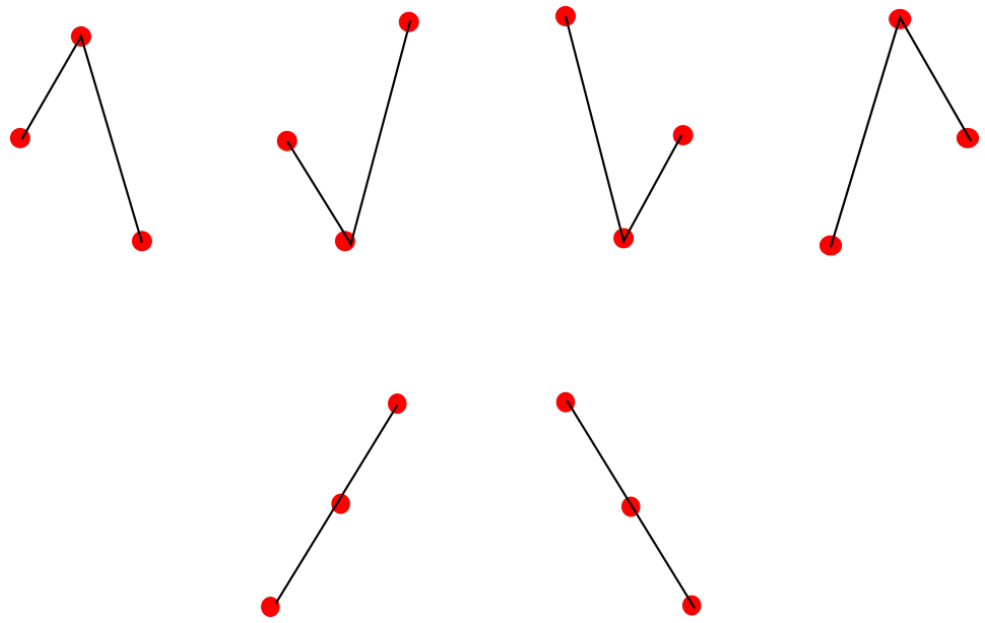


Abbildung 2.1: Visualisierung der möglichen Anordnungen von drei aufeinanderfolgenden Beobachtungen mit Turning-Point (oben) und ohne (unten)

gilt beispielhaft für $i = 2$: $E(T_2) = 4/3! = 2/3$. Für den Erwartungswert der Anzahl der Turning-Points $T = \sum_{i=2}^{N-1} T_i$ gilt deshalb:

$$E(T) = E\left(\sum_{i=2}^{N-1} T_i\right) = (N-2)E(T_2) = \frac{2(N-2)}{3}.$$

Die Varianz $Var(T)$ lässt sich ebenfalls mit einigen Überlegungen berechnen, wobei die Korrelation benachbarter Beobachtungen berücksichtigt werden muss. Für den Fall, dass $i = j$ ist, ergibt sich einfach $E(T_iT_j) = E(T_i)$.

Gilt wiederum $|i - j| = 1$, so gibt es genau 4! verschiedene Anordnungen von x_i, x_j und ihren zwei benachbarten Beobachtungen. Damit sowohl $T_i = 1$ als auch $T_j = 1$ gilt, muss die Sequenz alternierend sein. Solche alternierenden Permutationen treten bei genau 10 der 4! möglichen Anordnungen auf (s. Stanley, 2009). Somit ergibt sich hier $E(T_iT_j) = \frac{10}{4!} = \frac{5}{12}$.

Weiter ist der Fall zu betrachten, in dem $|i - j| = 2$ gilt. Hier gibt es 5! mögliche Anordnungen von x_i und x_j, den zwei außen liegenden, sowie der dazwischen liegenden Beobachtung. Damit hier $T_i = 1$ und $T_j = 1$ sein können, muss es sich entweder wieder um eine alternierende Permutation handeln, von denen es nach Stanley (2009) genau 32 gibt, oder die Beobachtung zwischen den beiden betrachteten Beobachtungen ist kein Turning-Point. Dann existieren sowohl für den Fall, dass $x_j < x_{j+1} < x_i$ ist, als auch für den Fall, dass $x_j > x_{j+1} > x_i$ ist jeweils 11 weitere Möglichkeiten, in denen bei x_i und x_j ein Turning-Point auftritt.

Insgesamt gibt es also 54 Anordnungen, bei denen $T_i = T_j = 1$ gilt. Somit gilt in diesem Fall letztendlich $E(T_iT_j) = \frac{54}{5!} = \frac{9}{20}$.

Im Fall, dass $|i - j| > 2$ ist, sind T_i und T_j unkorreliert und es gilt $E(T_iT_j) = E(T_i)^2 = \frac{4}{9}$.

Mit diesen Erkenntnissen, lässt sich die Varianz durch die Indikatorfunktionen berechnen als:

$$\begin{aligned}
\text{Var}[T] &= \text{Var}\left[\sum_{i=1}^{N} T_i\right] = \sum_{i=1}^{N} \text{Var}\left[T_i\right] + 2\sum_{i<j} \text{Cov}\left[T_i, T_j\right] \\
&= \sum_{i=1}^{N} \left(E\left[T_i^2\right] - E\left[T_i\right]^2\right) + 2\sum_{i<j} \left(E\left[T_i T_j\right] - E\left[T_i\right] E\left[T_j\right]\right) \\
&= \sum_{i=2}^{N-1} \left(\frac{2}{3} - \frac{4}{9}\right) + 2\sum_{i<j} \left(E\left[T_i T_j\right] - \frac{4}{9}\right) \\
&= (N-2)\cdot\frac{2}{9} + 2(N-3)\cdot\left(\frac{5}{12} - \frac{4}{9}\right) + 2(N-4)\cdot\left(\frac{9}{20} - \frac{4}{9}\right) \\
&= \frac{16N-29}{90}.
\end{aligned}$$

Nach Brockwell und Davis (2010, S. 37) ist die Anzahl der Turning-Points T approximativ normalverteilt mit einem Erwartungswert von $\mu_T = \frac{2(N-2)}{3}$ und der Varianz $\sigma_T^2 = \frac{16N-29}{90}$.

Ist T deutlich kleiner als μ_T, so deutet dies auf positive Korrelationen zwischen den Werten hin. Falls T deutlich größer ist, so ist dies ein Indiz für negative Korrelationen. Auf diese Weise kann die Nullhypothese der Zufälligkeit der Zeitreihe zum Niveau α abgelehnt werden, falls

$$\frac{|T - \mu_T|}{\sigma_T} \geq z_{1-\alpha/2}$$

gilt, wobei $z_{1-\alpha/2}$ wie in Kapitel 2.5 definiert ist.

In `R` kann der Turning-Point-Test mit der Funktion `turningpoint.test` aus dem Paket `spgs` von Hart und Martínez (2019) für einen numerischen Vektor oder eine univariate Zeitreihe durchgeführt werden.

2.7 Von-Neumann-Ratio-Rang-Test (VNRR-Test)

Der Von-Neumann-Ratio-Rang-Test stellt einen nichtparametrischen Test für die Zufälligkeit einer numerischen Zeitreihe $(x_1, ..., x_N)$ dar. Er basiert auf einer Rangversion der von von Neumann (1941) eingeführten Von-Neumann-Ratio (VNR). Sie ist definiert als:

$$VNR = \frac{\sum_{i=1}^{N-1} (x_{i+1} - x_i)^2 / (N-1)}{\sum_{i=1}^{N} (x_i - \bar{x})^2 / N}$$

und beschreibt das Verhältnis der mittleren quadrierten Differenzen aufeinanderfolgender Beobachtungen einer Zeitreihe zu ihrer Varianz. Dabei handelt es sich bei $\bar{x}$ um das arithmetische Mittel der Zeitreihe. Die Verteilung der VNR wurde von Hart und von Neumann (1942) unter der Annahme, dass es sich bei $(X_1, ..., X_N)$ um einen Gauß-Prozess handelt, tabelliert. Da die Approximation dieser Verteilung nach Bartels (1982) jedoch sehr anfällig für Abweichungen von der Normalitätsannahme ist, führte er eine rangbasierte und damit nichtparametrische Version der VNR ein, die definiert ist als:

$$T_{VNRR} = \frac{\sum_{i=1}^{N-1} (R_{i+1} - R_i)^2}{\sum_{i=1}^{N} \left(R_i - \bar{R}\right)^2}.$$

Dabei entspricht R_i dem Rang der i-ten Beobachtung und der Nenner dieser Statistik ist unter der Voraussetzung, dass keine Bindungen der Beobachtungen vorliegen, konstant. Beispielhaft gilt mit $\bar{R} = \frac{N+1}{2}$ unter Verwendung von Summenformeln sowie der Vertauschbarkeit der Ränge bei der Summierung:

$$\begin{aligned}
\sum_{i=1}^{N} \left(R_i - \bar{R}\right)^2 &= \sum_{i=1}^{N} i^2 + 2\left(\frac{N+1}{2}\right) \sum_{i=1}^{N} i + N\left(\frac{N+1}{2}\right)^2 \\
&= \frac{N(N+1)(2N+1)}{6} - \frac{2(N+1)N(N+1)}{4} + \frac{N(N^2+2N+1)}{4} \\
&= \frac{4N^3+6N^2+2N}{12} - \frac{3(N^3+2N^2+N)}{12} = \frac{N^3-N}{12}.
\end{aligned}$$

Aus diesem Grund wird häufig lediglich der Zähler $\left(T_{NM} = \sum_{i=1}^{N-1} (R_{i+1} \text{ - } R_i)^2\right)$ als Teststatistik herangezogen. Dieser kann Werte von (N - 1), falls die Ränge monoton fallend bzw. wachsend sind, bis maximal $\frac{1}{3}(N \text{ - } 1)(N^2 + N \text{ - } 3)$ annehmen, falls N gerade ist bzw. $\frac{1}{3}(N \text{ - } 1)(N^2 + N \text{ - } 3)$ - 1, falls es sich bei N um eine ungerade Zahl handelt. Die Werte von der VNRR bewegen sich dementsprechend asymptotisch zwischen 0 und 4, wie z. B. auch die der Durbin-Watson Statistik (s. Kap. 2.3). Bartels (1982) konnte weiterhin zeigen, dass

$$\frac{T_{VNRR} - 2}{4(N-2)(5N^2-2N-9)/(5N(N+1)(N-1)^2)}$$

asymptotisch standardnormalverteilt ist. Dafür berechnete er die Momente der Verteilung in Anlehnung an die von Young (1941) veröffentlichte wissenschaftliche Arbeit zur Bestimmung von den Momenten der herkömmlichen VNR.

In der Hoffnung, dass die Von-Neumann-Rang-Ratio als robustes Verfahren zur Erkennung von Abhängigkeiten mehr Anwendung in der Praxis finden würde, tabellierte Bartels (1982) ihre kritischen Werte (für $\alpha = 0.005, 0.01, 0.025, 0.05, 0.01$) für Stichprobenumfänge, in denen die asymptotische Normalität noch nicht greift. In kleinen Stichproben ($4 \leq N \leq 10$) wurden dabei die exakten Quantile anhand der gesamten Menge der Permutationen berechnet, für die jeweils die Teststatistik T_{NM} berechnet wurde. Für Stichprobengrößen zwischen 10 und 100 wurde die Verteilung der Teststatistik T_{VNRR} durch die Verteilung einer Beta-Verteilung auf dem Intervall $0 \leq x \leq 4$ mit der Dichtefunktion

$$f(x; p, q) = \frac{1}{B(p, q)} x^{p-1}(4 - x)^{q-1}/4^{p+q-1}$$

mit

$$B(p, q) = \int_0^1 u^{p-1}(1 - u)^{q-1}\, du$$

approximiert. Da der Erwartungswert einer solchen Beta-Verteilung $4p/(p + q)$ und die Varianz $16pq/(p + q)^2(p + q + 1)$ entspricht, führt eine Anpassung an die Momente von T_{VNRR} zu einer Wahl von $p = q = 5N(N + 1)(N - 1)^2/(2(N - 2))(5N^2 - 2N - 9) - \frac{1}{2}$ (s. Bartels, 1982). Für Stichprobenumfänge von $N > 100$ ist die Approximation der Normalverteilung hinreichend gut, sodass die entsprechenden kritischen Werte aus deren Quantilen entnommen werden können.

Bartels (1982) konnte zeigen, dass die Beta-Approximation eine sehr gute Anpassung an die exakte Verteilung liefert. So beträgt der berechnete Fehler der Approximation zum exakten Wert bei $N = 10$ maximal 0.001, wohingegen er bei der Normalverteilung bis zu 0.006 beträgt. Auch ein Vergleich der exakten kritischen Werte für Stichprobenumfänge von $N = 25$ und $N = 50$, mit den aus der Beta-Verteilung abgeleiteten Werten zeigte, dass sie in den meisten Fällen bis auf 2 Nachkommastellen übereinstimmten.

Im Fall, dass Bindungen in der Zeitreihe $(x_1, ..., x_N)$ auftreten, schlägt Bartels (1982) vor, gemittelte Ränge zu vergeben. Dadurch wird zwar nicht die asymptotische Normalität beeinflusst, wohl aber das Verhältnis zwischen Zähler und Nenner der VNRR. So ist insbesondere der Nenner der Teststatistik nicht mehr konstant. Aus diesem Grund sind die von Bartels (1982) bestimmten Quantile nicht mehr korrekt, können jedoch bei wenigen Bindungen trotzdem eine gute Approximation der kritischen Bereiche liefern.

In `R` kann der VNRR-Test mithilfe der Funktion `bartels.rank.test` aus dem Paket `randtests` von Caeiro und Mateus (2014) für einen numerischen Vektor von Beobachtungen durchgeführt werden. Die Alternativen (`two.sided`, `left.sided`, `right.sided`) können dabei über den Parameter `alternative` eingestellt werden. Die Art der Ermittlung der kritischen Werte kann über den Parameter `p.value` eingestellt werden. Als Möglichkeiten stehen die Approximation durch eine Normalverteilung (`normal`), eine Beta-Verteilung (`beta`) oder die exakten Werte (`exakt`) zur Verfügung. Aufgrund des Rechenaufwandes wird jedoch bei Stichprobengrößen > 10 von einer Berechnung der exakten Werte abgeraten.

2.8 Broock-Dechert-Schreinkman-Test (BDS-Test)

Beim Broock-Dechert-Schreinkman-Test handelt es sich um ein von Broock et al. (1996) vorgestelltes, nichtparametrisches Testverfahren zur Aufdeckung linearer sowie nichtlinearer Abhängigkeiten in einer Zeitreihe $(x_1, ..., x_N)$. Typischerweise dient es zur Analyse der Residuen eines angepassten statistischen Modells und damit als Kriterium für dessen Anpassungsgüte. Der Anwendungsbereich ähnelt damit z. B. dem des Durbin-Watson- (s. Kap. 2.3) und des Ljung-Box-Tests (s. Kap. 2.4). Ein wesentlicher Vorteil gegenüber anderen modelldiagnostischen Tests besteht in der universellen Anwendbarkeit beim Vorliegen diverser Abhängigkeitsstrukturen, die mit vielen herkömmlichen Tests nicht erfasst werden können (s. z. B. Broock et al., 1993). Eine Studie von Brooks (1999) bestätigte dabei zumindest für große Stichproben die gute Trennschärfe des Tests bei nichtlinearen Abhängigkeitsstrukturen.

Die Teststatistik des BDS-Tests beruht auf dem sogenannten Korrelationsintegral. Für die Definition dieses Integrals wird eine Familie unabhängiger, identisch verteilter Zufallsvariablen $(X_1, ..., X_N)$ mit zugehörigen Realisierungen $(x_1, ..., x_N)$ betrachtet. Weiter sei $I_\epsilon(x, y)$ für $x, y, \epsilon \in \mathbb{R}$ eine Indikatorfunktion, die definiert ist als:

$$I_\epsilon(x,y) = \begin{cases} 1 & \text{, falls } |x-y| < \epsilon \\ 0 & \text{, sonst.} \end{cases}$$

Die Indikatorfunktion zeigt also an, ob die beiden Werte x und y weiter als ϵ voneinander entfernt liegen. Das Korrelationsintegral ist damit definiert als:

$$c_{m,N}(\epsilon) = \frac{2}{(N-m+1)(N-m)} \sum_{s=1}^{N-m+1} \sum_{t=s+1}^{N-m+1} \prod_{j=0}^{m-1} I_\epsilon(x_{s+j}, x_{t+j}).$$

Mit seiner Hilfe wird also die Wahrscheinlichkeit geschätzt, dass 2 sogenannte m-Historien $(x_i, x_{i+1}, ..., x_{i+(m-1)})$ mit $i \in \{1, ..., N \text{ - } m+1\}$ einer Zeitreihe $(x_1, ..., x_N)$ bezüglich ihrer Supremumsnorm nicht weiter als ϵ auseinanderliegen. Der Parameter m wird dabei als Einbettungsdimension bezeichnet und der Vorfaktor des Termes entspricht dem Inversen der Anzahl der betrachteten m-Historien.

Die Grundidee des BDS-Tests im Fall $m = 2$ stellt dabei die Tatsache dar, dass unter der Nullhypothese der Unabhängigkeit der Zufallsvariablen $(X_1, ..., X_N)$ für die Erwartungswerte

$$E[I_\epsilon(X_t, X_s)] = P(I_\epsilon(X_t, X_s) = 1) \quad \text{und}$$
$$E[I_\epsilon(X_t, X_s) I_\epsilon(X_{t+1}, X_{s+1})] = P(I_\epsilon(X_t, X_s) = 1,\ I_\epsilon(X_{t+1}, X_{s+1}) = 1)$$

die Beziehung $E[I_\epsilon(X_t, X_s) I_\epsilon(X_{t+1}, X_{s+1})] = E[I_\epsilon(X_t, X_s)]^2$ für beliebige $s \neq \{t, (t+1)\} \in \{1, ..., N \text{ - } 1\}$ gilt. Für den Fall $s = t + 1$ kann allerdings keine derartige Aussage getroffen werden, da dafür die Unabhängigkeit der Zuwächse in der Zeitreihe gegeben sein müsste. Diese Diskrepanz könnte ein Grund dafür sein, dass die BDS-Statistik relativ langsam konvergiert,

sodass typischerweise große Beobachtungsumfänge gewählt werden müssen, damit der BDS-Test das angepeilte Signifikanzniveau einhalten kann. So wird der Anteil von 2 2-Historien mit $s = t + 1$ mit wachsendem Stichprobenumfang immer geringer.

Heuristisch soll also die Wahrscheinlichkeit, dass 2 m-Historien einer Zeitreihe einen größeren Abstand als ϵ zueinander aufweisen, nicht von ihrer Position in der Zeitreihe abhängen. Mit diesen Überlegungen kann geschlussfolgert werden, dass unter der Unabhängigkeit für das Korrelationsintegral zumindest annäherungsweise die Beziehung

$$E[c_{m,N}(\epsilon)] = E[c_{1,N}(\epsilon)]^m$$

gelten sollte (LeBaron, 1997). So gilt nach Caporale et al. (2004) für ein fixes $\epsilon > 0$ mit der Wahrscheinlichkeit 1: $c_{m,N}(\epsilon) \to c_{1,N}(\epsilon)^m$ für $N \to \infty$. Weiter konnten Broock et al. (1996) unter der Voraussetzung, dass der zugrunde liegende Prozess stationär und absolut regulär ist, zeigen, dass die Teststatistik

$$BDS(m, N, \epsilon) = \sqrt{N} \frac{c_{m,N}(\epsilon) - c_{1,N}^m(\epsilon)}{\sigma_{m,N}(\epsilon)}$$

asymptotisch standardnormalverteilt ist.

Bei der absoluten Regularität handelt es sich um ein von Volkonskii und Rozanov (1959) vorgestelltes Konzept, das Abhängigkeitsstrukturen in stationären Prozessen beschreibt. Ein Problem mit der stochastischen Unabhängigkeit, die ein ähnliches Konzept beschreibt, ist, dass es sich bei ihr um eine zu starke Eigenschaft handelt, deren Annahme vor allem in realen Anwendungen selten gerechtfertigt ist. Heuristisch gesehen stellt die absolute Regularität also eine schwächere Eigenschaft als die stochastische Unabhängigkeit dar, die es erlaubt, trotzdem theoretische Erkenntnisse über die Zeitreihe zu gewinnen und asymptotische Resultate zu erhalten (s. z. B. Wendler, 2011, S. 17 ff.).

Eine Testentscheidung kann somit anhand der kritischen Werte dieser Verteilung gefällt werden (s. z. B. Kap. 2.5). Die Varianz von $c_{m,N}(\epsilon)$ - $c_{1,N}^m(\epsilon)$ kann dabei nach LeBaron (1997) und Kim et al. (2003) konsistent geschätzt werden durch

$$\sigma_{m,N}^2(\epsilon) = 4 \left[k^m + 2 \sum_{j=1}^{m-1} k^{m-j} c^{2j} + (m-1)^2 c^{2m} - m^2 k c^{2m-2} \right],$$

wobei $c = E(c_{1,N}(\epsilon))$ konsistent durch $c_{1,N}(\epsilon)$ geschätzt werden kann. Die Konstante k kann weiterhin geschätzt werden durch:

$$k_N(\epsilon) = \frac{6}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{s=t+1}^{N} \sum_{r=s+1}^{N} b_\epsilon \left(X_t, X_s, X_r \right),$$

wobei gilt:

$$b_\epsilon(i, j, k) = \frac{1}{3} \left(I_\epsilon(i, j) I_\epsilon(j, k) + I_\epsilon(i, k) I_\epsilon(k, j) + I_\epsilon(j, i) I_\epsilon(i, k) \right).$$

Für die Wahl des Schwellenwertparameters ϵ und der Einbettungsdimension m orientieren sich viele Anwender an den von Broock et al. (1993) vorgeschlagenen Richtlinien (s. z. B. Kim et al., 2003). Dabei legen sie nahe, den BDS-Test bei Stichprobenumfängen von $N > 500$ zu verwenden und Einbettungsdimensionen $2 \leq m \leq 5$ zu betrachten. Als Wahl für die Schwellenwertdistanz ϵ wird die Hälfte der empirischen Standardabweichung der Beobachtungen vorgeschlagen.

In `R` kann der BDS-Test durch die Funktion `bds.test` aus dem Paket `tseries` von Trapletti und Hornik (2019) für einen numerischen Datenvektor oder eine Zeitreihe angewendet werden. Die Einbettungsdimension kann durch den Parameter `m` und die Schwellenwertdistanz durch den Parameter `eps` eingestellt werden. Als `default`-Wert wird dabei `m` $= 3$ gewählt und `eps` wird als verschiedene Vielfache (0.5, 1, 1.5 u. 2) der Standardabweichung der angegebenen Zeitreihe gesetzt, wobei für jede Option ein p-Wert ausgegeben wird. In dieser Arbeit wird der Test mit `m` $= 3$ und `eps` als die Hälfte der Standardabweichung durchgeführt.

2.9 K-Vorzeichentiefetests (K-VZ-Tests)

Die K-Vorzeichentiefetests stellen eine Reihe von nichtparametrischen und robusten statistischen Tests dar, die angewendet werden können, um die Residuen eines Modells auf Unabhängigkeit zu überprüfen und gleichzeitig sicherzustellen, dass ihr theoretischer Median 0 entspricht. Sind diese Eigenschaften erfüllt, so spricht dies für die Anpassungsgüte des aufgestellten Modells und insbesondere für die korrekte Spezifikation des Modellparameters.

Erstmals wurde diese Art von Hypothesentests von Leckey et al. (2020) beschrieben. Sie stellen eine Generalisierung des herkömmlichen Vorzeichentests dar, mit dem dieselbe Hypothese überprüft werden kann. Insbesondere kann gezeigt werden, dass der Vorzeichentest und der K-VZ-Test mit $K = 2$ asymptotisch äquivalent sind. Leckey et al. (2020) haben in Simulationsstudien zeigen können, dass die allgemeinen K-VZ-Tests mit $K \geq 3$ eine deutlich bessere Trennschärfe bei Alternativen besitzen, in denen der herkömmliche Vorzeichentests bei der Ablehnung der Nullhypothese scheitert.

Für diese Tests wird typischerweise ein stochastisches Modell mit unbekanntem Modellparameter $\theta \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$ der Form $y_n = g(x_n, \theta) + E_n$ mit $n \in \{1, ..., N\}$ betrachtet. Dabei handelt es sich bei $y_1, ..., y_N$ um Beobachtungen, $g(\bullet, \theta) : \mathbb{R} \to \mathbb{R}$ repräsentiert eine Regressionsfunktion in Abhängigkeit eines Regressorenvektors x_n^p und E_n bezeichnet additive Fehler. Weitere Beispiele, in denen die Tests Anwendung finden können, sind stochastische Prozesse wie AR(p)-Prozesse, die über den Zusammenhang $y_n = g(y_{n-1}, ..., y_{n-p}, \theta) + E_n$ modelliert werden können. Dabei wird stets vorausgesetzt, dass die Fehler unabhängig sind und die Voraussetzung

$$P_\theta(E_n > 0) = \frac{1}{2} = P_\theta(E_n < 0) \tag{2.1}$$

für $n \in \{1, ..., N\}$ erfüllen. Insbesondere bedeutet dies, dass die Wahrscheinlichkeit, dass ein Fehler den Wert 0 annimmt, also $P(E_n = 0)$, P-fast sicher 0 ist. Diese Voraussetzung wird z. B. von stetigen Fehlerverteilungen mit einem Median von 0 erfüllt. Falls derartige Modelle vorliegen, sind die aus ihnen resultierenden Residuen definiert als $R_n(\theta) = y_n$ - $g(x_n, \theta)$ oder im AR(p)-Fall als $R_n(\theta) = y_n$ - $g(y_{n-1}, ..., y_{n-p}, \theta)$. Handelt es sich bei θ um den wahren Modellparameter, so reduzieren sich die Residuen zu den oben erwähnten Fehlern und erfüllen somit ebenfalls die Eigenschaft 2.1.

In dem Zusammenhang fällt auf, dass die K-VZ-Tests stark abhängig von der Ordnung der Residuen ist. Im Fall von Zeitreihen ist dabei eine natürliche Ordnung vorgegeben. Auch im Fall einer einfachen Regression mit lediglich einem Regressor stellt die kanonische Ordnung der Residuen einen sinnvollen Ansatz dar. Schwieriger wird die Wahl einer solchen Ordnung im Fall einer multiplen Regression. Im Detail wird diese Problemstellung von Horn und Müller (2020) diskutiert. Auf der Grundlage von Simulationsstudien zeigte sich, dass der kürzeste Hammiltonpfad eine gute Wahl zur Ordnung der Residuen ist.

Historie

Allgemein geht das Konzept der Datentiefe auf die von Tukey (1975) eingeführte Halbraumtiefe zurück, die von der Verallgemeinerung des Medians auf einen multivariaten Datensatz motiviert wurde. Dabei wird die Halbraumtiefe eines Parameters $\mu \in \mathbb{R}^k$ in einem k-dimensionalen Datensatz als die minimale relative Anzahl derjenigen Punkte beschrieben, die in einem Halbraum liegen, der μ enthält. Der multivariate Median entspricht dabei den Datenpunkten mit der maximalen Halbraumtiefe.

Ein alternatives Konzept der Datentiefe wurde später durch Liu (1990) mit der Simplextiefe eingeführt, die für einen Parameter $\mu \in \mathbb{R}^k$ in einem k-dimensionalen Datensatz den relativen Anteil der durch $(k+1)$ Datenpunkte aufgespannten Simplizes entspricht, die μ enthalten.

Das Konzepte der Halbraumtiefe wurde dann von Rousseeuw und Hubert (1999) auf Regressionen übertragen, wodurch die sogenannte Regressionstiefe eingeführt wurde. Sie wurde als ausreißerrobustes Maß für die Anpassungsgüte eines Modells zu einem Parametervektor θ über das Konzept eines „Nonfits“ definiert. Im Kontext einer multiplen Regression (s. Kap. 2.2) wird der Parametervektor θ als Nonfit bezeichnet, falls eine affine Hyperebene V im Raum der Regressoren existiert, sodass V kein $x_i \in \mathbb{R}^p$ enthält. Weiterhin muss für alle x_i in einem der von V erzeugten Halbräume $r_i(\theta) > 0$ gelten und für alle x_i im anderen Halbraum $r_i(\theta) < 0$. Dies ist genau dann der Fall, wenn es einen anderen Parameter $\hat{\theta}$ gibt, sodass für die aus ihm resultierenden Residuen gilt:

$$|r_i(\hat{\theta})| \leq |r_i(\theta)| \quad \forall\, i \in \{1, ..., N\}.$$

Die Regressionstiefe eines Parametervektors in Bezug auf einen Datensatz entspricht dann der minimalen Anzahl an Beobachtungen, die entfernt werden muss, damit θ zum Nonfit wird.

Analog kann neben der Erweiterung der Halbraumtiefe auf die Regression die von Liu (1990) vorgestellte Simplextiefe verwendet werden. Dieses ebenfalls von Rousseeuw und Hubert (1999) vorgestellte Konzept, das auch als Simplex-Regressionstiefe bezeichnet wird, eignet sich nach Müller (2005) für robuste Testverfahren an Modellparametern eines Regressionsmodells. Kustosz et al. (2016b) konnten dabei zeigen, dass sich die Berechnung dieser Tiefe in vielen Fällen dahingehend vereinfachen lässt, dass lediglich diejenigen K-Tupel von geordneten Residuen gezählt werden, bei denen alternierende Vorzeichen auftreten. Diese Erkenntnis motivierten Leckey et al. (2020) dazu, die K-Vorzeichentiefe als die relative Anzahl der K-Tupel mit alternierenden Vorzeichen zu definieren, woraus im Endeffekt die K-VZ-Tests konstruiert wurden.

Teststatistik und Testentscheidung

Die K-Vorzeichentiefe eines Residuenvektors $(r_1(\theta), ..., r_N(\theta)) = (r_1, ..., r_N)$, die in den folgenden Ausführungen mit $d_K(r_1, ..., r_N)$ bezeichnet wird, entspricht dann der relativen Anzahl K-elementiger Teilmengen mit alternierenden Vorzeichen innerhalb des Residuenvektors.

Formal ist die K-Vorzeichentiefe für $K \geq 2$ definiert als:

$$d_K(r_1, ..., r_N) := \frac{1}{\binom{N}{K}} \sum_{1 \leq n_1 < n_2 < ... < n_K \leq N} \left(\prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \right).$$

In Simulationsstudien von Kustosz et al. (2016a) wies die K-Vorzeichentiefe im Fall von mit Ausreißern kontaminierten Modellfehlern robustes Verhalten auf, sodass nahe liegt, dass es sich bei ihr um eine ausreißerrobuste Kenngröße handelt. Im Hinblick auf die Tatsache, dass lediglich die Vorzeichen der Residuen zu ihrer Berechnung herangezogen werden, ist diese Eigenschaft auch heuristisch nachvollziehbar.

Die auf der K-Vorzeichentiefe basierende Teststatistik wird als $T_K(\theta)$ bezeichnet und ist definiert als:

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

Die Form dieser Teststatistik ermöglicht dabei die Berechnung ihrer asymptotischen Verteilung, aus der ihre asymptotischen Quantile abgeleitet werden können. Bei einer korrekten Spezifikation des Modells mit wahrem Parameter θ ist davon auszugehen, dass die Residuen unabhängig sind und einen Median von 0 aufweisen. In der zweiseitigen Version des Tests soll deshalb die Nullhypothese, dass der Parametervektor θ zur Menge der dem „wahren“ Modell zugrunde liegenden Parametervektoren Θ_0 gehört und der Median von den aus der Modellanpassung resultierenden Residuen $(r_1(\theta), ..., r_N(\theta))$ genau 0 entspricht, verworfen werden, wenn $d_k(r_1(\theta), ..., r_N(\theta))$ entweder zu groß oder zu klein ist. Dabei deuten zu wenige Vorzeichenwechsel typischerweise auf Abweichungen des Medians von 0 oder positive Korrelationen der Residuen hin. Auf der anderen Seite signalisieren zu viele Vorzeichenwechsel negativ korrelierte Residuen, können aber auch für eine gute Modellanpassung sprechen (s. Leckey et al., 2020).

Falls von der Unabhängigkeit der Residuen ausgegangen wird, bietet sich deshalb − auch in Abhängigkeit von dem betrachteten Modell − die einseitige Version des Tests an, bei der die Nullhypothese lediglich bei zu wenigen Vorzeichenwechseln verworfen wird. Um auf der anderen Seite die Unabhängigkeit der Residuen zu prüfen, stellt die zweiseitige Version des Tests, die in dieser Arbeit von zentralem Interesse ist, eine gute Wahl dar.

Bemerkenswerterweise konnten Leckey et al. (2020) in dem Zusammenhang zeigen, dass eine Vereinfachung der zweiseitigen Version des Tests, bei dem lediglich aufeinanderfolgende Residuen zur Berechnung der Tiefe herangezogen werden, dem Runs-Test nach Wald und Wolfowitz (1940) entspricht (s. Kap. 2.5). So definierten Kustosz et al. (2016b) eine vereinfachte Version des K-Vorzeichentiefe für $K \geq 2$ als:

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

Wird für diese Testgröße der Parameter $K = 2$ gewählt, so reduziert sie sich dahingehend, dass lediglich die relative Anzahl der Vorzeichenwechsel im Residualvektor − und damit die Anzahl der Runs − gezählt wird. In dem Zusammenhang konnten sie außerdem zeigen, dass es sich bei der Verteilung dieser vereinfachten Vorzeichentiefe unter der Voraussetzung 2.1 asymptotisch um eine Normalverteilung handelt und insbesondere gilt:

$$T_K^S(\theta) := \sqrt{N-K+1}\frac{d_K^S\left(R_1(\theta),\ldots,R_N(\theta)\right) - \left(\frac{1}{2}\right)^{K-1}}{\sqrt{\left(\frac{1}{2}\right)^{K-1}\cdot\left[3-\left(\frac{1}{2}\right)^{K-2}\cdot(K-1)-3\cdot\left(\frac{1}{2}\right)^{K-1}\right]}} \longrightarrow \mathcal{N}(0,1).$$

Nach Leckey et al. (2020) ist deshalb davon auszugehen, dass die herkömmlichen Versionen der K-VZ-Tests, die im Folgenden auch als vollständige Versionen bezeichnet werden, im Vergleich mit einer derartigen Vereinfachung, bei der lediglich (N - K + 1) anstelle von $\binom{N}{K}$ Teilmengen betrachtet werden, eine deutlich größere Macht aufweisen sollten.

Eine Ablehnung der Nullhypothese findet bei der zweiseitigen Version des K-VZ-Tests genau dann statt, wenn

$$\sup_{\theta\in\Theta_0} T_K(\theta) < q_{\alpha/2} \quad \text{oder} \quad \inf_{\theta\in\Theta_0} T_K(\theta) > q_{1-\alpha/2}$$

oder im Fall der vereinfachten Form des K-VZ-Tests

$$\sup_{\theta\in\Theta_0} T_K^S(\theta) < q_{\alpha/2} \quad \text{oder} \quad \inf_{\theta\in\Theta_0} T_K^S(\theta) > q_{1-\alpha/2}$$

gelten. Dabei entsprechen $q_{\alpha/2}$ und $q_{1-\alpha/2}$ dem (α/2)- bzw. dem (1 - α/2)-Quantil der Verteilung von $T_K(\theta)$ bzw. von $T_K^S(\theta)$.

Wie in diesem Kapitel noch diskutiert wird, ist die asymptotische Verteilung von $T_K(\theta)$ nicht symmetrisch. Aus diesem Grund liegt es nahe, den Ablehnungsbereich des Tests alternativ anders zu definieren, indem man ihn z. B. so wählt, dass seine Länge maximal ist. Wird die einseitige Version des Tests durchgeführt, so findet eine Ablehnung der Nullhypothese genau dann statt, wenn:

$$\sup_{\theta\in\Theta_0} T_K(\theta) < q_\alpha$$

gilt, wobei q_α das α-Quantil der asymptotischen Verteilung von $T_K(\theta)$ beschreibt. Welche Version des Tests verwendet werden sollte, hängt dabei mit oben ausgeführten Überlegungen und dem konkreten Anwendungsgebiet zusammen.

Für kleine Stichprobenumfänge kann die exakte Stichprobenverteilung der K-Vorzeichentiefe für jedes K ermittelt werden. Dies ist möglich, da die Annahme 2.1 garantiert, dass die Verteilung der Vorzeichentiefe invariant gegenüber der Verteilung der Fehler ist. Dazu werden alle 2^N Möglichkeiten eines binären Vektors mit Ausprägungen -1 und +1 betrachtet, die mit den Vorzeichen der Residuen identifiziert werden können und für jeden dieser Vektoren wird die entsprechende K-Vorzeichentiefe berechnet. Aus diesen Werten können dann die exakten Verteilungen sowie die entsprechenden kritischen Werte der K-Vorzeichentiefen abgeleitet werden.

Für größere N ist dieses Vorgehen aufgrund des immensen benötigten Zeitaufwandes allerdings äußerst impraktikabel.

Nachdem Kustosz et al. (2016a) einen Grenzwertsatz für die Teststatistik $T_3(\theta)$ beweisen konnten, gelang es Malcherczyk et al. (2020) dieses Resultat für $K \geq 3$ zu erweitern und zu verallgemeinern. Dadurch wurde es möglich, einen Algorithmus zur Ermittlung der K-Vorzeichentiefe in linearer Laufzeit anzuwenden, was eine bemerkenswerte Verbesserung im Vergleich zu einer naiven Laufzeit von N^K darstellt. In dem Zusammenhang ist zu bemerken, dass dabei eigentlich eine asymptotisch äquivalente Version der K-Vorzeichentiefe verwendet wird. Simulationsstudien zeigten jedoch, dass die Unterschiede zwischen den beiden Versionen vernachlässigbar klein sind.

Blockimplementation

Einen anderen Ansatz, um die Berechnungszeit der K-Vorzeichentiefe im Fall von wenigen Vorzeichenwechseln drastisch zu verkürzen und das Verständnis der Vorzeichentiefe zu vertiefen, stellt nach Leckey et al. (2020) die sogenannte Blockimplementation dar. Dazu wird ein Vektor der Residuen $r = (r_1, ..., r_N)$ betrachtet und in eine Menge von Blöcken zerlegt, wobei ein neuer Block genau dann beginnt, wenn sich ein Vorzeichenwechsel im Residualvektor ereignet. Die Funktion $\psi(x)$ wird dabei als Vorzeichen einer reellen Zahl x definiert, also $\psi(x) = \mathbb{1}\{x > 0\}$ - $\mathbb{1}\{x < 0\}$. Im Folgenden wird die Anzahl der Blöcke $B(r)$ mit zugehörigen Startpositionen $s_1(r), ..., s_{B(r)}(r)$ betrachtet und formal definiert als:

$$B(r) := 1 + \sum_{n=2}^{N} \mathbb{1}\{\psi(r_{n-1}) \neq \psi(r_n)\}$$
$$s_b(r) := \min\{l > s_{b-1}(r);\ \psi(r_l) \neq \psi(r_{l-1})\}, \quad b = 2, ..., B(r),$$

wobei für die Startposition des ersten Blockes als $s_1(r) := 1$ definiert wird. Der Vollständigkeit halber wird noch $s_{B(r)+1}$ als $(N+1)$ gesetzt. Des Weiteren wird die Größe des Blocks b definiert als:

$$q_b(r) := s_{b+1}(r) - s_b(r), \quad b = 1, ..., B(r).$$

Beispielsweise enthält der Vorzeichenvektor

$$r = (\underbrace{+, +, +}_{1.}, \underbrace{-, -}_{2.}, \underbrace{+}_{3.}, \underbrace{-, -}_{4.}, \underbrace{+, +}_{5.}, \underbrace{-}_{6.})$$

genau $B(r) = 6$ Blöcke mit Startpositionen 1, 4, 6, 7, 9 und 11 sowie die zugehörigen Blockgrößen $q_1(r) = 3$, $q_2(r) = 2$, $q_3(r) = 1$, $q_4(r) = 2$, $q_5(r) = 2$ und $q_6(r) = 1$.

Im Folgenden wird das n-te Residuum als dem j-ten Block mit $j \in \{1, ..., B(r)\}$ zugehörig bezeichnet, falls $s_j(r) \leq n < s_{j+1}(r)$ gilt. Weiterhin wird das Vorzeichen eines Blockes j als das Vorzeichen der in ihm enthaltenen Residuen verstanden. So werden die Blöcke $j_1 < ... < j_k$ als

alternierend bezeichnet, falls die Vorzeichen der Residuen in den entsprechenden Blöcken alternieren. Insbesondere können zwei Blöcke j_i und j_k mit $i, k \in \{1, ..., B(r)\}$ lediglich alternieren, falls i und k unterschiedliche Paritäten besitzen, also i gerade und k ungerade ist oder umgekehrt. Somit alternieren die Blöcke $j_1 < ... < j_k$ genau dann, wenn j_{i+1} - j_i für alle $i \in \{1, ..., (k$ - $1)\}$ ungerade ist.

Die Aufteilung des Residuenvektors in Blöcke erleichtert dabei die Identifikation der in ihm enthaltenen K-Tupel mit alternierenden Vorzeichen. So alternieren die Vorzeichen eines K-Tupels $(r_{n_1}, ..., r_{n_K})$ mit $1 \leq n_1 < ... < n_K \leq N$ genau dann, wenn die Residuen zu alternierenden Blöcken $j_1 < ... < j_K$ gehören. Mit dieser Erkenntnis kann eine alternative Repräsentation der K-Vorzeichentiefe des Residuenvektors r mit Hilfe der Menge der alternierenden Blöcke

$$\mathcal{A}_{K,B(r)} := \{(i_1, ..., i_K) \in \{1, ..., B(r)\}^K; i_k - i_{k-1} \text{ ist ungerade für } k = 2, ..., K\}$$

durch

$$d_K(r_1, ..., r_N) = d_{K,N,B(r)}(q_1(r), ..., q_{B(r)}(r)) := \frac{1}{\binom{N}{K}} \sum_{(i_1,...,i_K) \in \mathcal{A}_{K,B(r)}} \prod_{k=1}^{K} q_{i_k}$$

definiert werden, wobei q_{i_k} die Anzahl der Residuen im Block i_k beschreibt. Auf diese Weise kann die Rechenkomplexität erheblich verringert werden und nach Leckey et al. (2020) ist sogar noch eine weitere Reduktion durch das Speichern bereits berechneter Terme möglich. So kann die Berechnung der Tiefe linear in $N + B(r)$ durchgeführt werden.

Eigenschaften

Im Folgenden sollen die Eigenschaften der K-Vorzeichentiefe und der Teststatistik $T_K(\theta)$ in Anlehnung an Leckey et al. (2020) genauer beschrieben werden. Zunächst wird dabei auf die Asymptotik der K-Vorzeichentiefe eingegangen. Falls es sich bei $R_1(\theta) = R_1, ..., R_N(\theta) = R_N$ um unabhängige Zufallsvariablen handelt, die Voraussetzung 2.1 erfüllen, so gilt für den Erwartungswert der K-Vorzeichentiefe:

$$\begin{aligned}
&E_\theta\left[d_K(R_1, ..., r_N)\right] \\
&= E_\theta\left[\frac{1}{\binom{N}{K}} \sum_{1 \leq n_1 < n_2 < ... < n_K \leq N} \left(\prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k R_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k R_{n_k} < 0\right\}\right)\right] \\
&= \frac{1}{\binom{N}{K}} \cdot \binom{N}{K} \left(\prod_{k=1}^{K} E_\theta\left[\mathbb{1}\left\{(-1)^k R_{n_k} > 0\right\}\right] + \prod_{k=1}^{K} E_\theta\left[\mathbb{1}\left\{(-1)^k R_{n_k} < 0\right\}\right]\right) \\
&= \prod_{k=1}^{K} P_\theta\left[(-1)^k R_{n_k} > 0\right] + \prod_{k=1}^{K} P_\theta\left[(-1)^k R_{n_k} < 0\right] = \prod_{k=1}^{K} \frac{1}{2} + \prod_{k=1}^{K} \frac{1}{2} \\
&= \left(\frac{1}{2}\right)^{K-1}.
\end{aligned}$$

Die Konvergenz der K-Vorzeichentiefe gegen ihren Erwartungswert kann mithilfe einer alternativen Repräsentation durch das folgende Lemma nachvollzogen werden, wobei $i(j)$ statt i_j geschrieben wird, um dreifache Indizes zu vermeiden:

Lemma 1. *Seien $E_{n_1}, ..., E_{n_K}$ Zufallsvariablen mit $P(E_{n_i} \neq 0)$ für $i \in \{1, ..., K\}$ mit einem $K \in \mathbb{N} \setminus \{1\}$, dann gilt:*

$$\prod_{k=1}^{K} \mathbb{1}\left\{E_{n_k}(-1)^k > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{E_{n_k}(-1)^k < 0\right\} - \left(\frac{1}{2}\right)^{K-1}$$
$$= \frac{1}{2^{K-1}} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{1 \leq i(1) < ... < i(2L) \leq K} \prod_{j=1}^{2L} (-1)^{i(j)} \psi\left(E_{n_{i(j)}}\right), \quad P_\theta\text{-fast sicher.}$$

Damit konnten Leckey et al. (2020) zeigen, dass die Varianz der ganzen Vorzeichentiefe für $N \to \infty$ gegen 0 konvergiert, sodass folgendes Theorem für den Grenzwert der K-Vorzeichentiefe bewiesen werden kann:

Theorem 1. *Falls $R_1(\theta), ..., R_N(\theta)$ Voraussetzung 2.1 erfüllen, so gilt für $K \geq 2$:*

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

P_θ-fast sicher für $N \to \infty$.

Weiter ist von Interesse, den Wert der K-Vorzeichentiefe in Extremfällen zu betrachten, in denen sie Werte nahe ihres Maximums bzw. Minimums annimmt. Wie bereits erwähnt, deutet der Fall, dass die Residuen $r_1, ..., r_N$ alternierende Vorzeichen besitzen, also $\psi(r_n) = \text{-}\ \psi(r_{n+1})$ für jedes $n \in \{1, ..., N\text{ - }1\}$ gilt, auf eine gute Anpassung des Modells mit Modellparameter θ an die Daten hin. Beim zweiseitigen Test indizieren sie aber gleichermaßen eine starke negative Korrelation und die K-Vorzeichentiefe nimmt Leckey et al. (2020) zufolge in diesem Fall vermutlich ihr Maximum an. Um den Wert der Teststatistik in dieser Situation zu ermitteln, konnten Leckey et al. (2020) unter der Konvention, dass $\binom{n}{k} = 0$ für $n < k$ gilt, folgendes Theorem beweisen:

Theorem 2. *Handelt es sich bei $r_1, ..., r_N$ um Residuen mit alternierenden Vorzeichen, so gilt für $2 \leq K \leq N$:*

$$d_K(r_1, ..., r_N) = \frac{1}{\binom{N}{K}} \left(\binom{\lfloor (N+K)/2 \rfloor}{K} + \binom{\lceil (N+K-2)/2 \rceil}{K} \right).$$

Daraus lässt sich schlussfolgern, dass die K-Vorzeichentiefe im Fall von alternierenden Vorzeichen für $N \to \infty$ gegen ihren Erwartungswert $(1/2)^{K-1}$ konvergiert. Leckey et al. (2020) verallgemeinerten diese Erkenntnis für Fälle, in denen die Residuen in Blöcken der Größe M alternieren, wobei N ein Vielfaches von M ist. Mit den Überlegungen zur Blockimplementierung ist dies genau dann der Fall, wenn $q_j(r_1, ..., r_N) = M$ für alle $j = 1, ..., B(r)$ gilt. In dem Zusammenhang konnten sie folgendes Lemma beweisen:

Lemma 2. *Seien $M, N \in \mathbb{N}$ mit $B = N/M \in \mathbb{N}$ und bezeichne $\langle x \rangle_J = \prod_{j=0}^{J-1}(x-j)$ für $x, J \in \mathbb{N}$ und $x \geq J$. Falls $r_1, ..., r_N$ in Blöcken der Größe M alternieren und $B \geq K$ erfüllt ist, so gilt:*

$$(i) \quad d_K(r_1, \ldots, r_N) = \frac{\left\langle \frac{B+\kappa-2}{2} \right\rangle_{K-1}}{B^{K-1}} \cdot \frac{N^K}{\langle N \rangle_K} \qquad \textit{falls } K+B \textit{ gerade ist}$$

$$(ii) \quad d_K(r_1, \ldots, r_N) = \frac{2\left\langle \frac{B+K-1}{2} \right\rangle_K}{B^K} \cdot \frac{N^K}{\langle N \rangle_K} \qquad \textit{falls } K+B \textit{ ungerade ist.}$$

Diese Erkenntnisse ermöglichten es schließlich, den asymptotisch maximalen Grenzwert der Teststatistik zu ermitteln. Er ist durch folgendes Theorem gegeben:

Theorem 3. *Falls die Residuen $r_1, ..., r_N$ in Blöcken der Größe M für ein festes $M \in \mathbb{N}$ alternieren, so gilt:*

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

Daraus ergibt sich, dass der maximale Wert, den die Teststatistik $T_K(\theta)$ annehmen kann, asymptotisch $(K(K\text{ - }1))/2^K$ beträgt. Der zweiseitige K-VZ-Test würde in Situationen, in denen die Residuen in Blöcken der Größe M alternieren, die Hypothese ihrer Unabhängigkeit für hinreichend großes N also stets verwerfen.

Ihren minimalen Wert von 0 erreicht die Vorzeichentiefe genau dann, wenn es keine K-Tupel der Residuen gibt, die alternierende Vorzeichen besitzen. Dies ist genau dann der Fall, wenn die Anzahl der Blöcke $B(r)$ kleiner ist als das gewählte $K \geq 2$. Unter diesen Bedingungen verwerfen sowohl die einseitige, als auch die zweiseitige Version des K-VZ-Tests, sofern N groß genug ist, damit eine Verwerfung zum vorgegebenen Niveau α möglich ist. Für die Teststatistik bedeutet dies, dass sie einen Wert von $-N/2^{K-1}$ annimmt, der für $N \to \infty$ divergiert. Kustosz et al. (2016b) fanden in diesem Zusammenhang heraus, dass, falls es sich bei einem p-dimensionalen Parameter um einen Nonfit handelt, die daraus resultierenden Residuen typischerweise höchstens $(p\text{ - }1)$ Vorzeichenwechsel aufweisen. Aus diesem Grund empfiehlt es sich, die K-Vorzeichentiefe mit $K = p + 1$ zu wählen, obwohl auch kleinere Ks zu guten Trennschärfen führen können.

Insgesamt bedeuten diese Erkenntnisse, dass die asymptotische Verteilung der Statistik $T_K(\theta)$ zwar nach oben beschränkt, aber nach unten unbeschränkt ist, sodass es sich insbesondere um eine asymmetrische Verteilung handelt (Leckey et al., 2020).

In `R` kann die K-Vorzeichentiefe mithilfe der Funktion `calcDepth` aus dem Paket `GSignTest`, das von Horn (2020) implementiert wurde, berechnet werden. Dabei kann der Parameter K über die gleichnamige Parametereinstellung bestimmt werden und für die Berechnung in linearer Laufzeit nach Malcherczyk et al. (2020) kann die Parametereinstellung `linear=TRUE` gesetzt werden. Eine Durchführung des K-VZ-Tests ist dann möglich, indem die entsprechenden kritischen Werte aus demselben Paket durch die Funktion `qdepth` in Abhängigkeit vom Parameter K und dem Stichprobenumfang N ermittelt werden.

3 Statistische Auswertung

Im Folgenden werden die Ergebnisse der statistischen Auswertung vorgestellt. Sämtliche Simulationen, Berechnungen und grafische Darstellungen wurden mithilfe der Software R in der Version 3.6.2 (R Core Team, 2019) und den in dieser Arbeit aufgeführten Paketen durchgeführt.

3.1 AR(1)-Prozesse

In diesem Abschnitt werden die im Kapitel 2 beschriebenen Testverfahren auf simulierte Zeitreihen $(x_1, ..., x_N)$ mit $N \in \mathbb{N}$ angewendet, die einem stationären, autoregressiven Prozess 1. Ordnung (AR(1)-Prozess) folgen (s. Kap. 2.2). Es gilt dabei, dass ein AR(1)-Prozess genau dann stationär ist, wenn der autoregressive Parameter ρ_1 betragsmäßig kleiner als 1 ist, also wenn $|\rho_1| < 1$ erfüllt ist (Shumway und Stoffer, 2017, S. 77 f.). Im Fall, dass $|\rho_1|$ größer als 1 ist, spricht man von einem explosiven Prozess. Gilt $\rho_1 > 1$, so streben die Werte der Beobachtungen des Prozesses systematisch für $N \to \infty$ gegen -∞ oder ∞, abhängig davon, welche Parität die Zeitreihe in dem Moment der Explosion besitzt. Ist ρ_1 hingegen kleiner als -1, so alternieren die Werte und streben ebenfalls gegen ∞ bzw. -∞. Solche explosiven Prozesse kommen z. B. bei der Modellierung von Umsätzen bei Preisblasen oder dem Wachstum von Rissen zum Einsatz. Für $|\rho_1| = 1$ nennt man den entsprechenden Prozess eine Irrfahrt oder einen Random Walk, wobei solche Prozesse hauptsächlich bei der Beschreibung von Wertpapierpreisen oder Aktienkursen zur Anwendung kommen. Diese Art von Prozessen werden im aktuellen Kapitel aber nicht betrachtet.

Für die einzelnen Beobachtungen der simulierten Zeitreihen wird im Folgenden der Zusammenhang

$$x_t = \mu + \rho_1 x_{t-1} + w_t, \quad |\rho_1| < 1, \quad w_t \sim WN(0, \sigma^2_{WN})$$

für $t \in \{2, ..., N\}$ angenommen. Der Parameter μ, der als das Niveau der Zeitreihe aufgefasst werden kann, ist dabei nicht von Interesse, da sämtliche betrachtete Teststatistiken den Effekt einer Konstanten eliminieren. So werden die empirischen Autokorrelationskoeffizienten mithilfe der empirischen Kovarianz berechnet, die nicht von dem Parameter μ abhängt, sodass sein Wert für den Durbin-Watson-Test (DW-Test) (s. Kap. 2.3) sowie für den Ljung-Box-Test (LB-Test) (s. Kap. 2.4) keine Rolle spielt. Auch der Runs-Test (s. Kap. 2.5), der Turning-Point-Test (TP-Test) (s. Kap. 2.6), und der Von-Neumann-Ratio-Rang-Test (VNRR-Test) (s. Kap. 2.7) basieren lediglich auf dem sequenziellen Schema der Beobachtungen, das nicht von dem Niveau der Zeitreihe abhängt. Dies wird beim VNRR-Test durch die Rangbildung und beim

Runs-Test durch die Zentrierung mit dem empirischen Median erreicht. Beim TP-Test sind lediglich die Differenzen aufeinanderfolgender Beobachtungen von Interesse, die ebenfalls nicht von μ beeinflusst werden. Auch der Broock-Dechert-Schreinkman-Test (BDS-Test) (s. Kap. 2.8) basiert auf den Abständen zwischen Beobachtungen, auf die sich eine Konstante nicht auswirkt. Im Hinblick auf die K-Vorzeichentiefetests (K-VZ-Tests) (s. Kap. 2.9) muss zusätzlich bedacht werden, dass die Überprüfung der Nullhypothese stets auch ein Test dafür ist, ob der empirische Median der Daten 0 entspricht. Für ein $\mu \neq 0$ ist diese Voraussetzung in jedem Fall nicht erfüllt, sodass es sinnvoll erscheint, derartige Zeitreihen von vornherein nicht zu betrachten. Eine alternative Vorgehensweise wäre es z. B., eine Zentrierung der Daten durch den empirischen Median − wie es beim Runs-Test getan wird − vorzunehmen, wodurch zumindest der empirische Median in der Zeitreihe 0 entsprechen würde. Zunächst werden in diesem Kapitel aber lediglich Prozesse der obigen Form mit $\mu = 0$ betrachtet.

Die Simulationen der oben beschriebenen Zeitreihen wurden mithilfe der Funktion `arima.sim` aus dem R-Paket `stats`, das zur Basisversion der Software gehört, durchgeführt. Die Innovationen w_t wurden hier aus einer Standardnormalverteilung gezogen, sodass für $\sigma^2_{WN} = 1$ gilt. Um den Einfluss des Startwertes der Zeitreihe zu minimieren, wurde außerdem die Hälfte des zu betrachtenden Beobachtungsumfangs N am Anfang der Zeitreihe zusätzlich simuliert und nachträglich entfernt. Dieses Vorgehen wird auch als „Burn-In-Phase“ bezeichnet und ist für zuverlässige Simulationsstudien im Kontext von Zeitreihen unerlässlich.

Die Unabhängigkeit und die Zufälligkeit in der vorliegenden Zeitreihe ist dabei genau dann gegeben, wenn $\rho_1 = 0$ gilt. So kann für die K-VZ-Tests die Hypothese der Unabhängigkeit durch $H_0 : \theta_0 = (\mu, \rho_1)^T \in \Theta_0 = \{(0,0)^T\}$ überprüft werden. Als Alternativen werden hier Werte von ρ_1 auf einem Gitter von -0.99 − 0.99 mit einer Feinheit von 0.01 betrachtet. Es gilt zu bemerken, dass in dieser Arbeit lediglich mit den asymptotischen Quantilen der Verteilung der Vorzeichentiefe gearbeitet wird. Es handelt sich bei den dargestellten Ergebnissen also um die der asymptotischen Version des Tests. Die Trennschäfte der verschiedenen Testverfahren zum Niveau $\alpha = 0.05$ wurde dabei durch 100 wiederholte Simulationen für jeden der Gitterpunkte durch die relative Ablehnungsrate jedes Tests an denselben Zeitreihen empirisch ermittelt. Anschließend wurden diese Ablehnungsraten mithilfe der Funktion `levelplot` von Sarkar (2008) aufbereitet und farblich dargestellt. Die zu der konkreten Ablehnungsrate gehörige Farbe kann dabei der zusätzlich dargestellten Farblegende entnommen werden. Um beurteilen zu können, ob die jeweiligen Tests das angepeilte Testniveau unter der Unabhängigkeit einhalten, ist in den Abbildungen zur Trennschärfe für den Bereich mit Ablehnungsraten von ≤ 0.05 die Farbe schwarz gewählt worden. Die weiteren farblichen Abstufungen ereignen sich dann ebenfalls in Schritten von 0.05. Für die grafische Darstellung wurden außerdem die Pakete `magicaxis` von Robotham (2019), `latex2exp` von Meschiari (2015) sowie `viridis` von Garnier (2018) verwendet. Die relativ kleine Wiederholungszahl in dieser Simulationsstudie ist dabei der geringen zur Verfügung stehenden Rechenkapazität geschuldet. Allerdings zeigten diverse Vergleiche von unterschiedli-

chen Einstellungen bezüglich der Wiederholungszahl und Gitterfeinheit, dass die wesentlichen Systematiken zu den Trennschärfen auch bei einem Beobachtungsumfang von $N = 100$ hinreichend gut erkennbar sind. Jedoch sollte bedacht werden, dass eine leichte Überschreitung der gewünschten Ablehnungsrate kein eindeutiger Indikator für eine generelle Schwäche der Testverfahren bezüglich der Einhaltung des angepeilten Signifikanzniveaus ist.

Die Resultate einer ersten Simulation unter Normalbedingungen sind dabei für Stichprobenumfänge von $N = 20$, 50, 100 und 500 in Abbildung 3.1 dargestellt.

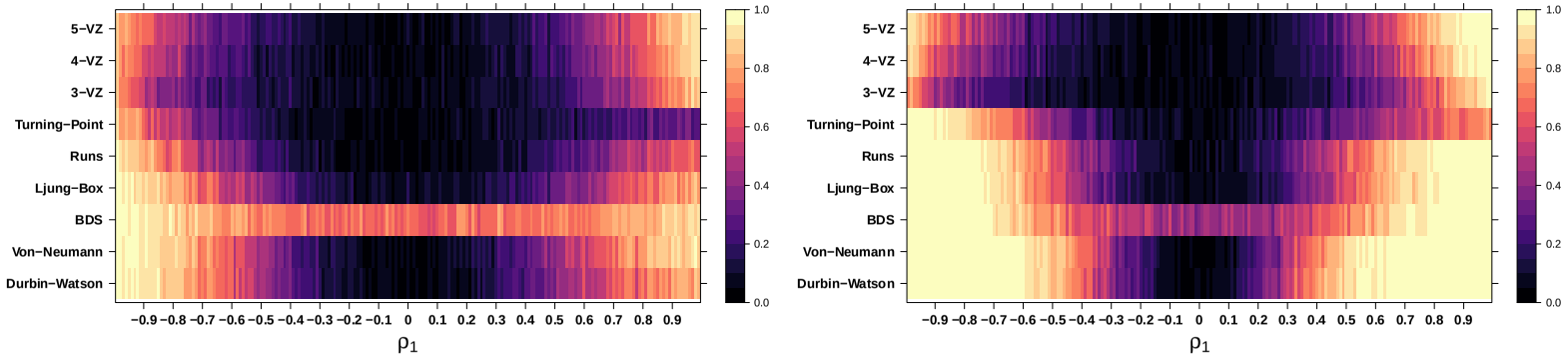
Die Ergebnisse zeigen, dass alle Verfahren − bis auf den BDS-Test − in der Lage sind, das Niveau bereits bei einem Stichprobenumfang von $N = 20$ einzuhalten. Vor allem für die asymptotischen Tests ist diese Erkenntnis von großer Bedeutung.

Die K-VZ-Tests haben unter den hier betrachteten Alternativen im Vergleich zu allen anderen Testverfahren deutlich mehr Schwierigkeiten, Abweichungen von der Nullhypothese zu erkennen. Allgemein scheinen hier außerdem Tests mit größerem K bessere Ergebnisse zu liefern. Während die Unterschiede bei einem Stichprobenumfang von $N = 20$ im Vergleich zu den anderen Verfahren noch nicht gravierend zu sein scheinen, gewinnen die K-VZ-Tests bei steigender Beobachtungszahl im Gegensatz zu den anderen Verfahren kaum an Trennschärfe dazu. Selbst bei $N = 500$ gelingt es ihnen lediglich bei extremen positiven Korrelationen mit $\rho_1 > 0.8$, die Nullhypothese in mehr als 95 % der Fälle korrekterweise zu verwerfen. Im Fall von negativen Korrelationen scheint lediglich der 5-VZ-Test eine Ablehnung der Nullhypothese bei Stichprobenumfängen von $N > 100$ und einem $\rho_1 <$ -0.9 zuverlässig zu erreichen. Damit weisen die Tests eine asymmetrische Güte auf und können positive Korrelationen leichter erkennen als negative.

Unter Normalbedingungen schneiden der DW-Test und der VNRR-Test am besten ab, da sie mit wachsendem Stichprobenumfang deutlich an Trennschärfe gewinnen und für eine Beobachtungszahl von $N = 500$ bereits gering ausgeprägte positive und negative Korrelationen mit $\rho_1 \notin$ [-0.15, 0.15] sicher erkennen. Der DW-Test hat dabei eine leicht bessere Trennschärfe.

Der LB-Test, der TP-Test sowie der Runs-Test liegen in diesem Szenario im Mittelfeld und weisen weniger Trennschärfe auf als die zuvor genannten Tests. Sie profitieren jedoch ebenso deutlich von einem wachsenden Stichprobenumfang. Der TP-Test scheint eine leicht asymmetrische Trennschärfe aufzuweisen und mehr Schwierigkeiten damit zu haben, die Nullhypothese bei positiven Korrelationen zu verwerfen. Für einen Stichprobenumfang von $N = 500$ können diese Verfahren die Nullhypothese bereits bei moderaten Korrelationen von $\rho_1 \notin$ [-0.2, 0.2] zuverlässig ablehnen, wobei der Runs-Test allgemein etwas besser abschneidet als die beiden anderen Testverfahren. Dabei scheint der LB-Test dem TP-Test noch etwas überlegen zu sein.

Beim BDS-Tests fällt vor allem auf, dass er große Stichprobenumfänge benötigt, um brauchbare Ergebnisse zu liefern. So scheint der Wert von ρ_1 bei einer Beobachtungszahl von $N = 20$ keine Rolle zu spielen, sodass die Ablehnungsrate auf dem gesamten Spektrum zufällig erscheint. Auch bei Stichprobenumfängen von $N \leq 100$ gelingt es dem Test nicht, das Niveau des Tests von $\alpha = 5\,\%$ unter der Unabhängigkeit der Zeitreihe bei $\rho_1 = 0$ einzuhalten. Erst ab einer großen



(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

Abbildung 3.1: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen unter Normalbedingungen, für unterschiedliche Stichprobenumfänge

Beobachtungszahl von $N = 500$ sind die Ergebnisse hinreichend gut und die Trennschärfe des Tests ist etwas besser als die des TP-Tests. Anhand dieser Erkenntnis kann die Wahl des in Kapitel 2.8 vorgeschlagenen Stichprobenumfangs von $N \geq 500$ nachvollzogen werden, da der Test in kleineren Stichproben offenbar deutliche Probleme bei der Einhaltung des angepeilten Signifikanzniveaus hat.

Ein Blick auf das Korrelogramm (s. Kap. 2.1) einer zufälligen Zeitreihe mit $\rho_1 = 0.7$ und einem Stichprobenumfang von $N = 100$ verdeutlicht, warum die Tests auf Grundlage der empirischen Autokorrelationskoeffizienten bei einer solchen Alternative so erfolgreich bei der Ablehnung der Nullhypothese sind (s. Abb. 3.2). So überschreiten hier sowohl $\hat{\rho}_1$ als auch $\hat{\rho}_2$ und $\hat{\rho}_3$ deutlich den kritischen Wert, der unter Annahme ihrer Normalverteilung und unter der Unabhängigkeit zu erwarten wäre. Er wird in der Grafik durch die blauen, gestrichelten Linien gekennzeichnet. Das zu erkennende abklingende Verhalten der empirischen Autokorrelationskoeffizienten ist typisch für autoregressive Prozesse und hängt damit zusammen, dass sich die Abhängigkeitsstrukturen zwischen den einzelnen Beobachtungen fortpflanzen. So sind beispielsweise x_t und x_{t-2} für $t \in \{3, ..., N\}$ über x_{t-1} indirekt voneinander abhängig, obwohl x_{t-2} nicht direkt in die Berechnung von x_t einfließt. Mathematisch kann dies durch sukzessives Einsetzen nachvollzogen werden:

$$x_t = \rho_1 x_{t-1} + w_t = \rho_1(\rho_1 x_{t-2} + w_{t-1}) + w_t = \rho_1^2 x_{t-2} + w_t + \rho_1 w_{t-1}.$$

Dabei klingen die Werte der empirischen Autokorrelationskoeffizienten exponentiell und damit für weiter auseinander liegende Beobachtungen umso schneller ab, je kleiner ρ_1 ist.

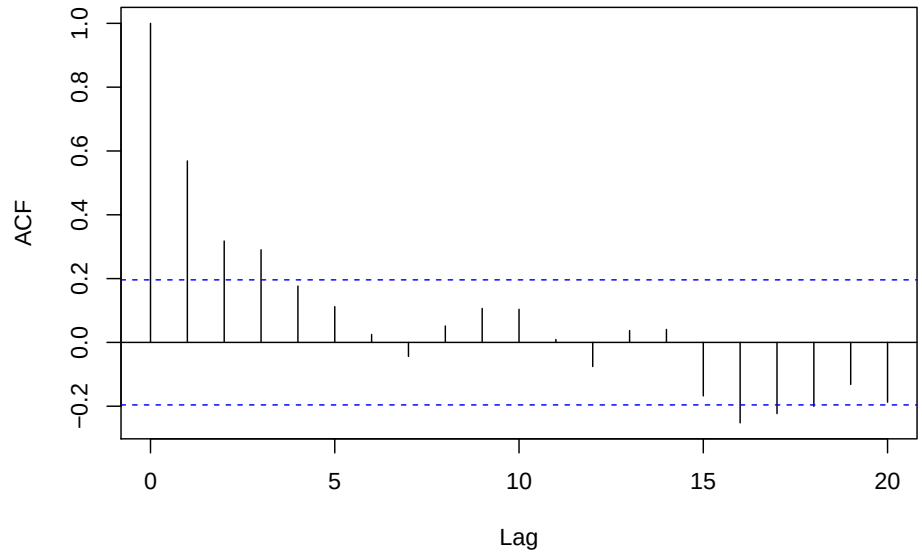
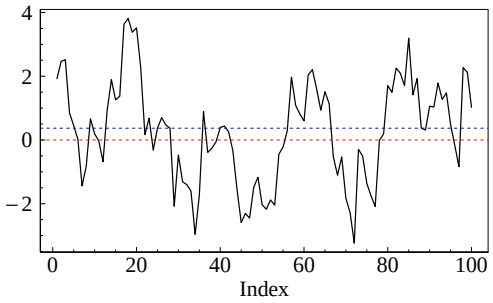


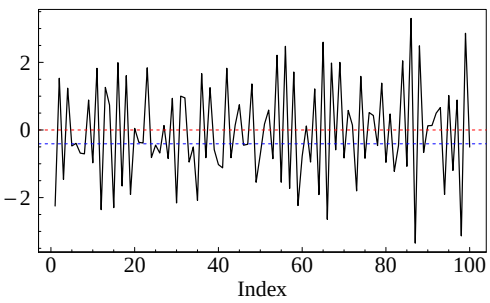
Abbildung 3.2: Korrelogramm eines AR(1)-Prozesses mit $\rho_1 = 0.7$ und $N = 100$ Beobachtungen, mit kritischen Werten (blaue Linien)

Die Tatsache, dass der DW-Test bessere Ergebnisse liefert als der LB-Test, lässt sich damit begründen, dass die Teststatistik des DW-Test lediglich auf $\hat{\rho}_1$ basiert und somit genau auf den hier betrachteten Fall ausgelegt ist. Der LB-Test zieht mit den vorgenommenen Einstellungen aber die ersten 15 empirischen Autokorrelationskoeffizienten mit in die Berechnung der Teststatistik ein und umfasst somit eine breitere Alternative von Abhängigkeitsstrukturen. Von den betrachteten Koeffizienten sind jedoch lediglich die ersten 3 signifikant von 0 verschieden. Die entsprechenden Quantile der Chi-Quadrat-Verteilung ab denen eine Verwerfung der Nullhypothese stattfindet, sind somit größer als bei der Betrachtung von weniger Lags. Dies wirkt sich offensichtlich negativ auf die Trennschärfe des Tests aus. Wird beim LB-Test ebenfalls lediglich $\hat{\rho}_1$ für die Testentscheidung herangezogen, also der Parameter H als 1 gewählt (vgl. Kap. 2.4), so zeigten Simulationen, dass die Verfahren eine vergleichbare Trennschärfe aufweisen.

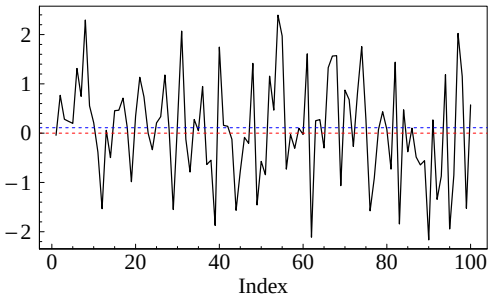
Um nachzuvollziehen, wie die nichtparametrischen Methoden eine Ablehnung der Nullhypothese erreichen, sind 3 zufällige Zeitreihen mit $N = 100$ Beobachtungen bei Autokorrelationskoeffizienten von $\rho_1 = 0.7$ und $\rho_1 = -0.7$ sowie unter der Nullhypothese der Unabhängigkeit (also mit $\rho_1 = 0$) in Abbildung 3.3 dargestellt.



(a) $\rho_1 = 0.7$



(b) $\rho_1 = -0.7$



(c) $\rho_1 = 0$

Abbildung 3.3: Zufällige Zeitreihen aus AR(1)-Prozessen mit $N = 100$ Beobachtungen und Autokorrelationskoeffizienten ρ_1 von 0.7 (a), -0.7 (b) und 0 (c), Nulllinie (rot) und empirischem Median (blau)

Beim Runs-Test dient die Anzahl der „Runs“ ober- und unterhalb des empirischen Medians als Teststatistik. Deren Anzahl ist, wie bereits in Kapitel 2.5 erläutert, äquivalent zur Anzahl der Durchgänge des Graphen durch das Niveau des empirischen Medians der Zeitreihe. Bei einem rein optischen Vergleich der beiden abhängigen Zeitreihen ((a) u. (b)) mit der unabhängigen Zeitreihe (c) werden die unterschiedlichen Anzahlen von Mediandurchgängen bereits offensichtlich. So kommen bei den abhängigen Zeitreihen deutlich weniger (a) bzw. mehr Durchgänge (b) und damit Runs vor als bei der unabhängigen. Auf diese Weise wird nachvollziehbar, warum in diesen Fällen eine Verwerfung der Nullhypothese stattfindet.

Gleichermaßen ist ersichtlich, dass die Anzahl der Turning-Points unter den Alternativen deutlich geringer (a) bzw. größer (b) ist als unter der Nullhypothese, sodass auch die Verwerfungen durch den TP-Test verständlich werden. Um zu verstehen, warum die Trennschärfe des TP-Tests eine leichte Asymmetrie aufweist − wobei eine Verwerfung der Nullhypothese bei negativen Korrelationen eher stattfindet als bei positiven Korrelationen − wurde eine kleine Simulation bezüglich der Anzahl von Turning-Points in korrelierten Zeitreihen der Länge $N = 100$ durchgeführt. Dazu wurden anhand des oben betrachteten Gitters von dem Parameter ρ_1 für jeden Gitterpunkt 1000 Zeitreihen simuliert. Für jede dieser Zeitreihen ist dann die Anzahl der Turning-Points erfasst worden und für jeden Gitterpunkt wurde anschließend die mittlere Anzahl an Turning-Points berechnet. Die Ergebnisse sind in Abbildung 3.4 dargestellt.

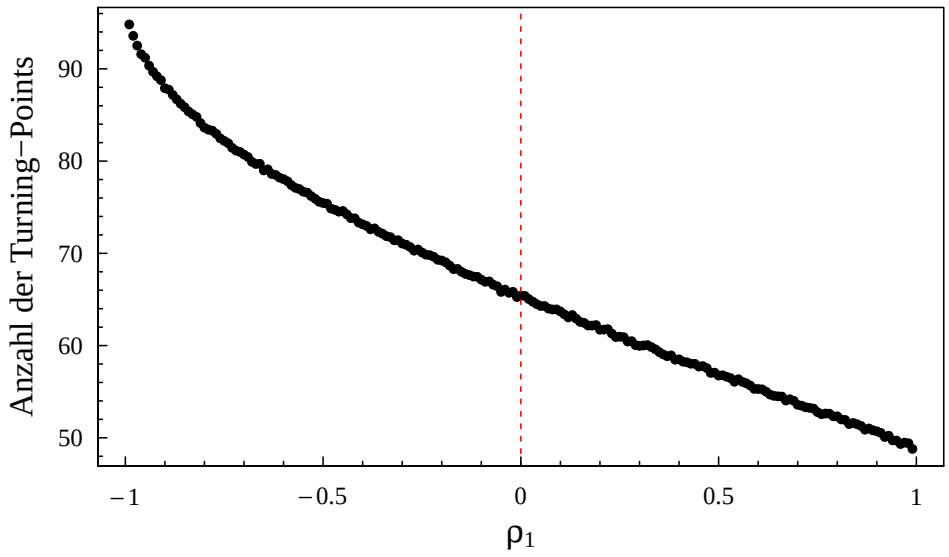


Abbildung 3.4: Simulierte mittlere Anzahl von Turning-Points in Zeitreihen mit $N = 100$ Beobachtungen von stationären AR(1)-Prozessen in Abhängigkeit von ρ_1

Aus dieser Grafik geht hervor, dass die Anzahl der Turning-Points für positive Autokorrelationen linear mit ρ_1 abzufallen scheint. Dabei entspricht die mittlere Anzahl von Turning-Points unter der Unabhängigkeit ungefähr seinem Erwartungswert von 66, während die minimale Anzahl an Turning-Points bei einer maximalen positiven Korrelation von 0.99 ungefähr 49 entspricht. Für negative Korrelationen hingegen steigt die Anzahl der Turning-Points mit kleiner werdendem ρ_1 zumindest polynomiell auf eine maximale Anzahl von ca. 95 an. Damit ist hier schon eine klare Asymmetrie zwischen dem Effekt von negativen und positiven Korrelationen auf die Anzahl der Turning-Points erkennbar. Insbesondere unterscheiden sich die mittleren Anzahlen der Turning-Points in den Extremfällen unterschiedlich stark vom Erwartungswert. Da die Anzahl der Turning-Points jedoch durch eine Normalverteilung approximiert wird, die einen symmetrischen Effekt unterstellt, wird auch klar, warum sich die Trennschärfe des TP-Tests asymmetrisch verhält.

Im Hinblick auf den VNRR-Test ist zu beachten, dass der Wert seiner Teststatistik maßgeblich vom Zähler $\sum_{i=1}^{N-1} (R_{i+1} - R_i)^2$ abhängt. Zu große oder zu kleine Werte dieses Terms führen dementsprechend zu einer Verwerfung der Nullhypothese. Solche Werte kommen zustande, wenn aufeinanderfolgende Beobachtungen systematisch ähnliche Ränge haben bzw. Ränge besitzen, die weit auseinanderliegen. Anhand von Abbildung 3.3 ist ersichtlich, dass benachbarte Beobachtungen im Fall von positiven Autokorrelationen (a) dazu neigen, ähnliche Werte anzunehmen, wohingegen negative Autokorrelationen (b) ein alternierendes Verhalten begünstigen. Dementsprechend kann nachvollzogen werden, dass die Teststatistik unter diesen Alternativen eher kleine bzw. große Werte annimmt und damit eine Verwerfung der Nullhypothese stattfinden kann. Die Überlegenheit gegenüber den anderen nichtparametrischen Tests rührt vermutlich aus der Tatsache, dass hier Ränge gebildet werden und somit mehr Information über die Zeitreihe in die Testentscheidung einfließen als lediglich das sequenzielle Schema.

In dieser ersten Simulationsreihe zeigt sich, dass die parametrischen Verfahren im Fall, dass die Bedingungen aller Testverfahren erfüllt sind, eine gute Wahl zur Überprüfung der Unabhängigkeit darstellen und hervorragende Trennschärfen aufweisen. Dabei ist der DW-Test dem LB-Test im AR(1)-Fall etwas überlegen, was auf die spezielle Struktur des zugrunde liegenden Prozesses zurückzuführen ist, die genau dem Anwendungsgebiet des DW-Testes entspricht. Erstaunlich ist, dass der VNRR-Test mit seinem Ansatz durch die Betrachtung der Ränge einen Vorteil gegenüber den anderen nichtparametrischen Verfahren zu haben scheint und unter Normalbedingungen bezüglich seiner Trennschärfe mit den parametrischen Verfahren mithalten kann. Unter den rein nichtparametrischen Verfahren schneidet der Runs-Test am besten ab, aber auch der TP-Test erzielt relativ gute Ergebnisse und die Einfachheit der Berechnung seiner Teststatistik macht ihn zu einer attraktiven Alternative zu den anderen Verfahren. Einen wesentlichen Nachteil stellt jedoch seine asymmetrische Güte dar. Die K-VZ-Tests schneiden hier am schlechtesten ab und können selbst bei großen Stichprobenumfängen lediglich extreme Korrelationen erkennen. Auch weisen sie eine leicht asymmetrische Trennschärfe auf.

3.1.1 Abweichungen von den Verteilungsannahmen der Innovationen

Im Folgenden soll untersucht werden, wie sich Abweichungen von den Voraussetzungen und Annahmen der einzelnen Tests auf deren Trennschärfe auswirken. Eine Möglichkeit hierbei ist es, die Innovationen w_t aus einer anderen Verteilung als der Normalverteilung zu generieren. Dabei wäre davon auszugehen, dass die Trennschärfen des LB-Tests sowie des DW-Tests, bei denen eine Normalverteilung der Innovationen vorausgesetzt wird, von dieser Modifikation beeinflusst werden. Im Hinblick auf die K-VZ-Tests wurden dafür zunächst Verteilungen genutzt, die die Voraussetzungen dieser Tests erfüllen, also einen Median von 0 besitzen.

Konkret wurden zum einen die Cauchy-Verteilung (t-Verteilung mit 1 Freiheitsgrad) sowie die t-Verteilung mit 5 Freiheitsgraden betrachtet. Dabei handelt es sich um Verteilungen mit schweren Rändern. Ein Vergleich der entsprechenden Dichtefunktionen mit der Dichte der Normalverteilung ist in Abbildung 3.5 dargestellt.

Die schweren Ränder der t-Verteilungen lassen sich in dieser Grafik anhand der im Vergleich zur Normalverteilung geringeren Wahrscheinlichkeitsmasse am Wert des Extrempunktes der Dichtefunktionen, die sich in den Bereich der Ränder verschiebt, erkennen. Für die Innovationen bedeutet das, dass extremere Werte mit größerer Häufigkeit zu erwarten sind, als es bei einer Normalverteilung der Fall ist. Bei der Cauchy-Verteilung gilt es weiterhin zu bemerken, dass ihre ersten beiden Momente nicht existieren, wodurch eine Anwendung des zentralen Grenzwertsatzes nicht möglich ist. Dadurch ist auch die Voraussetzung der parametrischen Testverfahren (also

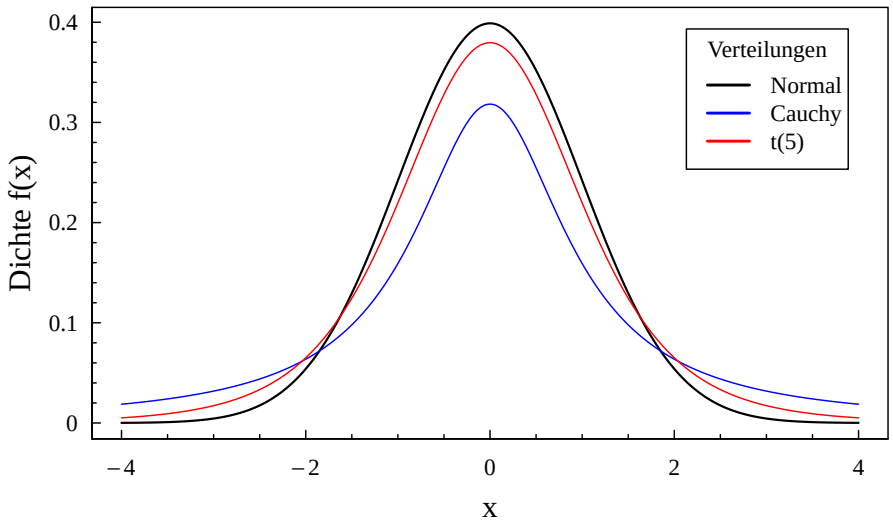


Abbildung 3.5: Vergleich der Dichtefunktionen der Normalverteilung, der Cauchy-Verteilung sowie der t-Verteilung mit 5 Freiheitsgraden

die Normalität der Autokorrelationskoeffizienten) nicht gegeben. Praktische Anwendung finden diese Verteilungen vor allem in Situationen, in denen regelmäßig mit extremen Veränderungen zwischen zwei Beobachtungen zu rechnen ist oder − wie in dieser Arbeit −, um zu untersuchen, inwieweit systematische Ausreißer einen Einfluss auf statistische Verfahren haben.

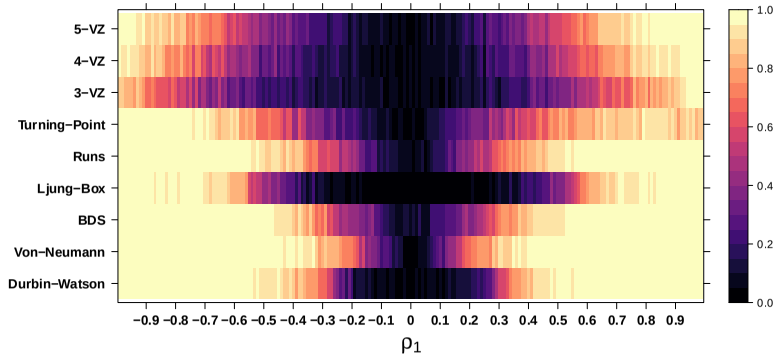
Wie sich diese Verteilungsänderungen auf die Trennschärfen der Testverfahren auswirken, zeigt Abbildung 3.6 für Stichprobenumfänge von $N = 50$ und $N = 500$.

Während im Vergleich mit den entsprechenden Grafiken zu normalverteilten Innovationen kaum ein Unterschied zwischen den dort erzielten Ergebnissen und denen mit t(5)-verteilten Innovationen zu erkennen ist, unterscheiden sie sich deutlich von den Ergebnissen bei Cauchy-verteilten Innovationen. Überraschenderweise gewinnen dabei alle nichtparametrischen Tests an Trennschärfe dazu und verwerfen somit die Nullhypothese bei kleineren Werten von $|\rho_1|$ öfter. Besonders deutlich wird dieser Effekt beim BDS-Test, der unter diesen Bedingungen bereits bei einem Stichprobenumfang von $N = 50$ ähnlich gute Ergebnisse erzielt wie der TP-Test oder der Runs-Test. Lediglich beim DW- und LB-Test führen die Verteilungsänderungen zu einer deutlichen Verschlechterung der Ergebnisse und veranlassen die Tests dazu, die Nullhypothese für Parameter in einem Bereich, in dem die Testentscheidungen nicht eindeutig sind, öfter beizubehalten. Ersichtlich ist dies anhand der entsprechenden Grafiken durch schärfere farbliche Abstufungen, als sie unter Normalbedingungen aufgetreten sind. Da es sich bei diesen Tests um diejenigen handelt, die eine Normalverteilung der Innovationen unterstellen, sind diese Ergebnisse wenig verwunderlich. Die Tatsache, dass der LB-Test wieder etwas schlechter als der DW-Test abschneidet, ist erneut auf die Anzahl der empirischen Korrelationskoeffizienten, die für eine Entscheidungsfindung herangezogen werden, zurückzuführen.

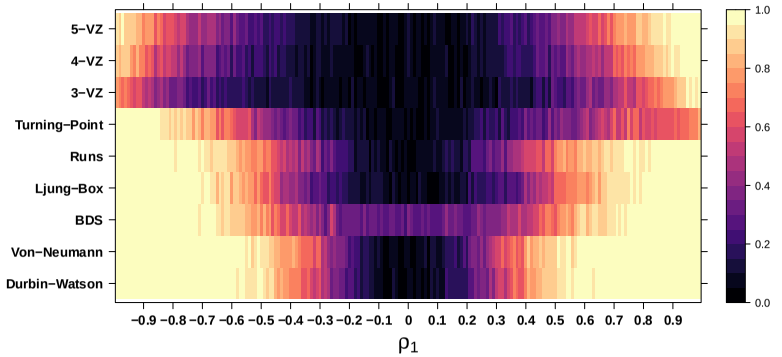
Als Konsequenz dieser Verteilungsänderung weisen die nichtparametrischen Tests, mit Ausnahme der K-VZ-Tests, bei allen Stichprobenumfängen eine bessere Trennschärfe als die parametrischen Tests auf. Spitzenreiter ist dabei der VNRR-Test, der aufgrund der Mehrinformation gegenüber den anderen nichtparametrischen Verfahren auch schon unter Normalbedingungen mit den parametrischen Tests mithalten konnte.

Die Tatsache, dass die t(5)-Verteilungen viel weniger starke Veränderungen der Testergebnisse herbeiführt als die Cauchy-Verteilung, macht Abbildung 3.5 deutlich. So unterscheidet sich die Dichtefunktion der t(5)-Verteilung an ihrem Peak und ihren Rändern nicht besonders stark von der Normalverteilung. Die Unterschiede zur Cauchy-Verteilung sind jedoch klar erkennbar. Außerdem ist der zentrale Grenzwertsatz bei den t(5)-verteilten Innovationen, anders als bei den Cauchy-verteilten, anwendbar, auch wenn zu erwarten ist, dass die Asymptotik langsamer greift, als es bei normalverteilten Innovationen der Fall ist.

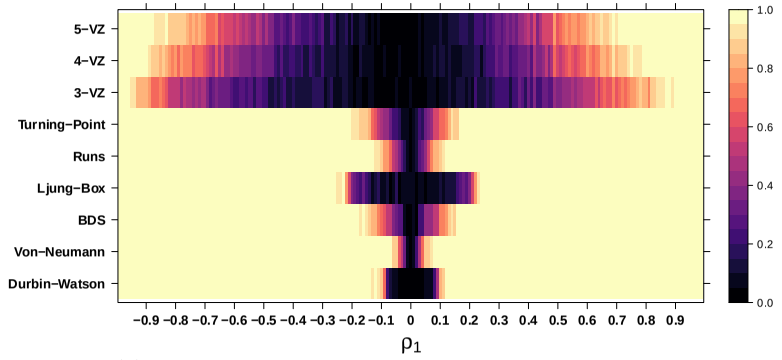
Weiter wird in Abbildung 3.7 deutlich, dass bei Cauchy-verteilten Innovationen viel häufiger extreme Werte zu beobachten sind, die sich deutlich auf das Verhalten der Zeitreihe auswirken. So wurde mithilfe der Funktion `pnorm` bzw. `pt` beispielhaft ermittelt, dass die Wahrscheinlichkeit des Auftretens einer Innovation mit einem Wert, der betragsmäßig größer als 5 ist, bei einer



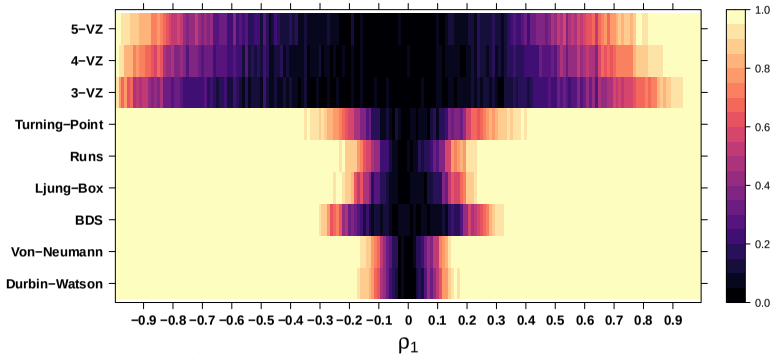
(a) Cauchy-verteilte Innovationen mit $N = 50$



(b) t(5)-verteilte Innovationen mit $N = 50$



(c) Cauchy-verteilte Innovationen mit $N = 500$



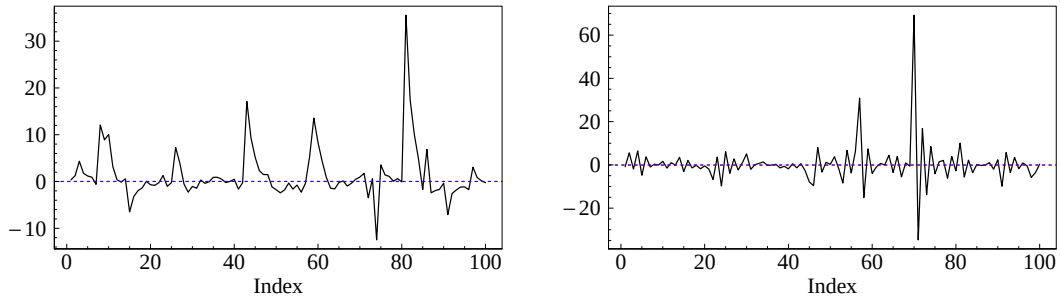
(d) t(5)-verteilte Innovationen mit $N = 500$

Abbildung 3.6: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit Cauchy- und t(5)-verteilten Innovationen, für unterschiedliche Stichprobenumfänge

Normalverteilung ca. $5.7 \cdot 10^{-7}$ beträgt. Bei einer Cauchy-Verteilung entspricht sie demgegenüber ca. 0.12 und bei der t(5)-Verteilung ca. 0.004.

Eine Begründung dafür, dass das Auftreten extremer Werte die Trennschärfe der meisten Tests sogar verbessert, ist, dass auf einen betragsmäßig sehr großen Wert − abhängig von $|\rho_1|$ − tendenziell betragsmäßig große Werte folgen, wodurch das Erkennen einer Abhängigkeitsstruktur erleichtert werden kann. So sind nach dem Auftreten extremer Innovationen im Fall, dass $\rho_1 > 0$ gilt, mehr Beobachtungen nötig, um wieder zum Niveau des Medians zurückzukehren bzw. es folgen bei einem $\rho_1 < 0$ mit großer Wahrscheinlichkeit ein oder mehrere Vorzeichenwechsel, bis der Schock ausgeschwungen ist. Dieses Verhalten lässt sich auf die Korrelationsfortpflanzung in autoregressiven Prozessen, wie sie z. B. in Abbildung 3.2 erkennbar ist, zurückführen. Durch diese Tendenz nimmt die Trennschärfe aller nichtparametrischen Testverfahren bei größeren Werten von $|\rho_1|$, also bei starken Korrelationen, zu.

Die verschiedenen nichtparametrischen Testverfahren reagieren in unterschiedlicher Weise auf das oben beschriebene Verhalten. Beim VNRR-Test führt es im Fall von positiven Korrelationen dazu, dass Beobachtungen nahe einer extremen Innovation ähnliche Ränge besitzen. Bei negativen Korrelationen führt dies wiederum dazu, dass die Ränge sehr unterschiedlich sind. In beiden Fällen wird dem Test eine Verwerfung der Nullhypothese erleichtert. Der TP-Test wird durch obiges Verhalten insoweit beeinflusst, dass nach extremen Werten mehr ($\rho_1 < 0$) bzw. weniger ($\rho_1 > 0$) Turning-Points auftreten, als es bei normalverteilten Innovationen zu erwarten wäre. Auch im Hinblick auf den Runs-Test führt dies bei negativen Korrelationen zu mehr − bzw. bei positiven Korrelationen zu weniger − Runs, als unter den gleichen Voraussetzungen bei Normalbedingungen. Die Tatsache, dass der BDS-Test so deutliche Verbesserungen zeigt, hängt vermutlich ebenfalls mit der Kompensationsphase nach extremen Innovationen zusammen. Durch sie werden die Abstände zwischen 2 m-Historien in diesem Bereich deutlich vergrößert.



(a) $\rho_1 = 0.5$ (b) ρ_1 = -0.5

Abbildung 3.7: Zufällige Zeitreihen aus AR(1)-Prozessen mit $N = 100$ Beobachtungen mit Cauchy-verteilten Innovationen bei $\rho_1 = 0.5$ (a) bzw. ρ_1 = -0.5 (b) mit Nulllinie (rot) und empirischem Median (blau)

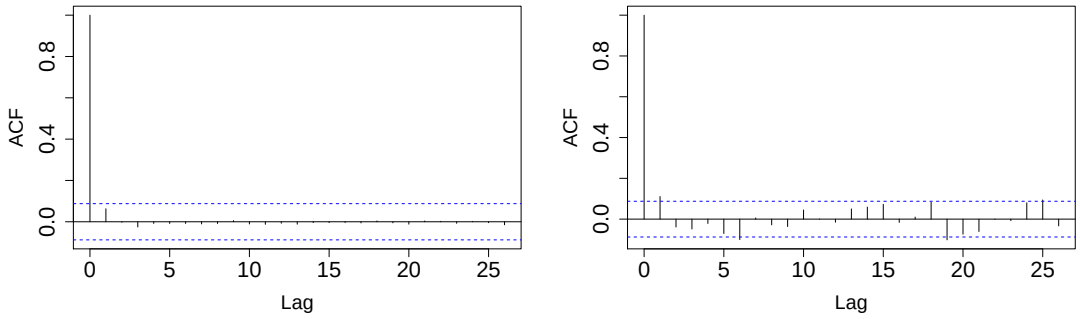
Das führt dazu, dass die Nullhypothese der Unabhängigkeit einfacher verworfen werden kann.

Um den Effekt auf die K-VZ-Tests zu verstehen, macht es Sinn, die Blockimplementation zu betrachten. So führt das Ausklingen der extremen Innovationen im Fall von negativen Korrelationen zu deutlich kleineren Blöcken, die im Fall eines alternierenden Verhaltens sogar lediglich die Länge 1 aufweisen. Die Teststatistik wird hier also schneller sehr klein. Bei positiven Korrelationen entstehen hingegen eher größere Blöcke, von denen, bei ähnlicher Korrelation, auch die in etwa selbe Größe erwartet werden könnte, wodurch die Teststatistik ebenfalls einen größeren Wert annimmt. In beiden Fällen wird eine Verwerfung der Nullhypothese erleichtert.

Auf der anderen Seite können extremwertige Innovationen aber auch Korrelationen bei moderat großen vorangegangenen Werten verschleiern, wodurch die empirischen Autokorrelationen unterschätzt werden und nicht mehr normalverteilt sind. Als Konsequenz neigen diejenigen Tests, die auf den Autokorrelationskoeffizienten und ihrer Normalität basieren, dazu, die Nullhypothese bei geringen Werten von ρ_1 weniger häufig zu verwerfen. Zur Veranschaulichung sind Korrelogramme von Zeitreihen mit normalverteilten und Cauchy-verteilten Innovationen bei einem Stichprobenumfang von $N = 500$ und $\rho_1 = 0.07$ in Abbildung 3.8 dargestellt.

Um diese Überlegungen zu untermauern, wurden zwei weitere Verteilungen betrachtet, die dazu neigen, extreme Werte anzunehmen − die Gumbel-Verteilung und die Laplace-Verteilung. In beiden Fällen wurde ein Skalen-Parameter von 2 gewählt. Ein Vergleich der Dichtefunktionen dieser beiden Verteilungen mit der Normalverteilung ist in Abbildung 3.9 dargestellt. Die Wahrscheinlichkeiten, dass eine Innovation einen Wert annimmt, der betragsmäßig größer als 5 ist, wurde dabei wie auf Seite 45 durch die Funktionen `pgumbel` bzw. `pdexp` ermittelt. Sie betragen bei der Gumbel-Verteilung ca. 0.177 und bei der Laplace-Verteilung ca. 0.08, sodass die entsprechende Wahrscheinlichkeit der Cauchy-Verteilung genau dazwischen liegt.

Bei der Gumbel-Verteilung gilt zu beachten, dass es sich bei ihr um eine asymmetrische Ver-



(a) $\rho_1 = 0.07$ mit Cauchy-Innovationen (b) $\rho_1 = 0.07$ mit normalen Innovationen

Abbildung 3.8: Korrelogramme von zufälligen Zeitreihen aus AR(1)-Prozessen mit $N = 500$ Beobachtungen mit Cauchy (a) und normalen Innovationen (b), bei $\rho_1 = 0.07$ und kritischen Werten (blau)

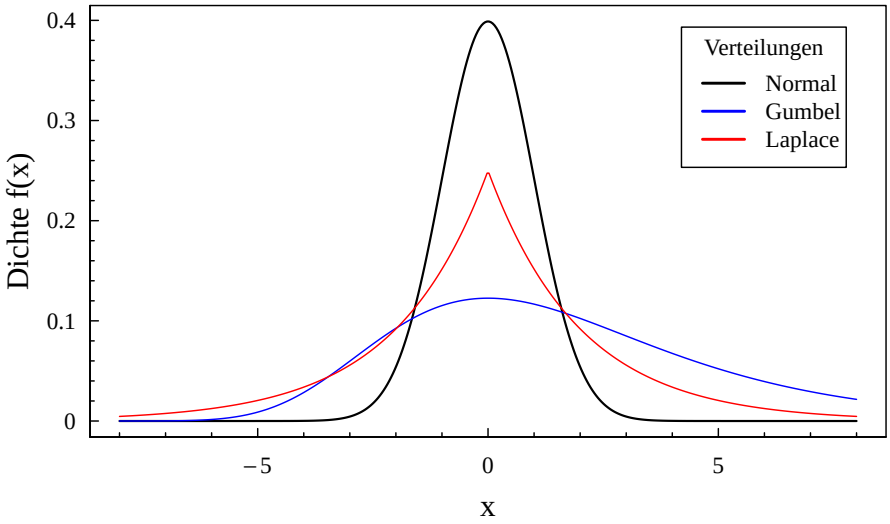
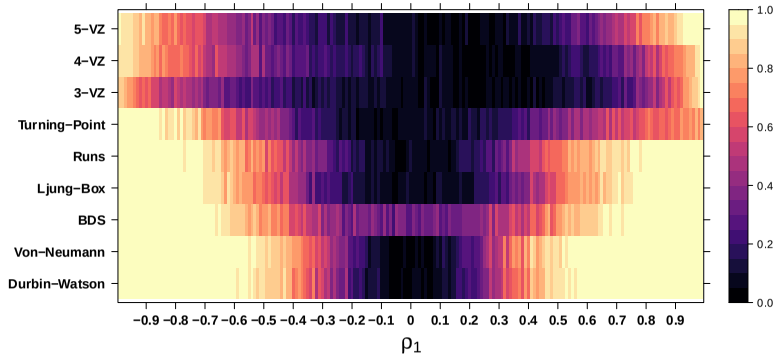


Abbildung 3.9: Vergleich der Dichtefunktionen der Normalverteilung, der Gumbel-Verteilung sowie der Laplace-Verteilung, mit Skalen-Parametern von 2

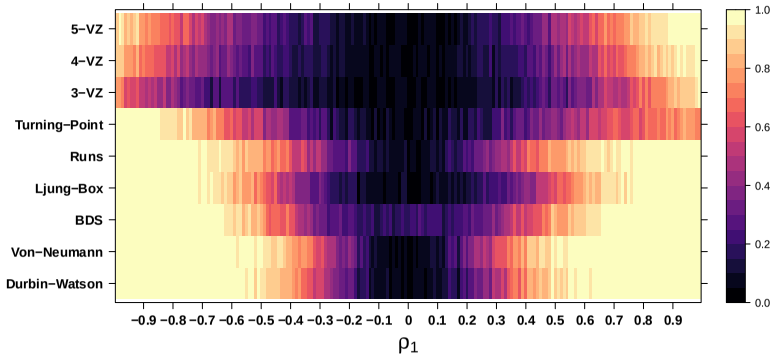
teilung handelt. Im Hinblick auf den obigen Vergleich beträgt die Wahrscheinlichkeit, einen Wert kleiner als -5 zu erhalten, 0.005 und einen Wert größer als 5 zu beobachten, 0.172. Insbesondere entspricht ihr Median deshalb nicht 0, sondern ca. 1.1. Da es hier aber nicht von Interesse ist, diese Abweichungen zu detektieren, wurde vor der Durchführung der K-VZ-Tests eine Zentrierung der Daten durch ihren empirischen Median vorgenommen. Damit wird im Folgenden außerdem untersucht, wie sinnvoll eine solche Zentrierung ist und welche Auswirkungen sie auf die Trennschärfe der K-VZ-Tests hat.

Die Simulationsergebnisse mit Innovationen aus diesen Verteilungen sind für Stichprobenumfänge von $N = 50$ und $N = 500$ in Abbildung 3.10 dargestellt.

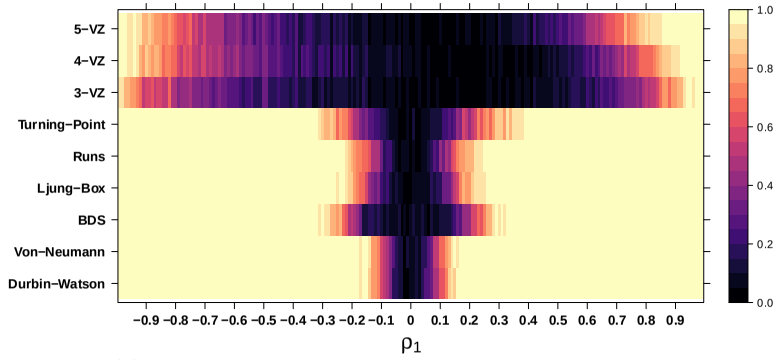
Wie auch schon bei den Cauchy-verteilten Innovationen ersichtlich, ist hier eine leichte Verbesserung der Trennschärfen der nichtparametrischen Tests gegenüber denen bei normalverteilten Innovationen zu erkennen. Dabei unterscheiden sich die Ergebnisse der beiden betrachteten Extremwertverteilungen aber nicht sehr stark. Insgesamt scheinen die Trennschärfen der Verfahren bei Laplace-verteilten Innovationen leicht besser zu sein. Dies macht sich besonders beim Runs-Test bemerkbar. Eine Erklärung dafür könnte die Neigung der Gumbel-Verteilung sein, eher positive extreme Werte anzunehmen sein. Das könnte dazu führen, dass die Anzahl der Runs trotz der Zentrierung durch den Median verzerrt ist. Eine weitere mögliche Ursache für die Überlegenheit der Testverfahren bei Cauchy- bzw. Laplace-verteilten Innovationen gegenüber dem Fall, dass sie aus einer Gumbel-Verteilung gezogen werden, könnte die Häufigkeit sein,



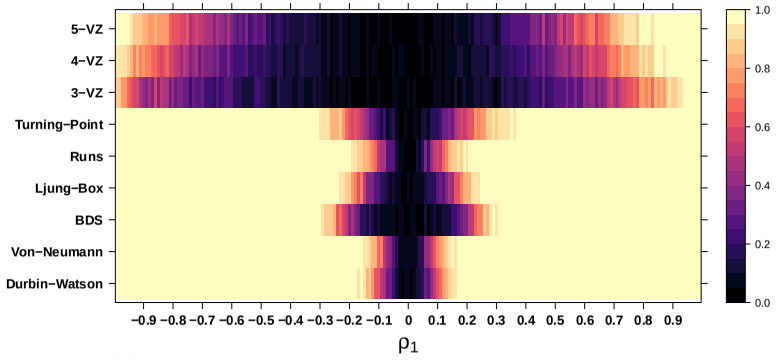
(a) Gumbel-verteilte Innovationen mit $N = 50$



(b) Laplace-verteilte Innovationen mit $N = 50$



(c) Gumbel-verteilte Innovationen mit $N = 500$



(d) Laplace-verteilte Innovationen mit $N = 500$

Abbildung 3.10: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit Gumbel- bzw. Laplace-verteilten Innovationen, mit Skalenparametern von 2, für unterschiedliche Stichprobenumfänge

mit der extreme Innovationen auftreten. So ist es z. B. bei der Gumbel-Verteilung wahrscheinlich, dass eine extreme Innovation schnell wieder durch eine weitere ausgeglichen wird, sodass das abklingende Verhalten dieser Schocks unterbrochen wird. Dies kann die Verbesserung der Trennschärfe durch extreme Innovationen abschwächen.

Im Hinblick auf die K-VZ-Tests fällt im Fall von Gumbel-verteilten Innovationen auf, dass der Bereich, in dem die Nullhypothese sicher angenommen wird (schwarz) im Vergleich zu den anderen Testverfahren und zu den Untersuchungen unter Normalbedingungen etwas verschoben zu sein scheint. Dies ist auf die Zentrierung der Beobachtungen zurückzuführen und deutet darauf hin, dass dieses Vorgehen keine optimale Lösung für Innovationsverteilungen zu sein scheint, bei denen die Annahme eines Medians von 0 nicht erfüllt ist.

Weiter ist ersichtlich, dass die Trennschärfen des DW-Tests und des LB-Tests hier weniger stark beeinflusst werden als bei der Cauchy-Verteilung. Ein Grund dafür könnte die geringe Dichte der Gumbel-Verteilung im Bereich von moderat großen Werten sein. Sie können dazu führen, die oben erwähnte Verschleierung von Korrelationen abgeschwächt wird. Aber auch die Tatsache, dass die Momente der Gumbel- und der Laplace-Verteilung existieren und die Anwendung des Zentralen Grenzwertsatzes somit möglich ist, könnte eine Erklärung für diese Beobachtungen sein.

Insgesamt deuten diese Simulationsergebnisse darauf hin, dass die nichtparametrischen Tests robust gegenüber Innovationen aus Verteilungen mit schweren Rändern reagieren. Anscheinend können sie sogar von solchen Bedingungen profitieren. Die Testentscheidung der parametrischen Testverfahren wird hingegen von den anders verteilten Innovationen beeinflusst, auch wenn ihre Trennschärfen nicht immer wesentlich darunter leiden. Insbesondere Verteilungen, deren Momente nicht existieren und damit eine Anwendung des zentralen Grenzwertsatzes unmöglich machen, führen hier aber zu deutlichen Verschlechterungen der Trennschärfe. Unter diesen Gesichtspunkten schneidet der VNRR-Test im Fall von anders verteilten Innovationen am besten ab, da er aufgrund der Betrachtung von Rängen einen Vorteil gegenüber den anderen nichtparametrischen Verfahren hat und gleichzeitig von deren robusten Eigenschaften profitiert. Außerdem wird deutlich, dass eine Asymmetrie der Verteilungen für den Runs-Test und die K-VZ-Tests, die sich am Median der Daten orientieren, zu Problemen führen kann.

3.1.2 Innovative Ausreißer

Das Auftreten von extremen Innovationen kann auch auf sogenannte innovative Ausreißer zurückzuführen sein. Das bedeutet, die Innovationen des betrachteten Prozesses entstammen zwar einer Normalverteilung, jedoch sind dafür einige davon deutlich zu groß und können als Ausreißer betrachtet werden. In der Praxis ist es möglich, dass diese Art von Ausreißern durch seltene, geplante oder ungeplante Ereignisse im Verlauf des Prozesses verursacht werden, die eine vorübergehende, abrupte Veränderung des Niveaus der Zeitreihe bewirken. Diese Ereignisse werden auch als Interventionen oder, im Zeitreihen-Kontext, als „Random Shocks“ bezeichnet.

Damit untersucht werden kann, wie sich die Intensität der Ausreißer auf die Testverfahren auswirkt, wird in dieser Arbeit mit einer festgesetzten Intensität gearbeitet. Dafür sollen die Innovationen an zufälligen Beobachtungszeitpunkten um einen Wert von 10 erhöht bzw. erniedrigt werden. Ein alternatives Vorgehen, um innovative Ausreißer zu erzeugen, stellt z. B. eine Vervielfachung des Wertes der Innovationen an den Zeitpunkten des Schocks dar. Dadurch kann jedoch nicht kontrolliert werden, wie extrem der Ausreißer ist. Durch den hier verfolgten Ansatz soll also eine bessere Beurteilung der Intensität von möglichen Interventionen erreicht werden, die es einem Anwender ermöglicht, mit gewissen Vorinformationen über den betrachteten Prozess den dafür bestmöglichen Test auszuwählen.

Mathematisch werden in diesem Abschnitt Prozesse der Form:

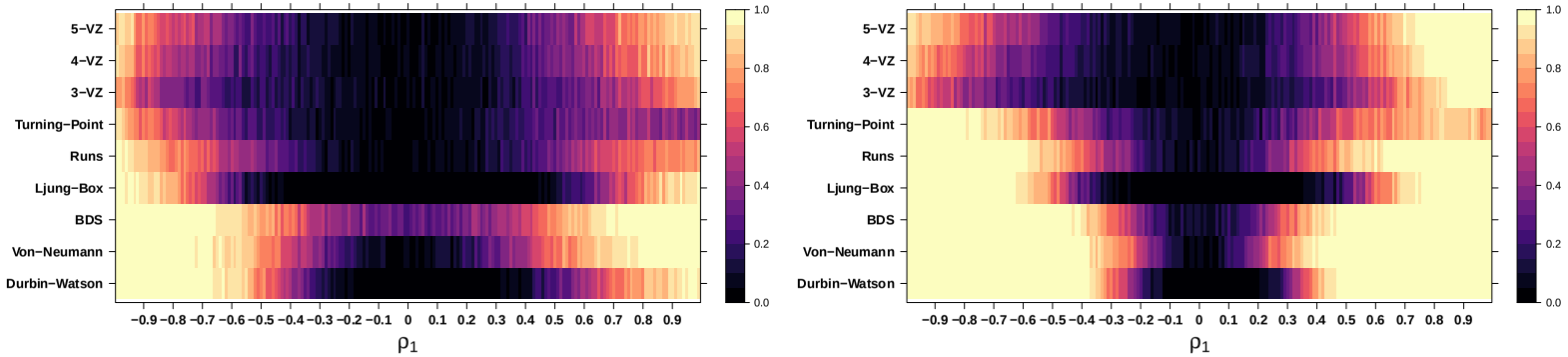
$$x_t = \rho_1 x_{t-1} + (w_t + \mathbb{1}_{InInd}\{t\} \cdot InInt)\,, \quad |\rho_1| < 1, \quad w_t \sim WN(0, \sigma^2_{WN})$$

betrachtet. Dabei bezeichnet $InInd$ die zufällig generierten Indizes, an denen der Prozess innovative Ausreißer enthält und $InInt$ deren Intensität, die hier als 10 bzw. -10 gewählt wurde. Die Parität von $InInt$ wurde dabei ebenfalls für jeden innovativen Ausreißer zufällig generiert.

Zunächst wurde untersucht, wie die verschiedenen Testverfahren in dem Fall reagieren, dass die betrachteten Zeitreihen 5 % innovative Ausreißer enthalten. Die entsprechenden Ergebnisse zeigt Abbildung 3.11.

Im Vergleich mit den Trennschärfen unter Normalbedingungen fällt auf, dass es wesentliche Unterschiede bezüglich der Reaktionen zwischen den parametrischen und nichtparametrischen Verfahren gibt. So gewinnen sämtliche nichtparametrischen Tests durch die innovativen Ausreißer sogar an Trennschärfe dazu, wie es auch schon bei anders verteilten Innovationen zu sehen war. Am auffälligsten ist dieser Effekt bei den K-VZ-Tests im Bereich positiver Korrelationen. Auch dem BDS-Test gelingt es in diesem Szenario deutlich häufiger, das Niveau des Tests einzuhalten. Anders als unter Normalbedingungen kann dieser bereits bei einem Stichprobenumfang von $N = 50$ mit den anderen nichtparametrischen Verfahren mithalten.

Die parametrischen Tests zeigen hier ein anderes Verhalten. So gelingt es ihnen zwar, die Nullhypothese sogar bei ähnlich starker Korrelation wie unter Normalbedingungen eindeutig abzulehnen, in Bereichen moderater Korrelation gelingt eine Verwerfung jedoch deutlich seltener. Diese Beobachtungen bedeuten, dass die parametrischen Tests in diesem Szenario dazu neigen, klarere Entscheidungen zu treffen: Die Bereiche, in denen diese Verfahren Unsicherheiten aufweisen − also die Nullhypothese für ca. 50 % der simulierten Zeitreihen ablehnen − werden auffällig stark verkleinert. Erkennbar ist diese Tendenz in den Grafiken anhand des deutlich breiteren, schwarzen Bereiches um den Wert $\rho_1 = 0$, in dem das Testniveau von 5 % eingehalten wird sowie an den schärferen Ränder der simulierten Trennschärfen. Eine weitere Auffälligkeit in diesem Szenario ist, dass die parametrischen Verfahren als Konsequenz der innovativen Ausreißer eine asymmetrische Trennschärfe entwickeln. Speziell gelingt es den Verfahren hier besser, negative Korrelationen zu erkennen als positive.



(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

Abbildung 3.11: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit 5 % innovativen Ausreißern der Intensität 10, für unterschiedliche Stichprobenumfänge

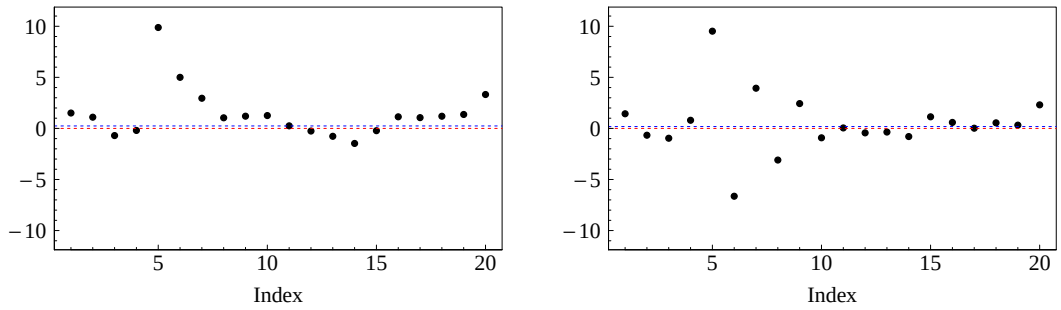
Der Runs- und der VNRR-Test weisen hier die besten Trennschärfen auf, da sie von den wünschenswerten robusten Eigenschaften nichtparametrischer Verfahren profitieren und nicht negativ durch die Ausreißer beeinflusst werden. Dabei ist der VNRR-Test dem Runs-Test aufgrund seiner Mehrinformationen durch die Betrachtung der Ränge wieder leicht überlegen.

Die Reaktionen der nichtparametrischen Verfahren lassen sich in diesem Szenario durch ähnliche Überlegungen begründen, wie es bei den Cauchy-verteilten Innovationen der Fall war. So folgt nach extremen Innovationen im AR(1)-Modell eine Kompensationsphase, durch die die Struktur der Zeitreihe deutlich beeinflusst wird. Anhand von Abbildung 3.12 wird dieses Verhalten noch einmal für starke Korrelationen von 0.9 bzw. -0.9 und moderaten Korrelationen von 0.6 bzw. -0.6 veranschaulicht. Dabei wurden Zeitreihen abgebildet, die zum Zeitpunkt 5 eine positive Innovation der Intensität 10 erfahren. Das Verhalten der 15 darauf folgenden Beobachtungen ist dargestellt.

Die erwähnte Systematik, dass das sequenzielle Schema der Beobachtung durch die Interventionen mit betragsmäßig größer werdenden Autokorrelationskoeffizienten zunehmend stark beeinflusst wird, ist hier klar erkennbar. Dies äußert sich konkret durch die deutlich größere Anzahl an Beobachtungen nach einem Schock, die benötigt wird, um wieder zum langfristigen Niveau der Zeitreihe zurückzukehren. Insbesondere führen negative Korrelationen nach dem Schockereignis zu mehr Runs, Turning-Points, Vorzeichenwechseln und dazu, dass benachbarte Beobachtungen weit auseinanderliegen und damit deutlich unterschiedliche Ränge aufweisen. Auf der anderen Seite verursachen positive Korrelationen im selben Kontext lange und damit weniger Runs, weniger Turning-Points, weniger Vorzeichenwechsel und sie führen dazu, dass aufeinanderfolgende Beobachtungen eine ähnliche Größenordnung aufweisen, wodurch auch ihre Ränge nah beieinander liegen. Auf diese Weise wird nachvollziehbar, weshalb die Trennschärfe der nichtparametrischen Verfahren von den innovativen Ausreißern profitiert. Auch werden die Überlegungen zu einer Verbesserung ihrer Trennschärfe im Fall, dass die Innovationen aus einer Verteilung stammen, in der extreme Werte häufiger vorkommen als bei der Normalverteilung, untermauert.

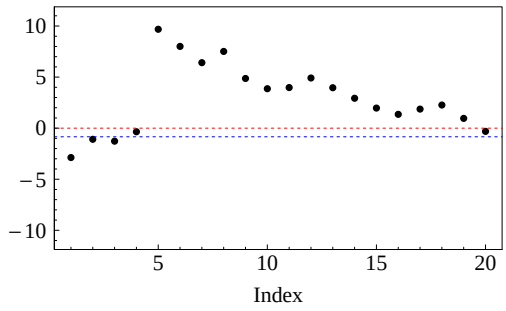
Um das Verhalten der parametrischen Verfahren zu verstehen, können erneut ähnliche Überlegungen wie z. B. im Fall von Cauchy-verteilten Innovationen angestellt werden. So können vorhandene moderate Korrelationen durch die innovativen Ausreißer verschleiert werden. Dabei scheint es einen Unterschied zu machen, ob eine positive oder eine negative Korrelation vorliegt. Um diese Tendenz noch einmal zu veranschaulichen, wurde das Korrelogramm einer Zeitreihe mit $\rho_1 = 0.2$ sowie mit $\rho_1 = -0.2$ dem Korrelogramm derselben Zeitreihe mit nachträglich eingefügten innovativen Ausreißern in Abbildung 3.13 gegenübergestellt.

Dabei gilt es zu beachten, dass Korrelationen der Stärke 0.2 bei dem betrachteten Stichprobenumfang von $N = 100$ für den DW-Test in einem Bereich liegen, in dem eine korrekte Ablehnung der Nullhypothese generell schwierig ist. So entspricht der kritische Wert des empirischen Autokorrelationskoeffizienten hier ziemlich genau der vorhandenen Korrelation in der

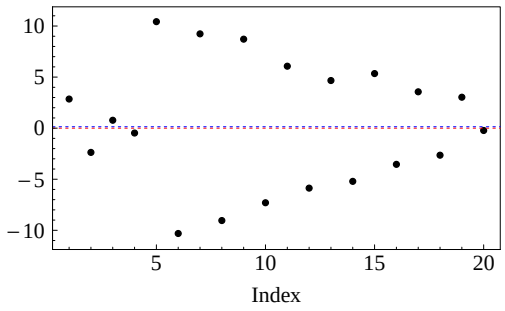


(a) $\rho_1 = 0.6$

(b) ρ_1 = -0.6



(c) $\rho_1 = 0.9$

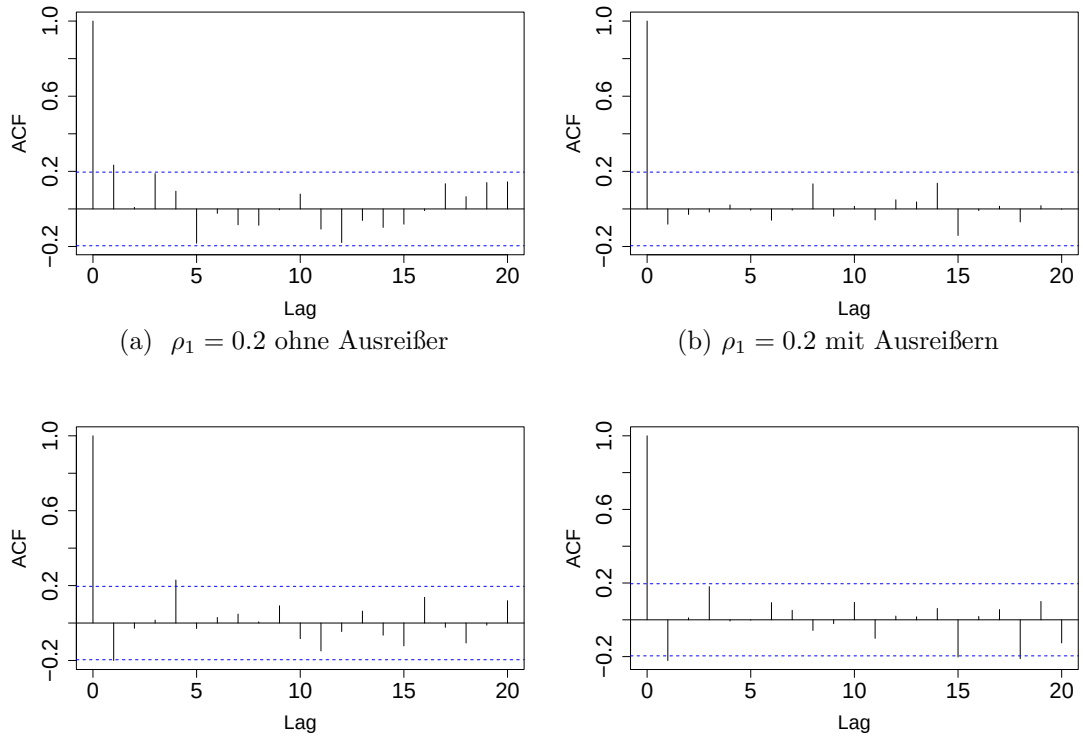


(d) ρ_1 = -0.9

Abbildung 3.12: Ausschnitte aus Zeitreihen mit innovativen Ausreißern beim Index 5, mit der Intensität 10, zu unterschiedlichen Autokorrelationskoeffizienten, Nulllinie (rot) und empirischer Median (blau)

Zeitreihe. Damit kann die Nullhypothese unter Normalbedingungen ungefähr nur in der Hälfte der Fälle stattfinden, nämlich wenn die tatsächliche Korrelation nicht unterschätzt wird. Anhand der Korrelogramme kann wieder erkannt werden, dass die innovativen Ausreißer zu einer Unterschätzung des Autokorrelationskoeffizienten $\hat{\rho}_1$ führen können und die auf ihm basierenden Tests damit nicht in der Lage sind, die Nullhypothese zu verwerfen. Auf der anderen Seite bewirken die innovativen Ausreißer bei einem Autokorrelationskoeffizienten von -0.2, dass die empirische Korrelation 1. Grades als zu klein eingeschätzt wird, wodurch eine Verwerfung der Nullhypothese wahrscheinlicher wird.

Um diese Systematiken genauer zu erörtern, wurde mit Hilfe einer kleinen Simulationsstudie untersucht, wie sich die Schätzer der empirischen Autokorrelationskoeffizienten beim Vorhandensein von innovativen Ausreißern in der Zeitreihe verhalten. Dafür wurden auf einem Gitter der Feinheit 0.01 Autokorrelationskoeffizienten von -0.99 bis 0.99 betrachtet, wobei für jeden der Gitterpunkte 1000 Zeitreihen mit $N = 100$ Beobachtungen und 5 % Interventionen der Intensität 10 simuliert worden sind. Anschließend wurde die mittlere Abweichung des empirischen Autokorrelationskoeffizienten von der tatsächlichen Korrelation, also der Wert ($\hat{\rho}_1$ - ρ_1),



(a) $\rho_1 = 0.2$ ohne Ausreißer

(b) $\rho_1 = 0.2$ mit Ausreißern

(c) ρ_1 = -0.2 ohne Ausreißer

(d) ρ_1 = -0.2 mit Ausreißern

Abbildung 3.13: Korrelogramm zufälliger Zeitreihen aus AR(1)-Prozessen mit $N = 100$ Beobachtungen ohne ((a) und (c)) und mit 5 % innovativen Ausreißern ((b) und (d)), bei $\rho_1 = 0.2$ und ρ_1 = -0.2 und kritischen Werten (blau)

für jede der oben beschriebenen Korrelationen berechnet. Die entsprechenden Ergebnisse dieser Simulationen sind in Abbildung 3.14 dargestellt.

Dabei bedeuten negative Differenzen, dass der Schätzer die tatsächliche Korrelation unterschätzt. Werte, die größer als 0 sind, deuten hingegen auf eine systematische Überschätzung der Korrelation hin. Im Idealfall, unter Normalbedingungen, sollten sich die mittleren Abweichungen unabhängig von dem Wert ρ_1 gleichmäßig um 0 verteilen.

Aus dieser Darstellung geht hervor, dass tatsächlich eine Systematik erkennbar ist. So führen Ausreißer bei positiven Werten von ρ_1 zu einer systematischen Unterschätzung des Autokorrelationskoeffizienten, die sich mit zunehmender Intensität der Korrelation verstärkt. Auf diese Weise lässt sich die Verschleierung moderater positiver Korrelationen nachvollziehen. Weiter ist auffällig, dass kleine negative Korrelationen ebenfalls zu einer Unterschätzung durch $\hat{\rho}_1$ führen, wodurch die Nullhypothese in diesem Fall mit größerer Wahrscheinlichkeit verworfen wird. Ab einem Wert des Parameters ρ_1 von -0.4 scheinen die tatsächlichen Korrelationen jedoch eher überschätzt zu werden. Anhand dieser Erkenntnisse kann auch das asymmetrische Verhalten der parametrischen Verfahren im Fall von innovativen Ausreißern nachvollzogen werden.

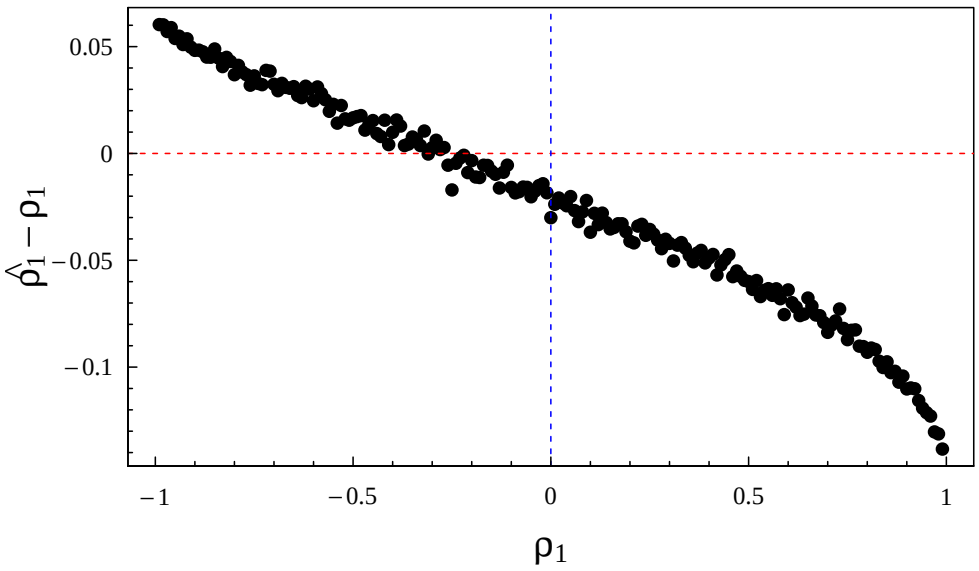
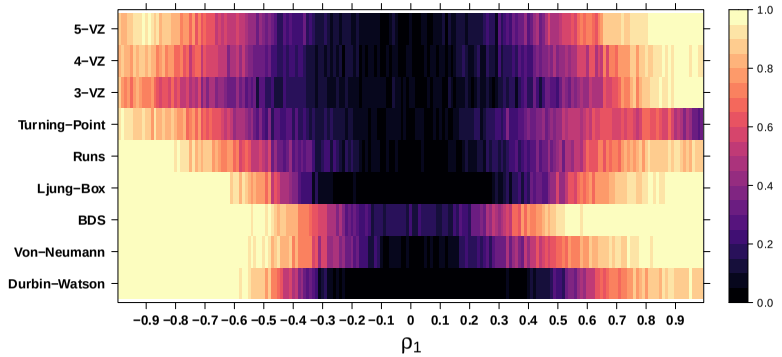


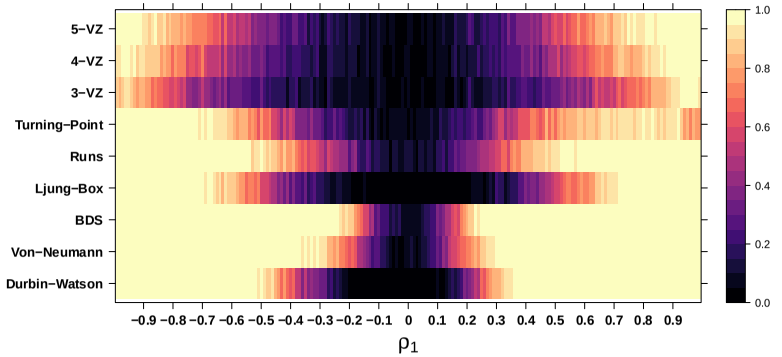
Abbildung 3.14: Simulierte mittlere Differenz des Autokorrelationsschätzers und der Autokorrelation bei Vorhandensein von 5 % Interventionen der Intensität 10 in Zeitreihen mit $N = 100$ Beobachtungen, mit Nullniveau (rot) und Symmetrieachse (blau)

Weiter ist es interessant zu untersuchen, wie die Testverfahren reagieren, falls der Anteil an innovativen Ausreißern erhöht wird. Dazu wurden obige Simulationen mit einem Innovationsanteil von 10 % bzw. 20 % wiederholt. Die entsprechenden Ergebnisse sind in Abbildung 3.15 und 3.16 dargestellt.

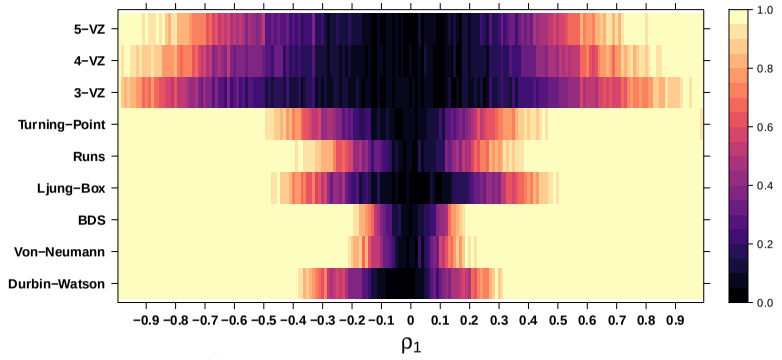
Die Abbildungen zeigen, dass sich die Tendenzen der unterschiedlichen Verfahren bei Stichprobenumfängen von $N \geq 50$ mit größerem Anteil an innovativen Ausreißern verstärken. Im Fall von $N = 20$ Beobachtungen tritt allerdings beim Sprung von 10 %- auf 20 %-Anteil an Ausreißern eine leichte Verschlechterung der Trennschärfen auf. Dies könnte auf die zu kleine Stichprobe zurückzuführen sein, in der die Kompensationsphase nach einem Schock nicht hinreichend lang werden kann, um weiterhin davon zu profitieren. Außerdem scheint der BDS-Test sehr positiv von den Ausreißern beeinflusst zu werden, sodass er bei einem Ausreißeranteil von 20 % bereits bei $N = 20$ Beobachtungen eine ausgezeichnete Trennschärfe aufweist. Allerdings wird ersichtlich, dass der BDS-Test in diesem Szenario kaum noch von einem wachsenden Stichprobenumfang profitieren kann. So ist keine deutliche Verbesserung seiner Trennschärfe beim Sprung von $N = 50$ auf $N = 500$ Beobachtungen erkennbar. Damit scheint seine Testentscheidungen und vor allem seine Konsistenzeigenschaften stark von den Ausreißern beeinflusst zu werden, was eine Beurteilung der Testentscheidung in der Praxis erschweren kann.



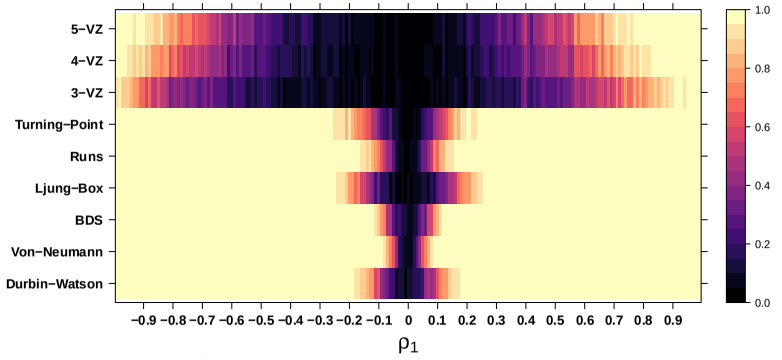
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$

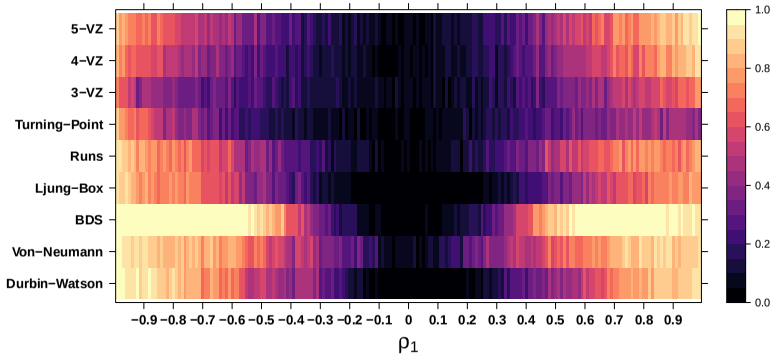


(c) Stichprobenumfang $N = 100$

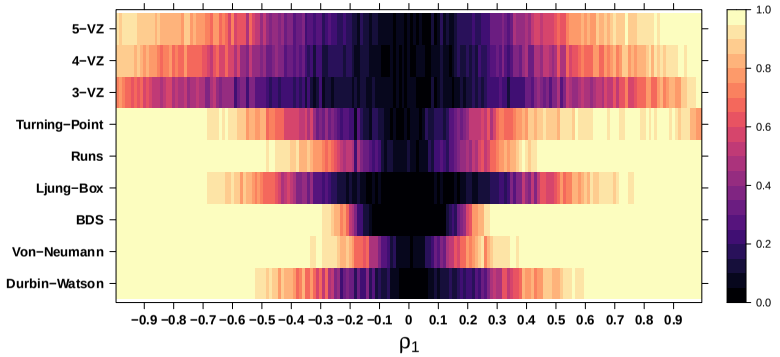


(d) Stichprobenumfang $N = 500$

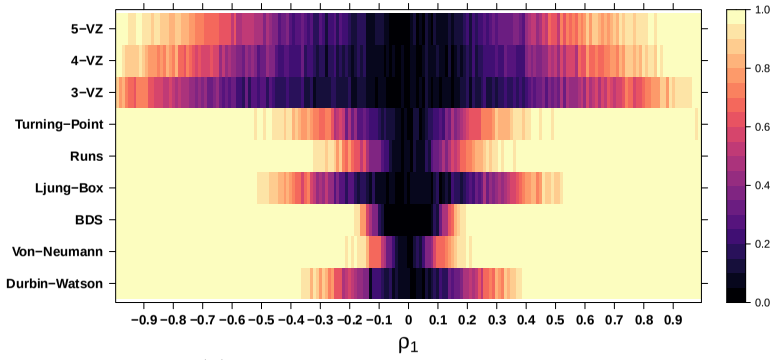
Abbildung 3.15: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 10 % innovativen Ausreißern der Intensität 10, für unterschiedliche Stichprobenumfänge



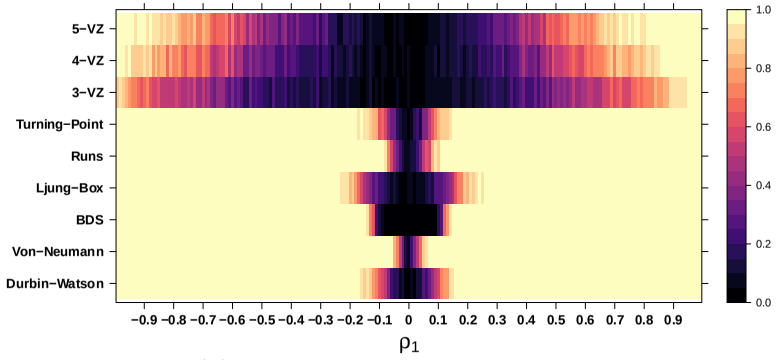
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



(d) Stichprobenumfang $N = 500$

Abbildung 3.16: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 20 % innovativen Ausreißern der Intensität 10, für unterschiedliche Stichprobenumfänge

Auffällig ist auch, dass die parametrischen Verfahren bei zunehmender Anzahl von innovativen Ausreißern, genau wie die nichtparametrischen Verfahren, ein wenig an Trennschärfe zu gewinnen scheinen. Allerdings scheint der Annahmebereiches des DW-Tests im Fall von mehr Ausreißern eine deutlichere Asymmetrie in Abhängigkeit von Stichprobenumfang zu entwickeln.

Unabhängig vom Interventionsanteil sticht der VNRR-Test hier erneut als Test mit der besten Trennschärfe heraus. Dabei zeichnet er sich vor allem durch seine Robustheit aus, da er, anders als z. B. der BDS-Test, der seinerseits dem VNRR-Test in einigen Fällen überlegen zu sein scheint, bei allen betrachteten Ausreißeranteilen ein beständiges und berechenbares Ablehnungsverhalten zeigt.

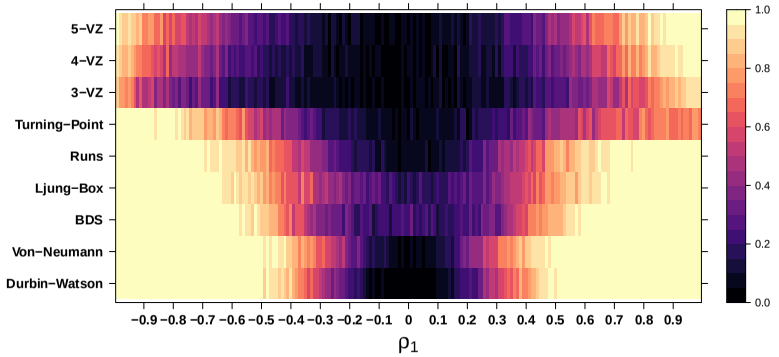
Schließlich gilt es noch zu untersuchen, inwieweit sich die Intensität der Interventionen auf die Trennschärfe der verschiedenen Testverfahren auswirkt. Dazu wurde der Ausreißeranteil auf 5 % beschränkt, da es sich dabei um einen in der Praxis realistischen Anteil handelt und die Auswirkungen der Ausreißer auf die Testentscheidungen auch hier schon deutlich ersichtlich waren. Im Folgenden werden konkret Ausreißer der Intensität 5, 20, 50 und 100 für Stichprobenumfänge von $N = 50$ bzw. $N = 500$ betrachtet. Die entsprechenden Simulationsergebnisse sind in Abbildung 3.17 und 3.18 dargestellt.

Betrachtet man die Entwicklung der Trennschärfen für zunehmende Ausreißerintensitäten, so wird deutlich, dass sich die oben beschriebenen Tendenzen verstärken. Dies wird zudem durch die größere Ähnlichkeit der Trennschärfen im Fall von Interventionen der Stärke 5 zu den Ergebnissen unter Normalbedingungen untermauert.

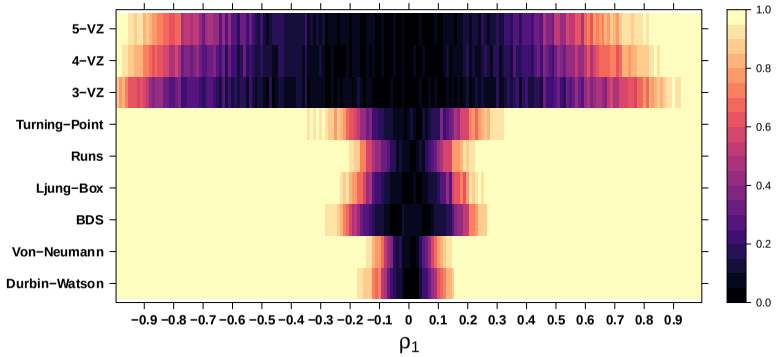
Diese Beobachtungen bei den nichtparametrischen Verfahren sind mit der Tatsache zu erklären, dass größere Werte des Ausreißers zu einer längeren Abklingphase des Schocks führen. Somit wird das sequenzielle Schema der Zeitreihe mit größerer Intensität zunehmend stark beeinflusst, was diesen Tests eine Verwerfung der Nullhypothese erleichtert.

Die Reaktion der parametrischen Verfahren ist dabei etwas diffuser und folgt keinem eindeutigen Schema. So gibt es Fälle, in denen eher die Asymmetrie ihrer Trennschärfen verstärkt wird und wieder andere, in denen die Testverfahren Probleme bei der Einhaltung des Niveaus bekommen. Allgemein kann jedoch die Tendenz beobachtet werden, dass eine zunehmende Ausreißerintensität die Trennschärfe dieser Testverfahren verschlechtert − und vor allem unberechenbarer macht.

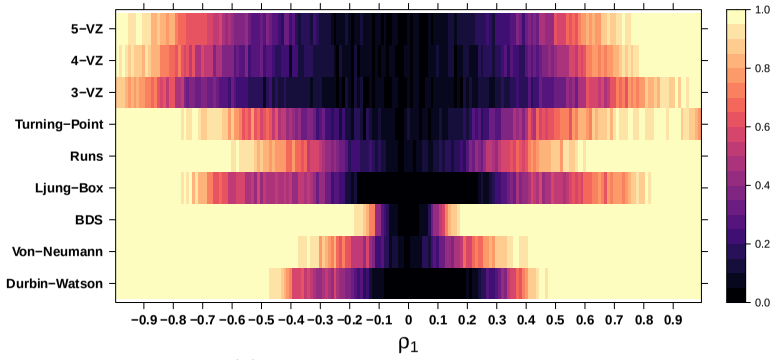
Insgesamt legen die Simulationsergebnisse nah, dass die Testverfahren bei innovativen Ausreißern ähnliche Reaktionen zeigen, wie im Fall von Innovationsverteilungen mit schweren Ränder: Parametrische Verfahren scheinen deutlich unter innovativen Ausreißern zu leiden, während nichtparametrische Verfahren sogar davon profitieren können. So verbessert sich die Trennschärfe dieser Tests − und vor allem der K-VZ-Tests − mit zunehmendem Ausreißeranteil und deren Intensität. Die K-VZ-Tests erreichen trotz der deutlichen Verbesserungen gegenüber dem Fall einer Normalverteilung in größeren Stichproben immer noch keine Güte, die mit anderen Verfahren mithalten kann. Die parametrischen Verfahren entwickeln bei größerer Abweichung von



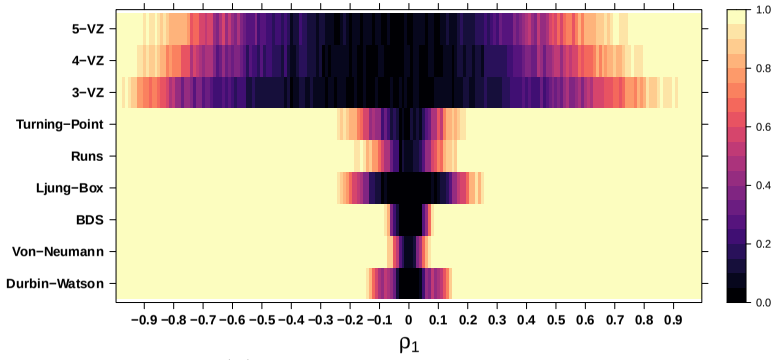
(a) $N = 50$, Intensität= 5



(b) $N = 500$, Intensität= 5

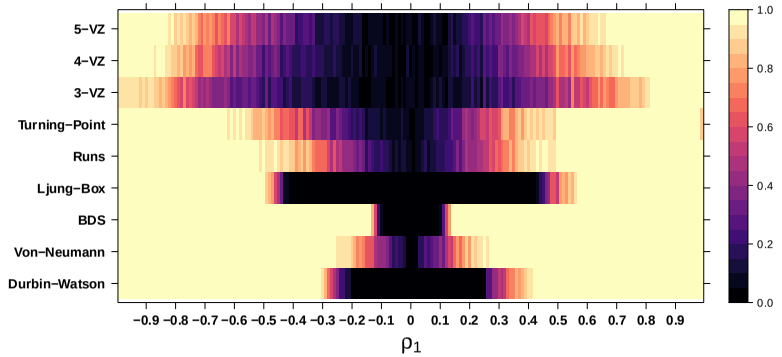


(c) $N = 50$, Intensität= 20

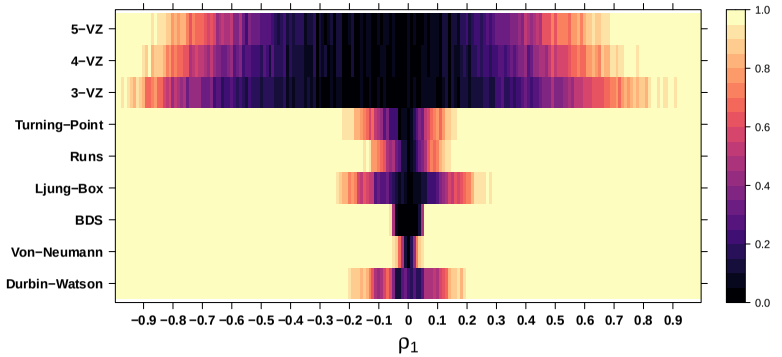


(d) $N = 500$, Intensität= 20

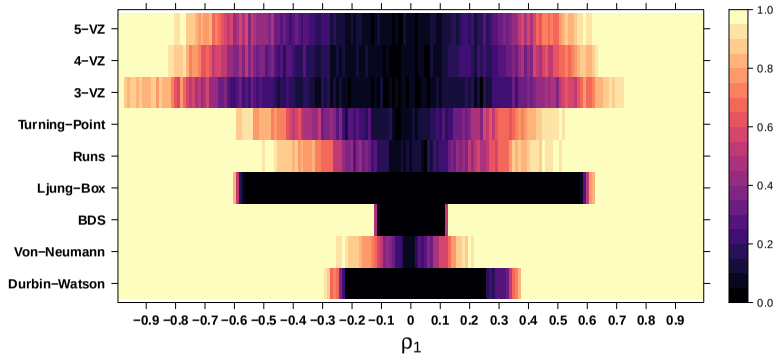
Abbildung 3.17: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 5 % innovativen Ausreißern der Intensität 5 bzw. 20, für unterschiedliche Stichprobenumfänge



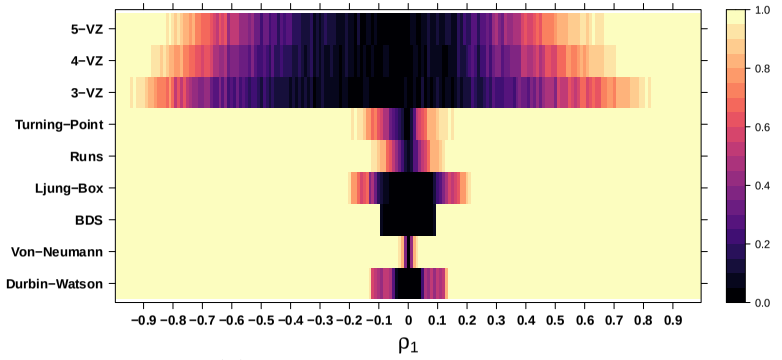
(a) $N = 50$, Intensität= 50



(b) $N = 500$, Intensität= 50



(c) $N = 50$, Intensität= 100



(d) $N = 500$, Intensität= 100

Abbildung 3.18: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 5 % innovativen Ausreißern der Intensität 50 bzw. 100, für unterschiedliche Stichprobenumfänge

den Normalbedingungen asymmetrische Ablehnungsbereiche und bekommen, unabhängig vom Stichprobenumfang, zunehmende Probleme bei der Ablehnung der Nullhypothese in Bereichen kleiner bis moderater Korrelationen. Eine Sonderstellung nimmt in diesem Szenario der BDS-Test ein, der deutlich von den Interventionen zu profitieren scheint. Allerdings leiden seine Konsistenzeigenschaften und die Verbesserung seiner Trennschärfe folgt keinem klaren Schema. Der am besten geeignetstes Testverfahren in diesem Abschnitt ist der VNRR-Test, der von den robusten Eigenschaften nichtparametrischer Verfahren profitiert und durch die Betrachtung der Ränge eine leicht bessere Trennschärfe aufweist als die übrigen nichtparametrischen Verfahren. Dicht gefolgt wird er vom Runs-Test, der ebenfalls sehr gute Ergebnisse liefert. Generell sollte vor allem auf parametrische Verfahren verzichtet werden, falls davon auszugehen ist, dass der zu untersuchende Prozess Interventionen beinhalten kann.

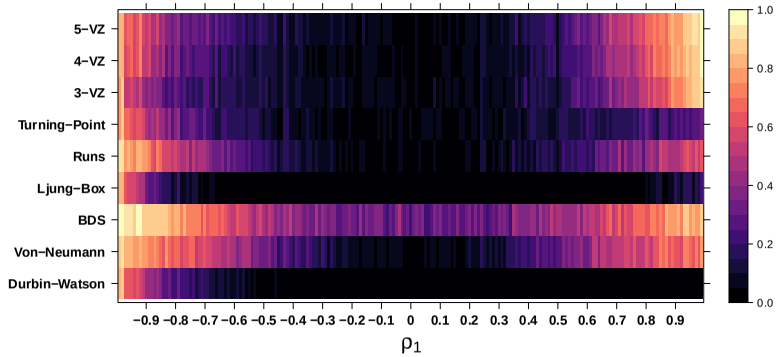
3.1.3 Kontaminationen

Ein weiteres Szenario stellt das Vorhandensein von Kontamination in den Zeitreihen dar, die auch als additive Ausreißer bezeichnet werden. Dabei handelt es sich heuristisch um Verunreinigungen des Prozesses, die sich anders als bei extremen Werten oder innovativen Ausreißern, wie sie bis jetzt betrachtet wurden, im AR(1)-Modell nicht fortpflanzen. Eine Situation, in der solche Kontaminationen einer Zeitreihe in der Praxis vorkommen, sind Fehler, die bei der Messung der Zeitreihe aufgetreten sind und sich nicht auf den zugrunde liegenden Prozess zurückführen lassen. Diese Fehler machen sich aber bei bloßer Betrachtung der Zeitreihe oft ebenfalls als Ausreißer bemerkbar und die Unterscheidung, um welche Art von Ausreißern es sich handelt, ist in der Regel nicht trivial. Deshalb ist es von großer Wichtigkeit, dass kontaminierte Messwerte nicht überinterpretiert werden, sondern Testverfahren in der Lage sind, sie als nicht essenziellen Teil des zugrunde liegenden Prozesses zu erkennen. Mathematisch werden im Folgenden also Prozesse der Form

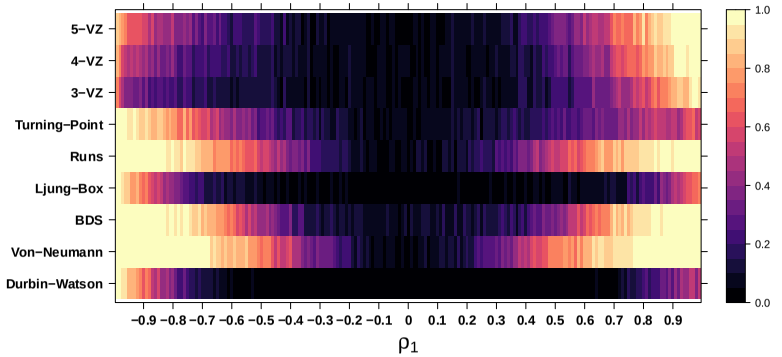
$$x_t + \mathbb{1}_{KontInd}\{t\} \cdot KontInt = \rho_1 x_{t-1} + w_t, \quad |\rho_1| < 1, \quad w_t \sim WN(0, \sigma^2_{WN})$$

betrachtet. Dabei entspricht $KontInd$ den zufällig generierten Indices, an denen Kontaminationen auftreten und $KontInt$ ihrer Intensität. Hier wurden bei jeder Zeitreihe 5 % der Daten zufällig kontaminiert, was einem in der praktischen Anwendung als realistisch anzunehmenden Anteil entspricht. Die Intensität der Kontamination wurde dabei auf 10 gesetzt und die entsprechende Parität ist für jede Kontamination zufällig generiert worden. Dies entspricht einer deutlichen Über- bzw. Unterschätzung des wahren Wertes der Zeitreihe. Zur Orientierung: Die Wahrscheinlichkeit eine standardnormalverteilte Innovation der Größe 10 oder höher bzw. -10 oder niedriger zu erhalten, beträgt nahezu 0. Die Simulationsergebnisse in diesem Szenario sind in Abbildung 3.19 dargestellt.

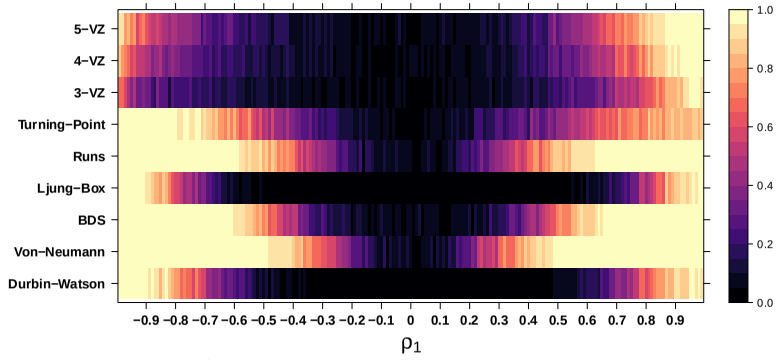
Auffällig ist hier vor allem, dass die parametrischen Testverfahren sichtlich unter den Kontaminationen leiden. So gelingt es bei einem kleinen Stichprobenumfang von $N = 20$, selbst bei



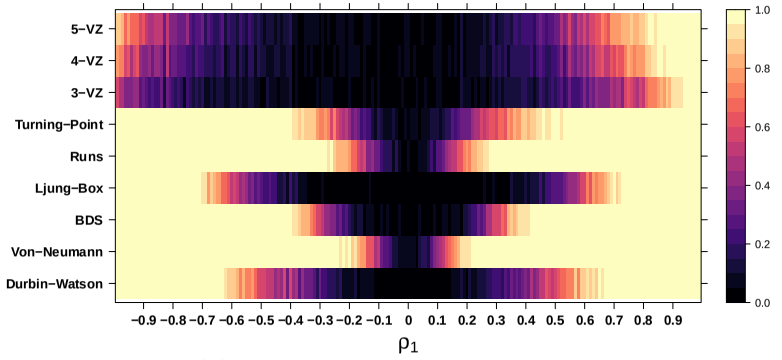
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



(d) Stichprobenumfang $N = 500$

Abbildung 3.19: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 5 % Kontaminationsanteil der Intensität 10, für unterschiedliche Stichprobenumfänge

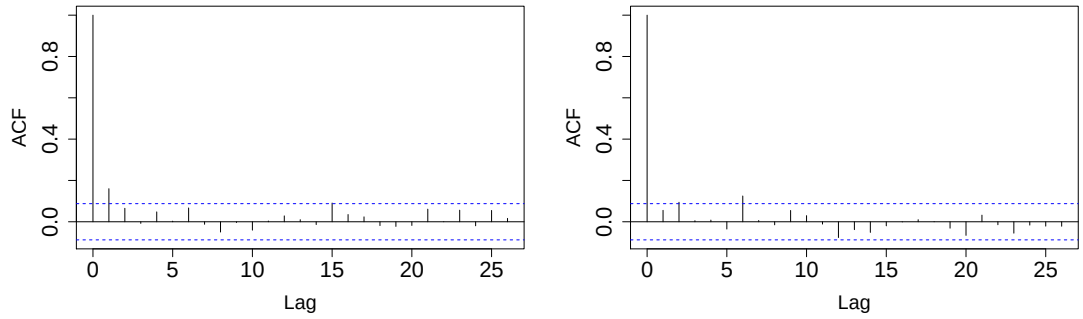
starken Korrelationen, weder dem DW-Test noch dem LB-Test, die Nullhypothese zuverlässig zu verwerfen. Auch bei größeren Beobachtungszahlen von bis zu $N = 500$ sind die Ergebnisse sehr unbefriedigend. So sind die Tests hier zwar dazu in der Lage, das Niveau unter der Unabhängigkeit einzuhalten, allerdings gelingt es ihnen weiterhin nicht, relativ starke Korrelationen von $|\rho_1| \approx 0.4$ zu erkennen. Damit erzielen diese Verfahren deutlich schlechtere Ergebnisse als die nichtparametrischen − mit Ausnahme der K-VZ-Tests (bei $N \geq 100$) − die erneut am wenigsten Trennschärfe aufweisen. Das größte Problem der parametrischen Verfahren besteht jedoch darin, dass sie die Nullhypothese selbst unter deutlichen Abweichungen des Korrelationskoeffizienten von 0 nicht eindeutig verwerfen können.

Am besten gelingt es dem VNRR-Test und dem Runs-Test, mit den Kontaminationen umzugehen, aber auch die Trennschärfen der K-VZ-Tests und des TP-Tests werden durch sie im Vergleich zur Simulation unter Normalbedingungen kaum verschlechtert.

Um nachvollziehen zu können, wie die Kontaminationen den DW-Test und den LB-Test beeinflussen, sind die Korrelogramme einer zufälligen Zeitreihe mit einem Beobachtungsumfang von $N = 500$ sowie derselben Zeitreihe mit nachträglich eingefügten Kontaminationen in Abbildung 3.20 dargestellt.

Hieraus wird ersichtlich, dass die empirische Korrelation zum Lag 1 durch die Kontaminationen deutlich abgeschwächt wird. Der Grund dafür liegt in der Berechnung der empirischen Autokorrelationskoeffizienten (s. Kap. 2.1). So werden ihr Zähler und Nenner durch die Kontaminationen verzerrt, sodass diese Normierung die tatsächliche Korrelation als zu gering einschätzt.

Die nichtparametrischen Verfahren werden durch die einzelnen Kontaminationen hingegen nur marginal beeinflusst, da die durch sie verursachten Ausreißer nicht in die Berechnung zukünftiger Werte eingehen. Aus diesem Grund findet kein Abklingen nach dem Auftreten extremer Werte statt und das sequenzielle Schema wird nur in geringem Maße und nur zum konkreten Zeit-



(a) normale Zeitreihe mit $\rho_1 = 0.2$ (b) kontaminierte Zeitreihe mit $\rho_1 = 0.2$

Abbildung 3.20: Korrelogramm einer zufälligen Zeitreihe mit $N = 500$ Beobachtungen ohne (a) und mit (b) 5 % Kontaminationsanteil der Intensität 10, bei $\rho_1 = 0.2$ und kritischen Werten (blau)

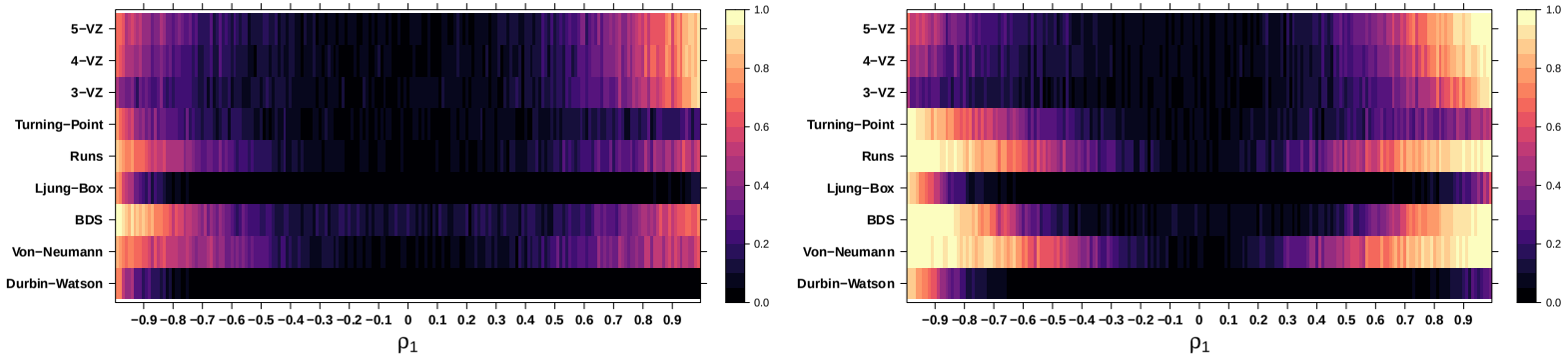
punkt der Kontamination verändert. Außerdem spielt die Höhe oder Intensität des Ausschlags für sie keine Rolle. Trotzdem kann es sein, dass durch die Kontaminationen z. B. ein Run vorzeitig beendet oder ein Turning-Point fälschlicherweise erzeugt wird. Da aber lediglich 5 % Kontaminationsanteile in den Zeitreihen vorhanden sind und solche Ereignisse auch unter der Nullhypothese erfolgen könnten, sind diese Effekte zu vernachlässigen.

Theoretisch müsste der VNRR-Test am stärksten durch die Kontaminationen beeinflusst werden, da den Beobachtungen zu den entsprechenden Zeitpunkten entweder sehr hohe oder sehr niedrige Ränge zugeordnet werden müssten. Das könnte zu einer Verzerrung der Teststatistik führen. Aber in diesem Szenario ist der Anteil an Kontaminationen offenbar zu gering, um eine systematische Verschlechterung der Ergebnisse zu verursachen.

Mit diesen Erkenntnissen ist es weiter interessant, zu untersuchen, wie die verschiedenen Verfahren reagieren, wenn der Anteil an Kontaminationen in den Zeitreihen erhöht wird. Im Folgenden werden deshalb Szenarien betrachtet, in denen 10 bzw. 20 % der Beobachtungen in den Zeitreihen kontaminiert worden sind. Die entsprechenden Ergebnisse zeigen die Abbildungen 3.21 und 3.22.

Wie bereits zu erwarten war, verschlechtert sich die Trennschärfe der parametrischen Verfahren mit zunehmendem Anteil an Kontaminationen immer weiter. Dabei sind die Unterschiede vor allem bei Stichprobenumfängen von $N \geq 100$ sichtbar, da die Nullhypothese in kleinen Stichproben schon bei 5 % beinahe für jeden Wert von ρ_1 nicht abgelehnt werden konnte. Aber auch die übrigen, nichtparametrischen Verfahren scheinen unter einem größeren Anteil von Kontaminationen zu leiden. So verschlechtert sich auch ihre Trennschärfe mit steigendem Anteil an Kontaminationen. Beim Beobachtungsumfängen von $N \leq 50$ und einem Kontaminationsanteil von 20 % gelingt es dabei keinem der Verfahren, moderate bis starke Korrelationen von $|\rho_1| < 0.8$ zu detektieren. Insgesamt können der VNRR-Test und der Runs-Test am besten mit den Kontaminationen umgehen, wobei diesmal der Runs-Test eine leicht bessere Trennschärfe aufweist. Besonders auffällig ist in diesem Szenario der BDS-Test, dessen Trennschärfe unter den nichtparametrischen Verfahren mit Abstand am drastischsten verschlechtert wird. Dabei gelingt es ihm aber überraschenderweise, anders als unter Normalbedingungen, das Niveau des Tests einzuhalten. Mit steigender Kontaminationsanzahl nähert sich seine Trennschärfe jedoch der von den parametrischen Verfahren schnell an.

Während die Simulationsergebnisse für die meisten Verfahren mit den generellen, vorangegangenen Überlegungen zu begründen sind und sich die beschriebenen Effekte bei einer höheren Anzahl von Kontaminationen lediglich verstärken, stellt sich nun noch die Frage, warum der BDS-Test so außergewöhnlich reagiert. Der Grund dafür liegt vermutlich in der Beschaffenheit seiner Teststatistik. So spielen bei ihm, anders als bei den übrigen nichtparametrischen Verfahren, die Intensitäten der Kontaminationen eine entscheidende Rolle. Im Hinblick auf die in Kapitel 2.8 beschriebene heuristische Veranschaulichung seiner Teststatistik, verringert sich die Wahrscheinlichkeit, dass 2 m-Historien weiter als ϵ auseinander liegen, durch die Kontaminatio-



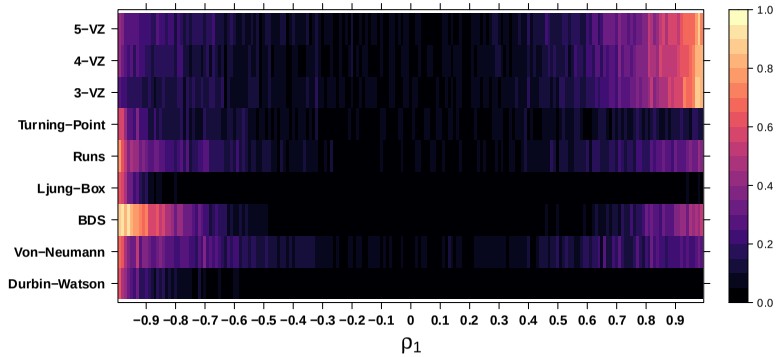
(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

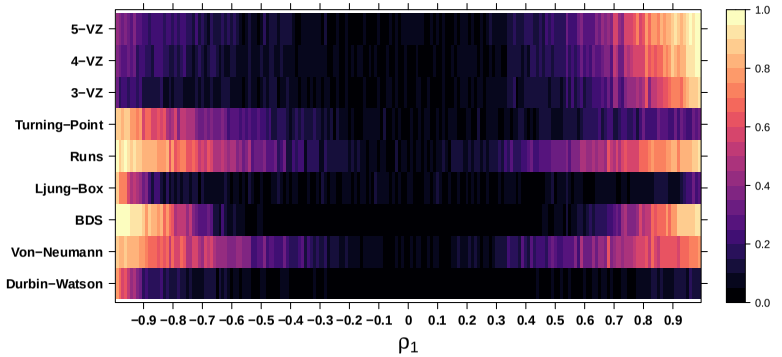
(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

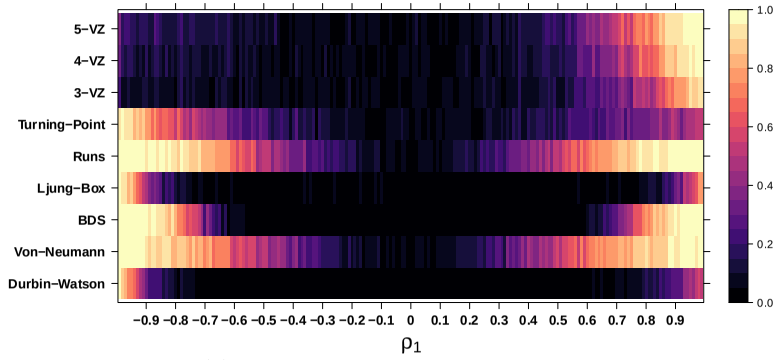
Abbildung 3.21: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 10 % Kontaminationsanteil der Intensität 10, für unterschiedliche Stichprobenumfänge



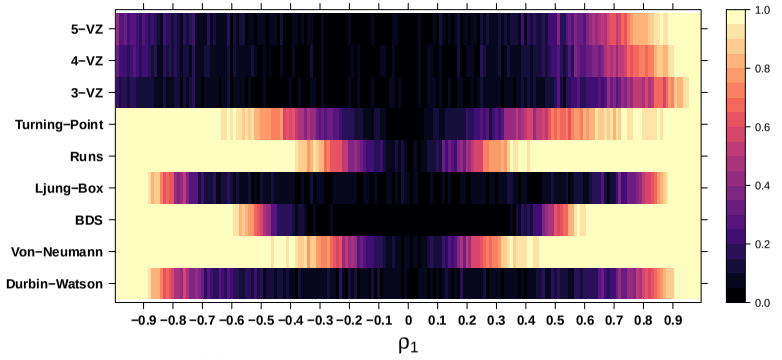
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



(d) Stichprobenumfang $N = 500$

Abbildung 3.22: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 20 % Kontaminationsanteil der Intensität 10, für unterschiedliche Stichprobenumfänge

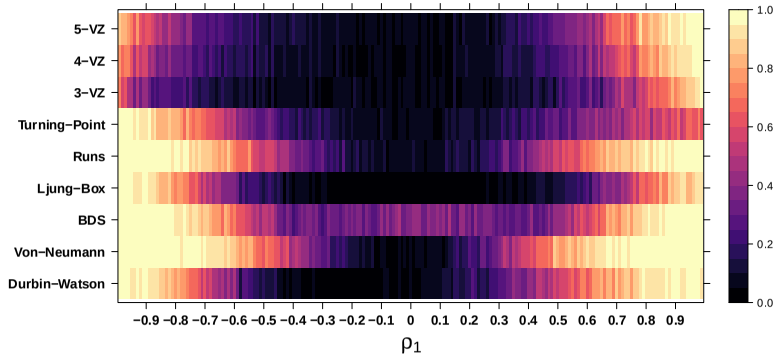
nen drastisch. Das lässt sich damit begründen, dass additive Ausreißer dafür sorgen, dass die Standartabweichung und damit auch ϵ hier wesentlich größer ist als unter Normalbedingungen. Auch kann der Test nicht − wie im Fall von Interventionen erläutert − von einer Abklingphase nach extremen Werten profitieren. Auf diese Weise könnte dem Test die Erkennung von Korrelationsstrukturen erschwert werden.

Weiter interessiert es, inwieweit die Intensität der Kontamination für die verschiedenen Testverfahren eine Rolle spielt. Exemplarisch wurden deshalb Zeitreihen mit Beobachtungsumfängen von $N = 50$ und $N = 500$ sowie einem Kontaminationsanteil von 5 % simuliert. Die Intensität der Kontaminationen ist dabei auf 5, 20, 50 und 100 gesetzt worden. Die entsprechenden Simulationsergebnisse sind in Abbildung 3.23 und 3.24 dargestellt.

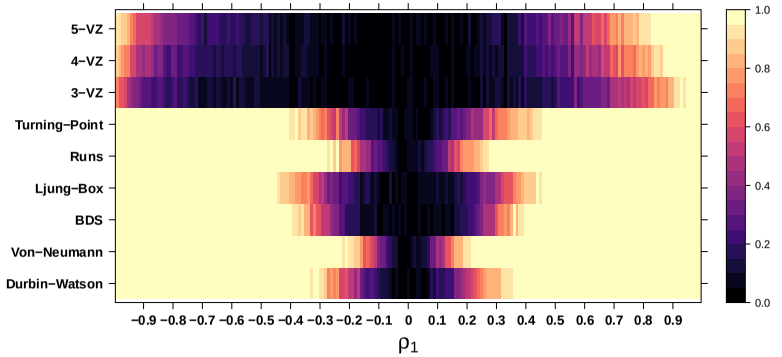
Es fällt auf, dass die Trennschärfen des VNRR-Tests, des Runs-Tests, des TP-Tests sowie der K-VZ-Tests durch eine Veränderung der Kontaminationsintensität nicht sichtbar beeinflusst werden. Da diese Testverfahren lediglich das sequenzielle Schema der Beobachtungen bzw. ihre Ränge betrachten und somit die konkreten Werte der Beobachtungen für die Tests keine Rolle spielen, sind diese Ergebnisse auch gut nachvollziehbar. Auf der anderen Seite reagieren die parametrischen Verfahren sowie der BDS-Test deutlich auf die oben beschriebenen Änderungen. Ihre Trennschärfen leiden mit größer werdender Intensität zunehmend stark unter den Kontaminationen, sodass sie unabhängig von Stichprobenumfang bei Intensitäten größer 50 quasi nicht mehr in der Lage sind, die Nullhypothese der Unabhängigkeit zu verwerfen. Bis zu einer Intensität von 50 können in großen Stichproben von $N = 500$ zumindest noch extreme Korrelationen mit $|\rho_1| \geq 0.9$ relativ zuverlässig detektiert werden.

Diese Beobachtungen zu den parametrischen Verfahren lassen sich dabei mit den oben diskutierten, durch die Kontaminationen verursachten Verschleierungen von Korrelationen, begründen. Im Fall des BDS-Tests lässt sich die drastische Verschlechterung erneut auf die Beschaffenheit seiner Teststatistik zurückführen, die im Wesentlichen auf den Abständen zwischen den Beobachtungen und dem Schwellenwert ϵ basiert. Anders als bei den übrigen nichtparametrischen Testverfahren spielen hier die konkreten Werte der Beobachtungen also eine entscheidende Rolle und starke Ausreißer, die die Standartabweichung der Stichprobe verändern, verzerren die Testentscheidungen des BDS-Tests.

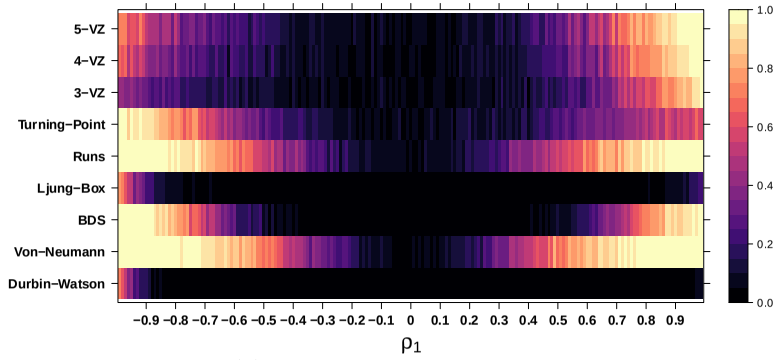
Insgesamt zeigt sich eine deutliche Überlegenheit nichtparametrischer Testverfahren − mit Ausnahme des BDS-Tests − im Fall, dass die betrachteten Zeitreihen Kontaminationen enthalten. Bei bis zu 5 % Kontaminationsanteil leiden diese Verfahren kaum unter den Verunreinigungen und deren Intensität spielt für sie keine Rolle. Die Trennschärfen des DW-Tests und des LB-Tests hingegen verschlechtern sich in solchen Szenarien so stark, dass sämtliche nichtparametrischen Verfahren − mit Ausnahme der K-VZ-Tests − die Nullhypothese unabhängig vom Stichprobenumfang bei geringeren Korrelationen verwerfen können. Auch profitieren sie weniger von einem wachsenden Stichprobenumfang, sodass davon ausgegangen werden kann, dass Kontaminationen die Konsistenzeingenschaften dieser Tests verschlechtern. Trotzdem gelingt es



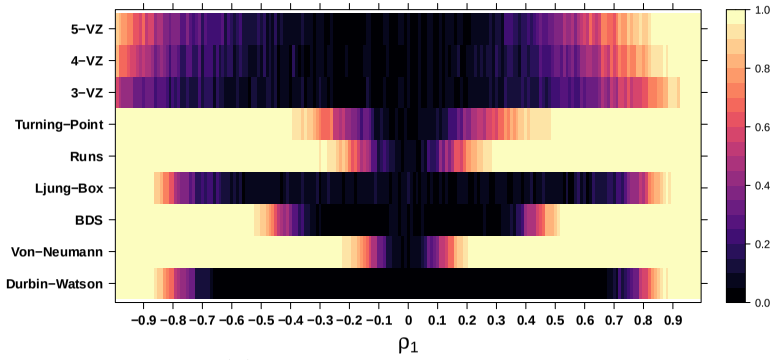
(a) $N = 50$, Intensität= 5



(b) $N = 500$, Intensität= 5

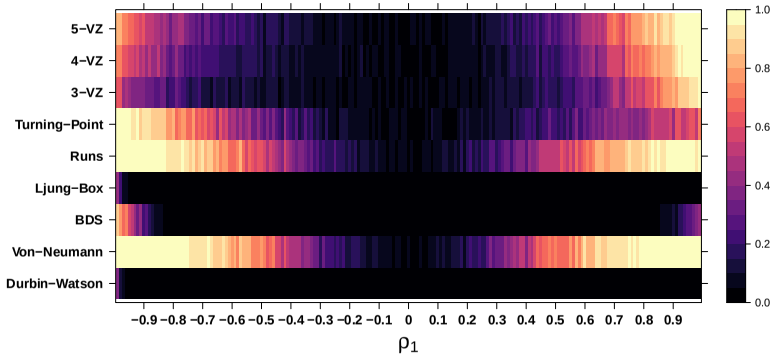


(c) $N = 50$, Intensität= 20

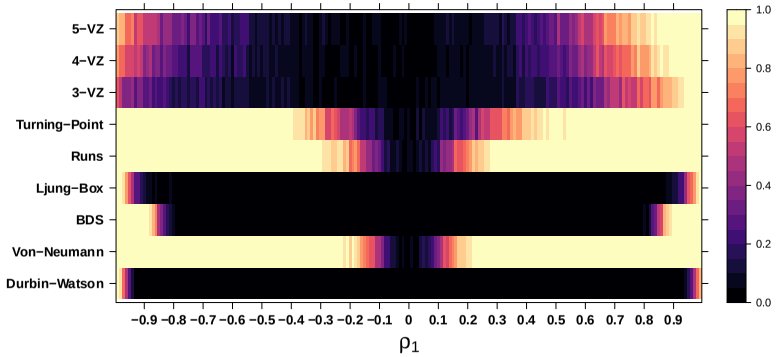


(d) $N = 500$, Intensität= 20

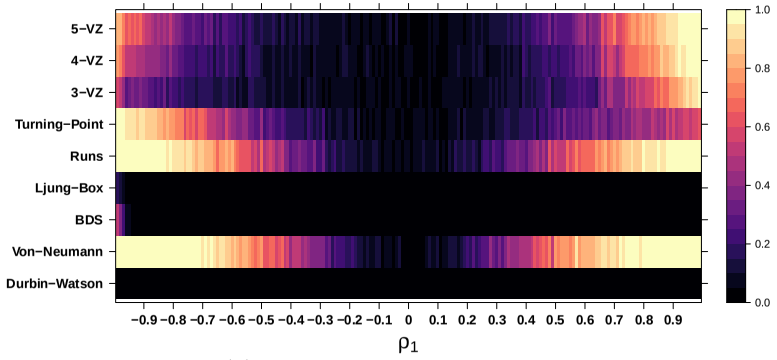
Abbildung 3.23: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 5 % Kontaminationsanteil und Intensitäten von 5 und 20, für unterschiedliche Stichprobenumfänge



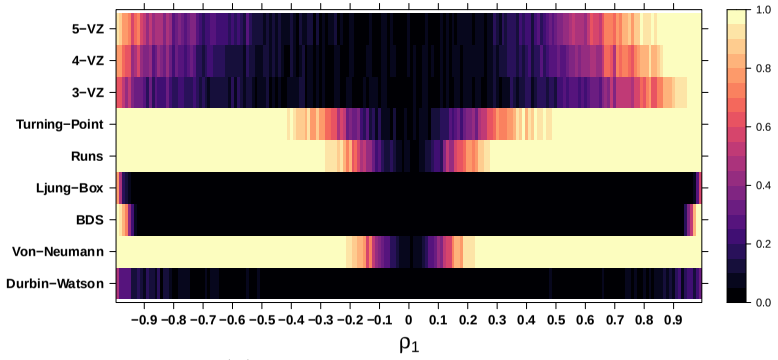
(a) $N = 50$, Intensität= 50



(b) $N = 500$, Intensität= 50



(c) $N = 50$, Intensität= 100



(d) $N = 50$, Intensität= 100

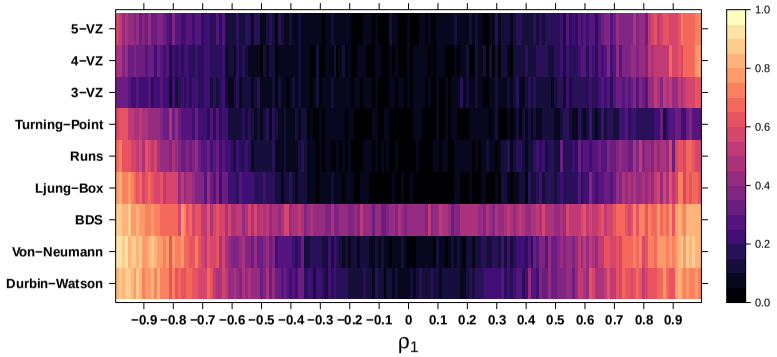
Abbildung 3.24: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit 5 % Kontaminationsanteil und Intensitäten von 50 und 100, für unterschiedliche Stichprobenumfänge

den parametrischen Verfahren weiterhin, das Niveau unter der Unabhängigkeit einzuhalten. Mit steigender Anzahl von Verunreinigungen fällt auf, dass auch die nichtparametrischen Verfahren ebenfalls geringere Trennschärfen aufweisen. Diese Verschlechterungen fallen jedoch − mit Ausnahme von der des BDS-Tests − deutlich geringer aus als bei den parametrischen Verfahren, sodass sie auch hier die bessere Wahl darstellen. Eine Erhöhung der Intensität von den Kontaminationen führt in diesem Zusammenhang bei den parametrischen Verfahren sowie beim BDS-Test zu einer deutlichen Verschlechterung der Trennschärfe, während die übrigen nichtparametrischen Verfahren in diesem Szenario von ihren robusten Eigenschaften profitieren. Als nichtparametrischer Test mit der besten Trennschärfe stellen der VNRR-Test und der Runs-Test also gute Wahlen zur Überprüfung von Korrelationsstrukturen dar, falls davon auszugehen ist, dass die betrachteten Zeitreihen Kontaminationen enthalten. Bei einem zu erwartenden Kontaminationsanteil von mehr als 5 % bietet sich dabei der Runs-Test aufgrund seiner leicht stärkeren Robustheit an.

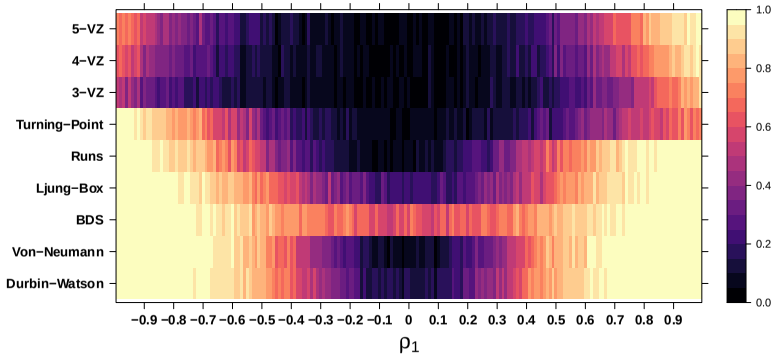
3.1.4 Abweichungen von der Varianzhomogenität

Eine weitere Annahme, die vielen der Testverfahren zugrunde liegt, ist die der Varianzhomogenität in der Zeitreihe. Es wird dabei davon ausgegangen, dass die Innovationen in einer Zeitreihe stets dieselbe Varianz aufweisen. Situationen, in denen diese Voraussetzung nicht erfüllt ist, stellen z. B. Überwachungsmessungen an Maschinenelementen dar, deren Verhalten sich in Zuge allmählicher Veränderungen des zu überwachenden Vorgangs verändert. Solche Ereignisse können z. B. sich anbahnende Überhitzungen bzw. Überlastungen der Maschine oder Anlaufvorgänge sein. Typisch dabei ist, dass die Varianz des Prozesses vor solchen Events zunimmt.

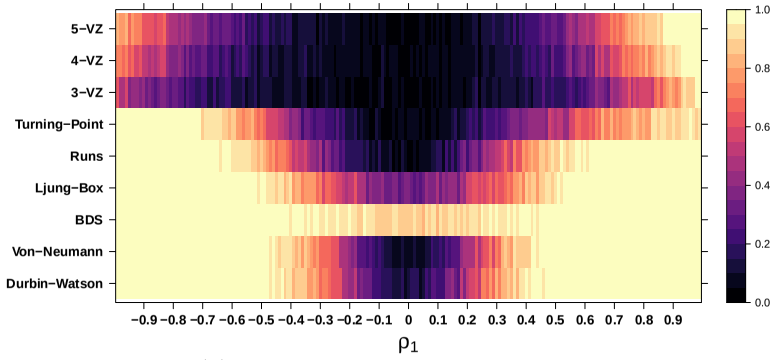
Um zu überprüfen, inwieweit Verletzungen dieser Voraussetzung die Trennschärfe der Testverfahren beeinflussen, werden im Folgenden genau solche Szenarien untersucht. Dafür wurden Zeitreihen erzeugt, in denen die Varianz über die Zeit hinweg zunimmt. Konkret wurde die Varianz am Anfang der Zeitreihe auf 1 gesetzt und am Ende der Zeitreihe auf $N/10$, wobei sie gleichmäßig mit steigender Beobachtungszahl zunimmt. Es gilt zu beachten, dass die Beobachtungen in diesem Fall nie zufällig sind. So handelt es sich nicht um identisch verteilte Zufallsvariablen, selbst wenn $\rho_1 = 0$ gilt. Es liegen dann lediglich keine Autokorrelationen vor − wohl aber eine Abhängigkeitsstruktur über die Zeit. Da alle der betrachteten Tests − bis auf den BDS-Test − hauptsächlich Autokorrelationen des Prozesses als Art von Abhängigkeitsstruktur aufzudecken versuchen, ist es von Interesse, wie sie in diesem Szenario reagieren. So wäre es einerseits wünschenswert, dass die Erkennung von Autokorrelationen nicht negativ durch die wachsende Varianz beeinflusst wird, andererseits kann es auch von Vorteil sein, wenn ein Verfahren in der Lage ist, die unterschiedliche Verteilung der Beobachtungen als Abweichung von der Zufälligkeit der Zeitreihe zu detektieren. Die Ergebnisse der Simulation unter den oben beschriebenen Voraussetzungen sind in Abbildung 3.25 für unterschiedliche Stichprobenumfänge dargestellt.



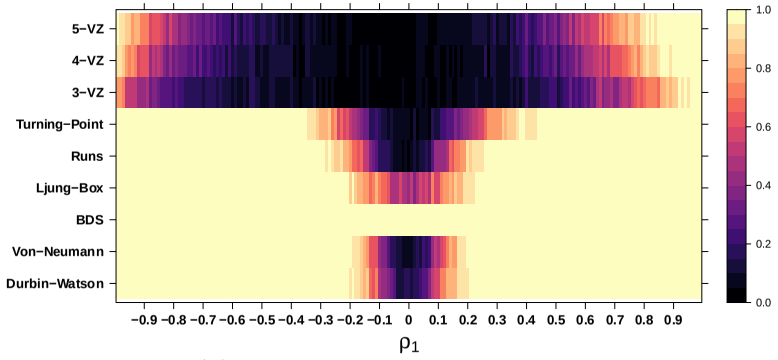
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



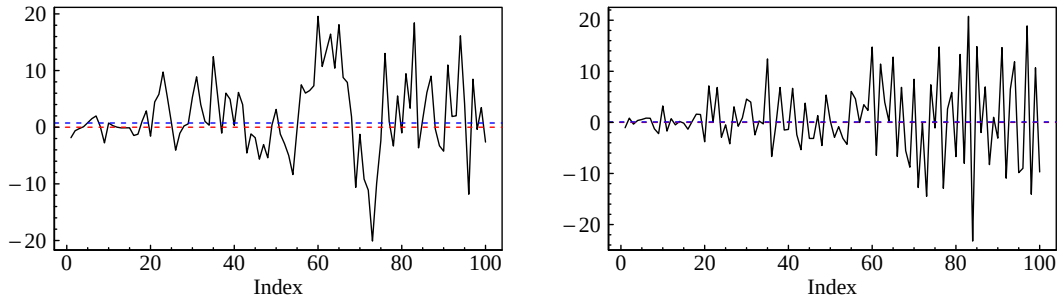
(d) Stichprobenumfang $N = 500$

Abbildung 3.25: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit wachsender Varianz, für unterschiedliche Stichprobenumfänge

Im Vergleich zu den Ergebnissen unter Normalbedingungen fällt auf, dass die K-VZ-Tests, der Runs-Test sowie der TP-Test, die lediglich auf dem sequenziellen Schema der Beobachtungen basieren, durch die wachsende Varianz kaum beeinflusst werden. Der DW-Test sowie der VNRR-Test reagieren ebenfalls nicht besonders stark, lehnen die Nullhypothese jedoch, unabhängig von Stichprobenumfang, in Bereichen sehr geringer oder fehlender Korrelation seltener ab. Der DW-Test scheint dadurch leichte Probleme mit der Einhaltung des Niveaus zu entwickeln. In diesem Szenario stechen besonders der BDS-Test und der LB-Test hervor, da sie die wachsende Varianz, bei einem für den BDS-Test geeigneten Stichprobenumfang von $N = 500$, als Abhängigkeitsstruktur zu erkennen scheinen. Der LB-Test schafft es dann nicht mehr, das Niveau unter der Unabhängigkeit einzuhalten und verwirft die Nullhypothese im Bereich um $\rho_1 = 0$ in ungefähr 50 % der Fälle. Beim BDS-Test ist das Verhalten so stark ausgeprägt, dass er die Nullhypothese auf dem gesamten Spektrum deutlich ablehnt.

Um die Robustheit des K-VZ-Tests, des Runs-Tests und des TP-Tests bei wachsender Varianz nachvollziehen zu können, ist es sinnvoll, einen Blick auf die Struktur der daraus resultierenden Zeitreihen zu werfen. Zwei solcher Zeitreihen sind zur Veranschaulichung in Abbildung 3.26 dargestellt. Dort ist die Varianzänderung klar erkennbar. Bei einem Vergleich mit Abbildung 3.3 wird jedoch deutlich, dass das sequenzielle Schema durch die sich ändernde Varianz nicht stark beeinflusst wird. Dieses Verhalten ist wenig verwunderlich, da die Varianz, solange der Erwartungswert der Innovationen 0 entspricht, keinen Einfluss auf die Wahrscheinlichkeiten $P(w_t < 0)$ bzw. $P(w_t > 0)$ hat, die nach wie vor 0.5 entspricht. Weiter sind die Höhen der Ausschläge, die eine wesentliche Konsequenz einer wachsenden Varianz darstellen, für die betrachteten, nichtparametrischen Verfahren irrelevant.

Die Reaktionen des DW-Tests und des LB-Tests lassen sich auf Abweichungen der empirischen Autokorrelationskoeffizienten von der ihnen durch die Verfahren unterstellten Normalität zurückführen. Diese Abweichung wird durch die Varianzheterogenität der Beobachtungen ver-



(a) $\rho_1 = 0.5$ (b) ρ_1 = -0.5

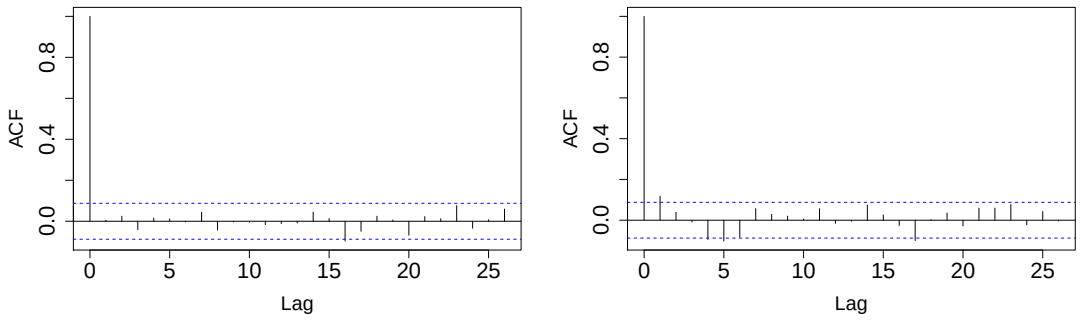
Abbildung 3.26: Zeitreihen mit $N = 100$ Beobachtungen mit wachsender Varianz für $\rho_1 = 0.5$ (a) und ρ_1 = -0.5 (b), mit Nulllinie (rot) und empirischem Median (blau)

ursacht und erschwert das Annehmen der Nullhypothese bei fehlender Korrelation. Zur Veranschaulichung sind Korrelogramme zweier zufälliger Zeitreihen mit einem Beobachtungsumfang von $N = 500$ für $\rho_1 = 0$ mit und ohne wachsende Varianz in Abbildung 3.27 dargestellt.

Im direkten Vergleich der Korrelogramme kann festgestellt werden, dass die wachsende Varianz dazu führt, dass größere und damit signifikante empirische Korrelationen in der Zeitreihe vorkommen. So überschreiten 4 der ersten 15 betrachteten Autokorrelationskoeffizienten trotz fehlender Korrelation die unter Normalbedingungen geltenden kritischen Werte (blaue Linien). Dies führt vor allem beim LB-Test dazu, dass er deutliche Schwierigkeiten hat, die Nullhypothese unter solchen Umständen korrekterweise beizubehalten. Das Verhalten lässt sich dadurch erklären, dass nahe beieinanderliegende Beobachtungen − abhängig von der Position in der Zeitreihe − dazu neigen, betragsmäßig moderate (am Anfang) oder große Werte (am Ende) anzunehmen, wodurch die empirischen Korrelationskoeffizienten die tatsächliche Korrelation in der Zeitreihe überschätzen.

Beim VNRR-Test wirkt sich die wachsende Varianz insofern auf die Teststatistik aus, dass sehr kleine bzw. große Ränge systematisch häufiger am Ende der Zeitreihe vergeben werden. Aus diesem Grund ist es dort auch wahrscheinlicher, dass selbst bei $\rho_1 = 0$ nur geringe Rangunterschiede am Anfang bzw. große am Ende der betrachteten Zeitreihe auftreten. Das ermöglicht dem Test, Abweichungen von der Unabhängigkeitsannahme zu erkennen und die Nullhypothese in diesen Bereichen öfter zu verwerfen, als es unter Normalbedingungen der Fall ist.

Die außergewöhnliche Sensitivität des BDS-Tests in Bezug auf diese Abweichung von der Zufälligkeit ist auf die Beschaffenheit seiner Teststatistik zurückzuführen (s. Kap. 2.8). So wird beim BDS-Test im Prinzip überprüft, ob der Abstand zwischen Beobachtungen unabhängig von ihrer zeitlichen Ordnung − und damit ihrer Position in der Zeitreihe ist. Da eine Zeitreihe mit wachsender Varianz offenbar genau diese Eigenschaften nicht erfüllt, sondern dazu führt, dass Abstände zwischen Beobachtungen im Verlauf der Zeitreihe immer größer werden, ist die Verwer-



(a) Normalbedingungen (b) mit wachsender Varianz

Abbildung 3.27: Korrelogramme von Zeitreihen mit $N = 500$ Beobachtungen ohne (a) und mit (b) wachsender Varianz, für $\rho_1 = 0$ und kritischen Werten (blau)

fung der Nullhypothese unabhängig von ρ_1 bei hinreichend großer Beobachtungszahl nachvollziehbar. Da der BDS-Test so sensibel auf die wachsende Varianz reagiert, ist es von Interesse, wie er sich bei unterschiedlichen maximalen Varianzen der Innovationen verhält. Aus diesem Grund ist die Trennschärfe des Tests in Abbildung 3.28 für einen Stichprobenumfang von $N = 500$ in Abhängigkeit von ρ_1 und der Varianz am Ende der Zeitreihe illustriert. Dabei wurden maximale Varianzen von $1 - 50$ in Schritten von einer Einheit und zusätzlich schrumpfende Varianzen von $1/10 - 9/10$ der Ausgangsvarianz betrachtet.

Dabei zeigt sich, dass der Test äußerst sensibel auf Varianzänderungen dieser Art regiert und bereits bei gering wachsender Varianz mit bemerkenswerter Treffsicherheit erkennt, wann eine Heterogenität der Varianzen vorliegt.

Im Vergleich dazu verdeutlicht der selbige Simulationsaufbau für den LB-Test in Abbildung 3.29, dass hier lediglich die Spezifität unter der Nullhypothese bei dem Vorliegen einer wachsenden Varianz abnimmt und damit das Niveau des Tests nicht mehr eingehalten werden kann. Dabei ist der konkrete maximale Wert der Varianz irrelevant und es scheint, als würde der LB-Test durch sie niemals eine klare Ablehnung der Nullhypothese erreichen können. Diese Ergebnisse sind mit obigen Überlegungen durch die Abweichungen von der Normalität der empirischen Autokorrelationskoeffizienten zu erklären.

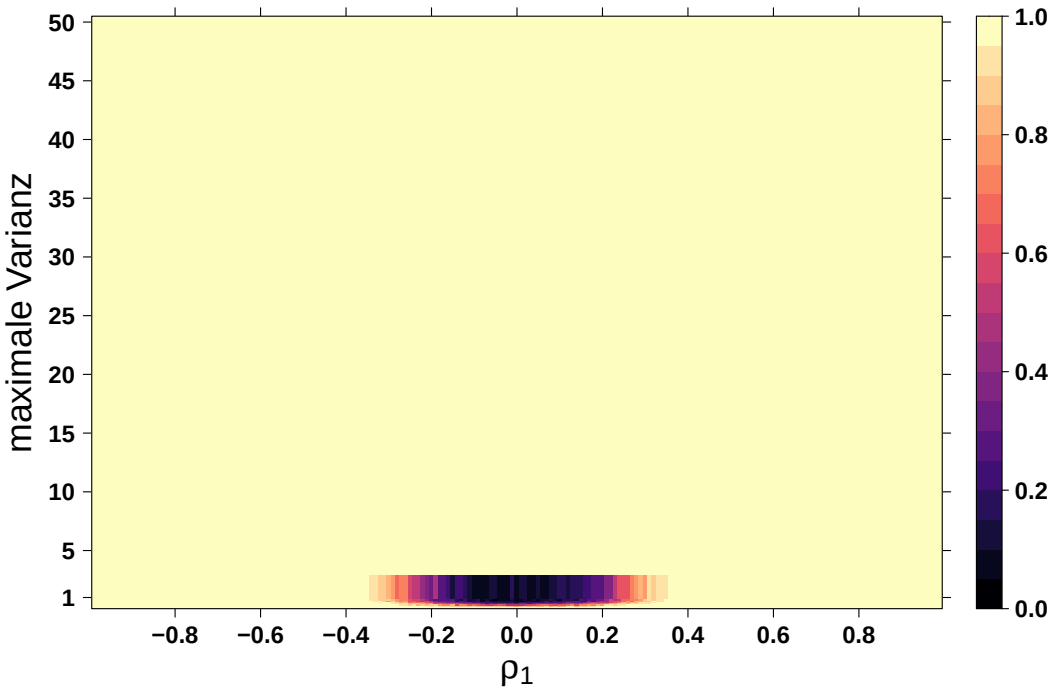


Abbildung 3.28: Simulierte Trennschärfen des BDS-Tests bei wachsender Varianz der Innovationen in Abhängigkeit von ρ_1 und der Varianz am Ende der Zeitreihe, bei $N = 500$ Beobachtungen

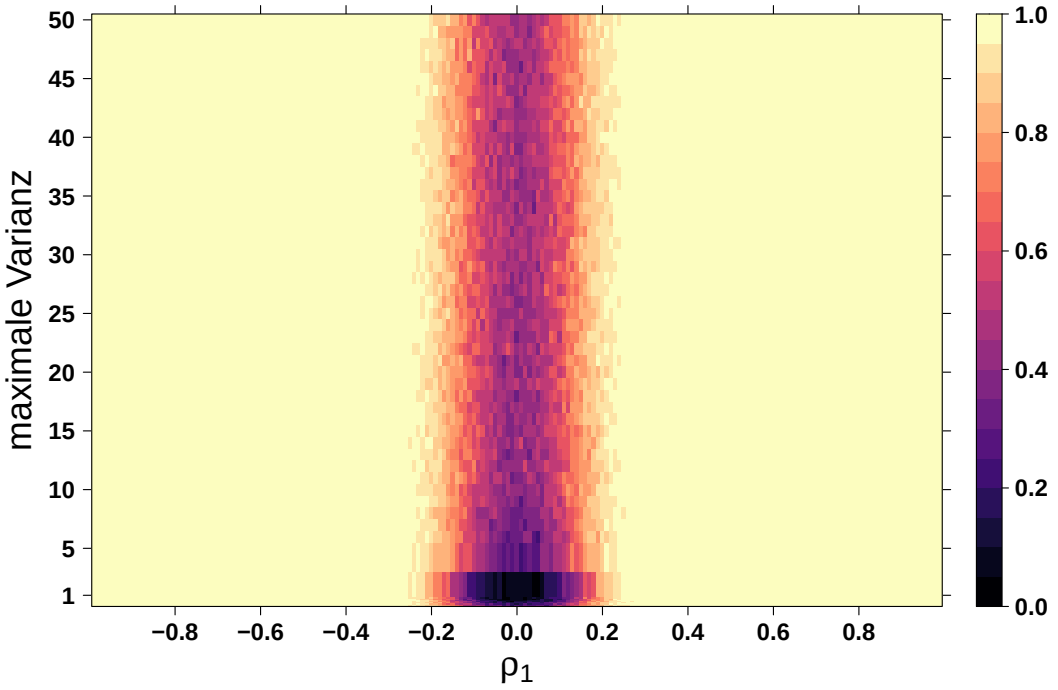


Abbildung 3.29: Simulierte Trennschärfen des LB-Tests bei wachsender Varianz der Innovationen in Abhängigkeit von ρ_1 und der Varianz am Ende der Zeitreihe, bei $N = 500$ Beobachtungen

Insgesamt zeigen diese Simulationsergebnisse, dass die Robustheit derjenigen Verfahren, die lediglich das sequenzielle Schema der Beobachtungen heranziehen, dazu führt, dass Korrelationen ungeachtet einer sich ändernden Varianz in der Zeitreihe sicher erkannt werden können. Parametrische Tests dagegen leiden unter den Verletzungen der Annahmen, sodass sie Schwierigkeiten haben, das Niveau einzuhalten. Es gelingt ihnen aber auch nicht, eine klare Entscheidung gegen die Unabhängigkeit aufgrund der Varianzinhomogenität zu treffen. Besonders der LB-Test wird durch sie in großem Maße beeinflusst und stellt in diesem Szenario keine gute Wahl zur Überprüfung auf Autokorrelationen oder auf Varianzheterogenitäten in den Zeitreihen dar. Auch der VNRR-Test, bei dem zumindest die Ränge der Beobachtungen eine Rolle spielen, wird von einer wachsenden Varianz beeinflusst und hat Schwierigkeiten damit, das Niveau des Tests einzuhalten. Ist es von Interesse, zusätzlich eine Abweichung von der Homogenitätsannahme der Varianzen zu erkennen, so überzeugt der BDS-Test bei hinreichend großem Stichprobenumfang durch seine bemerkenswerte Sensibilität unter solchen Alternativen. Falls lediglich untersucht werden soll, ob Autokorrelationen in einer Zeitreihe vorliegen, ist der Runs-Test eine gute Wahl, da er eine gute Trennschärfe aufweist und – anders als der VNRR-Test – nicht durch die Varianzheterogenität beeinflusst wird.

3.1.5 Änderungen des Niveaus

Eine wichtige Voraussetzung für die Modellierung und Prognose von Zeitreihen stellt die Eigenschaft eines konstanten Niveaus dar. Oft wird dafür die Stationarität des zugrunde liegenden Prozesses vorausgesetzt, die einen zeitunabhängigen, konstanten Erwartungswert der einzelnen Beobachtungen unterstellt (s. Kap. 2.1). In der Praxis ist diese Voraussetzung wegen ihrer Stärke jedoch sehr selten erfüllt, sodass kurz- oder langfristige Änderungen des Niveaus diverser Formen häufig zu beobachten sind. Im Weiteren soll deshalb untersucht werden, wie die einzelnen Testverfahren auf verschiedenartige Änderungen des Niveaus der Zeitreihe reagieren.

Sprünge

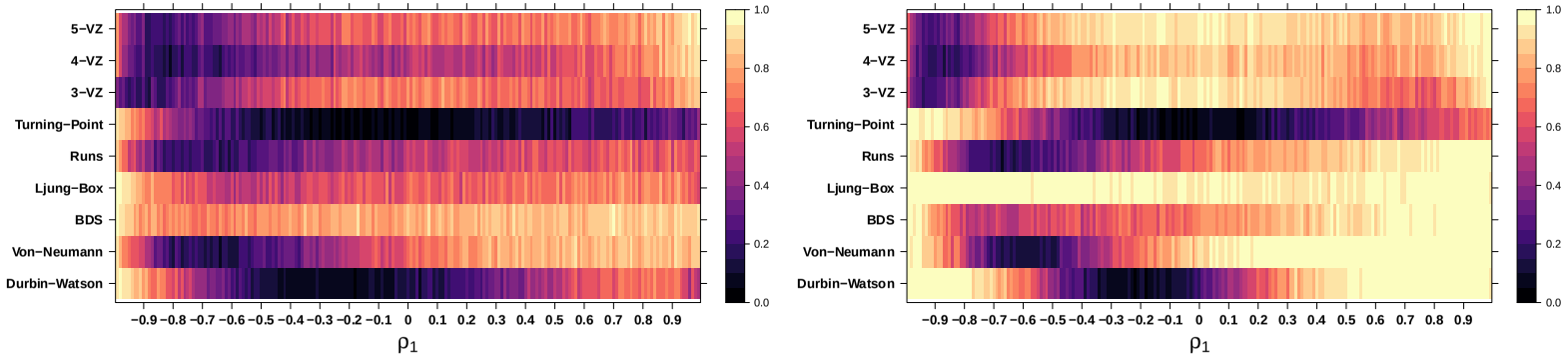
Zunächst soll der Fall betrachtet werden, dass Sprünge in der Zeitreihe auftreten. Solche Abhängigkeitsstrukturen sind in Situationen von Interesse, in denen zeitliche Prozesse überwacht werden, deren Niveau sich infolge bestimmter Ereignisse abrupt ändert. Ein Beispiel dafür sind Hochwasserstatistiken oder Drucksondierungen in geringen Zeitabständen, in denen solche Niveauänderungen infolge von Starkregenereignissen oder Schichtwechseln auftreten können. Ähnlich wie im Fall von wachsenden Varianzen ist es dabei einerseits − abhängig von Anwendungsgebiet − von Interesse, welche der Verfahren robust reagieren, also Abweichungen des Parameters ρ_1 von 0 trotz des Sprunges ähnlich zuverlässig wie unter Normalbedingungen erkennen. Andererseits stellt das Springen einer Zeitreihe auch eine Abweichung von ihrer Zufälligkeit dar, sodass sich die Frage stellt, welche Verfahren in der Lage sind, diese Systematik zu erkennen.

Um zu untersuchen, in welcher Art die hier betrachteten Testverfahren reagieren, wurden Zeitreihen der Länge N konstruiert, die bis zur $(N/2)$-ten Beobachtung einen Mittelwert von -1.96/2 aufweisen und danach einen Sprung der Höhe 1.96 nach oben vollziehen. Dieser Wert entspricht dabei ungefähr dem 95-%-Quantil der Standartnormalverteilung, sodass sich unter der Annahme der Unabhängigkeit vor und nach dem Sprung approximativ nur 32.7 % der Werte überlappen sollten. Dieser Wert kann berechnet werden, indem das Integral der Normalverteilung mit dem größeren Erwartungswert bis zum Schnittpunkt der beiden Verteilungen (der genau in der Mitte zwischen beiden Erwartungswerte − also im Nullpunkt liegt) verdoppelt wird. Dies ist möglich, da es sich um Verteilungen mit identischer Varianz handelt und die konkrete Überlappung kann in `R` also berechnet werden durch:

$$2 \cdot \texttt{pnorm}(0, mean = 1.96/2, sd = 1).$$

Diese Beschaffenheit garantiert zumindest theoretisch einen Median von 0 in der vorliegenden Zeitreihe, sodass keine Zentrierung für die K-VZ-Tests durchgeführt werden muss. Die Simulationsergebnisse für dieses Szenario wurden dabei in Abbildung 3.30 dargestellt.

Hier ist besonders auffällig, dass die Annahmebereiche sämtlicher Tests − mit Ausnahme des TP-Tests − in den Bereich negativer Korrelationen verschoben sind. Am stärksten ist diese Verschiebung bei den K-VZ-Tests ausgeprägt und am schwächsten beim DW-Test. Weiter scheinen



(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

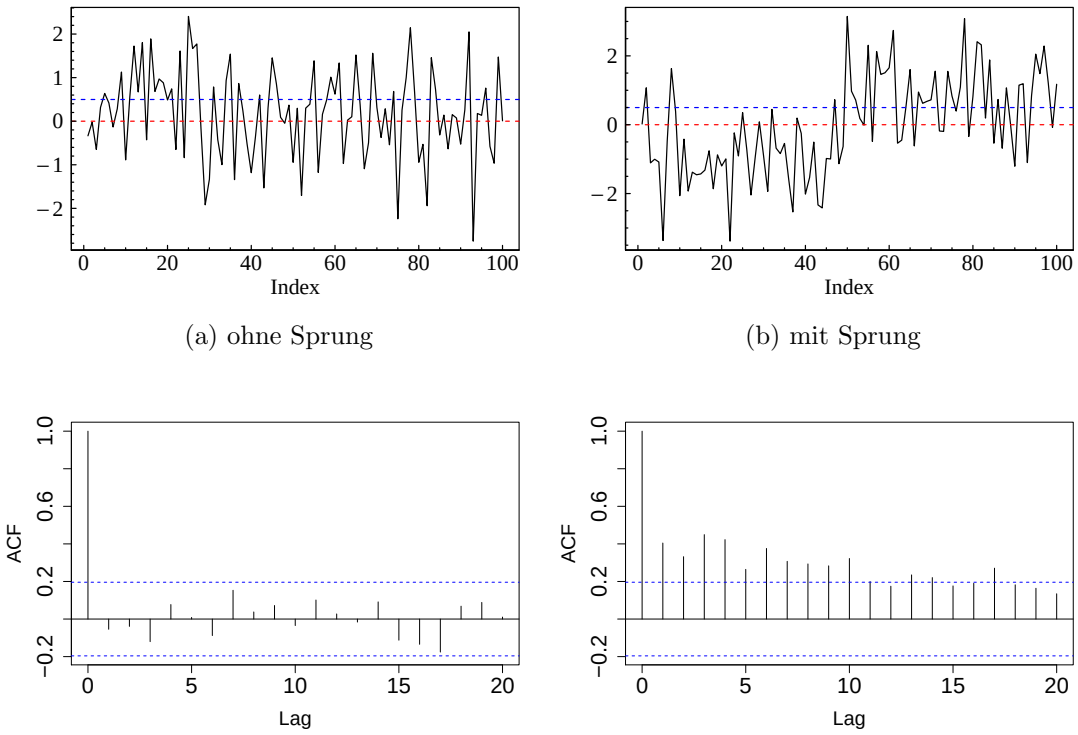
(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

Abbildung 3.30: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit einem Sprung der Höhe 1.96 nach der Hälfte der Beobachtungen, für unterschiedliche Stichprobenumfänge

der BDS-Test und der LB-Test die Abweichung von der Zufälligkeit, ähnlich wie im vorherigen Abschnitt, bei hinreichend großem Stichprobenumfang zu erkennen. Der LB-Test reagiert dabei deutlich sensitiver, sodass dieser die Nullhypothese bereits bei einem Beobachtungsumfang von $N = 100$ für jegliche Werte von ρ_1 verwerfen kann. Auch für die K-VZ-Tests kann die Nullhypothese durch die extreme Verschiebung − bis auf den Bereich sehr extremer Korrelationen − fast auf dem gesamten Spektrum von ρ_1 verworfen werden. Somit liegt auch hier die Vermutung nahe, dass sich diese Tests für die Detektion von Sprüngen eignen.

Die Tatsache, dass der TP-Test unter diesen Voraussetzungen so robust reagiert, ist wenig verwunderlich, da lediglich die Anzahl der Turning-Points in die Teststatistik einfließt. Diese wird von dem Sprung − zumindest bis auf den konkreten Zeitpunkt in dem er stattfindet − nicht beeinflusst. Um das Verhalten der übrigen Tests nachvollziehen zu können, scheint es sinnvoll, wieder typische simulierte Zeitreihen dieser Form sowie die dazugehörigen Korrelogramme zu betrachten. Zunächst wurde dafür eine Zeitreihe mit $N = 100$ Beobachtungen unter der Nullhypothese sowie dieselbe Zeitreihe mit einem nachträglich modellierten Sprung gemeinsam mit den zu ihnen gehörenden Korrelogrammen in Abbildung 3.31 dargestellt.



(a) ohne Sprung (b) mit Sprung

(c) ohne Sprung (d) mit Sprung

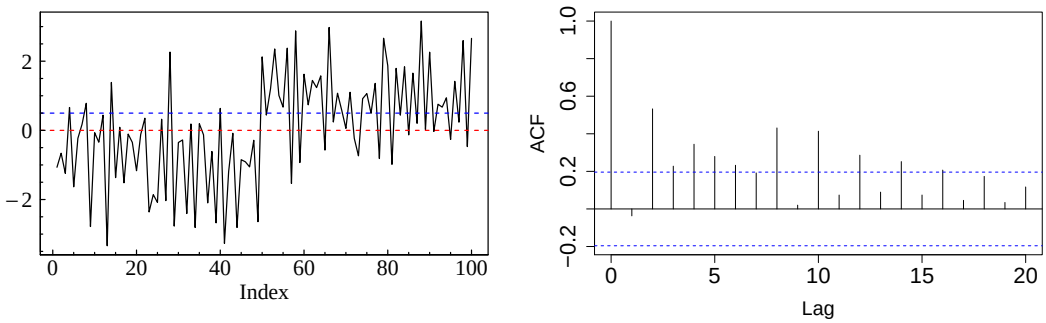
Abbildung 3.31: Zeitreihen mit $N = 100$ Beobachtungen ohne (a) und mit (b) Sprung der Höhe 1.96 beim Index 50, für $\rho_1 = 0$, mit zugehörigen Korrelogrammen ((c) und (d)), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

Hieraus wird deutlich, dass das Springen die empirischen Autokorrelationskoeffizienten stark beeinflusst. So überschreiten fast sämtliche, vom LB-Test betrachteten Korrelationsschätzer den kritischen Wert unter der Nullhypothese. Dieses Verhalten lässt sich auf die Berechnung der empirischen Autokorrelationskoeffizienten zurückführen (s. Kap. 2.1). So entspricht der Mittelwert $\bar{x}$ in den hier betrachteten Zeitreihen aufgrund ihrer Konstruktion zwar annähernd 0, er unterscheidet sich jedoch faktisch für die erste und zweite Hälfte der Beobachtungen. Konkret beträgt er in der ersten Hälfte ca. -1.96/2 und in der zweiten ca. 1.96/2. Die Summanden im Zähler des Schätzers sind dementsprechend in beiden Teilen immer deutlich größer als bei einer unabhängigen Zeitreihe ohne Sprung, sodass die empirischen Autokorrelationskoeffizienten die wahre Autokorrelation stark überschätzen. Aus diesem Grund ist nachvollziehbar, dass diejenigen Testverfahren, die auf den empirischen Autokorrelationskoeffizienten basieren, die Nullhypothese trotz fehlender Korrelationen ablehnen.

Weiter lässt sich anhand von Abbildung 3.31 erkennen, dass das sequenzielle Schema sowohl in Bezug auf den Median, als auch auf das Nullniveau schwerwiegend verändert wird. So gibt es zum einen deutlich längere und damit weniger Runs, sodass die Testentscheidung des Runs-Tests stark beeinflusst wird und eine Verwerfung der Nullhypothese im Fall $\rho_1 = 0$ offensichtlich werden lässt. Im Hinblick auf die K-VZ-Tests wird anhand dieses Beispiels deutlich, dass die Teststatistik sehr kleine Werte annimmt, da die Anzahl der Vorzeichenblöcke sehr gering ist und jeder Block dementsprechend viele Beobachtungen enthält. Für noch größere Sprünge ist zu erwarten, dass sich der Wert der Teststatistik seinem Minimum annähern würde, da irgendwann mit hoher Wahrscheinlichkeit lediglich zwei Vorzeichenblöcke vorhanden sind und alle K-VZ-Tests für $K \geq 3$ hier stets eine Tiefe von 0 aufweisen würden. Weiterhin hat der Sprung einen Einfluss auf die Ränge der Beobachtungen, die in den jeweiligen Abschnitten deutlich enger zusammenliegen als es unter der Unabhängigkeit zu erwarten wäre. So werden am Anfang der Zeitreihe systematisch niedrigere − und am Ende der Zeitreihe hohe Ränge vergeben, sodass auch die Ergebnisse des VNRR-Tests durch dem Sprung verzerrt werden.

Um nachzuvollziehen, warum die Annahmebereiche der meisten Tests in den Bereich negativer Korrelationen verschoben werden, scheint es sinnvoll, erneut typische Zeitreihen und Korrelogramme unter Alternativen zu betrachten, bei denen eine Verwerfung der Nullhypothese fälschlicherweise nicht stattfinden kann. Eine solche Zeitreihe ist in Abbildung 3.32 für einen Autokorrelationskoeffizienten von ρ_1 = -0.65 illustriert.

Im Vergleich zu Abbildung 3.31 fällt vor allem auf, dass die Werte der Autokorrelationsfunktion insgesamt geringer ausfallen als bei fehlender Korrelation. Besonders auffällig ist, dass $\hat{\rho_1}$ einen sehr kleinen Wert aufweist und nicht mehr als signifikant erhöht eingestuft wird. Die negative Korrelation scheint also die Mittelwertunterschiede bei der Berechnung der empirischen Autokorrelationskoeffizienten auszugleichen, sodass der DW-Test die Nullhypothese unter solchen Alternativen fälschlicherweise beibehält. Da lediglich der erste Autokorrelationskoeffizient so stark beeinflusst wird, ist auch verständlich, warum der LB-Test die Nullhypothese hier trotz-



(a) Zeitreihe mit ρ_1 = -0.65 (b) Korrelogramm der Zeitreihe

Abbildung 3.32: Zeitreihe mit $N = 100$ Beobachtungen mit Sprung der Höhe 1.96 nach der Hälfte der Beobachtungen, für ρ_1 = -0.65 (a), mit zugehörigem Korrelogramm (b), Nulllinie (rot) und empirischem Median bzw. kritischen Werten (blau)

dem erfolgreich verwerfen kann. Auch die Anzahl der Runs fällt durch die negative Korrelation deutlich größer aus als bei einem Sprung falls $\rho_1 = 0$ gilt. Auf der anderen Seite ist sie deutlich kleiner als im Fall ρ_1 = -0.65, wenn kein Sprung in der Zeitreihe vorhanden ist. Auf diese Weise wird der Effekt des Sprungs durch die negative Korrelation kompensiert und der Wert der Runs-Statistik nähert sich bei bestimmten Alternativen dem unter der Nullhypothese zu erwartenden Wert an. Dadurch ist deren Verwerfung nicht mehr möglich. Ähnliche Überlegungen können auch für die Ränge der Beobachtungen angestellt werden, deren sukzessive Abstände sich bei einer bestimmten negativen Korrelation denen unter der Nullhypothese annähern. Die Tatsache, dass der DW-Test die Nullhypothese bei einer deutlich weniger stark ausgeprägten negativen Korrelation nicht ablehnt, obwohl der Wert $\hat{\rho}_1$ in Abbildung 3.32 hinreichend klein zu sein scheint, liegt an der Beschaffenheit seiner Teststatistik (vgl. Kap. 2.3). Wie dort erläutert, beruht die Umformung des Wertes T_{DW} lediglich auf der Annahme, dass jede der Beobachtungen einen Mittelwert von 0 aufweist. Dies ist bei den vorliegenden Zeitreihen mit Sprung offensichtlich nicht der Fall. Somit ist hier eine Rückführung auf den empirischen Autokorrelationskoeffizienten ρ_1 nicht gerechtfertigt. Trotzdem scheint sich die Teststatistik bei einem bestimmten Ausmaß an negativer Autokorrelation dem Wert der Teststatistik unter der Nullhypothese anzunähern. Als Konsequenz ist der Test dann nicht in der Lage, diese zu verwerfen, obwohl Korrelationen vorliegen.

Besonders stechen in diesem Szenario die K-VZ-Tests heraus, bei denen die Nullhypothese lediglich im Bereich sehr extremer Korrelationen beibehalten wird. Aber auch der BDS-Test sowie der LB-Test fallen auf, da sie, ähnlich wie bei einer wachsenden Varianz, Verletzungen der Zufälligkeitsannahmen durch den Sprung zu detektieren scheinen. Um ein genaueres Bild über dieses Verhalten zu bekommen, wurden die Trennschärfen dieser Tests in den Abbildungen 3.33, 3.34 und 3.35 in Abhängigkeit von der Sprunghöhe und von ρ_1 dargestellt.

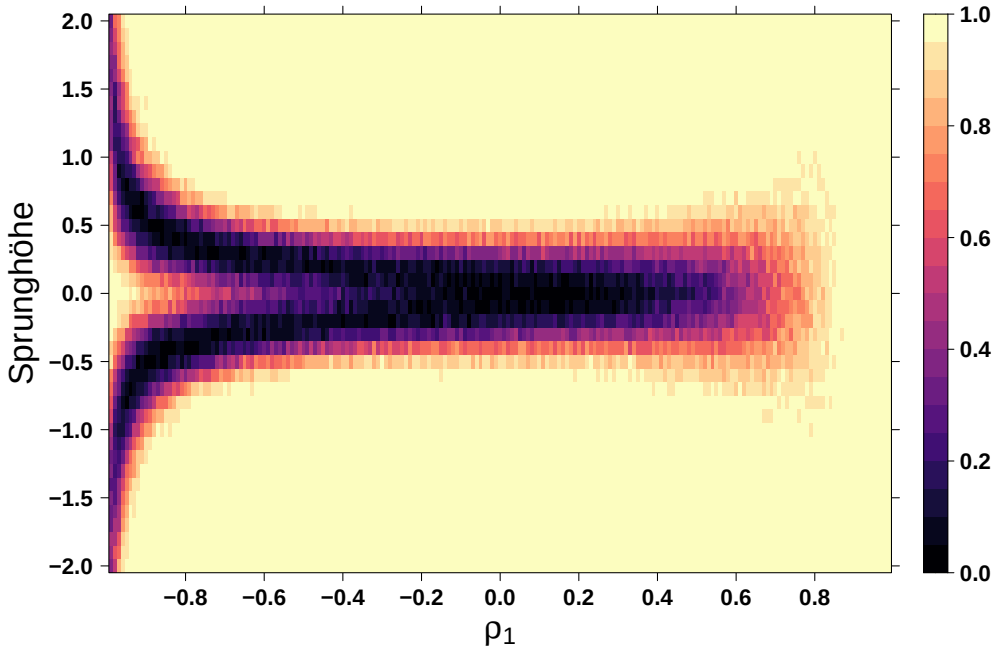


Abbildung 3.33: Simulierte Trennschärfen des 5-VZ-Tests bei einem Sprung in Abhängigkeit von ρ_1 und der Sprunghöhe, bei einer Beobachtungszahl von $N = 500$

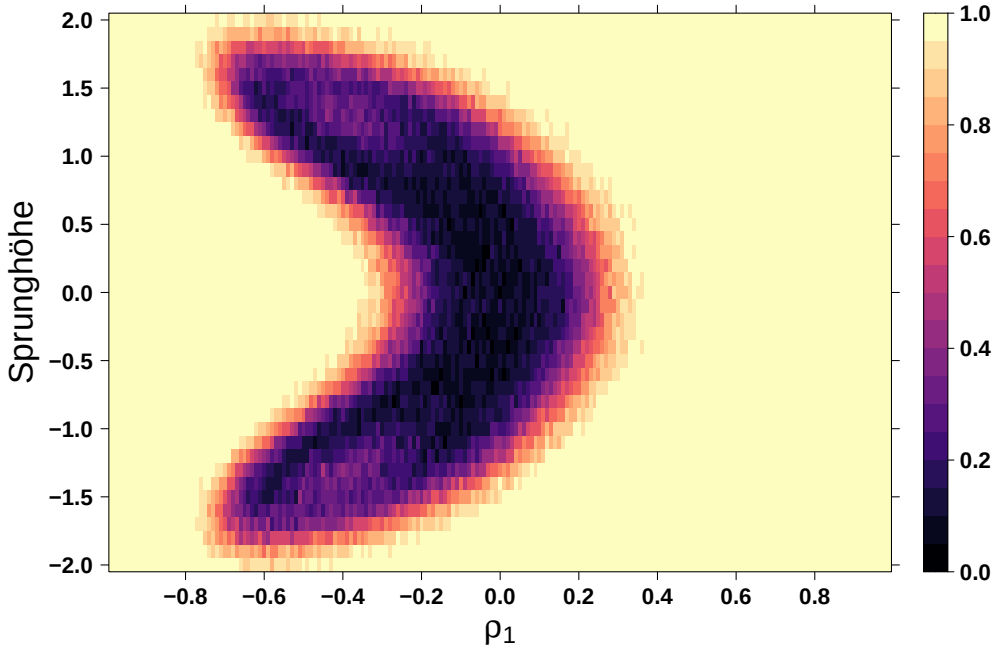


Abbildung 3.34: Simulierte Trennschärfen des BDS-Tests bei einen Sprung in Abhängigkeit von ρ_1 und der Sprunghöhe, bei einer Beobachtungszahl von $N = 500$

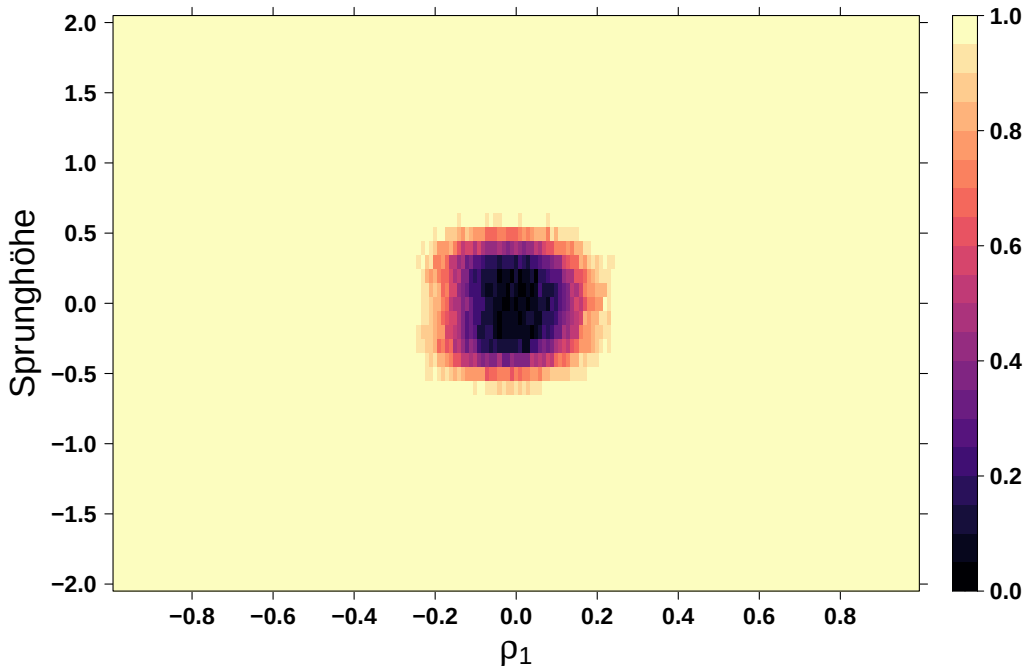


Abbildung 3.35: Simulierte Trennschärfen des LB-Tests bei einen Sprung in Abhängigkeit von ρ_1 und der Sprunghöhe, bei einer Beobachtungszahl von $N = 500$

Zunächst ist dabei zu erkennen, dass der Effekt der Sprunghöhe offensichtlich unabhängig davon ist, ob die Zeitreihe einen Sprung nach unten oder oben vollzieht − lediglich die Höhe des Sprunges spielt eine Rolle für die betrachteten Testverfahren. Beim 5-VZ-Test wird deutlich, dass die Sprunghöhe einen starken Einfluss auf seine Testentscheidung hat. So verschiebt sich der Annahmebereich mit zunehmender Sprunghöhe schnell und deutlich in den Bereich von negativen Korrelationen und schon ab einer betragsmäßigen Sprunghöhe von 0.6 werden lediglich Alternativen mit einer Korrelation von ca. -0.9 bis -0.8 nicht verworfen. Der Grund hierfür liegt wieder in der Beschaffenheit seiner Teststatistik. So führt der Sprung − wie es beispielsweise in Abbildung 3.31 zu erkennen ist − dazu, dass deutlich weniger Blöcke mit alternierenden Vorzeichen vorhanden sind, als es unter der Nullhypothese zu erwarten wäre. Außerdem weisen sie stark differierende Blockgrößen auf, wodurch die Teststatistik dahingehend verändert wird, dass eine Beibehaltung der Nullhypothese nicht mehr möglich ist. Ähnlich wie bei den anderen Tests scheint es hier aber zu jeder Sprunghöhe einen Autokorrelationskoeffizienten zu geben, der dafür sorgt, dass die Teststatistik einen Wert annimmt, der nah an einem unter der Nullhypothese zu erwartenden liegt. Somit kann eine Verwerfung der Nullhypothese in diesem Fall nicht stattfinden. Auffällig im Vergleich zu den anderen Testverfahren ist außerdem, dass dieser Annahmebereich mit zunehmender Sprunghöhe deutlich kleiner wird und bei einer Sprunghöhe von ca. 2 eine sehr extreme negative Korrelation von ca. -0.99 nötig ist, um die Wahrscheinlichkeit von einer Verwerfung der Nullhypothese zu verringern.

In der entsprechenden Grafik für den BDS-Test (Abb. 3.34) lässt sich ein ähnliches Verhalten erkennen. So verschiebt sich auch hier der Annahmebereich mit zunehmender Sprunghöhe in den Bereich negativer Korrelationen. Dies geschieht jedoch deutlich weniger schnell und extrem als beim 5-VZ-Test, sodass die Nullhypothese zum Vergleich bei einer Sprunghöhe von 0.6 in einem Bereich von -0.2 bis 0.1 selten verworfen wird. Insbesondere erkennt der Test bei einer hinreichend großen Sprunghöhe von ca. 2 die Abweichungen von der Nullhypothese unabhängig vom Wert des Parameters ρ_1 und verwirft sie für mehr als 95 % der simulierten Zeitreihen.

Der LB-Test scheint hier am besten geeignet zu sein, um Sprünge in der Zeitreihe zu detektieren. So gelingt es ihm bereits ab einer Sprunghöhe von 0.5 die Nullhypothese eindeutig zu verwerfen. Bei kleineren Sprunghöhen bleibt derjenige Bereich, in dem die Nullhypothese beibehalten wird, relativ konstant und verschiebt sich − anders als bei den übrigen Verfahren − nicht in den Bereich negativer Korrelationen. Der Grund dafür, dass der Test die Abweichungen von der Unabhängigkeit detektiert, liegt dabei in dem Verhalten der empirischen Autokorrelationskoeffizienten, das in Abbildung 3.31 dargestellt wird. Mit den oben gewonnenen Erkenntnissen lässt sich auch die fehlende Verschiebung in den Bereich negativer Korrelationen erklären. So kann eine negative Korrelationen 1. Grades anscheinend lediglich dazu führen, dass $\hat{\rho_1}$ nicht mehr als signifikant erhöht erkannt wird. Sie kann aber den Effekt des Sprunges auf die übrigen Autokorrelationskoeffizienten nicht verschleiern.

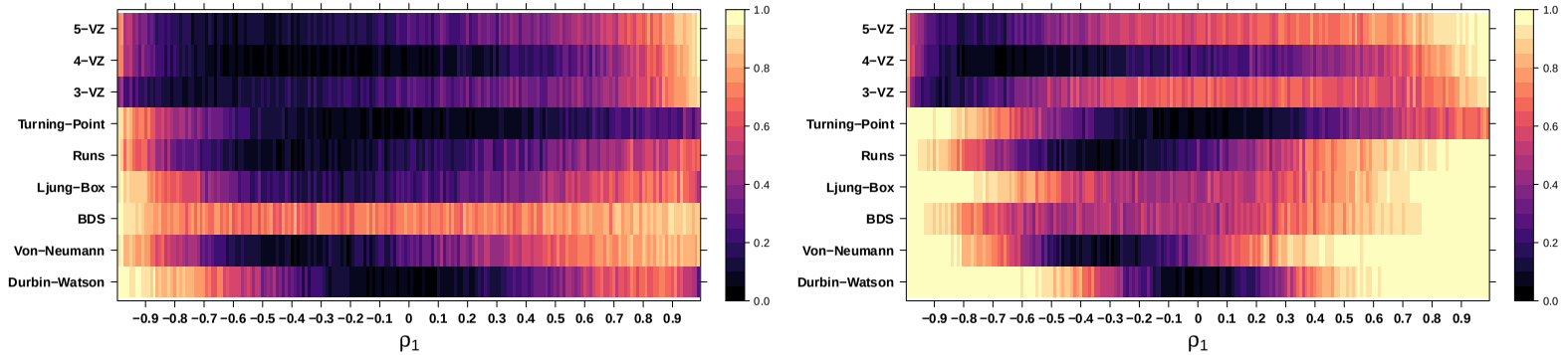
Trend

Änderungen des Niveaus einer Zeitreihe können − anders als durch einen Sprung − auch schrittweise und allmählich erfolgen. Im Folgenden wird deshalb untersucht, in welcher Weise sich ein Trend in der Zeitreihe auf die Testverfahren auswirkt. Dazu sind Zeitreihen zu unterschiedlichen Autokorrelationskoeffizienten und Beobachtungsumfängen N mit einem Trend der Steigung $2/N$ simuliert worden. Das entspricht einem Szenario, in dem eine steigende Beobachtungszahl mit einer erhöhten Abtastrate desselben Abschnittes vom zugrunde liegenden Prozess einhergeht. Der Gesamtanstieg des Niveaus entspricht also für jeden Beobachtungsumfang 2. Im Hinblick auf die K-VZ-Tests wird dabei sichergestellt, dass die Zeitreihen einen Median von 0 aufweisen. Hier ist wieder von Interesse, welche Verfahren ungeachtet des Trends Autokorrelationen erkennen können. Andererseits sind Trends in Zeitreihen mit dem bloßen Auge bisweilen nicht erkennbar bzw. eine optische Inspektion ist nicht möglich. Dann ist es ebenfalls sinnvoll, zu untersuchen, welche Testverfahren den Trend als Abweichung von der Zufälligkeit detektieren können.

Mathematisch werden im Folgenden Modelle der Form:

$$x_t = \frac{2}{N} \cdot (t - \frac{N}{2}) + \rho_1 x_{t-1} + w_t, \quad |\rho_1| < 1, \quad w_t \sim WN(0, \sigma^2_{WN})$$

für $t \in \{1, ..., N\}$ betrachtet. Die entsprechenden Simulationsergebnisse für die einzelnen Testverfahren sind in Abbildung 3.36 dargestellt.



(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

Abbildung 3.36: Simulierte Trennschärfe der Testverfahren bei stationären AR(1)-Alternativen mit Trend der Steigung $2/N$, für unterschiedliche Stichprobenumfänge

Beim Betrachten der Grafik zeigt sich ein ähnliches Verhalten wie im Fall, dass ein Sprung in die Zeitreihe modelliert wurde. So verschieben sich die Annahmebereiche der meisten Testverfahren in den Bereich von negativen Korrelationen. Insgesamt fällt aber auf, dass die Verschiebungen weniger extrem sind, obwohl die Höhe des gesamten Niveauanstieges in etwa der des Sprunges entspricht. Weiter ist zu bemerken, dass sowohl der DW-Test als auch der TP-Test trotz des Trends keinerlei Verschiebungen zeigen und der Annahmebereich dem der Simulation unter Normalbedingungen entspricht. Ein weiterer bemerkenswerter Unterschied betrifft den BDS-Test. Im Gegensatz zum Fall des Sprunges scheint er hier, auch bei großem Stichprobenumfang, keine Abweichungen von der Nullhypothese durch den Trend zu detektieren. Sein Annahmebereich ist dabei ebenfalls verschoben und von allen Verfahren weist er zu jedem Stichprobenumfang die geringste Trennschärfe auf. Der LB-Test verhält sich ähnlich wie beim Sprung und verwirft die Nullhypothese der Unabhängigkeit bereits bei einem Beobachtungsumfang von $N = 100$ zuverlässig.

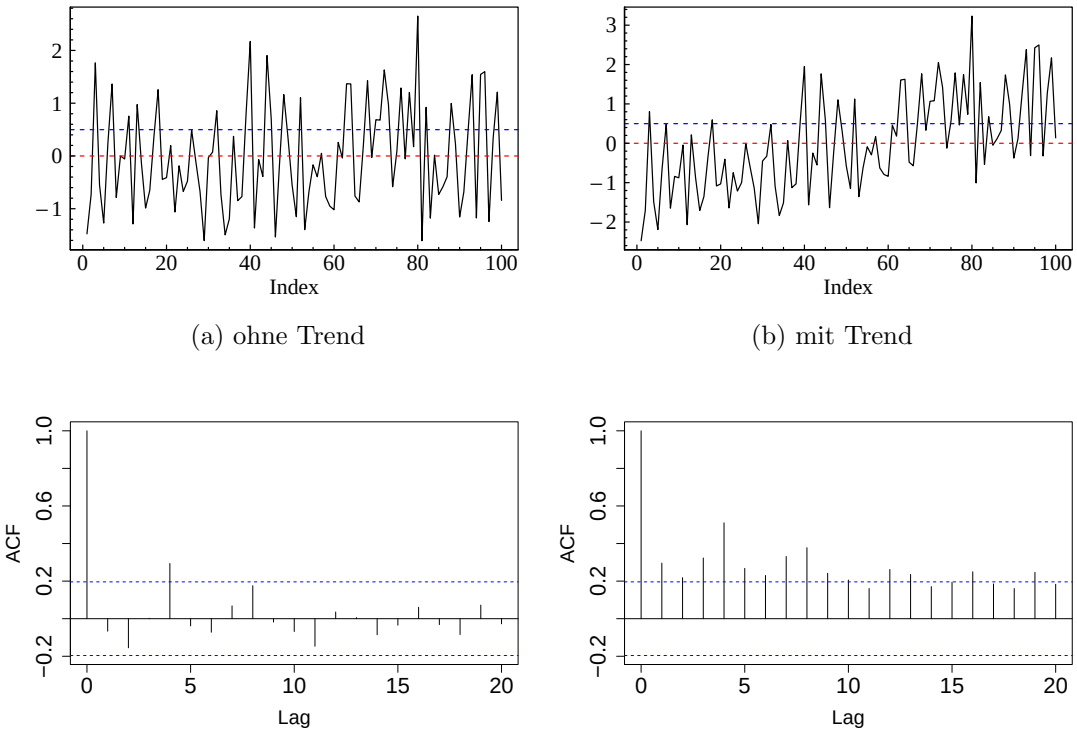
Um zu ergründen, wie diese Reaktionen der Testverfahren zu erklären sind, wurden erneut eine Zeitreihe unter der Nullhypothese sowie die gleiche Zeitreihe mit nachträglich eingefügtem Trend gemeinsam mit den zugehörigen Korrelogrammen in Abbildung 3.37 dargestellt.

Dabei ist eine ähnliche Systematik wie im Fall eines Sprunges erkennbar. So werden die empirischen Autokorrelationskoeffizienten bei vorhandenem Trend erhöht, wodurch die Fähigkeit des LB-Tests, den Trend zu detektieren, nachvollziehbar wird. Im Vergleich mit der Zeitreihe des Sprunges (s. Abb. 3.31) sind die empirischen Autokorrelationskoeffizienten aber deutlich schwächer ausgeprägt. Dieses Verhalten könnte eine Erklärung für die weniger starke Verschiebung vom Annahmebereich des DW-Tests in den Bereich von negativen Korrelationen sein. Des Weiteren ist ersichtlich, dass die Anzahl der Median- bzw. Nulldurchgänge aufgrund des Trends abnimmt. Dieser Effekt äußert sich jedoch weniger stark als bei einem Sprung. Das ist damit zu erklären, dass der lokale Mittelwert der Zeitreihe beim Sprung stets ca. eine Einheit vom Nullniveau bzw. empirischen Median von 0 entfernt ist, wobei dies beim Trend nur bei der ersten und letzten Beobachtung der Fall ist. Konkret beträgt der erwartete Abstand von Beobachtung $t \in \{1, .., N\}$ zum Mittelwert im Fall von N Beobachtungen lediglich

$$\left| \left(t - \frac{N}{2} \right) \cdot \frac{2}{N} \right| = \left| \frac{2t}{N} - 1 \right| \leq |1| \,.$$

Die Wahrscheinlichkeit, dass eine Innovation die Zeitreihe veranlasst, den empirischen Median bzw. die Nulllinie zu überspringen, ist damit für jede Beobachtung mindestens so groß wie im Fall eines Sprunges und die oben beschriebenen Beobachtungen sind damit nachvollziehbar.

Auch im Hinblick auf den VNRR-Test legen diese Überlegungen nahe, dass sich die Ränge der Beobachtungen nicht so stark von denen unter der Nullhypothese unterscheiden, wie bei einem Sprung. Die Tatsache, dass der TP-Test wieder robust reagiert, ist mit ähnlichen Argumenten wie im Fall eines Sprunges zu begründen − so ist der Trend nicht stark genug ausgeprägt, um einen Einfluss auf die Anzahl der Turning-Points in der Zeitreihe zu haben.



(a) ohne Trend

(b) mit Trend

(c) ohne Trend

(d) mit Trend

Abbildung 3.37: Zeitreihen mit $N = 100$ Beobachtungen ohne (a) und mit (b) Trend der Steigung 0.02, für $\rho_1 = 0$, mit zugehörigen Korrelogrammen ((c) und (d)), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

Für den 5-VZ-Test soll stellvertretend für alle K-VZ-Tests untersucht werden, wie er sich bei unterschiedlichen Steigungen in Abhängigkeit von dem Autokorrelationskoeffizienten verhält. In Abbildung 3.38 ist dieser Zusammenhang grafisch dargestellt.

Dabei fällt sofort die Ähnlichkeit der Form mit der bei einem Sprung auf. Jedoch wird auch deutlich, dass der Annahmebereich durch den Trend weniger stark in den Bereich negativer Korrelationen verschoben wird, als es bei der entsprechenden Sprunghöhe der Fall ist. Ein Grund dafür ist, dass hier das sequenzielle Schema der Vorzeichen durch den Trend weniger stark beeinflusst wird. Das ist erneut auf die oben aufgeführten Überlegungen zu den Nulldurchgängen zurückzuführen. Zum Vergleich ist die entsprechende Trennschärfe der LB-Tests bei demselben Simulationsaufbau in Abbildung 3.39 dargestellt.

Hieraus ist ersichtlich, dass der LB-Test in der Lage ist, den Trend ab einer gewissen Steigung zuverlässig zu erkennen. Im Bereich von geringen Anstiegen, die nicht als Abweichung von der Unabhängigkeit der Zeitreihe erkannt werden können, reagiert er robust und schafft es, vorhandene Korrelationen in der Zeitreihe zu entdecken und gleichzeitig das Niveau bei $\rho_1 = 0$ einzuhalten.

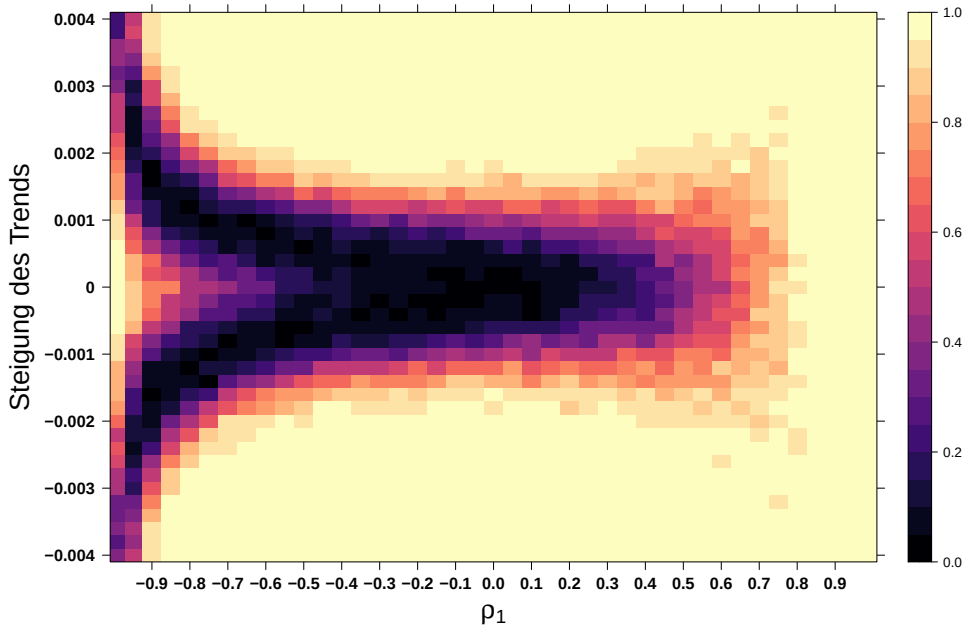


Abbildung 3.38: Simulierte Trennschärfen des 5-VZ-Tests in Abhängigkeit von der Steigung des Trends und von ρ_1 bei $N = 500$ Beobachtungen

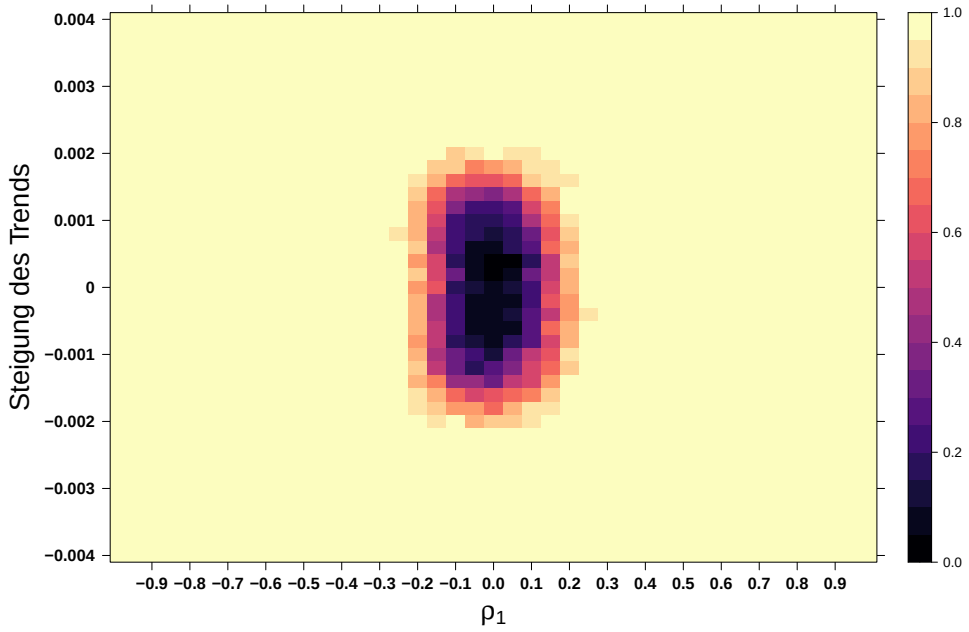


Abbildung 3.39: Simulierte Trennschärfen des LB-Tests in Abhängigkeit von der Steigung des Trends und von ρ_1 bei $N = 500$ Beobachtungen

Osziallationen

Eine weitere Alternative zur Zufälligkeit einer Zeitreihe im Kontext von Niveauänderungen stellt das Vorhandensein von Oszillationen dar. Dabei soll der Fokus in dieser Arbeit auf harmonischen Schwingungen liegen, also auf solchen, die mithilfe von einer Sinus- bzw. Cosinusfunktion beschrieben werden können. Statistische Anwendungsgebiete, in denen derartige Schwingungen zu beobachten sind, stellen z. B. die Musikdatenanalyse, in der Schwingungsdaten von Musikinstrumenten betrachtet werden, oder die Elektrotechnik dar, wobei unerwünschte Effekte wie Mikrofonie eine Rolle spielen können. Ähnlich wie bei den Simulationen mit einem Trend ist es hier von Interesse, welche der Verfahren die Oszillationen als Abweichung von der Zufälligkeitsannahme erkennen. Auch soll wieder untersucht werden, welche Testverfahren in diesem Fall robust reagieren und ungeachtet der Schwingungen zuverlässig Korrelationen detektieren können. Für die K-VZ-Tests ist in den Fällen, in denen der Median von den Beobachtungen nicht 0 entspricht, eine Zentrierung mit dem empirischen Median vorgenommen worden.

Feste Anzahl von Oszillationen

Im Folgenden sind deshalb Zeitreihen aus AR(1)-Prozessen zu verschiedenen Autokorrelationskoeffizienten generiert worden, die im Nachhinein mit einer Cosinusfunktion überlagert wurden. Dabei ist, unabhängig von der Beobachtungszahl, jeweils der Cosinus in dem Bereich [0,10] für unterschiedliche Amplituden auf die entsprechende Zeitreihe addiert worden, was in etwa einer und einer halben Cosinusschwingung entspricht. Damit simuliert eine größere Beobachtungszahl eine feinere Abtastrate desselben Signals. Denkbar wäre es auch, die Länge des Cosinussignals mit wachsender Beobachtungszahl zu erhören, was einer größeren Beobachtungsdauer bei konstanter Abtastrate entsprechen würde. Die Ergebnisse des TP-Tests und der K-VZ-Tests unterscheiden sich bei beiden Ansätze drastisch, während die anderen Tests nahezu identische Ergebnisse liefern. Zunächst soll aber der Fall untersucht werden, dass eine höhere Beobachtungszahl mit einer feineren Abtastrate einhergeht.

Die Unabhängigkeit der Zeitreihe ist in diesem Szenario genau dann gegeben, wenn die Amplitude des Cosinus 0 ist und der Autokorrelationskoeffizient ρ_1 ebenfalls 0 entspricht. Die Trennschärfen der einzelnen Verfahren sind für Stichprobenumfänge von $N = 20$ bis $N = 500$ in den Abbildungen $3.40 - 3.43$ dargestellt.

Beim Betrachten dieser Abbildungen fällt zunächst auf, dass die 3 verschiedenen K-VZ-Tests bei einer geringen Beobachtungszahl von $N = 20$ deutlich unterschiedlich auf die Oszillationen reagieren. So hat der 3-VZ-Test auf dem gesamten Spektrum Schwierigkeiten damit, die Nullhypothese zu verwerfen. Insbesondere wird der Wert von ρ_1, ab dem ihm eine Ablehnung gelingt, mit zunehmender absoluter Amplitudenhöhe immer größer, sodass seine Trennschärfe abnimmt. Beim 4-VZ-Test hingegen deutet sich an, dass eine Verwerfung der Nullhypothese mit wachsender Amplitudenhöhe wahrscheinlicher wird, während dieses Verhalten beim 5-VZ-Test deutlich zu erkennen ist. Bei hinreichend großer Beobachtungszahl scheinen die Unterschiede

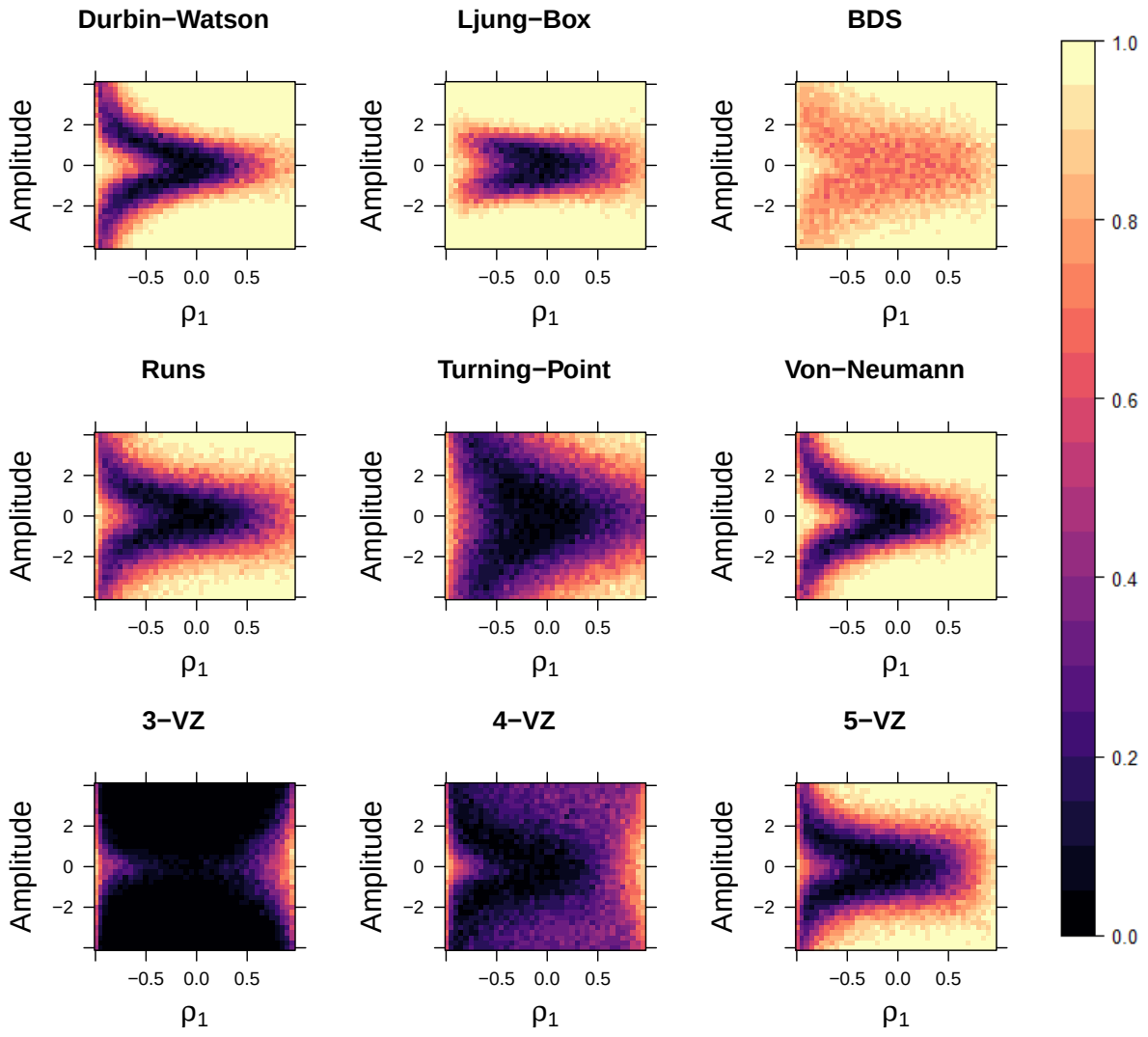


Abbildung 3.40: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 20$ Beobachtungen und einer festen Anzahl von Oszillationen

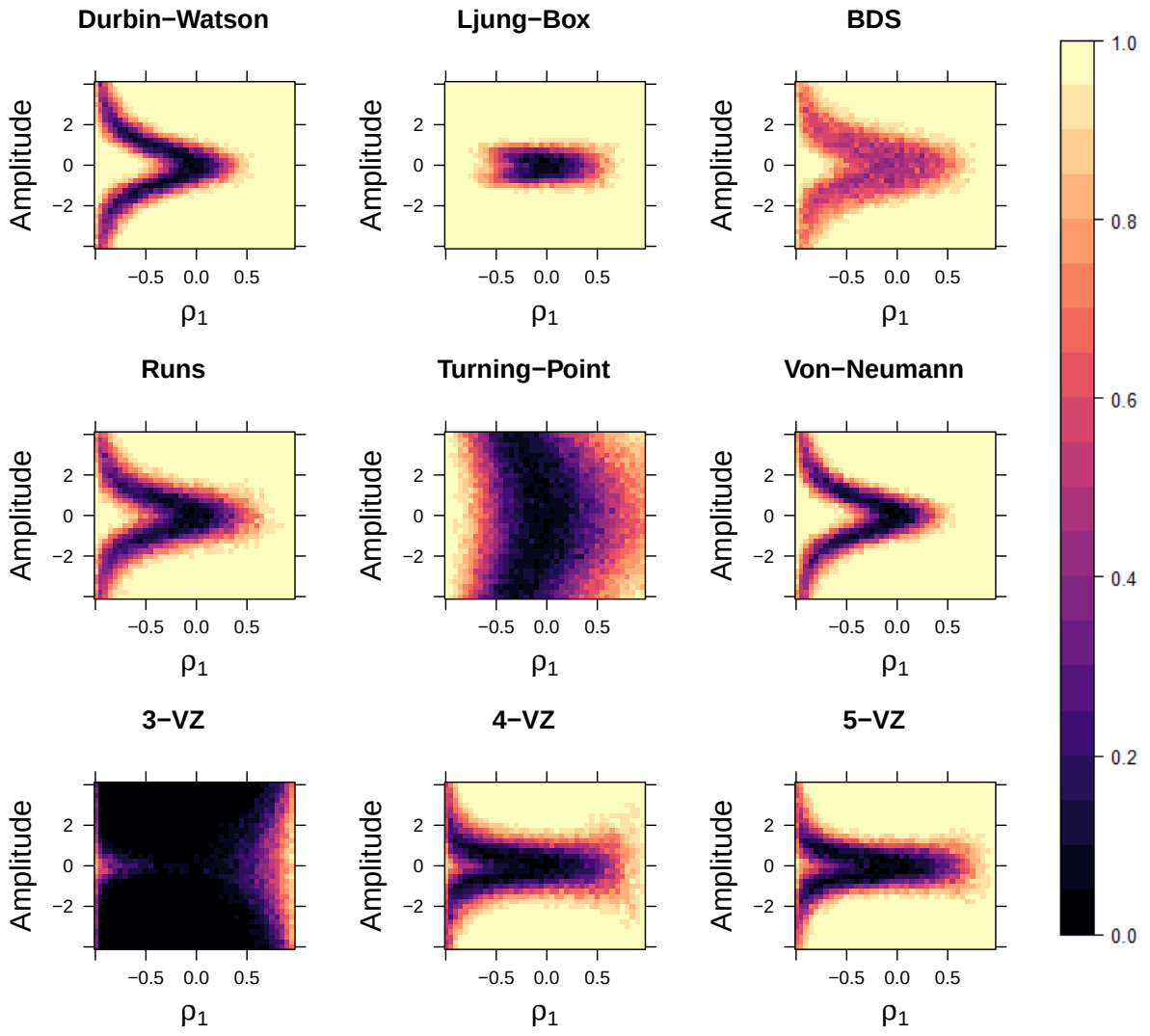


Abbildung 3.41: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 50$ Beobachtungen und einer festen Anzahl von Oszillationen

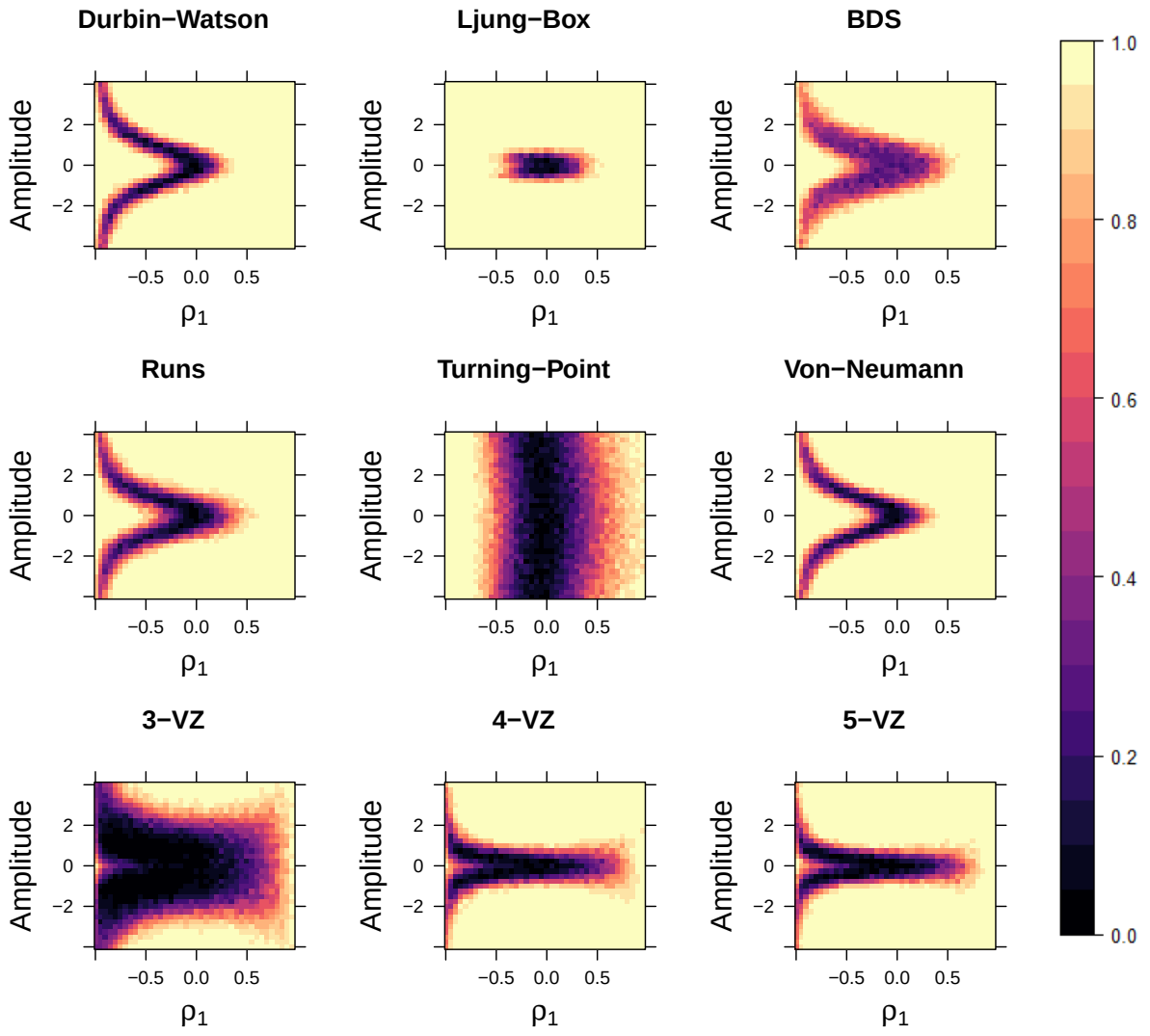


Abbildung 3.42: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 100$ Beobachtungen und einer festen Anzahl von Oszillationen

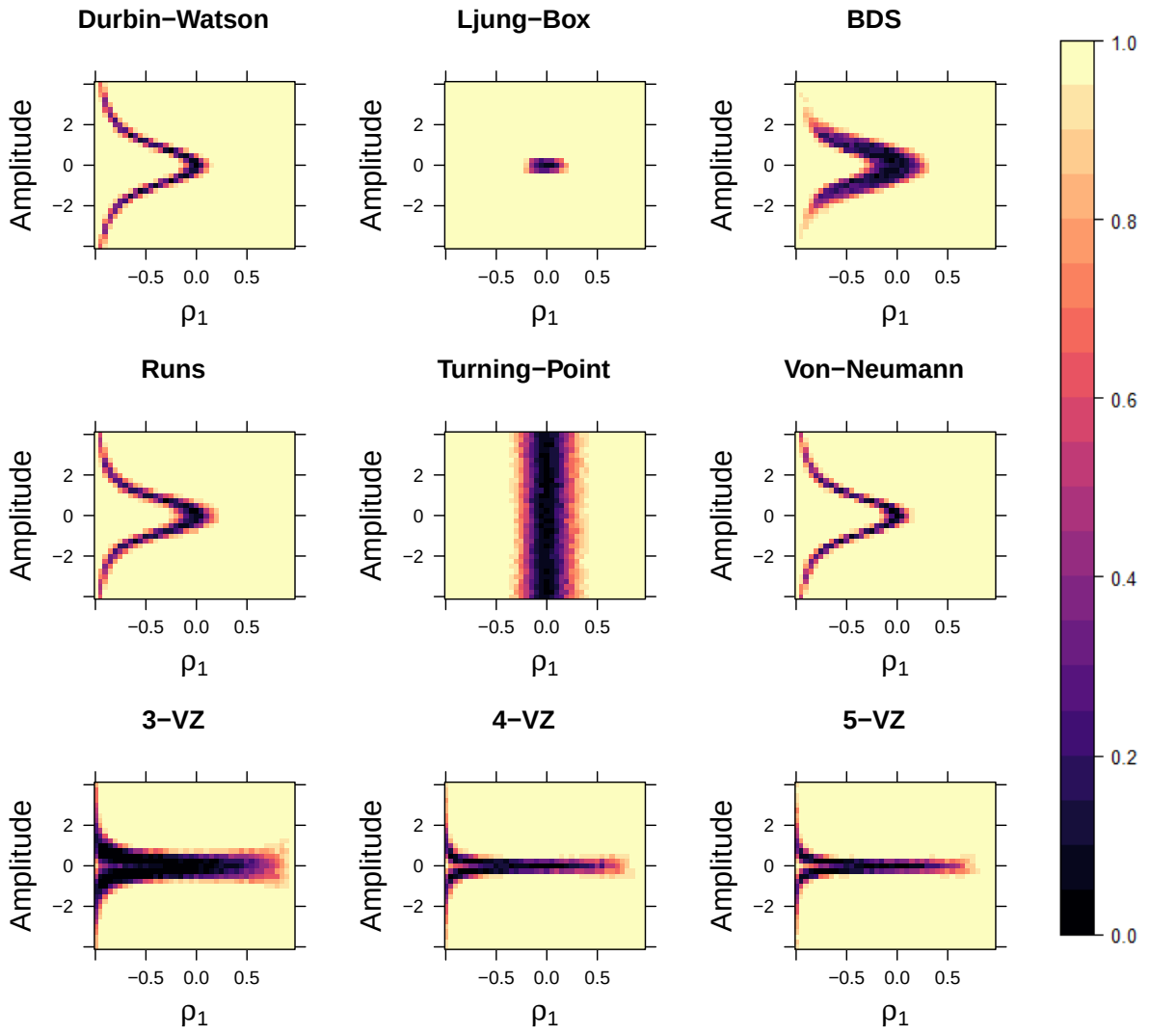


Abbildung 3.43: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 500$ Beobachtungen und einer festen Anzahl von Oszillationen

zwischen den verschiedenen K-VZ-Tests jedoch immer unwesentlicher zu werden, sodass ihre Trennschärfen bei $N = 500$ bereits sehr ähnlich aussehen.

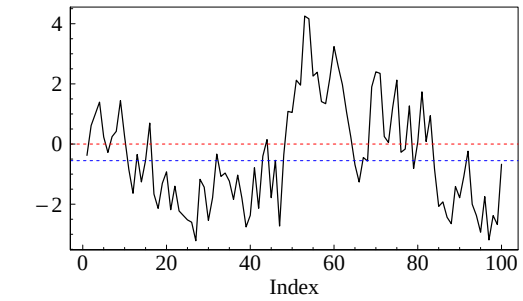
Allgemein ist zu beobachten, dass sich die Annahmebereiche des DW-Tests, des Runs-Tests des VNRR-Tests, der K-VZ-Tests sowie des BDS-Tests mit zunehmender Amplitudenhöhe in den Bereich negativer Korrelationen verschieben, wie es schon bei einem Sprung und einem Trend der Fall war. Dabei fällt auf, dass dies bei den K-VZ-Tests deutlich abrupter geschieht als bei den anderen Verfahren. So kann die Nullhypothese bei einem Stichprobenumfang von $N = 500$ bereits bei einer Amplitudenhöhe von ca. 1 lediglich im Bereich sehr starker Korrelationen von ca. -0.95 in wenigen Fällen nicht abgelehnt werden. Im Vergleich beträgt der Autokorrelationskoeffizient zu der entsprechenden Amplitudenhöhe, bei dem keine Verwerfung stattfinden kann, für die anderen Verfahren ungefähr ρ_1 = -0.5.

Der TP-Test und der LB-Test tun sich in diesem Szenario erneut besonders hervor. So zeigt der TP-Test − ähnlich wie in den Szenarien mit einem Trend oder Sprung − wieder ein robustes Verhalten. Seine Testentscheidung scheint bei einer Beobachtungszahl von $N \geq 100$ nicht von der Höhe der Amplitude sondern lediglich von der Korrelation abzuhängen. Die Trennschärfe des LB-Tests ist in diesem Szenario mit Abstand am besten. So gelingt es ihm als einziger Test, sowohl Korrelationen als auch eine vorhandene Oszillation separat voneinander zu detektieren. Bei einer Beobachtungszahl von $N = 500$ hält er das Testniveau im Fall der Unabhängigkeit − wie sie oben definiert wurde − ein und lehnt die Nullhypothese ab einer Amplitudenhöhe von 0.5 und einer Korrelation von $|\rho_1| > 0.1$ sicher ab. Von den anderen Verfahren hebt er sich insbesondere darin ab, dass der Annahmebereich bei vorhandener Oszillation nicht verschoben wird. Damit sind wieder Parallelen zu seiner Trennschärfe bei den anderen betrachteten Niveauänderungen erkennbar, in denen der Test ein ähnliches Verhalten gezeigt hat.

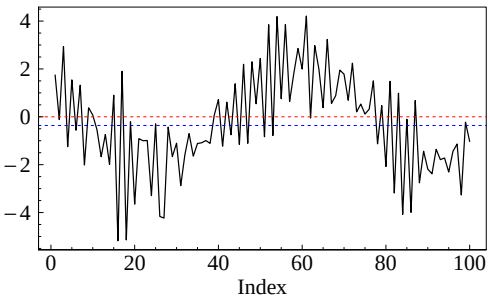
Um die Ergebnisse der verschiedenen Testverfahren im Folgenden nachvollziehen zu können, sind zwei Zeitreihen mit Oszillationen bei einer Amplitudenhöhe von 2 für den Fall ρ = 0.7 sowie ρ = -0.7 mit $N = 100$ Beobachtungen und den zu ihnen gehörigen Korrelogrammen in Abbildung 3.44 dargestellt.

Die Ergebnisse der Simulationsstudie des Runs-, des BDS-, des DW- sowie des VNRR-Tests können mit ähnlichen Überlegungen nachvollzogen werden, wie es bei einem Trend der Zeitreihe der Fall war. Dort wurde bereits diskutiert, wie ein gewisses Maß an negativen Korrelationen in Abhängigkeit mit der Steigung des Trends die Teststatistiken dahingehen beeinflussen kann, dass sie Werte annehmen, die unter der Nullhypothese zu erwarten wären. Dieses Verhalten lässt sich im Fall einer Oszillation in der Zeitreihe, wie anhand von Abbildung 3.44 erkennbar ist, ebenfalls beobachten.

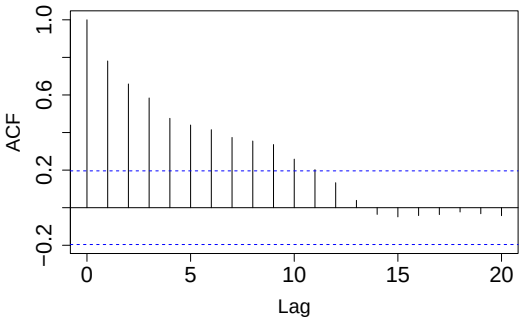
Die Tatsache, dass der TP-Test hier erneut so robust reagiert, ist auf den geringen Einfluss, den die langsamen Oszillationen auf die Anzahl der Turning-Points haben, zurückzuführen. Die Überlegenheit des LB-Tests gegenüber den anderen Tests lässt sich erneut damit begründen, dass er so viele Autokorrelationskoeffizienten für seine Testentscheidung heranzieht, dass eine



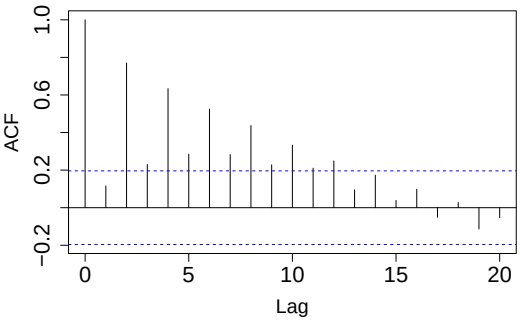
(a) Zeitreihe mit $\rho_1 = 0.7$



(b) Zeitreihe mit $\rho_1 = -0.7$



(c) ACF der Zeitreihe mit $\rho_1 = 0.7$



(d) ACF der Zeitreihe mit $\rho_1 = -0.7$

Abbildung 3.44: Zeitreihen mit $N = 100$ Beobachtungen mit Oszillationen und Amplitudenhöhe 2, für $\rho_1 = 0.7$ (a) und $\rho_1 = -0.7$ (b), mit zugehörigen Korrelogrammen (c) und (d), Nulllinie (rot) und empirischem Median bzw. kritischen Werten (blau)

vorhandene Korrelation in der Zeitreihe es nicht vermag, die durch die Oszillation verursachten Erhöhungen der Autokorrelationskoeffizienten zu verschleiern.

Die Beobachtung, dass der 3-VZ-Test bei geringer Beobachtungszahl so viel schlechter abschneidet als der 5-VZ-Test, hängt hier mit dem Schema der Vorzeichenblöcke, wie es durch die Oszillation verursacht wird, zusammen. So verläuft die Schwingung des Cosinus im Intervall $[0, \pi/2]$ im positiven Bereich, im Intervall $[\pi/2, 3\pi/2]$ im negativen Bereich, im Intervall $[3\pi/2, 5\pi/2]$ wieder im positiven Bereich und das restliche Intervall liegt erneut im negativen Bereich. Dieses Verhalten führt vor allem bei kleinen Stichprobenumfängen dazu, dass tendenziell 3 größere Vorzeichenblöcke entstehen. Da im Kapitel 2.9 bereits erwähnt wurde, dass es sich bei (p - 1) Vorzeichenwechseln generell anbietet, das K für die Vorzeichentiefe mindestens als $(p+1)$ zu wählen, lässt sich nachvollziehen, warum ein größeres K hier bessere Ergebnisse liefert. Der Effekt wird bei einer höheren Abtastrate − und damit tendenziell mehr Vorzeichenblöcken − immer schwächer, sodass sich die Trennschärfen der 3 verschiedenen K-VZ-Tests immer weiter annähern. Weiter kann mit den obigen Überlegungen aber auch nachvollzogen werden, warum die Tests so sensibel auf eine zunehmende Amplitudenhöhe reagieren. So begünstigt sie das Auf-

treten weniger, ähnlich großer Vorzeichenblöcke, die dazu führen, dass die Teststatistiken Werte nahe ihres Maximums annehmen.

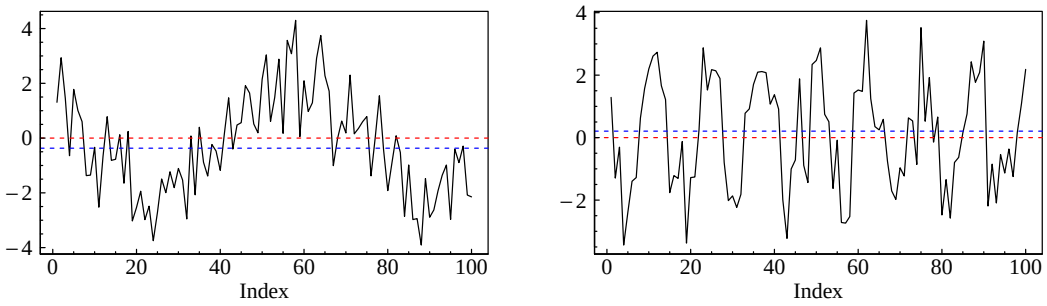
Wachsende Anzahl an Oszillationen

Wie zu Beginn dieses Abschnittes erwähnt, ist es weiterhin von Interesse, wie die Verfahren reagieren, falls mit einer erhöhten Beobachtungszahl größere Ausschnitte der harmonischen Schwingung beobachtet werden. Dazu wurden die simulierten Zeitreihen wie im oben betrachteten Fall von einer Cosinusfunktion überlagert. Diesmal wurde aber der Cosinus im Intervall [0,N/2] auf die Zeitreihe addiert. Damit sind die beiden Szenarien für den Fall $N = 20$ identisch, unterscheiden sich aber für größere Beobachtungszahlen. Dies wird in Abbildung 3.45 anhand zweier Zeitreihen für den Fall $N = 100$ veranschaulicht.

Die Simulationsergebnisse in diesem Szenario sind in Abbildungen 3.46 und 3.47 für Stichprobenumfänge von $N = 50$ und $N = 500$ dargestellt.

Hier fällt − wie bereits erwähnt − auf, dass lediglich der TP-Test und die K-VZ-Tests sichtbar unterschiedliche Ergebnisse als im Fall einer zunehmenden Abtastrate liefern. So wird der TP-Test diesmal deutlich durch die Oszillationen beeinflusst und zeigt ein ähnliches Verhalten wie der Runs-, der DW- und der VNRR-Test, bei denen sich der Annahmebereich mit zunehmender Amplitudenhöhe in Bereiche negativer Korrelationen verschiebt. Die K-VZ-Tests weisen in diesem Szenario, unabhängig von der Wahl des Parameters K, eine deutlich schlechtere Trennschärfe auf als bei einer festen Anzahl von Oszillationen. So gelingt es ihnen selbst bei einer Beobachtungszahl von $N = 500$ auf dem gesamten Spektrum kaum, Abweichungen von der Nullhypothese der Unabhängigkeit zu erkennen.

Das Verhalten des TP-Tests lässt sich anhand von Abbildung 3.45 nachvollziehen. So verhindert die Oszillation durch die viel größere lokale Steigung in einigen Fällen das Auftreten von Turning-Points, sodass unter der Nullhypothese weniger von ihnen entstehen, als es bei einer



(a) mit erhöhter Abtastrate (b) mit längerer Beobachtungsdauer

Abbildung 3.45: Zeitreihen mit $N = 100$ Beobachtungen bei fester und wachsender Oszillationsanzahl und Amplitudenhöhe 2, Nulllinie (rot) und empirischem Median (blau)

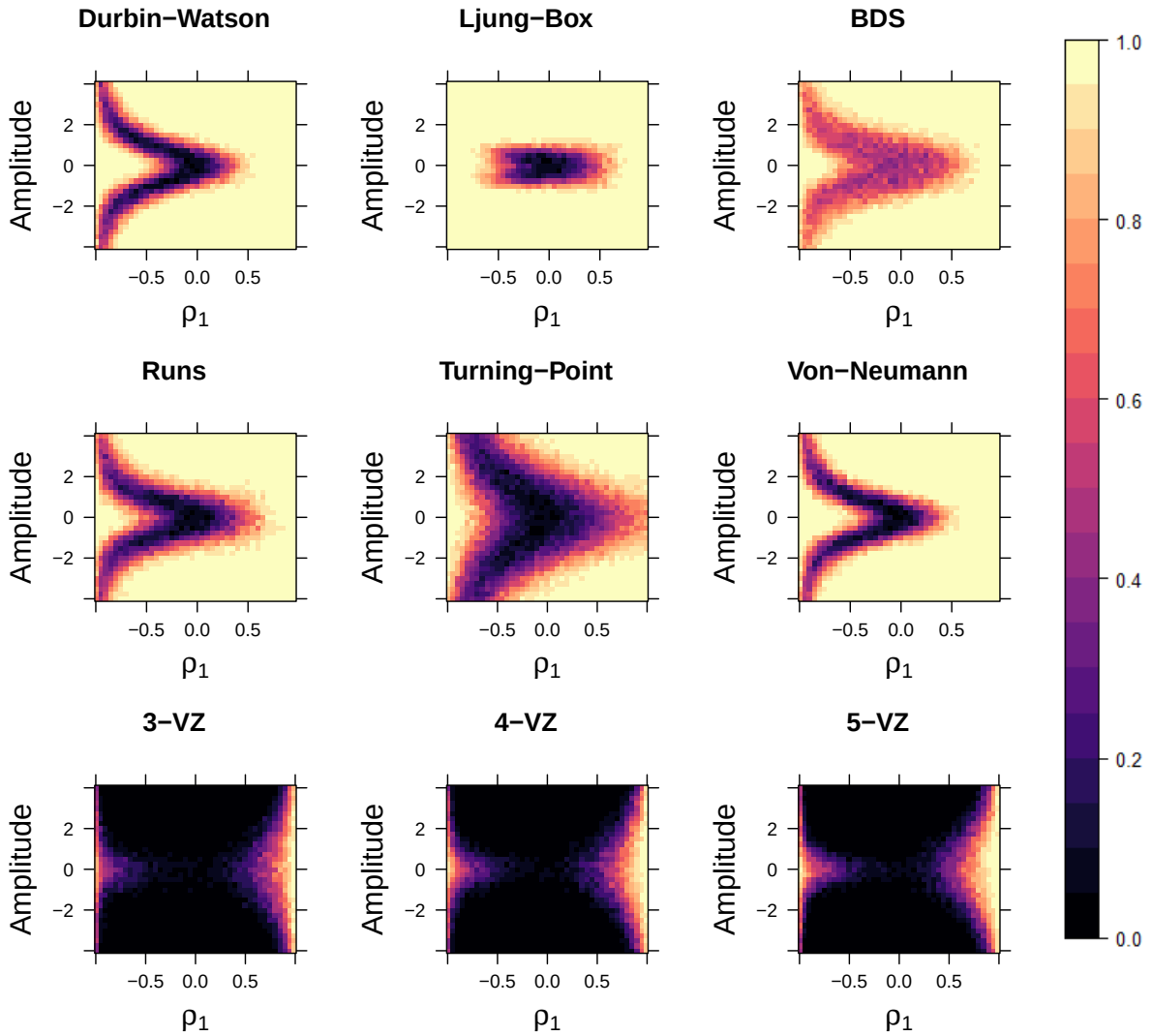


Abbildung 3.46: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 50$ Beobachtungen und einer wachsenden Anzahl von Oszillationen

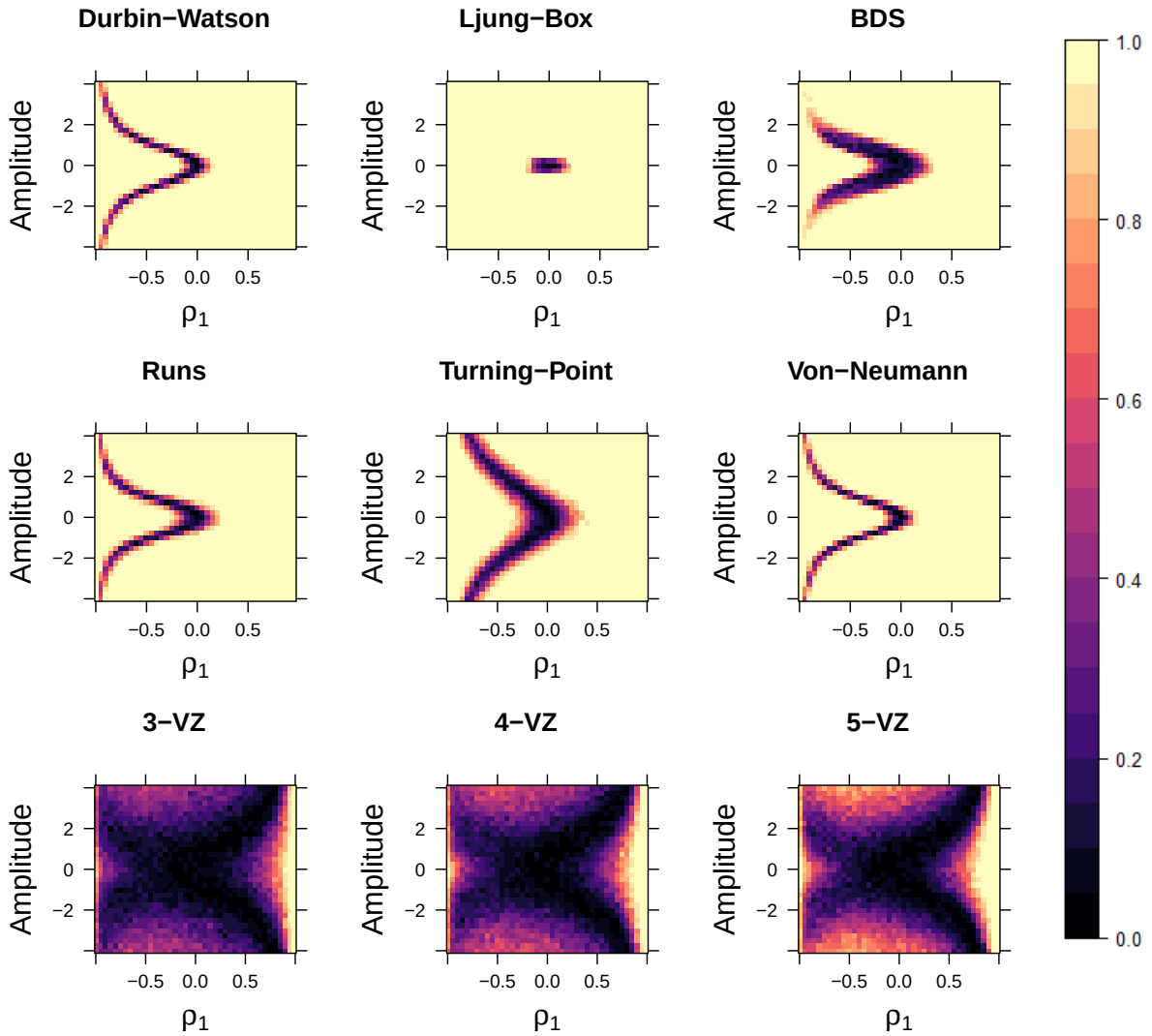


Abbildung 3.47: Simulierte Trennschärfen der Testverfahren bei stationären AR(1)-Alternativen mit Oszillationen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 500$ Beobachtungen und einer wachsenden Anzahl von Oszillationen

zufälligen Zeitreihe zu erwarten wäre. Dieses Phänomen ist mit zunehmender Amplitudenhöhe stärker ausgeprägt. Negative Korrelationen erzielen einen gegenteiligen Effekt und erhöhen die Anzahl der Turning-Points gegenüber der unter der Nullhypothese zu erwartenden Anzahl. Deshalb können negative Korrelationen den Effekt einer Oszillation kompensieren. Somit ist auch nachvollziehbar, warum sich der Annahmebereich mit zunehmender Amplitudenhöhe in den Bereich negativer Korrelationen verschiebt.

Um nachzuvollziehen, warum die Trennschärfe der K-VZ-Tests so viel schlechter ist als im vorherigen Szenario, macht es erneut Sinn, sich auf die Vorzeichenblock-Struktur zu beziehen und die im Kapitel 2.9 erläuterten Vorschläge zur Wahl des Parameters K zu berücksichtigen. Anders als im Fall einer erhöhten Abtastrate, geht hier eine erhöhte Beobachtungszahl mit einer höheren Anzahl an Vorzeichenblöcken einher (s. Abb. 3.45). Während im Fall $N = 20$ (der äquivalent zu dem der sich erhöhenden Abtastrate ist) ein K von 5 zur Überprüfung der Nullhypothese gerechtfertigt zu sein scheint, sind im Fall $N = 50$ bereits deutlich mehr Vorzeichenblöcke zu erwarten. Hier wäre also die Verwendung eines größeren Ks nötig, um die Struktur der Zeitreihe als Abweichung von der Unabhängigkeit zu erfassen. Da die Anzahl der zu erwartenden Vorzeichenblöcke mit größerer Beobachtungszahl immer weiter zunimmt, ist es auch nicht verwunderlich, dass auch bei $N = 500$ noch keiner der Tests in der Lage ist, gute Ergebnisse bei der Aufdeckung der Abhängigkeitsstrukturen zu erzielen. Damit zeigt sich außerdem, dass die K-VZ-Tests durch sich wiederholende Muster in der Zeitreihe kaum an Trennschärfe gewinnen.

Insgesamt legen die Simulationsergebnisse nah, dass alle der betrachteten Testverfahren − mit Ausnahme des TP-Tests − tendenziell in der Lage sind, Niveauänderungen in einer Zeitreihe zu detektieren. Dabei muss allerdings beachtet werden, dass viele der Verfahren gewisse Alternativen mit negativer Autokorrelation nicht als Abhängigkeitsstruktur erkennen können. Dieses Verhalten ist bei den K-VZ-Tests am stärksten ausgeprägt. Der Annahmebereich verschiebt sich bei ihnen also am schnellsten in den Bereich negativer Autokorrelationen, sodass bei moderaten Niveauänderungen bereits extreme negative Korrelationen nötig sind, damit die Nullhypothese beibehalten werden kann. Somit scheint er sich gut für die Detektion dieser Struktur zu eignen. Allerdings sind die K-VZ-Tests, zumindest für diejenigen Ks, die gegenwärtig für die Testdurchführung verfügbar sind, nicht in der Lage, Oszillationen, die zu vielen Vorzeichenblöcken führen, zu erkennen. Spitzenreiter im Fall, dass Niveauänderungen in der Zeitreihe vorhanden sind, ist der LB-Test, der unter sämtlichen Alternativen zur Zufälligkeit mit Abstand am besten abschneidet und Niveauänderungen und Autokorrelationen gleichermaßen zu detektieren vermag. Auch der BDS-Test ist bei hinreichend großem Beobachtungsumfang und ausreichend stark ausgeprägten Niveauänderungen in der Lage, diese als Abhängigkeitsstruktur zu erkennen. Eine Sonderstellung nimmt der TP-Test ein, der in allen Szenarien deutlich robuster als die übrigen Tests reagiert und ungeachtet der Niveauänderung zuverlässig Autokorrelationen identifizieren kann. Lediglich bei Oszillationen mit hoher Frequenz konnte eine Verschiebung in den Bereich negativer Korrelationen festgestellt werden.

3.1.6 Resümee

Zusammenfassend lässt sich für den Fall, dass Zeitreihen einem stationären AR(1)-Prozess folgen, feststellen, dass die beste Wahl für einen Test auf Unabhängigkeit bzw. Zufälligkeit von den Eigenschaften der Zeitreihe und den konkreten Bedürfnissen des Anwenders abhängt. So weisen die parametrischen Verfahren sehr gute Trennschärfen auf, solange ihre Anwendungsvoraussetzungen erfüllt sind. Ist dies allerdings nicht der Fall oder kann deren Erfüllung im Vorhinein nicht beurteilt werden, so stellen nichtparametrische Verfahren, die in solchen Fällen robust reagieren, eine sinnvolle Alternative dar. Während Verletzungen von den Verteilungsannahmen der Innovationen und innovative Ausreißer keine gravierende Verschlechterung der parametrischen Testverfahren bewirken, solange die Momente der Innovationsverteilungen existieren und damit eine Anwendung des zentralen Grenzwertsatzes möglich ist, verschlechtert sich ihre Trennschärfe bei anderen Verteilungen deutlich. Besonders stark werden diese Verfahren auch von Kontaminationen in der Zeitreihe beeinflusst und ihre Trennschärfe leidet in solchen Fällen drastisch.

Als nichtparametrischer Test mit der besten Trennschärfe sticht der VNRR-Test hervor, der durch die Betrachtung von Rängen mehr Informationen über die vorliegende Zeitreihe verwertet. Allerdings bewirken diese Mehrinformationen eine weniger starke Robustheit des Tests, sodass z. B. im Fall von Kontaminationen und wachsenden Varianzen festgestellt werden konnte, dass der Runs-Test leicht bessere Ergebnisse liefert.

Im Hinblick auf die K-VZ-Tests zeigt sich, dass diese deutliche Probleme damit haben, Korrelationen im Rahmen von AR(1)-Prozessen zu detektieren. So ist ihre Trennschärfe bei der Erkennung solcher Abhängigkeitsstrukturen unter allen betrachteten Verfahren am schlechtesten. In kleinen Stichproben ist die Trennschärfe der K-VZ-Tests noch mit den anderen Verfahren vergleichbar, aber sie profitieren kaum von einem wachsenden Stichprobenumfang. Eine Begründung dafür könnte in der zunehmend großen Menge von weit auseinanderliegenden K-Tupeln liegen, die für eine Testentscheidung herangezogen werden. So klingen die Autokorrelationen in AR(1)-Prozessen exponentiell ab und Vorzeichenwechsel von weit entfernten Tupeln liefern kaum Informationen über diese Art von Abhängigkeitsstruktur.

Weisen die Zeitreihen Strukturen auf, die neben der hier primär interessierenden Korrelation eine Abweichung von der Zufälligkeit der Zeitreihe darstellen, ist es für die Wahl eines geeigneten Testverfahrens wichtig, dass der Anwender sich darüber bewusst ist, was für Strukturen in der Zeitreihe zu erwarten sind und was genau er durch das Testen aufdecken möchte. Eine Sonderstellung nimmt der TP-Test ein, der über Trends, Oszillationen sowie Varianzheterogenitäten hinwegzusehen und das reine Vorhandensein einer Korrelation in der Zeitreihe detektieren kann. Anders zeichnen sich der LB-Test sowie der BDS-Test aus, indem sie Varianzheterogenitäten, Oszillationen und Sprünge als Strukturen, die eine Abweichung von der Zufälligkeit der Zeitreihe darstellen, neben den Korrelationen erkennen können. Dabei ist die Güte des LB-Tests der des BDS-Tests, zumindest für kleine Stichprobenumfänge, deutlich überlegen und ihm gelingt es sogar, Trends in der Zeitreihe aufzudecken. Der relativ große erforderliche Stichprobenumfang

des BDS-Tests von $N \geq 500$, ab dem das Niveau des Tests eingehalten werden kann, stellt damit einen wesentlichen Nachteil dieses Verfahrens dar. Allerdings profitiert der LB-Test nicht von den robusten Eigenschaften nichtparametrischer Verfahren. Bei den übrigen Testverfahren ergibt sich das Problem, dass Abhängigkeitsstrukturen bei deren Durchführung untereinander interagieren können, sodass die Nullhypothese unter gewissen Alternativen nicht abgelehnt werden kann. Dabei zeichnen sich die K-VZ-Tests dadurch aus, das dieser Alternativenbereich deutlich kleiner ist als bei den anderen Verfahren und sich viele dieser Alternativen im Rahmen von unrealistischen Abhängigkeitsstrukturen bewegen. So sind sehr extreme negative Korrelationen nötig, um z. B. mittelmäßig stark ausgeprägte Oszillationen, Sprünge oder Trends zu kompensieren.

3.2 AR(2)-Prozesse

In diesem Abschnitt werden die Trennschärfen der verschiedenen Testverfahren im Kontext von stationären, autoregressiven Prozessen 2. Ordnung (AR(2)-Prozesse) untersucht. Anders als im vorherigen Kapitel hängt der Wert einer Beobachtung bei solchen Prozessen zu einem gewissen Anteil vom Wert der vorangegangenen Beobachtung sowie von dem Wert der Beobachtung, die 2 Zeitschritte in der Vergangenheit liegt, ab. Schematisch handelt es sich also um Prozesse der Form:

$$x_t = \mu + \rho_1 x_{t-1} + \rho_2 x_{t-2} + w_t, \quad w_t \sim WN(0, \sigma^2_{WN})$$

für $t \in \{3, ..., N\}$, wobei ρ_1 und ρ_2 als autoregressive Parameter 1. und 2. Ordnung bezeichnet werden. Die Unabhängigkeit des Prozesses liegt dabei genau dann vor, wenn $\rho_1 = \rho_2 = 0$ gilt. Es kann gezeigt werden, dass AR(2)-Prozesse genau dann stationär sind, wenn die Parameter ρ_1 und ρ_2 den folgenden Gleichung genügen (s. Schlittgen, 2001, S. 128):

(i) $\rho_1 + \rho_2 < 1$
(ii) ρ_1 - $\rho_2 < 1$
(iii) -1 $< \rho_2 <$ 1.

Sie definieren dabei das sogenannte Stationaritätsdreieck für AR(2)-Prozesse, das in Abbildung 3.48 dargestellt ist.

Zu Beginn der Untersuchungen werden Prozesse betrachtet, bei deren Innovationen es sich um standardnormalverteilte Zufallsgrößen mit $\sigma^2_{WN} = 1$ handelt und die auch sonst alle Voraussetzungen für die betrachteten Testverfahren erfüllen. Auch wird der Mittelwert der Zeitreihen μ als 0 gesetzt. Dieses Szenario wird dabei in Anlehnung an Kapitel 3.1 als „Normalbedingungen“ bezeichnet. Um die Trennschärfe der Tests zu beurteilen, wurden innerhalb des Stationaritätdreiecks jeweils 100 Zeitreihen für jeden Gitterpunkt simuliert, wobei ein Gitter der Feinheit 0.1 für beide autoregressiven Parameter gewählt worden ist. Die Länge der Burn-In-Phase wurde

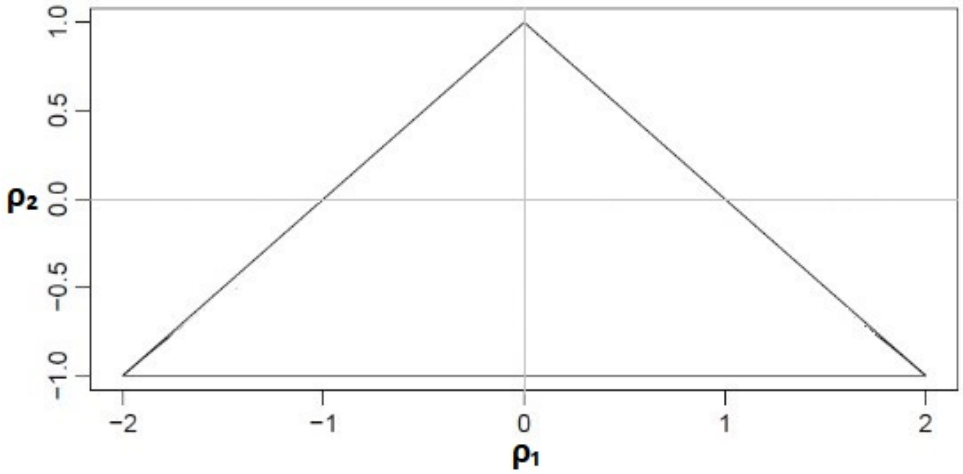


Abbildung 3.48: Das Stationaritätsdreieck für AR(2)-Prozesse

dabei wie im Kapitel 3.1 gewählt. Im Anschluss sind die relativen Ablehnungsraten der jeweiligen Testverfahren für die einzelnen Alternativen berechnet worden. Die Ergebnisse dieses Vorgehens sind in den Abbildungen 3.49 bis 3.52 grafisch dargestellt.

Dabei fällt zunächst auf, dass der DW-Test, der VNRR-Test sowie der Runs-Test ähnliche Ablehnungsmuster aufweisen. So können sie ausschließlich Abweichungen des Parameters ρ_1 von 0 detektieren. Der Bereich, ab dem Abweichungen von ρ_1 zuverlässig erkannt werden, wird dabei mit zunehmender Größe des Parameters ρ_2 systematisch kleiner. Konkret findet eine Verwerfung der Nullhypothese bei einem Stichprobenumfang von $N = 500$ und Parameterwerten von $\rho_2 > 0.5$ und $|\rho_1| > 0.1$ für all diese Verfahren in mehr als 95 % der Fälle statt. Dabei gelingt es diesen Tests jedoch nicht, selbst extreme Autokorrelationen 2. Ordnung zuverlässig zu detektieren, solange $\rho_1 = 0$ gilt. Von den oben genannten Methoden schneidet der Runs-Test dabei am schlechtesten ab, während sich der VNRR-Test und der DW-Test bezüglich ihrer Trennschärfe kaum unterscheiden. Weiterhin ist auffällig, dass der DW-Test die Nullhypothese bei kleinen Beobachtungszahlen und einem stark negativen ρ_2 für negative Werten von ρ_1 öfter zu verwerfen scheint als für positive − er weist also eine leichte Asymmetrie auf.

Beim BDS-Test ist eine ähnliche Systematik wie bei den oben erwähnten Testverfahren erkennbar. Allerdings benötigt er − wie im AR(1)-Fall − eine große Beobachtungszahl um zufriedenstellende Ergebnisse zu liefern. Dabei ist er den oben diskutierten Verfahren, selbst bei einem Stichprobenumfang von $N = 500$ in Bezug auf seine Trennschärfe noch deutlich unterlegen. Anders als die anderen Methoden scheint es dem BDS-Test jedoch bei hinreichend großer Beobachtungszahl möglich zu sein, extreme negative Korrelationen 2. Ordnung zu erkennen, sodass die Nullhypothese in diesen Fällen zumindest bei mehr als 70 % der Simulationen verworfen werden kann.

Der TP-Test weist ein deutlich anderes Muster als die übrigen Verfahren auf und scheint die Nullhypothese vor allem im Bereich positiver Werte von ρ_1 in Kombination mit negativen Werten von ρ_2 abzulehnen. Insgesamt ist die Trennschärfe für negative ρ_1 und positive ρ_2 hier eher schlecht. Ab einer Beobachtungszahl von $N = 500$ wird die Systematik erkennbar, dass eine Verwerfung der Nullhypothese in den Fällen, in denen nahezu der Zusammenhang $\rho_1 = \rho_2$ gilt, nicht stattfinden kann.

Die K-VZ-Tests zeigen im AR(2)-Fall eine sehr schlechte Trennschärfe. So gelingt es ihnen die Nullhypothese unabhängig vom Stichprobenumfang nur für wenige Alternativen mit extremen Werten von ρ_1 und ρ_2 am „Rand“ des Stationaritätsdreiecks zu verwerfen. Auffällig ist allerdings, dass die K-VZ-Tests bei einer Beobachtungszahl von $N = 500$ extreme negative Werte von ρ_2 relativ zuverlässig detektieren können, während viele der andere Verfahren in diesem Bereich systematische Schwachstellen aufweisen.

Absoluter Spitzenreiter in diesem Szenario ist der LB-Test, da er für jeden betrachteten Stichprobenumfang sowohl Autokorrelationen 1. als auch 2. Ordnung sehr zuverlässig erkennen kann. Die Trennschärfe bei einem Stichprobenumfang von $N = 500$ gehört für beide Parameter zu den

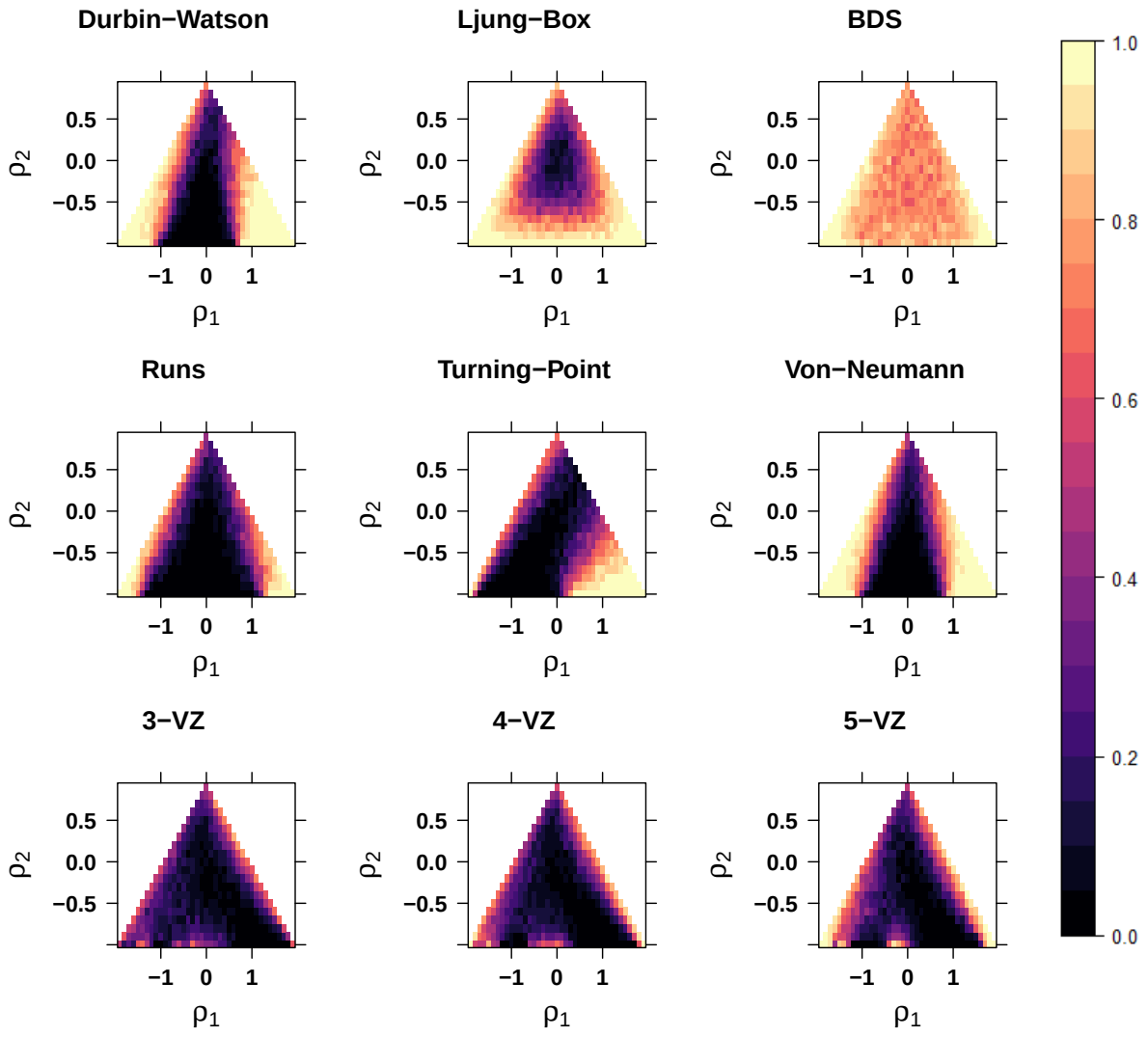


Abbildung 3.49: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen unter Normalbedingungen, in Abhängigkeit von ρ_1 und ρ_2 bei einer Beobachtungszahl von $N = 20$

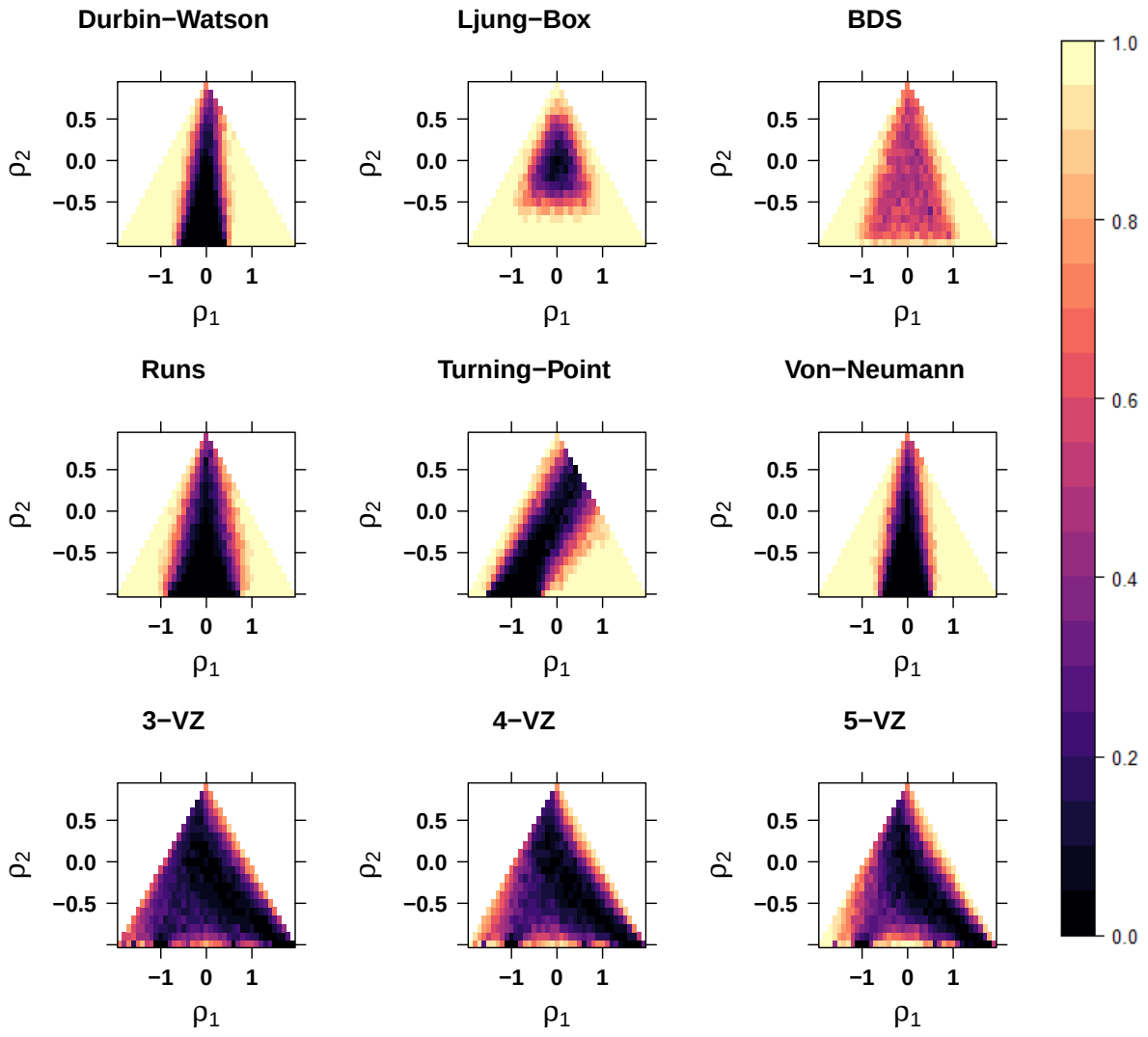


Abbildung 3.50: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen unter Normalbedingungen, in Abhängigkeit von ρ_1 und ρ_2 bei einer Beobachtungszahl von $N = 50$

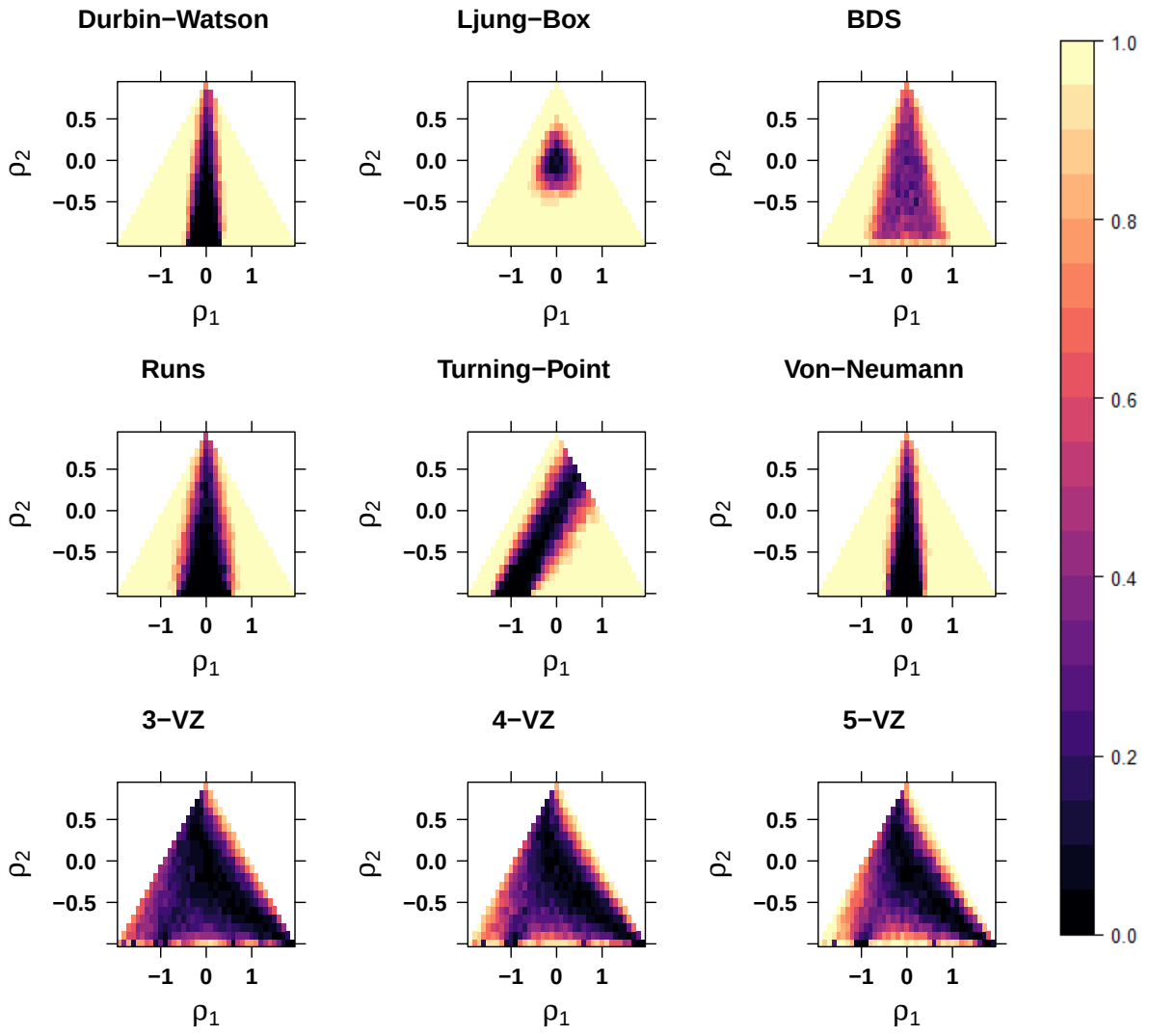


Abbildung 3.51: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen unter Normalbedingungen, in Abhängigkeit von ρ_1 und ρ_2 bei einer Beobachtungszahl von $N = 100$

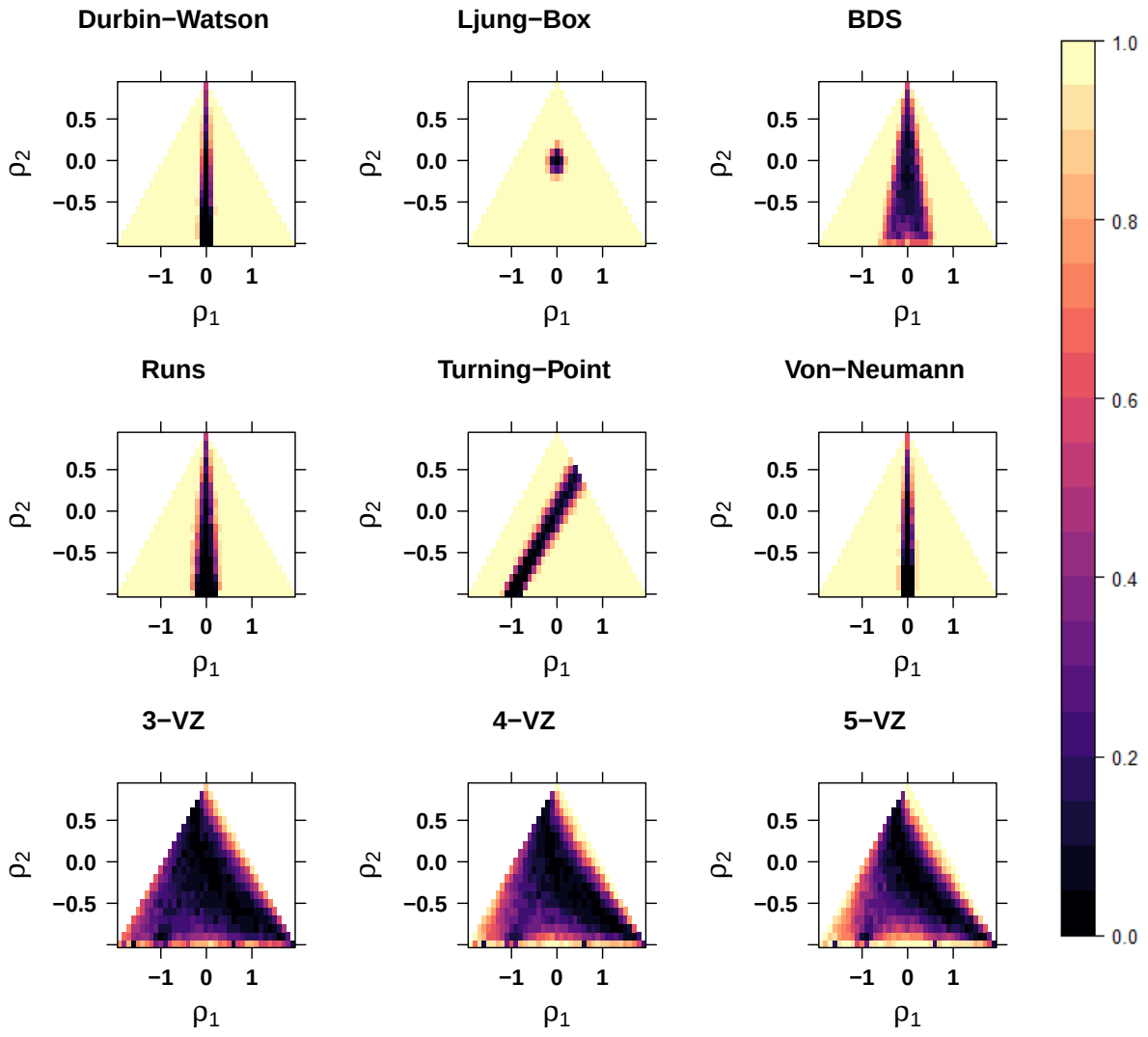


Abbildung 3.52: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen unter Normalbedingungen, in Abhängigkeit von ρ_1 und ρ_2 bei einer Beobachtungszahl von $N = 500$

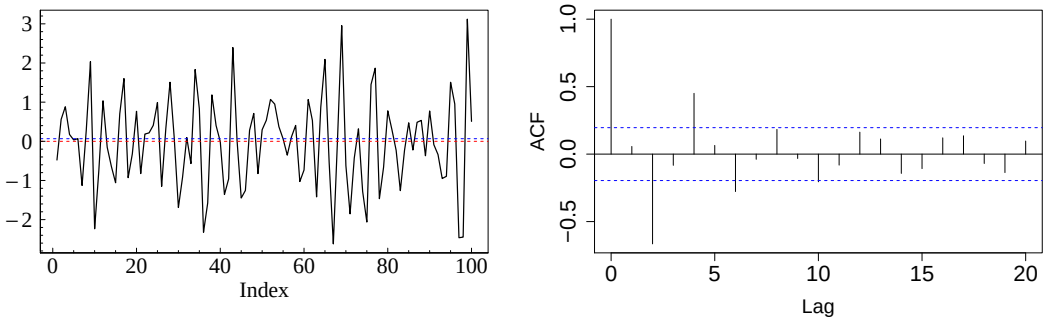
besten und bereits geringe Abweichungen von der Nullhypothese können zuverlässig erkannt werden.

Um die Entscheidungen der verschiedenen Testverfahren nachvollziehen zu können, werden im Folgenden wieder einige exemplarische Zeitreihen betrachtet. Zunächst ist dafür eine Zeitreihe eines AR(2)-Prozesses mit $\rho_1 = 0$ und $\rho_2 = -0.7$ gemeinsam mit dem zugehörigen Korrelogramm, in Abbildung 3.53 dargestellt. Mit diesen Parametern fällt der zugrunde liegende Prozess in einen Bereich, in dem der DW-, der BDS-, der Runs- sowie der VNRR-Test Schwierigkeiten mit der Verwerfung der Nullhypothese der Unabhängigkeit haben.

Anhand des Korrelogramms kann nachvollzogen werden, warum der DW-Test, der im Wesentlichen auf dem empirischen Autokorrelationskoeffizienten 1. Ordnung beruht, in so einem Fall nicht in der Lage ist, die Nullhypothese zu verwerfen. So weist $\hat{\rho}_1$ hier einen sehr geringen Wert auf und ist insbesondere nicht signifikant erhöht. Außerdem setzt sich die durch ρ_2 erzeugte Korrelation lediglich zu jedem zweiten Lag fort, sodass umgekehrt jeder zweite empirische Autokorrelationskoeffizient ebenfalls sehr gering und nicht signifikant ausfällt. Dieses Verhalten wird schon durch die Modellspezifikation nahegelegt und kann − ähnlich wie bereits für den AR(1)-Prozess dargelegt − durch sukzessives Ersetzen der Autokorrelationskoeffizienzen auf folgende Weise verdeutlicht werden:

$$x_t = \rho_2 x_{t-2} + w_t = \rho_2(\rho_2 x_{t-4} + w_{t-2}) + w_t = \rho_2^2 x_{t-4} + \rho_2 w_{t-2} + w_t = \dots .$$

Mit diesen Überlegungen wird auch deutlich, warum der LB-Test hier viel bessere Ergebnisse liefert. So sind hier 4 der 15 betrachteten empirischen Autokorrelationskoeffizienten signifikant von 0 verschieden, wodurch dem LB-Test eine Verwerfung der Nullhypothese ermöglicht wird. Damit zeigt sich, dass die Alternative des LB-Tests durch die Betrachtung mehrerer Autokorrelationskoeffizienten eine größere Menge von Abhängigkeitsstrukturen enthält als die des DW-Tests. Es sollte jedoch bedacht werden, dass damit auch − wie in Kapitel 3.1 erörtert − ge-



(a) Zeitreihe mit $\rho_2 = -0.7$ und $\rho_1 = 0$ (b) Korrelogramm der Zeitreihe

Abbildung 3.53: Zeitreihe eines AR(2)-Prozesses bei einer Beobachtungszahl von $N = 100$, mit $\rho_1 = 0$ und $\rho_2 = -0.7$ (a), mit zugehörigem Korrelogramm (b), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

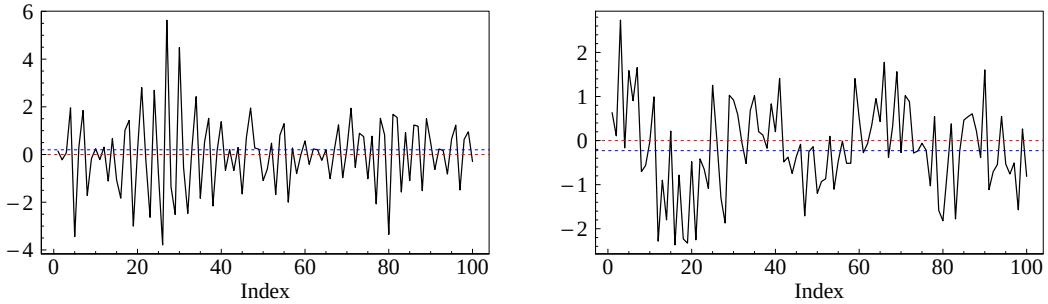
ringe Einbußen bezüglich der Trennschärfe bei AR(1)-Prozessen einhergehen, in denen die vielen betrachteten Autokorrelationskoeffizienten wenig zusätzlichen Informationsgewinn liefern.

Die Tatsache, dass der Runs- und der VNRR-Test in diesem Szenario nicht in der Lage sind, Abweichungen von der Unabhängigkeit zu erkennen, lässt sich anhand der in Abbildung 3.53 dargestellten Zeitreihe im Vergleich mit der Zeitreihe in Abbildung 3.3 verstehen. Dabei ist zunächst ersichtlich, dass die Anzahl der Durchgänge des Graphen durch den Median deutlich geringer ist, als es im Fall von einer Autokorrelation 1. Grades der gleichen Größenordnung zu erwarten wäre. Dies ist damit zu erklären, dass eine Beobachtung und die darauf folgende in einem Prozess, wie er hier betrachtet wird, unkorreliert sind. Somit tritt ein alternierendes Verhalten erst bei jeder zweiten Beobachtung auf. Dadurch wird die Anzahl der Runs deutlich kleiner als im AR(1)-Fall, wenn ρ_1 = -0.7 gilt und entspricht in etwa einer, die unter der Nullhypothese zu erwarten wäre.

Auch im Hinblick auf den VNRR-Test führt dieses Verhalten dazu, dass aufeinanderfolgende Ränge weder systematisch sehr nah beieinander noch weit entfernt voneinander, sondern nahezu unabhängig sind.

Für die Anzahl der Turning-Points bedeutet das alternierende Verhalten zum Lag 2, dass ein Turning-Point bei jeder zweiten Beobachtung zu erwarten ist, was zu weniger Turning-Points führt, als es unter der Nullhypothese der Unabhängigkeit zu erwarten wäre (nämlich bei 2 von 3 Beobachtungen; vgl. Kap. 2.6), sodass ihre Verwerfung in einem solchen Fall nachvollziehbar erscheint. Um das Verhalten des TP-Tests weiter zu verstehen, wurden zwei Zeitreihen unter Alternativen, in denen der TP-Test die Nullhypothese nicht verwerfen kann, in Abbildung 3.54 dargestellt.

Im Fall, dass $\rho_1 = \rho_2$ = -0.7 gilt, ist eine ähnliche Systematik wie in Abbildung 3.3 zu beobachten. So sorgt der Parameter ρ_1 zusätzlich zu den in Abbildung 3.53 erzeugten Turning-Points



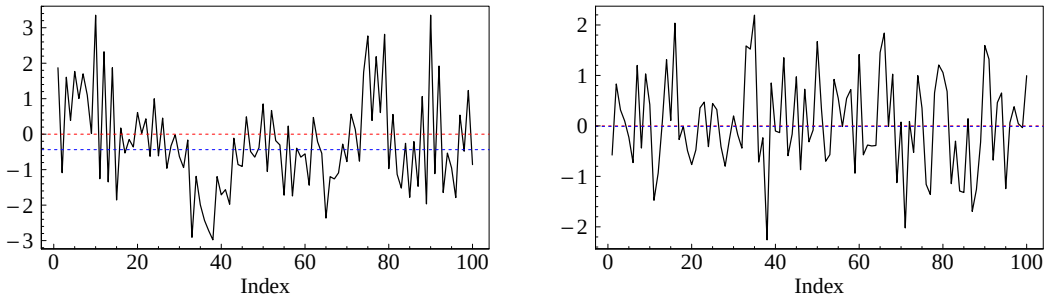
(a) Zeitreihe mit $\rho_1 = \rho_2$ = -0.7

(b) Zeitreihe mit $\rho_1 = \rho_2 = 0.3$

Abbildung 3.54: Zeitreihen eines AR(2)-Prozesses bei einer Beobachtungszahl von $N = 100$, mit unterschiedlichen, Parameterwerten für $\rho_1 = \rho_2$, mit Nulllinien (rot) und empirischem Median (blau)

dafür, dass die Wahrscheinlichkeit eines alternierenden Verhaltens von aufeinanderfolgenden Beobachtungen der Zeitreihe vergrößert wird. Somit gleichen sich der negative Wert von ρ_2 − durch den weniger Turning-Points in der Zeitreihe erzeugt werden, als unter der Unabhängigkeit zu erwarten wären − und der Wert von ρ_1 − der alternierendes Verhalten begünstigst − aus. Dadurch entspricht die Anzahl der Turning-Points in diesem Fall einer, die auch unter der Nullhypothese zu erwarten wäre. Auf ähnliche Weise wird durch einen positiven Wert des Parameters ρ_2 alternierendes Verhalten aufeinanderfolgender Beobachtungen begünstigt, was in Abbildung 3.55 verdeutlicht wird. Die größere Anzahl von Turning-Points, die durch diesen Parameter erzeugt wird, kann dann durch einen positiven Wert ähnlicher Intensität von ρ_1 ausgeglichen werden, durch den einem alternierenden Verhalten entgegenwirkt wird.

Insgesamt zeigt sich unter Normalbedingungen, dass der LB-Test mit Abstand die beste Wahl darstellt, falls es von Interesse ist, sowohl Korrelationen 1. als auch 2. Ordnung im Rahmen eines AR(2)-Prozesses zu erkennen. So existieren bei sämtlichen anderen Verfahren Alternativen, bei denen eine Verwerfung der Nullhypothese trotz deutlicher Korrelationsstrukturen nicht möglich ist. Im Hinblick auf den DW-, Runs-, BDS- und den VNRR-Test zeigt sich, dass diese Verfahren bei Stichprobenumfängen von bis zu $N = 500$ nicht in der Lage sind, Abhängigkeitsstrukturen 2. Ordnung zu erkennen, falls keine Korrelation 1. Ordnung vorliegt. Der TP-Test nimmt hier eine Sonderposition ein, da er Alternativen, bei denen $\rho_1 \approx \rho_2$ gilt, nicht verwerfen kann. Die K-VZ-Tests schneiden in diesem Szenario wieder am schlechtesten ab und es gelingt ihnen nur sehr wenige extreme Alternativen zur Unabhängigkeit im Stationaritätsdreieck zu detektieren. Allerdings fällt auf, dass sie, anders als die meisten anderen Verfahren, in der Lage sind, Prozesse mit extremen Werten des Parameters ρ_2 als Abweichungen von der Unabhängigkeit zu erkennen. Dies legt die Vermutung nahe, dass sie eine breitere Alternative umfassen als viele der gängigen Verfahren.



(a) Zeitreihe mit $\rho_1 = 0$ und $\rho_2 = 0.7$

(b) Zeitreihe mit $\rho_1 = \rho_2 = 0$

Abbildung 3.55: Zeitreihen zweier AR(2)-Prozesse mit $N = 100$ Beobachtungen, mit unterschiedlichen Parametern für ρ_1 und ρ_2, mit Nulllinie (rot) und empirischem Median (blau)

Für die weiteren Untersuchungen ist es − wie in Kapitel 3.1 durchgeführt − von Interesse, wie sich Abweichungen von den Voraussetzungen der verschiedenen Testverfahren auf deren Trennschärfe auswirken. So stellt sich im Besonderen die Frage, ob diese Effekte von dem zugrunde liegenden Prozess abhängen oder ähnlich ausfallen wie im AR(1)-Fall. Exemplarisch sollen dabei die Auswirkungen von Abweichungen von den Verteilungsannahmen der Innovationen, Kontaminationen und einem Trend in der Zeitreihe betrachtet werden.

3.2.1 Abweichungen von den Verteilungsannahmen der Innovationen

Im Folgenden wird untersucht, wie sich die Tests im AR(2)-Fall bei Abweichungen von der Normalität der Innovationen verhalten. In Anlehnung an die entsprechenden Untersuchungen im AR(1)-Fall wurden dafür zum einen Prozesse mit Cauchy-verteilten und zum anderen mit Laplace-verteilten Innovationen betrachtet. Die Simulationsergebnisse sind in den Abbildungen 3.56 bis 3.59 für Beobachtungszahlen $N = 50$ und $N = 500$ dargestellt.

Beim Betrachten der Testergebnisse fällt zunächst auf, dass die nichtparametrischen Verfahren, ähnlich wie im AR(1)-Fall, im Allgemeinen von den Abweichungen der Verteilungsannahmen profitieren. So verbessern sich ihre Trennschärfen für beide Verteilungen und Stichprobenumfänge im Vergleich zu derselben Situation unter Normalbedingungen. Dabei scheinen die Verfahren im Fall der Cauchy-Verteilung nochmals eine bessere Trennschärfe aufzuweisen als bei der Laplace-Verteilung. Am deutlichsten fallen die durch die Verteilung bedingten Unterschiede beim BDS-Test und bei den K-VZ-Tests aus. Während der BDS-Test − abgesehen von den K-VZ-Tests − unter Normalbedingungen am schlechtesten abgeschnitten hat, weist er bei der Cauchy-Verteilung bereits bei $N = 50$ eine der besten Trennschärfen unter allen Tests auf. Bei einem Stichprobenumfang von $N = 500$ schafft der Tests es dann bereits, minimale Abweichungen des Parameters ρ_1 von 0 äußerst zuverlässig zu detektieren. Den K-VZ-Tests gelingt es, die Nullhypothese durch die Verteilungsabweichung dabei unter deutlich mehr Alternativen zu verwerfen und speziell der 5-VZ-Test kann bei einer Beobachtungszahl von $N = 500$ auch negative Ausprägungen von ρ_2 relativ zielsicher erkennen. Trotzdem gibt es noch diverse Alternativen mit starken Abhängigkeitsstrukturen, in denen die Verfahren keine Abweichung von der Nullhypothese feststellen können.

Im Hinblick auf den Runs-Test und den VNRR-Test fällt auf, dass die Verteilungsänderung die Tendenz der Verfahren, Abweichungen des Parameters ρ_1 von 0 bei größer werdenden positiven Werten von ρ_2 besser zu erkennen, verstärkt. Dies macht sich in den entsprechenden Abbildungen durch einen „spitzer“ zulaufenden Annahmebereich im Stationaritätsdreieck bemerkbar. Der TP-Test wird − genau wie im AR(1)-Fall − mit denselben Überlegungen nicht durch die Verteilungsänderung beeinflusst.

Bei den parametrischen Verfahren zeigt sich ebenfalls ein ähnliches Verhalten wie im AR(1)-Fall. So wird die Nullhypothese innerhalb des „unsicheren“ Bereichs, in dem sie unter Normalbedingungen in ca. 50 % der Fälle verworfen wird, deutlich öfter abgelehnt. Das heißt, die

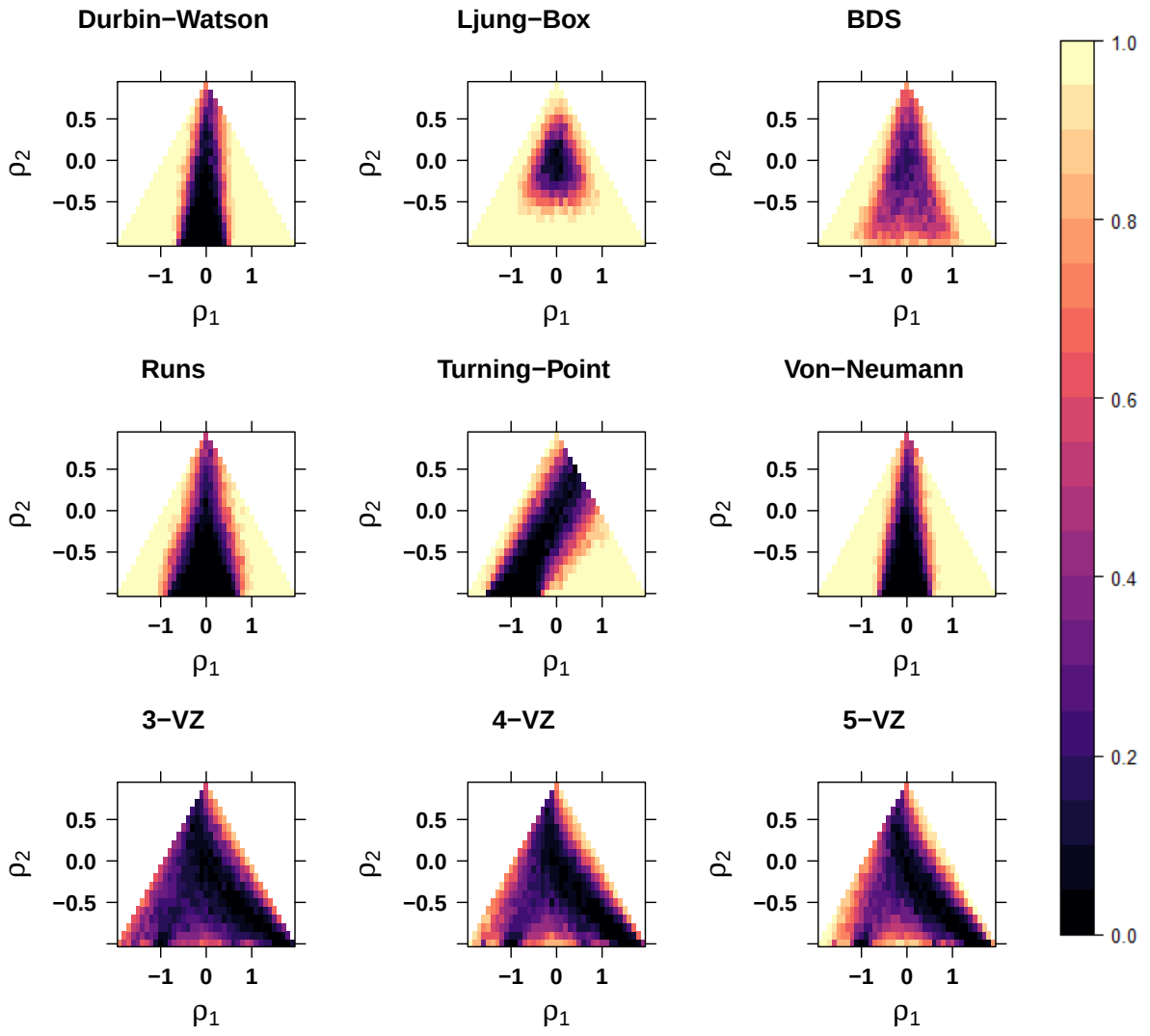


Abbildung 3.56: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei Laplace-verteilten Innovationen und $N = 50$ Beobachtungen

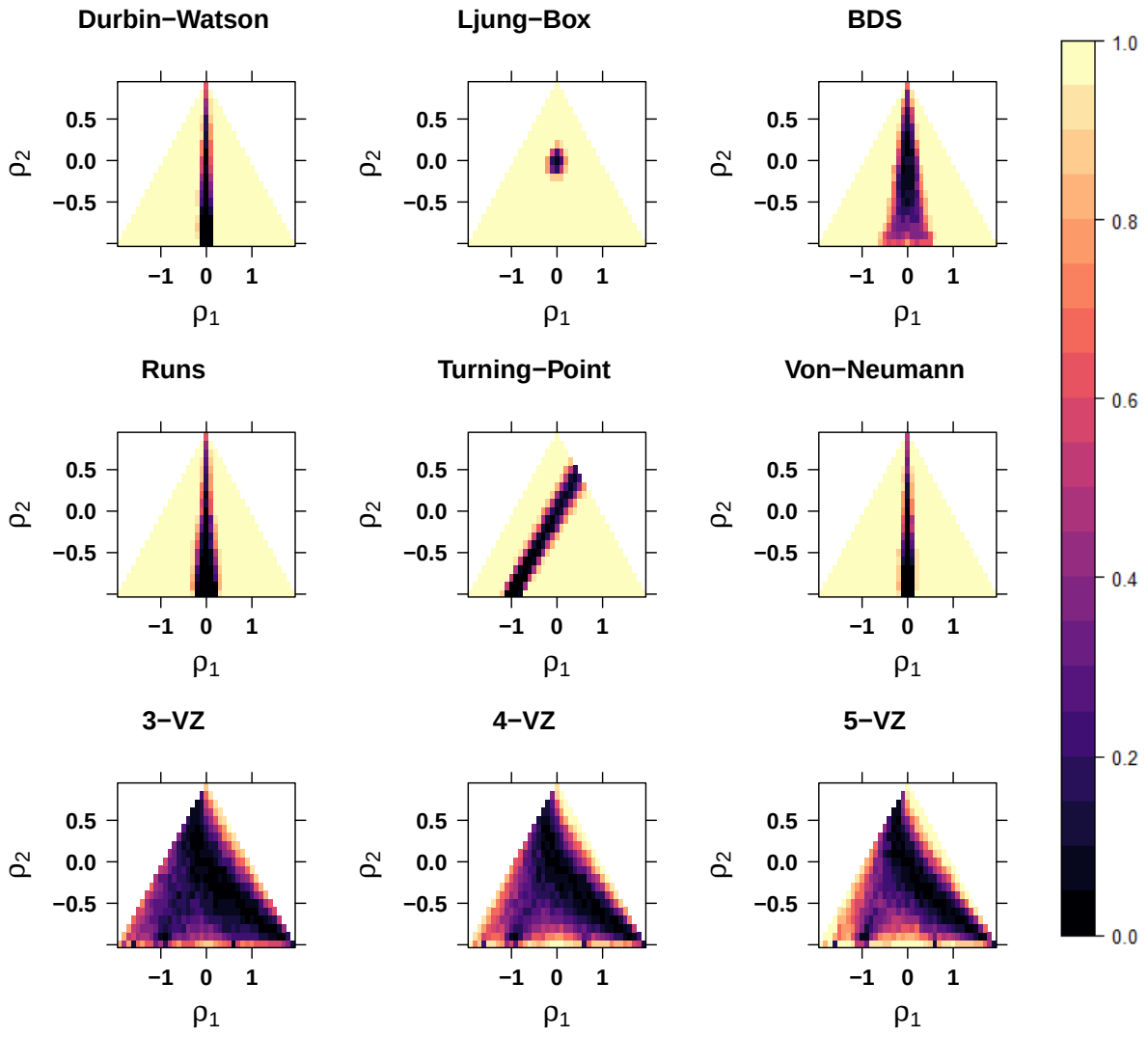


Abbildung 3.57: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei Laplace-verteilten Innovationen und $N = 500$ Beobachtungen

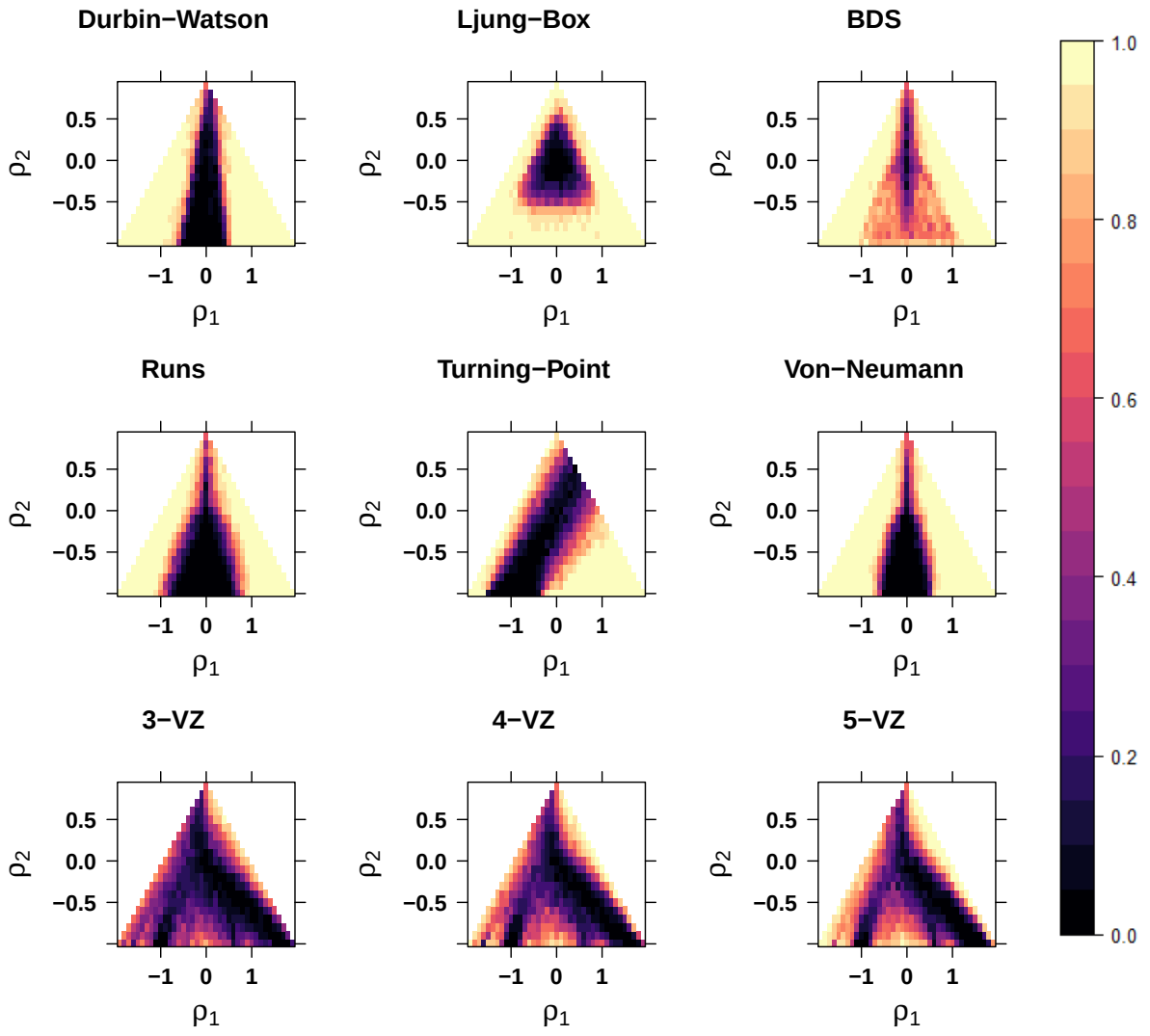


Abbildung 3.58: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei Cauchy-verteilten Innovationen und $N = 50$ Beobachtungen

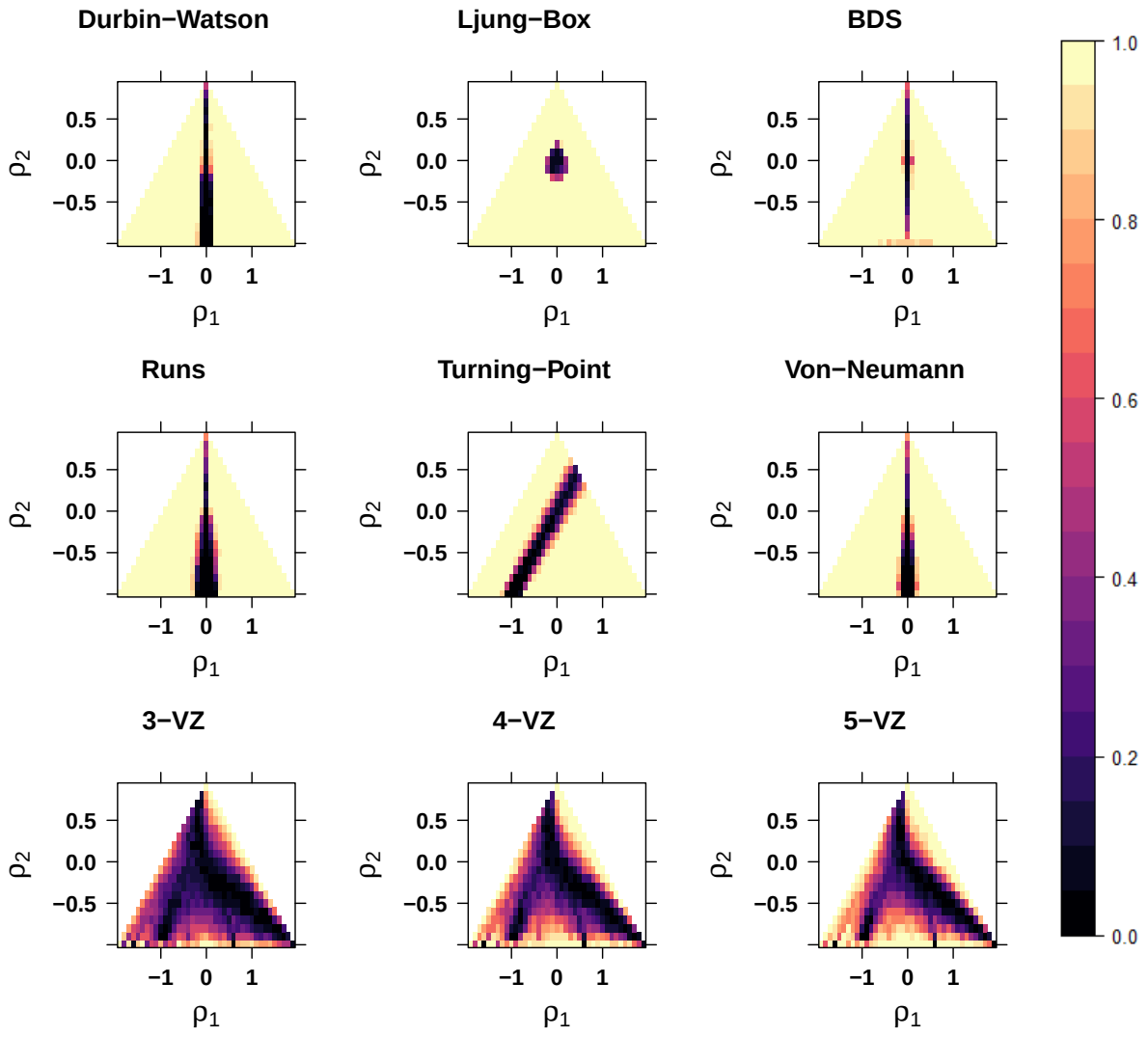


Abbildung 3.59: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei Cauchy-verteilten Innovationen und $N = 500$ Beobachtungen

Nullhypothese wird von diesen Tests eindeutiger verworfen bzw. beibehalten, was sich durch schärfere Ränder des Annahmebereiches bemerkbar macht. Dieses Verhalten lässt sich erneut durch die Abweichungen der empirischen Autokorrelationskoeffizienten von der Normalität erklären. Der LB-Test reagiert hier stärker auf die Verteilungsänderung, was sich durch die Menge der betrachteten empirischen Autokorrelationskoeffizienten erklärt.

Um zu verstehen, warum der BDS-Test in diesem Szenario auch in kleineren Stichproben deutlich mehr Trennschärfe aufweist als die anderen Tests, ist es sinnvoll, sich die Berechnung seiner Teststatistik vor Augen zu führen. So lehnt der Test die Nullhypothese genau dann ab, wenn der Anteil der Supremumsnormen zweier m-Historien, die größer als die Hälfte der Standartabweichung der Simulationsdaten sind, größer oder kleiner ist, als es unter der Nullhypothese der Unabhängigkeit zu erwarten wäre. Da extreme Innovationen und die darauf folgenden Ausschwingphasen, wie sie bei den betrachteten Verteilung mit schweren Rändern zu erwarten sind, dazu führen, dass 2-Historien tendenziell größere Abstände aufweisen, ist die bessere Trennschärfe unter solchen Alternativen nachvollziehbar. Dabei ist es auch einleuchtend, dass die Cauchy-Verteilung, bei der anteilig weniger extreme Innovationen vorkommen, dazu führt, dass es mehr von diesen 2-Historien gibt, da insbesondere ϵ bei der Laplace-Verteilung als Resultat ihrer größeren, zu erwartenden, empirischen Standardabweichung deutlich größer ausfällt.

In AR(2)-Prozessen ist also im Fall, dass die Innovationen aus Verteilungen mit schweren Rändern generiert werden, ein ähnliches Verhalten wie in AR(1)-Prozessen (s. Kap. 3.1) beobachtbar. So neigen die parametrischen Verfahren dazu, die Nullhypothese unter Alternativen mit geringen Abhängigkeiten ein wenig häufiger zu verwerfen. Die nichtparametrischen Verfahren können hingegen von der Verteilungsänderung profitieren. Dieses Verhalten ist beim BDS-Test am deutlichsten ausgeprägt, da er in diesem Szenario bereits ab einem Stichprobenumfang von $N = 50$ die beste Trennschärfe unter allen nichtparametrischen Verfahren aufweist und sogar dem DW-Test deutlich überlegen ist. Der TP-Test nimmt erneut eine Sonderstellung ein und wird durch die anders verteilten Innovationen kaum merklich beeinflusst. Aus diesen Beobachtungen kann geschlussfolgert werden, dass sich das Auftreten extremerer Innovationen, als sie bei der Normalverteilung erwartet werden, unabhängig vom konkreten Prozess in ähnlicher Weise auf die Trennschärfen der Testverfahren auswirkt. Dasselbe Verhalten kann somit auch für das Vorhandensein von innovativen Ausreißern angenommen werden.

3.2.2 Kontaminationen

Da es weiterhin interessant ist, welchen Effekt Kontaminationen bei einem autoregressiven Prozess 2. Ordnung auf die betrachteten Testverfahren haben, werden sie im Folgenden auf simulierte Zeitreihen mit 5 % Kontaminationsanteil der Intensität 10 angewendet (vgl. Kap. 3.1). Die entsprechenden Ergebnisse sind in den Abbildungen 3.60 bis 3.63 dargestellt.

Im Vergleich zu den Simulationsergebnissen unter Normalbedingungen fällt auf, dass die Trennschärfen sämtlicher Verfahren − genau wie im AR(1)-Fall − unter den Kontaminatio-

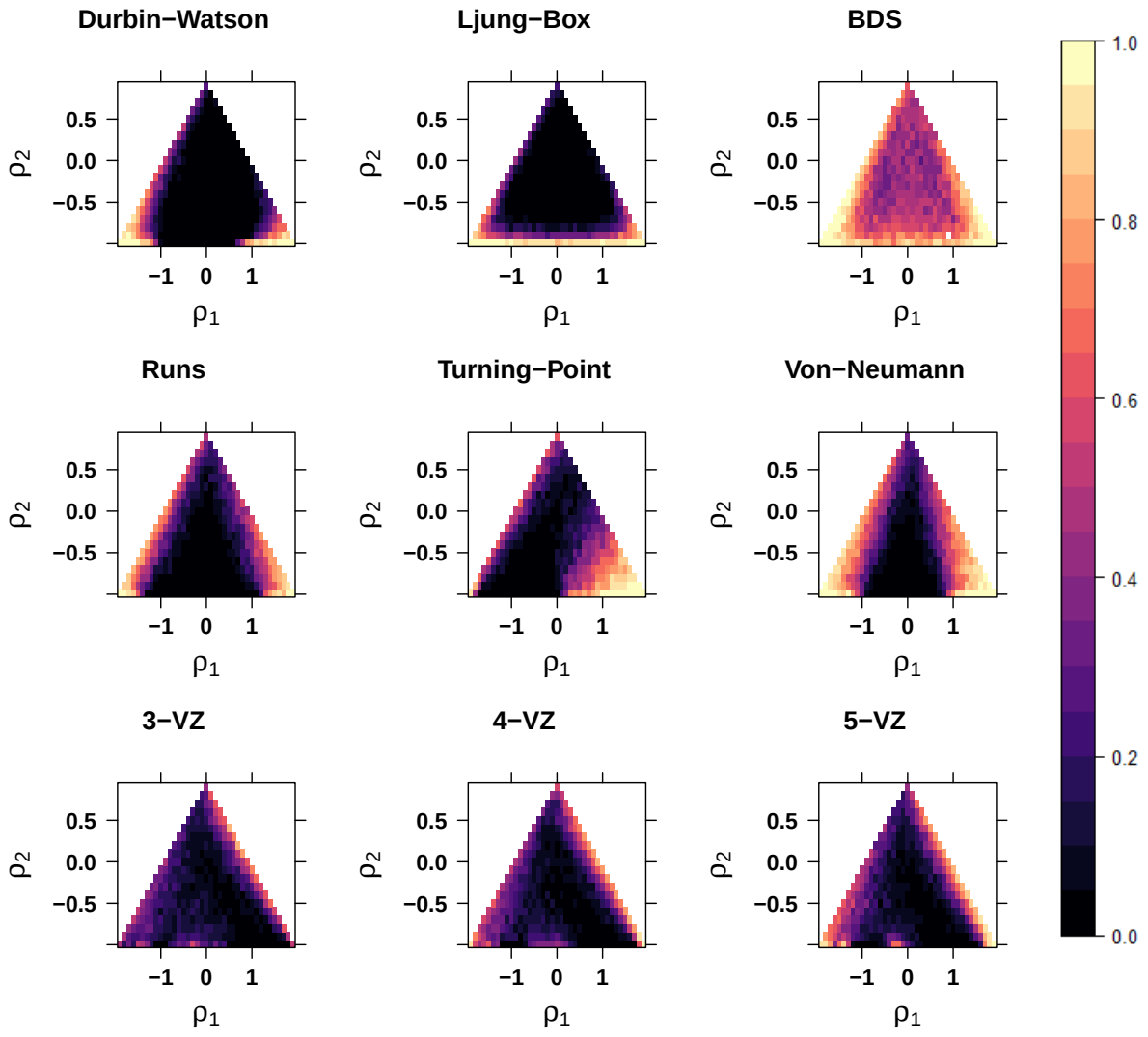


Abbildung 3.60: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei 5 % Kontaminationen der Intensität 10 und $N = 20$ Beobachtungen

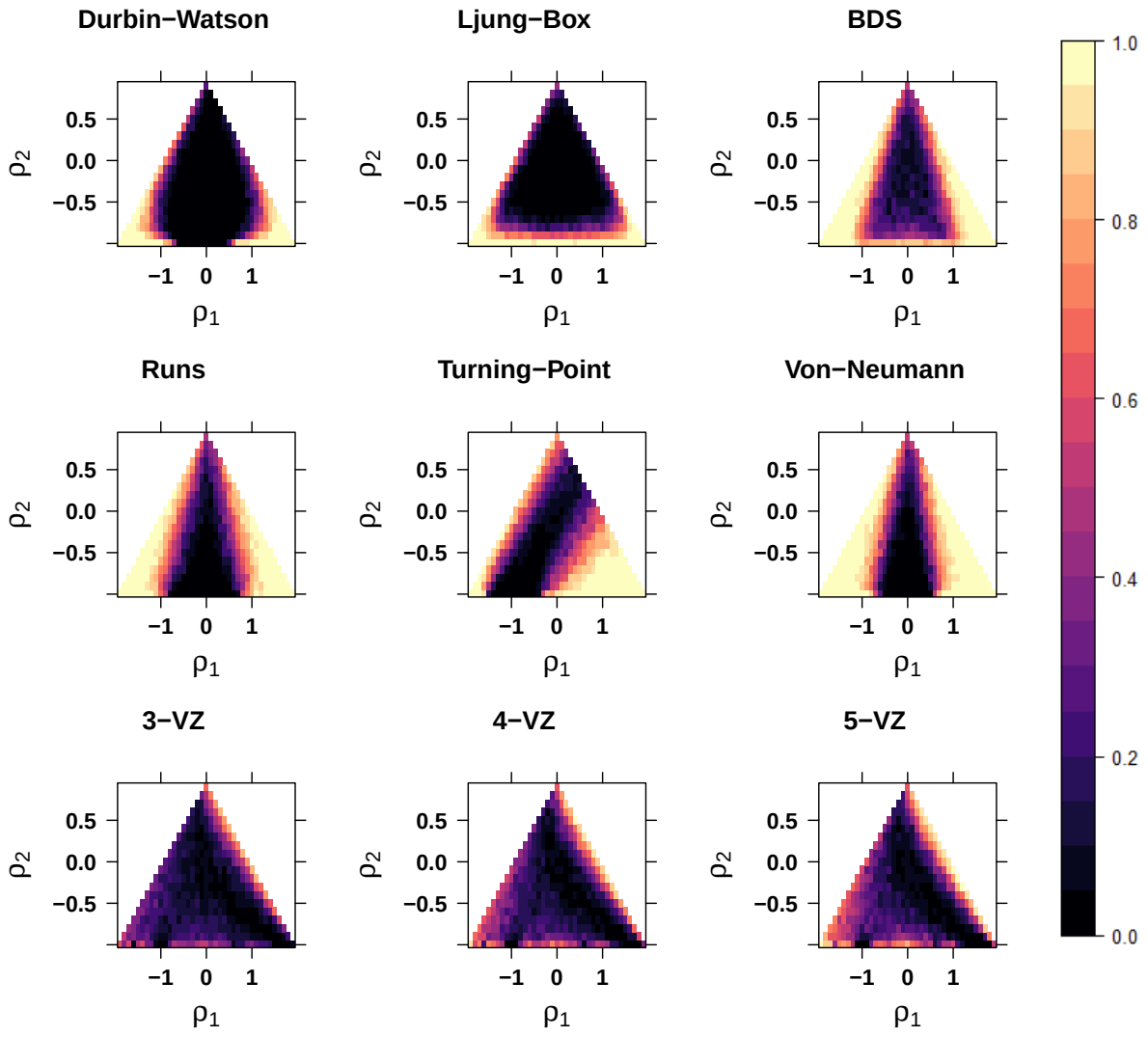


Abbildung 3.61: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei 5 % Kontaminationen der Intensität 10 und $N = 50$ Beobachtungen

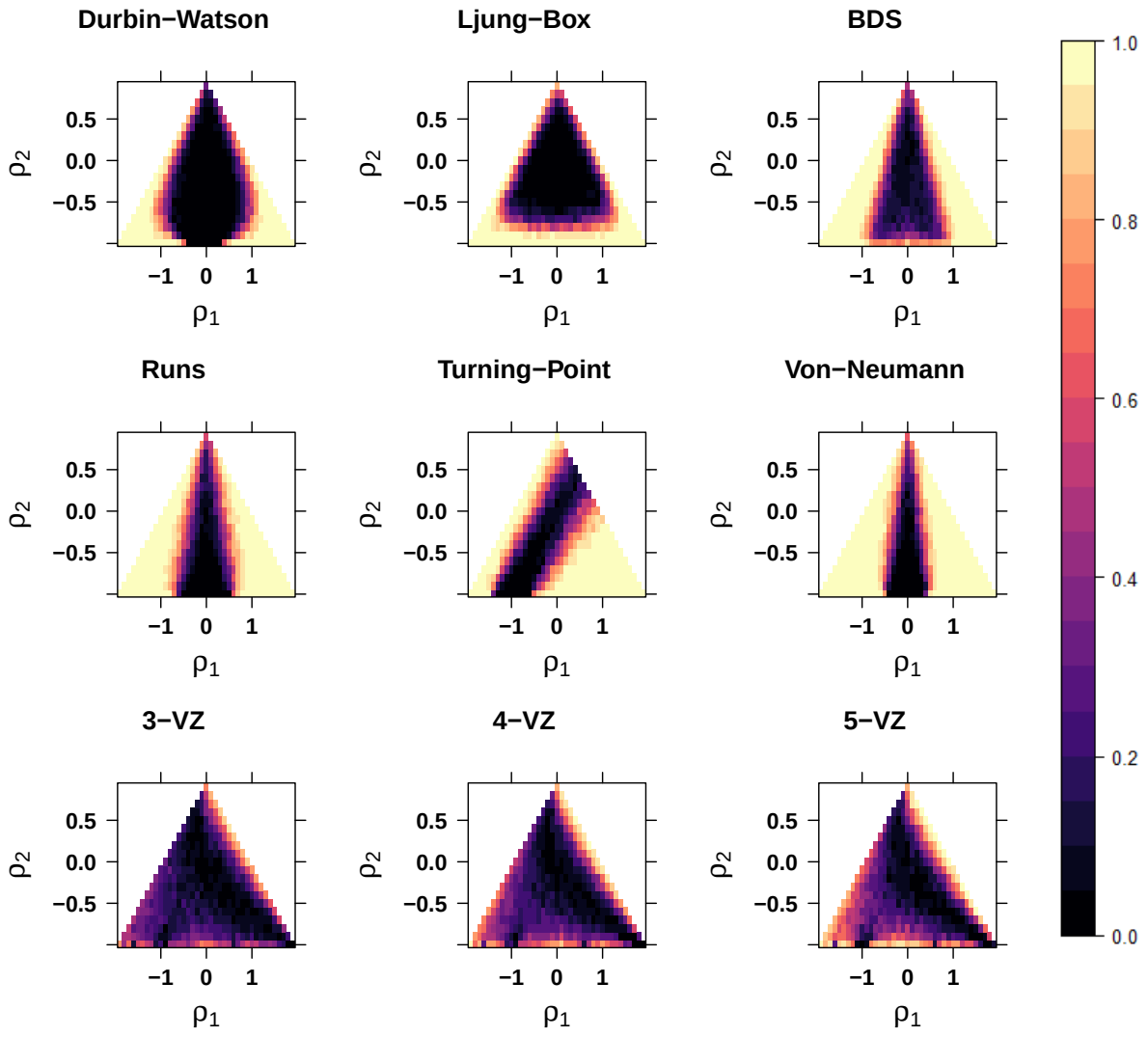


Abbildung 3.62: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei 5 % Kontaminationen der Intensität 10 und $N = 100$ Beobachtungen

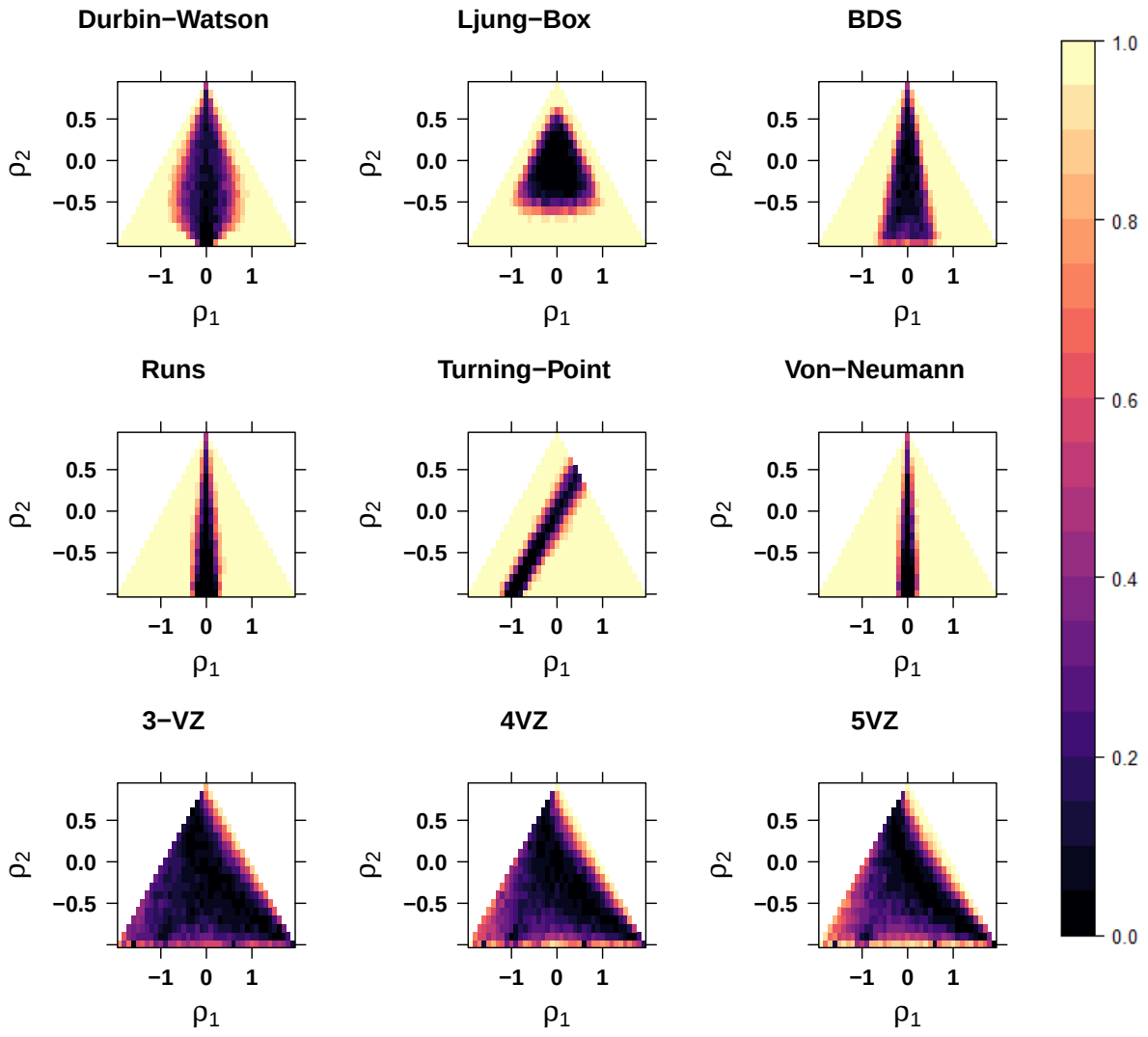


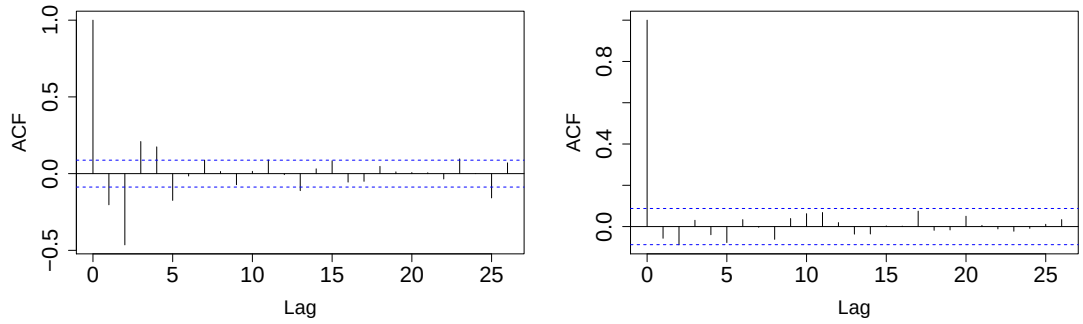
Abbildung 3.63: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, bei 5 % Kontaminationen der Intensität 10 und $N = 500$ Beobachtungen

nen leiden. Während die Unterschiede der nichtparametrischen Verfahren bei sämtlichen Beobachtungszahlen sehr gering ausfallen, weisen die parametrischen Tests deutlich sichtbare Verschlechterungen auf. So gelingt es den Verfahren bei niedrigen Stichprobenumfängen kaum, die Nullhypothese zu verwerfen. Auch bei einer großen Beobachtungszahl von $N = 500$ sind die Annahmebereiche noch sehr ungenau und die Trennschärfen entsprechen in etwa solchen, die unter Normalbedingungen bei einer Beobachtungszahl von $20 - 50$ zu erwarten wären. Einen Sonderfall unter den nichtparametrischen Verfahren stellt der BDS-Test dar, der die Nullhypothese vor allem bei kleinen Beobachtungszahlen deutlich häufiger verwirft, als unter Normalbedingungen.

Um das Verhalten der parametrischen Verfahren nachvollziehen zu können, wurden die Korrelogramme einer Zeitreihe mit und ohne Kontaminationen verglichen. Diese wurden für eine Zeitreihe mit $N = 100$ Beobachtungen beispielhaft in Abbildung 3.64 für Autokorrelationskoeffizienten von $\rho_1 =$ -0.3 und $\rho_2 =$ -0.5 dargestellt. Die Parameter wurden dabei aus einem Bereich gewählt, in dem der DW-Test und der LB-Test die Unabhängigkeit unter Normalbedingungen klar ablehnen können, während sie bei Kontaminationen nicht dazu in der Lage sind.

Anhand dieses Beispiels lässt sich − wie bereits im Fall von AR(1)-Prozessen − erkennen, dass Kontaminationen in der Zeitreihe vorhandene Korrelationen verschleiern können. So sind in der Ausgangszeitreihe 6 Korrelationskoeffizienten bis zu einem Lag von 10 teilweise stark signifikant von 0 verschieden, während in derselben Zeitreihe mit nachträglicher Kontaminierung keiner der empirische Autokorrelationskoeffizienten noch signifikant erhöht ist. Die Tatsache, dass Kontaminationen die Trennschärfe derjenigen Tests, die auf den Autokorrelationskoeffizienten basieren, deutlich verschlechtert, ist somit nachvollziehbar.

Auch die Tatsache, dass die nichtparametrischen Verfahren kaum durch die Kontaminationen beeinflusst werden, ist mit ähnlichen Überlegungen wie in Kapitel 3.1 zu erklären. So wirken sie sich nur in geringer Weise auf das sequenzielle Schema der Zeitreihe aus und lassen somit



(a) Korrelogramm ohne Kontaminationen (b) Korrelogramm mit Kontaminationen

Abbildung 3.64: Korrelogramme einer Zeitreihe mit $N = 500$ Beobachtungen eines AR(2)-Prozesses mit (a) und ohne Kontaminationen (b) zu den Parametern $\rho_1 =$ -0.3 und $\rho_2 =$ -0.5

die Anzahl der Vorzeichenwechsel, Runs, Turning-Points sowie die Ränge der Beobachtungen weitestgehend unbeeinflusst.

Insgesamt lässt sich erneut − wie im AR(1)-Fall − feststellen, dass vor allem parametrische Verfahren unter Kontaminationen in der Zeitreihe leiden, sodass es sich in solchen Situationen anbietet, auf nichtparametrische Verfahren zurückzugreifen. Auffällig ist, dass die Trennschärfe des BDS-Tests ebenfalls stark von den Kontaminationen beeinflusst zu werden scheint. Zu erklären ist das damit, dass der konkrete Abstand zwischen zwei Beobachtung, anders als bei den übrigen nichtparametrischen Verfahren, für diesen Test eine Rolle spielt. In diesem Szenario stellt der VNRR-Test in Hinblick auf die Menge der Alternativen, in denen er eine Ablehnung der Nullhypothese erreichen kann, eine sinnvolle Wahl dar. Jedoch ist mit Ausnahme des LB-Test weiterhin kein Verfahren erfolgreich bei der Ablehnung von Autokorrelationen 2. Ordnung. Damit ist der LB-Test, falls mit solchen Abhängigkeitsstrukturen zu rechnen ist, trotz der gravierenden Verschlechterung im Vergleich zu den Normalbedingungen, in diesem Fall am geeignetsten. Aus diesen Erkenntnissen lässt sich außerdem schließen, dass Kontaminationen in einer Zeitreihe, unabhängig von der konkreten Art des zugrunde liegenden Prozesses, ähnliche Auswirkungen auf die Trennschärfen der betrachteten Tests hat.

3.2.3 Abweichungen von der Varianzhomogenität

Weiterhin ist interessant, wie sich die Testverfahren im AR(2)-Fall verhalten, falls Abweichungen von der Homogenität der Varianzen in den Zeitreihen vorliegen. In Anlehnung an dieselben Untersuchen in Kapitel 3.1 werden die Testverfahren deshalb im Folgenden auf simulierte Zeitreihen mit einer wachsenden Varianz von 1 auf $N/10$ am Ende der Zeitreihe angewendet. Die Simulationsergebnisse sind in den Abbildungen 3.65 bis 3.68 dargestellt.

Dabei fällt auf, dass diejenigen Tests, die lediglich auf dem sequenziellen Schema der Beobachtungen beruhen, im Vergleich zu den Ergebnissen unter Normalbedingungen keine deutliche Veränderung zeigen. In kleinen Stichproben von $N \leq 50$ sind dabei leichte Verschlechterungen der Trennschärfe erkennbar, für größere Beobachtungszahlen sind jedoch kaum noch Unterschiede zwischen den beiden Szenarien festzustellen. Auch der VNRR-Test zeigt ein ähnliches Verhalten. Da die Größe der Beobachtungen, die im Wesentlichen durch die Varianzveränderungen beeinflusst wird, für diese Verfahren keine Rolle spielt, ist dieses Verhalten verständlich.

Wie bereits in Kapitel 3.1 erörtert, sind der BDS- und der LB-Test die einzigen, die eine wachsende Varianz als Abweichung von der Zufälligkeit erkennen können. Dabei gelingt es dem BDS-Test bei einem Stichprobenumfang von $N = 500$, die Nullhypothese auf dem gesamten Spektrum eindeutig abzulehnen. Der LB-Test hingegen bekommt lediglich deutliche Schwierigkeiten, das Niveau unter der Nullhypothese einzuhalten, erreicht aber zu den betrachteten Stichprobenumfängen nie eine eindeutige Ablehnung. Die Begründungen für diese Verhaltensweisen wurden im Rahmen von AR(1)-Prozessen bereits ausgiebig diskutiert (s. Kap 3.1.4).

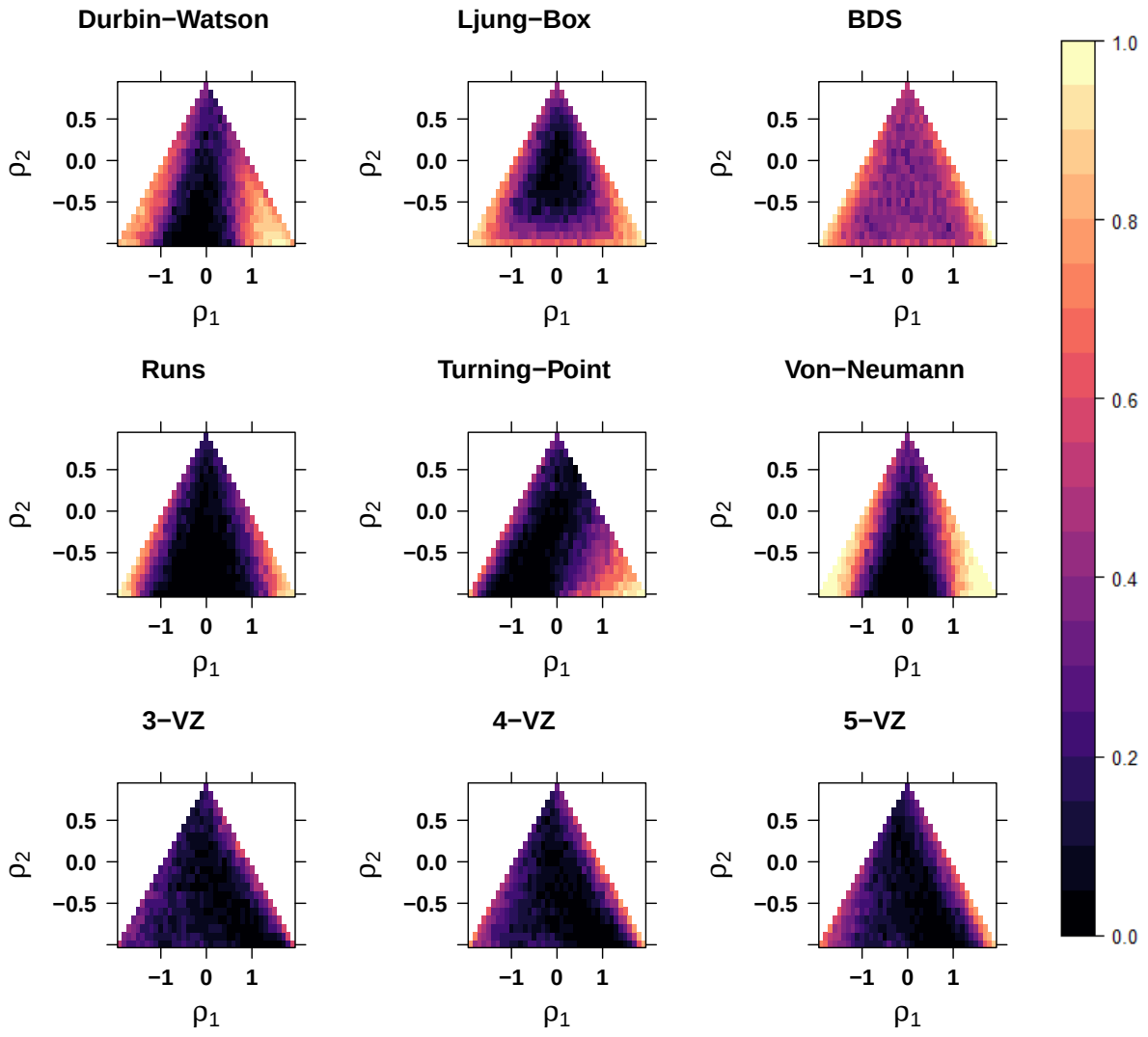


Abbildung 3.65: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2 für $N = 20$ Beobachtungen und einer wachsenden Varianz

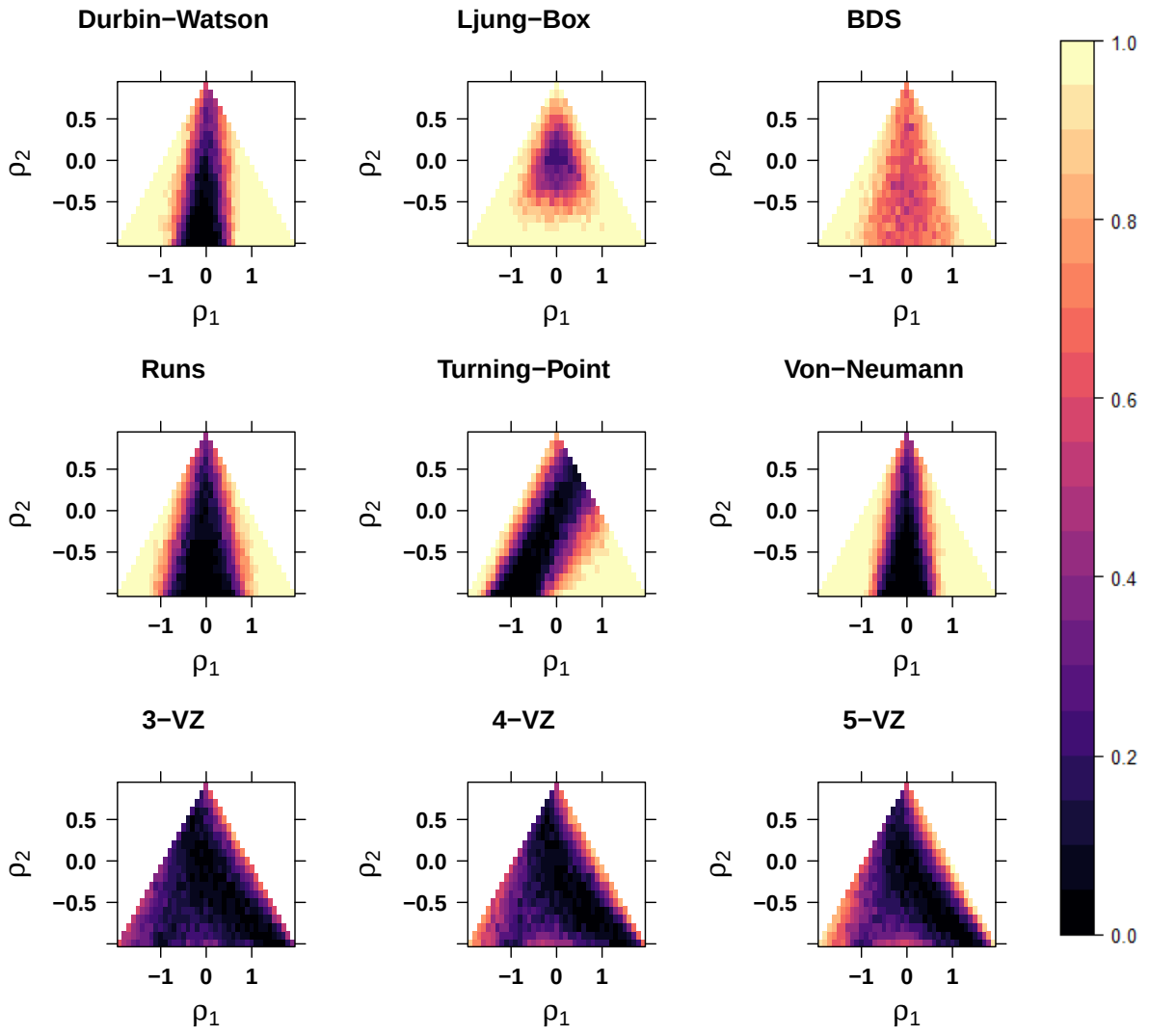


Abbildung 3.66: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2 für $N = 50$ Beobachtungen und einer wachsenden Varianz

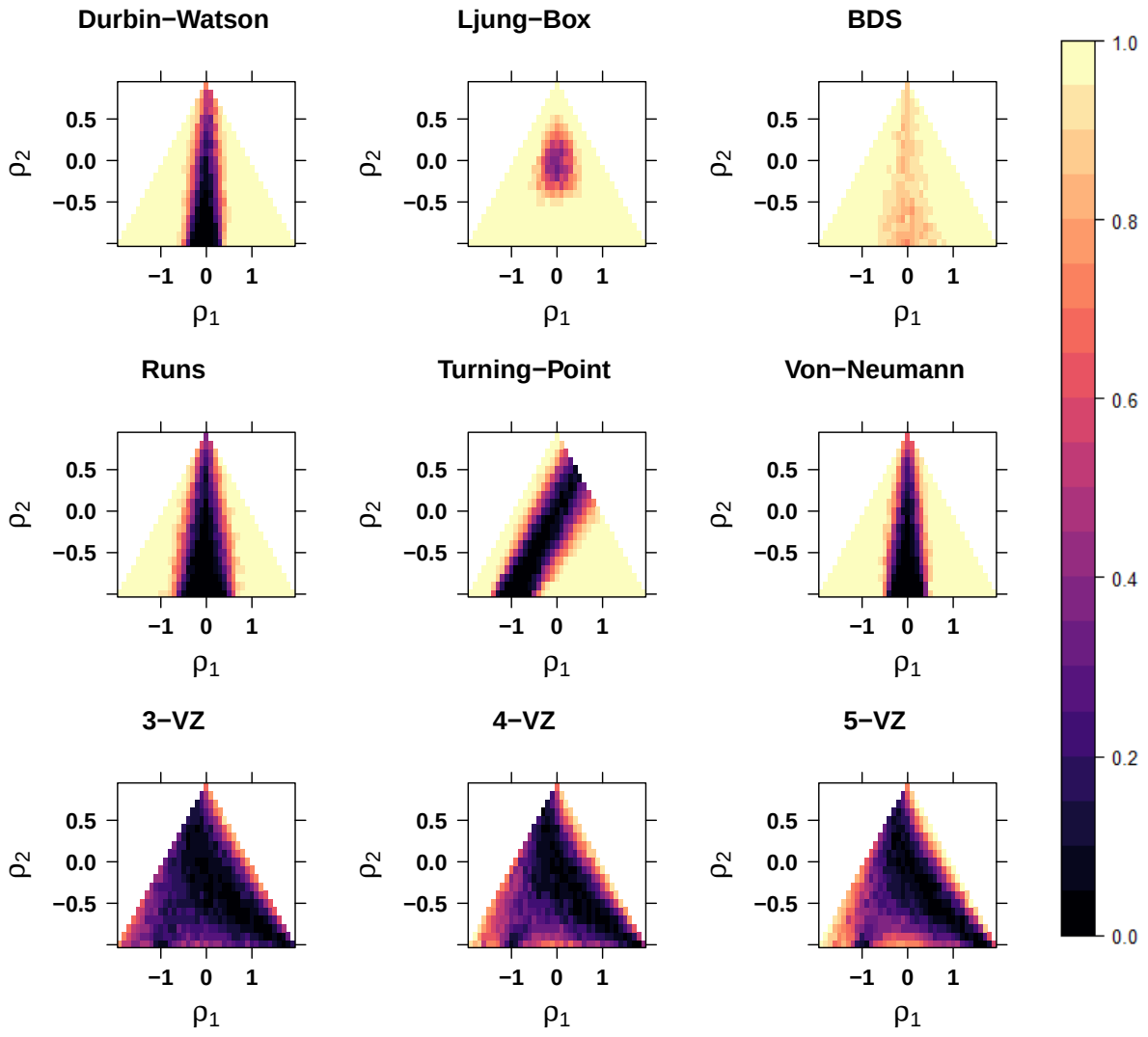


Abbildung 3.67: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2 für $N = 100$ Beobachtungen und einer wachsenden Varianz

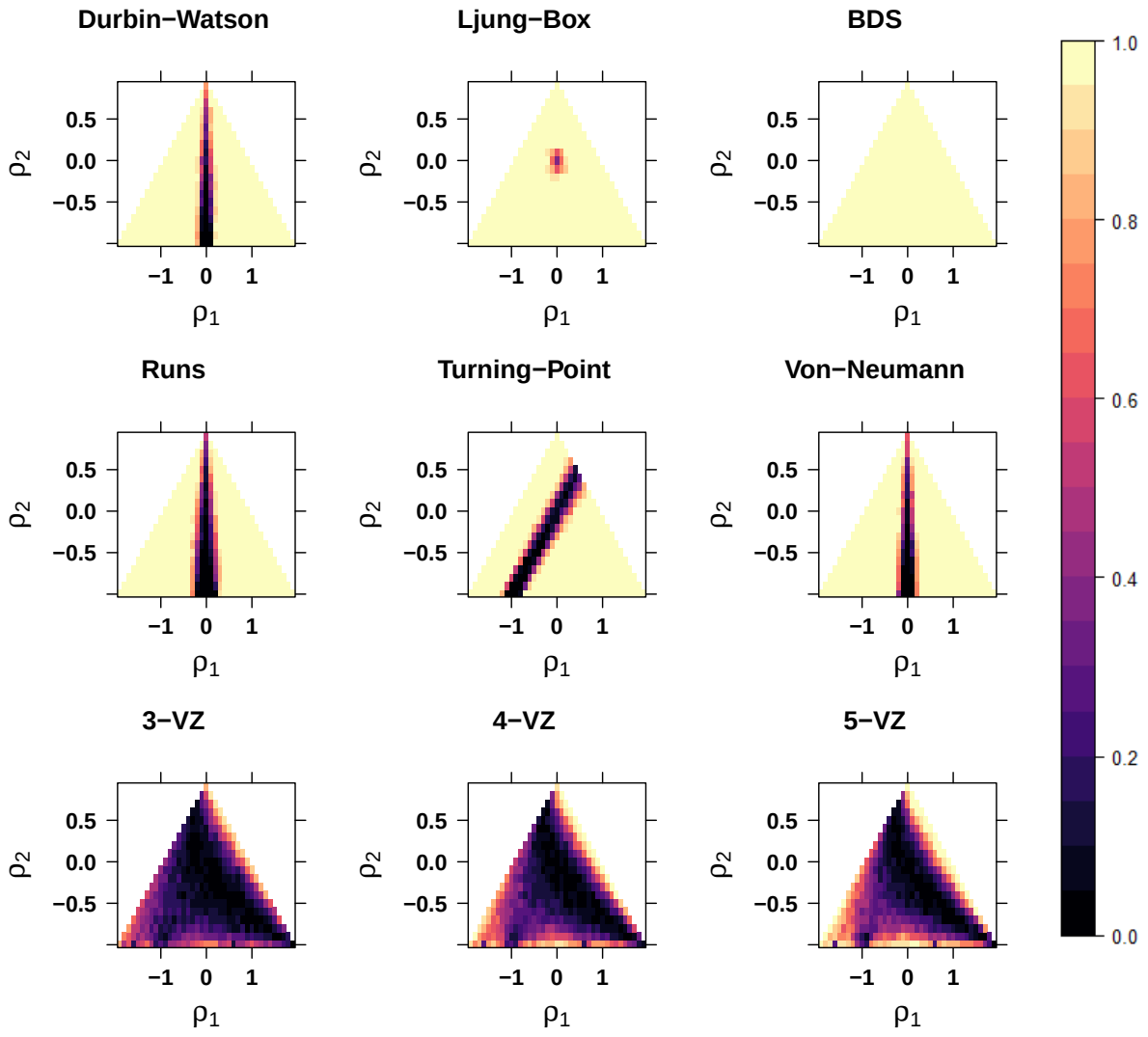


Abbildung 3.68: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2 für $N = 500$ Beobachtungen und einer wachsenden Varianz

Diese Simulationsergebnisse legen also nahe, dass eine wachsende Varianz stets die gleichen Auswirkungen auf die Testverfahren hat. So gelingt es lediglich dem BDS- und dem LB-Test, Varianzänderungen in der Zeitreihe zu detektieren. Die übrigen Verfahren reagieren robust.

3.2.4 Trend

Weiter ist von Interesse, wie die Testverfahren im AR(2)-Fall reagieren, wenn eine Niveauänderung in der Zeitreihe vorhanden ist. Dazu soll hier exemplarisch untersucht werden, wie sich ein Trend der Steigung $2/N$ in den einzelnen Zeitreihen auf die Trennschärfen der verschiedenen Testverfahren auswirkt (vgl. Kap. 3.1). Die Ergebnisse dieser Simulation sind für unterschiedliche Stichprobenumfänge in den Abbildungen 3.69 bis 3.72 dargestellt.

Bei der Betrachtung der Trennschärfen werden wieder deutliche Parallelen zu den Auswirkungen eines Trends bei einem AR(1)-Prozess sichtbar. So bleiben die Trennschärfen des DW-Tests und des TP-Tests von dem Trend weitestgehend unbeeinflusst, wohingegen der Runs-Test, der VNRR-Test und der BDS-Test Verschiebungen in den Bereich negativer Korrelationen zeigen. Auffällig ist, dass nicht nur eine Wechselwirkung zwischen ρ_1 und dem Trend vorhanden zu sein scheint, sondern auch der konkrete Wert von ρ_2 eine Rolle für die Testentscheidungen spielt. So ist die Verschiebung für sehr kleine Werte von ρ_2 schwächer ausgeprägt als für Werte zwischen -0.5 und 0, bei denen die Verschiebung am stärksten ist. Anhand der Abbildungen macht sich diese Systematik durch die nach links gekrümmten Formen der Annahmebereiche bemerkbar.

Der LB-Test und die K-VZ-Tests nehmen in diesem Szenario, ähnlich wie im AR(1)-Fall, eine Sonderstellung ein. So sind sie in der Lage, den Trend als eine Abhängigkeitsstruktur über die Zeit in der Zeitreihe zu erkennen und die Nullhypothese für eine hinreichend große Stichprobe abzulehnen. Dabei gelingt es dem LB-Test bereits ab einem Stichprobenumfang von $N = 100$, die Nullhypothese für alle Werte von ρ_1 und ρ_2 mit großer Zuverlässigkeit zu verwerfen. Bei den K-VZ-Tests ist dies bei derselben Beobachtungszahl lediglich für Kombinationen von betragsmäßig kleinen Werten in der Mitte des Stationaritätsdreiecks möglich. Ab einem Stichprobenumfang von $N = 500$ erreichen sie dann − abgesehen von Bereichen mit stark negativen Werten von ρ_2 − eine ähnliche Trennschärfe wie der LB-Test.

Die Veränderungen der Trennschärfen sind dabei mit ähnlichen Überlegungen zu erklären, wie bei AR(1)-Prozessen. So reagiert der TP-Test wieder robust, da die Anzahl der Turning-Points durch einen geringen Trend kaum beeinflusst wird. Die Verschiebungen in den Bereich negativer Korrelationen der anderen nichtparametrischen Verfahren sind erneut mit Wechselwirkungen des Trends mit den autoregressiven Parametern zu erklären, die dazu führen, dass die Teststatistiken unter Alternativen Werte annehmen, die unter der Nullhypothese zu erwarten wären. Die Fähigkeit des LB-Tests und der K-VZ-Tests, den Trend zu detektieren, sind dabei auf die breiteren Alternativen, die sie umfassen, zurückzuführen. So werden ihre Teststatistiken nicht im Wesentlichen durch eine Autokorrelation 1. Ordnung beeinflusst, dessen Schätzer durch den Trend unter gewissen Alternativen verzerrt werden kann.

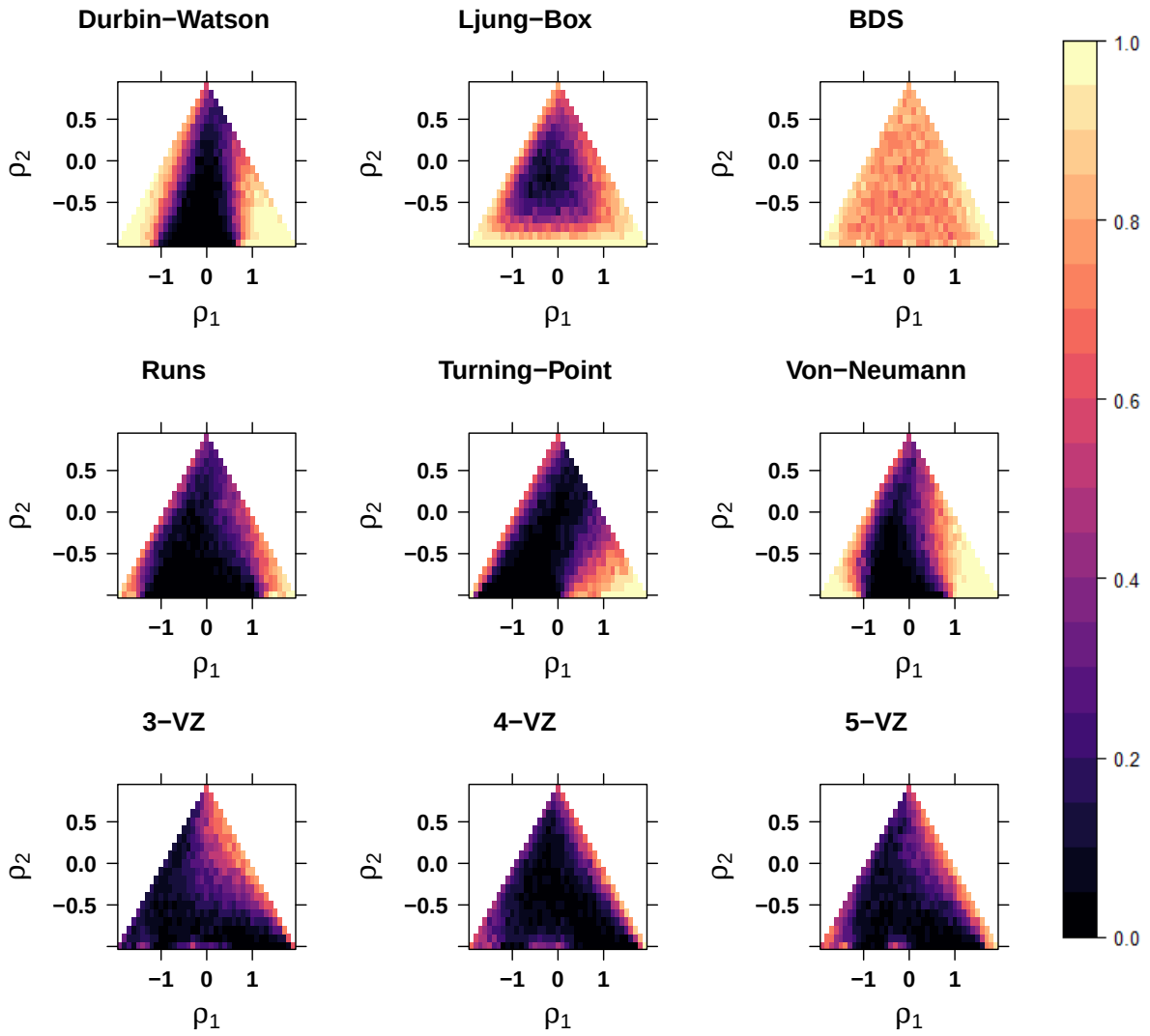


Abbildung 3.69: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, für $N = 20$ Beobachtungen und einem Trend der Steigung $2/N$

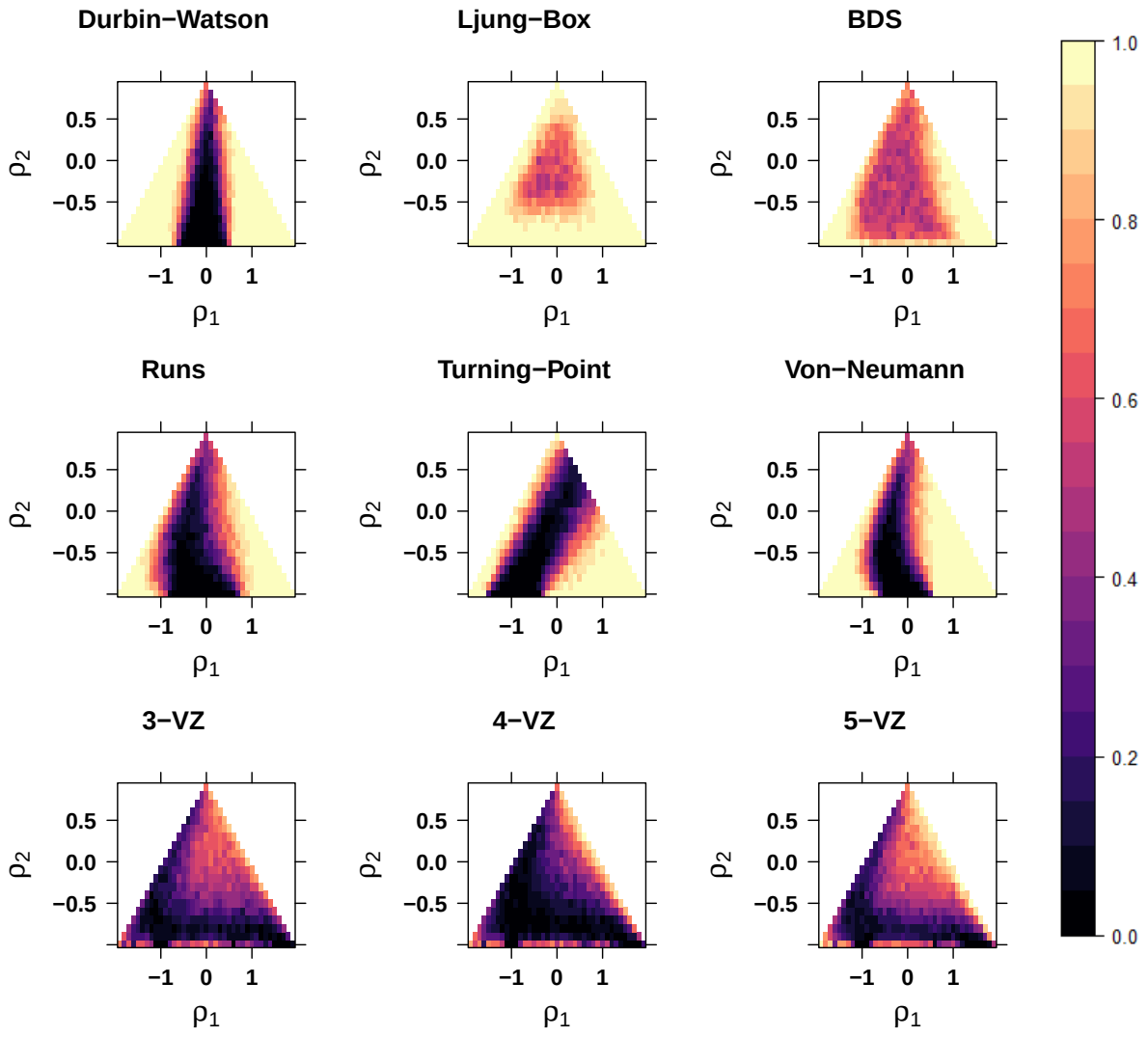


Abbildung 3.70: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, für $N = 50$ Beobachtungen und einem Trend der Steigung $2/N$

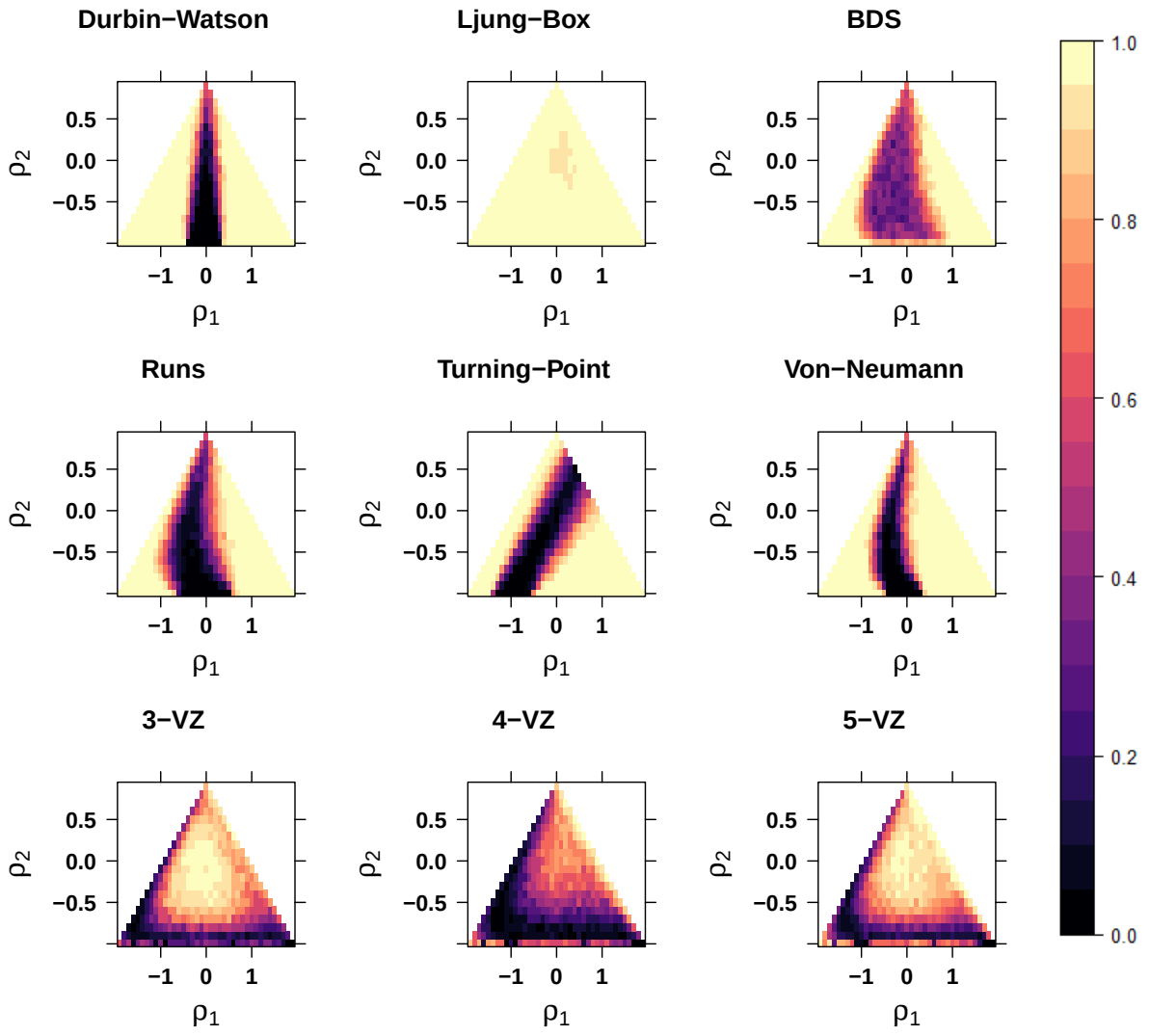


Abbildung 3.71: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, für $N = 100$ Beobachtungen und einem Trend der Steigung $2/N$

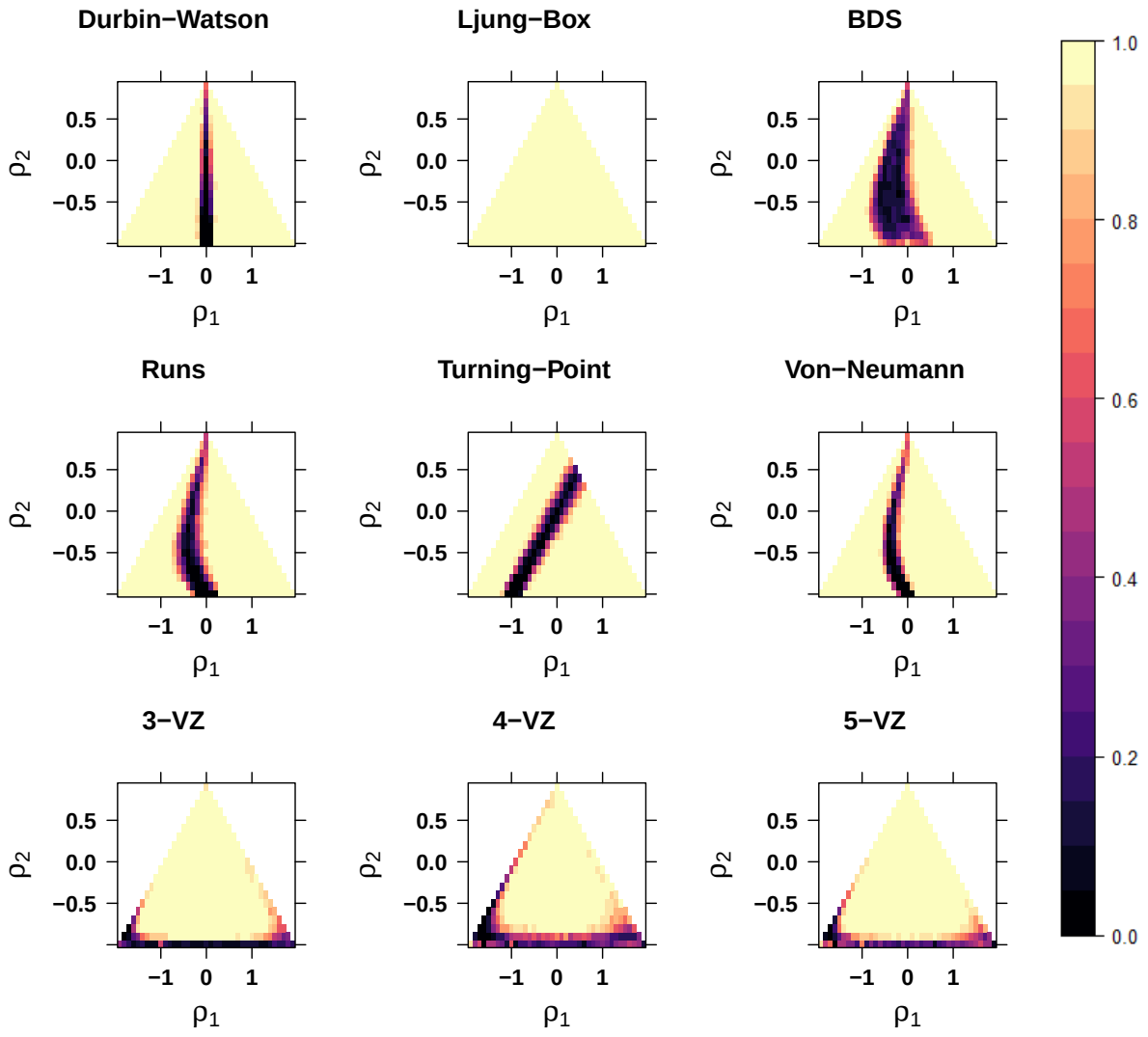


Abbildung 3.72: Simulierte Trennschärfen der Testverfahren bei stationären AR(2)-Alternativen in Abhängigkeit von ρ_1 und ρ_2, für $N = 500$ Beobachtungen und einem Trend der Steigung $2/N$

Insgesamt legen die Ergebnisse der Testverfahren bei Zeitreihen mit einem Trend nahe, dass sich auch die Auswirkungen von Niveauveränderungen auf ihre Trennschärfen für unterschiedliche, zugrunde liegende Prozesse nicht deutlich unterscheiden. So sind hier bei allen Testverfahren ähnliche Effekte erkennbar wie im AR(1)-Fall. Während der TP-Test und der DW-Test robust regieren, zeichnen sich der LB-Test und die K-VZ-Tests dadurch aus, dass sie den Trend als eine Abweichung von der Zufälligkeit identifizieren können. Die Annahmebereiche der übrigen Verfahren verschieben sich erneut in den Bereich negativer Korrelationen.

3.2.5 Resümee

Bei der Betrachtung der Trennschärfen in AR(2)-Prozessen zeigt sich deutlich, dass viele der betrachteten Verfahren lediglich für die Aufdeckung von Autokorrelationen 1. Ordnung geeignet sind. So hat der Parameter ρ_2 beim DW-, BDS-, Runs- und VNRR-Test unter Normalbedingungen kaum einen Einfluss auf deren Testentscheidungen. Der TP-Test nimmt eine Sonderstellung ein und ist im Prinzip auch in der Lage, Autokorrelationen 2. Ordnung zu erkennen. Allerdings gelingt ihm eine Ablehnung der Nullhypothese unter Alternativen nicht, in denen zumindest annäherungsweise $\rho_1 = \rho_2$ gilt. Die K-VZ-Tests erreichen in diesem Szenario für sehr wenige Alternativen überhaupt eine Ablehnung der Nullhypothese. Sie zeichnen sich jedoch dadurch aus, dass die ebenfalls extreme Korrelationen 2. Grades zu erkennen scheinen. Mit Abstand am besten schneidet hier der LB-Test ab, der für eine hinreichend große Anzahl an Beobachtung eine hervorragende Trennschärfe aufweist und bereits geringe Abweichungen der Parameter ρ_1 und ρ_2 von 0 erkennen kann.

Die Betrachtung von Szenarien, in denen Abweichungen von den Voraussetzungen der Testverfahren vorhanden sind, legt nahe, dass sie die gleichen Effekte haben wie im Fall von AR(1)-Prozessen. So bewirken Abweichungen von den Verteilungsannahmen bei den nichtparametrischen Verfahren sogar eine Verbesserung der Trennschärfe und Kontaminationen verschlechtern die Trennschärfe der parametrischen Verfahren wieder deutlich. Auch wachsende Varianzen haben lediglich einen Einfluss auf den LB- sowie den BDS-Test, die sie als Abweichung von der Zufälligkeit der Zeitreihen erkennen können. Das Vorhandensein eines Trends auf die verschiedenen Testverfahren hat ebenfalls ähnliche Effekte wie im AR(1)-Fall. Damit liegt die Vermutung nahe, dass die Effekte dieser Szenarien nicht von der Art des zugrunde liegenden Prozesses abhängen. Da die Auswirkungen dieser Szenarien bereits umfassend in Kapitel 3.1 diskutiert wurden, sollen die Trennschärfen für die folgenden Untersuchungen, bei denen noch weitere Prozesse betrachtet werden, nur noch unter Normalbedingungen diskutiert werden.

3.3 Saisonale AR-Prozesse (SAR-Prozesse)

Im Kapitel 3.2 zu AR(2)-Prozessen wurde deutlich, dass viele der Testverfahren Probleme damit haben, Autokorrelationen zum Lag 2 zu detektieren, während eine Erkennung von Abweichungen des Parameters ρ_1 von 0 für sie kein Problem darstellt. Somit stellt sich die Frage, wie die verschiedenen Testverfahren auf Autokorrelationen zu einem höheren Lag reagieren und welche Verfahren in der Lage sind, derartige Korrelationsstrukturen über weitere Distanzen als Abweichung von der Unabhängigkeit in der Zeitreihe zu erkennen.

Im Folgenden sollen deshalb Prozesse betrachtet werden, in denen Abhängigkeitsstrukturen zwischen Beobachtungen zu Vielfachen eines saisonalen Lags S zu beobachten sind. Beispiele, in denen solche Abhängigkeitsstrukturen zu erwarten wären, sind z. B. Quartals- oder Monatsdaten, bei denen Beobachtungen aus denselben Quartalen (Lag 4) bzw. Monaten (Lag 12) dazu neigen, ähnliche Werte annehmen, während aufeinanderfolgende Werte typischerweise keine oder eine vergleichsweise geringe Korrelationen aufweisen. Ein Beispiel dafür sind Umsatzdaten, die maßgeblich von der Anzahl der Arbeitstage im entsprechenden Monat und von einer monatsspezifischen Nachfrage abhängen. Prozesse dieser Art werden gemeinhin als saisonale autoregressive Prozesse zur Saisonalität S bezeichnet. In Anlehnung an Brockwell und Davis (2010, S. 218) werden im Folgenden also Modelle der Form:

$$x_t = \mu + \rho_S x_{t-S} + w_t, \quad |\rho_S| < 1, \quad w_t \sim WN(0, \sigma^2_{WN})$$

mit der Saisonalität $S \in \{1, ..., 6\}$ und $t \in \{S+1, ..., N\}$ mit $\mu = 0$ betrachtet. Ein solcher Prozess ist dabei, ähnlich wie im AR(1)-Fall, genau dann stationär, wenn $|\rho_S| < 1$ gilt.

Die Simulationsergebnisse für Zeitreihen mit $N = 50$ und $N = 500$ Beobachtungen sind in Abhängigkeit von der Saisonalität S und dem zugehörigem Autokorrelationskoeffizienten ρ_S in Abbildung 3.73 und 3.74 dargestellt. Dabei ist zu beachten, dass die Ergebnisse für $S = 1$ denen des AR(1)-Falls aus Kapitel 3.1 entsprechen und die Ergebnisse für $S = 2$ denen im AR(2)-Fall aus Kapitel 3.2, falls der Parameter ρ_1 auf 0 fixiert wird.

Die Ergebnisse verdeutlichen, dass die meisten der betrachteten Testverfahren lediglich für die Saisonalität $S = 1$ zufriedenstellende Ergebnisse liefern. So gelingt es bei $S = 2$ und einem Beobachtungsumfang von $N = 500$ lediglich dem TP-Test und dem LB-Test, die Nullhypothese bei moderaten betragsmäßigen Abweichungen von ca. 0.3 erfolgreich zu verwerfen. Alle anderen Verfahren erreichen eine Verwerfung erst bei sehr extremen Abweichungen von ca. $|0.9|$ in mehr als 80 % der Fälle. In diesem Zusammenhang ist besonders auffällig, dass der DW-, der Runs- sowie der VNRR-Test − wie es bereits im Kapitel 3.2 zu sehen war − bei einer Saisonalität von $S = 2$ selbst extreme negative Korrelationen nicht zu erkennen scheinen. Ein ähnliche Asymmetrie kann für den TP-Test bei einer Saisonalität von $S = 3$ im Bereich von stark positiven Korrelationen festgestellt werden. Allgemein ist die Trennschärfe des TP-Tests in Abhängigkeit von der betrachteten Saison S erneut leicht asymmetrisch und es scheint ein alternierendes

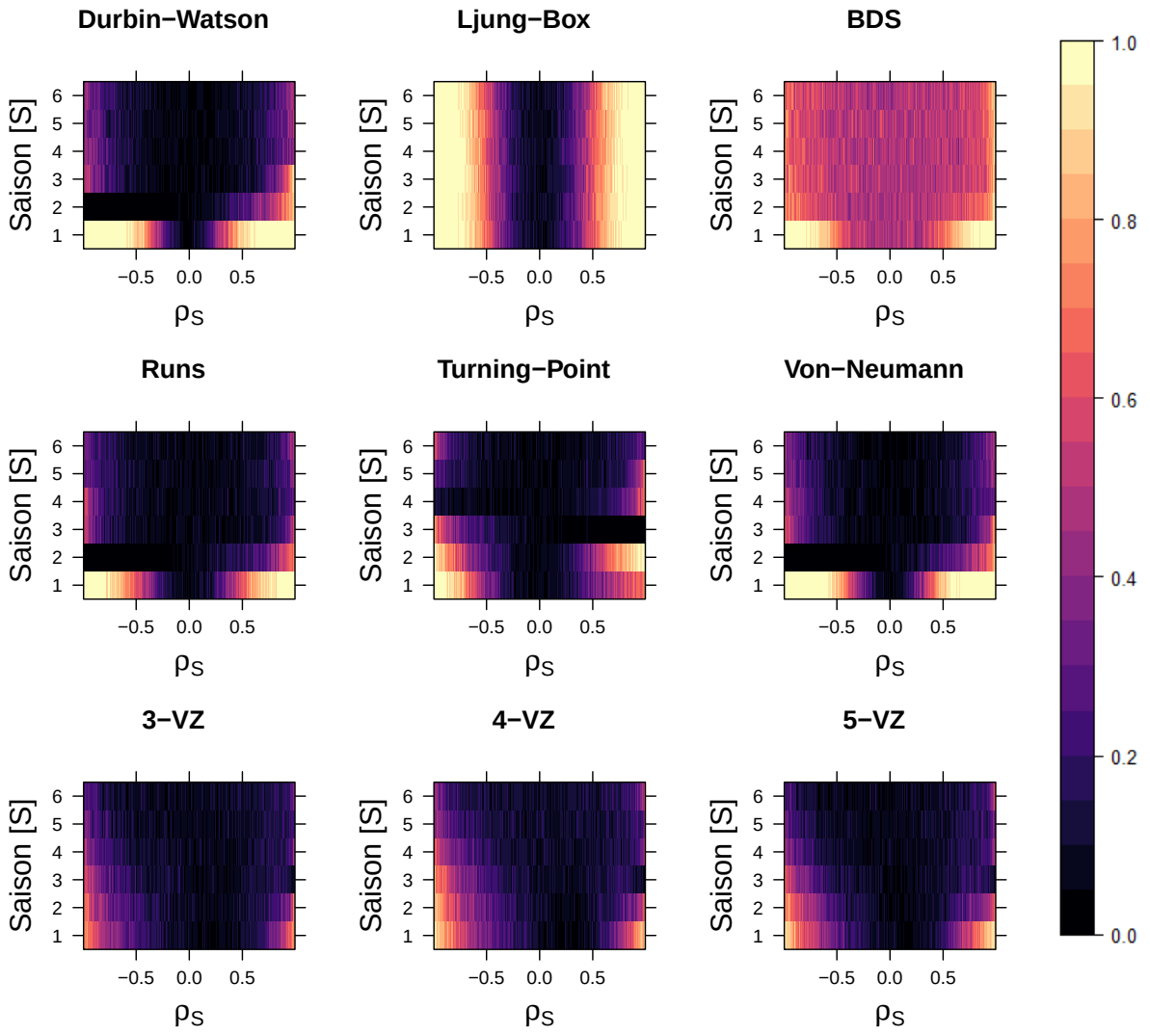


Abbildung 3.73: Simulierte Trennschärfen der Testverfahren bei stationären saisonalen AR-Prozessen in Abhängigkeit von ρ_S und der Saisonalität S, bei $N = 50$ Beobachtungen

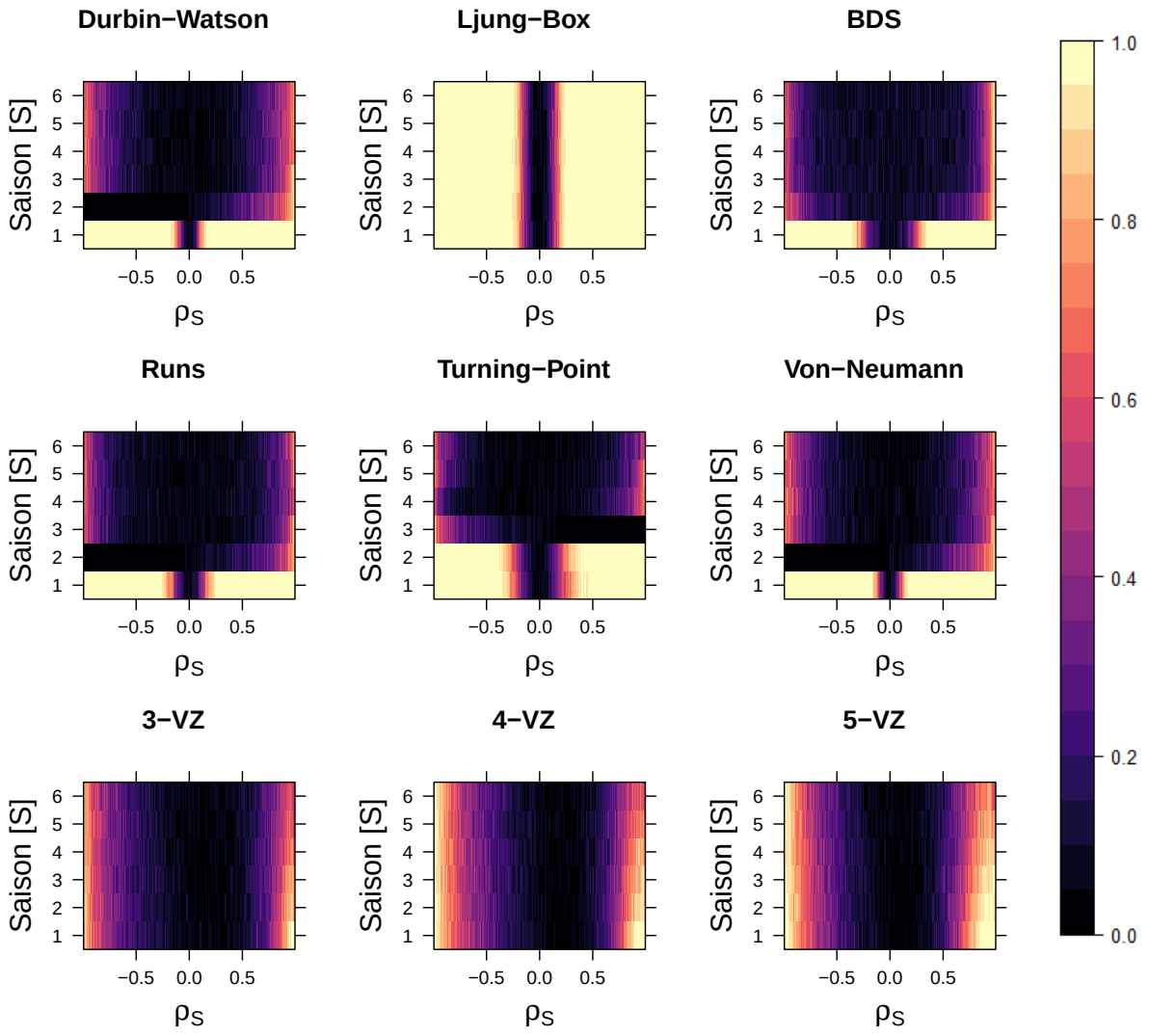


Abbildung 3.74: Simulierte Trennschärfen der Testverfahren bei stationären saisonalen AR-Prozessen in Abhängigkeit von ρ_S und der Saisonalität S, bei $N = 500$ Beobachtungen

Verhalten vorzuliegen. Konkret kann der TP-Test für ungerade S negative Korrelationen besser erkennen als positive, während es bei geraden S umgekehrt zu sein scheint.

Im Hinblick auf die K-VZ-Tests wird deutlich, dass sie ebenfalls die Tendenz aufweisen, bei kleineren Lags eine bessere Trennschärfe zu erreichen. Insgesamt zeigen sie jedoch für verschiedene Saisonalitäten eine relativ konstante Trennschärfe. So ist kein klarer „Abriss“ der Trennschärfe ab einem bestimmten S, wie es für viele der anderen Verfahren der Fall ist, erkennbar. Insbesondere schneidet die Trennschärfe des 5-VZ-Tests bei $S \geq 3$ unter allen Verfahren, mit Ausnahme des LB-Tests, am besten ab. Auffällig ist weiterhin, dass die K-VZ-Tests zu jedem Lag wieder eine asymmetrische Trennschärfen aufweisen, wobei positive Korrelationen immer leichter erkannt werden können als negative.

Der LB-Test ist in diesem Szenario mit Abstand am besten geeignet und erreicht eine Verwerfung der Nullhypothese zu sämtlichen, hier betrachteten Saisons mit einer sehr guten und konstanten Trennschärfe. Dabei gelingt es ihm, bereits kleine Abweichungen der saisonalen Autokorrelationsparameter ρ_S von 0 ähnlich zuverlässig zu erkennen wie im AR(1)-Fall.

Weiter ist es nun interessant herauszufinden, inwieweit sich die Wahl des Parameters H der zu betrachtenden empirischen Autokorrelationskoeffizienten beim LB-Test (vgl. Kap. 2.4) auf seine Trennschärfe im Rahmen von saisonalen AR-Modellen auswirkt. Insbesondere wird untersucht was passiert, falls sie Saison S den Wert des Parameters H überschreitet. Dazu wurde $H = 3$ gesetzt und die Trennschärfe des LB-Tests ist für $S \in \{1, ..., 6\}$ in Abbildung 3.75 dargestellt.

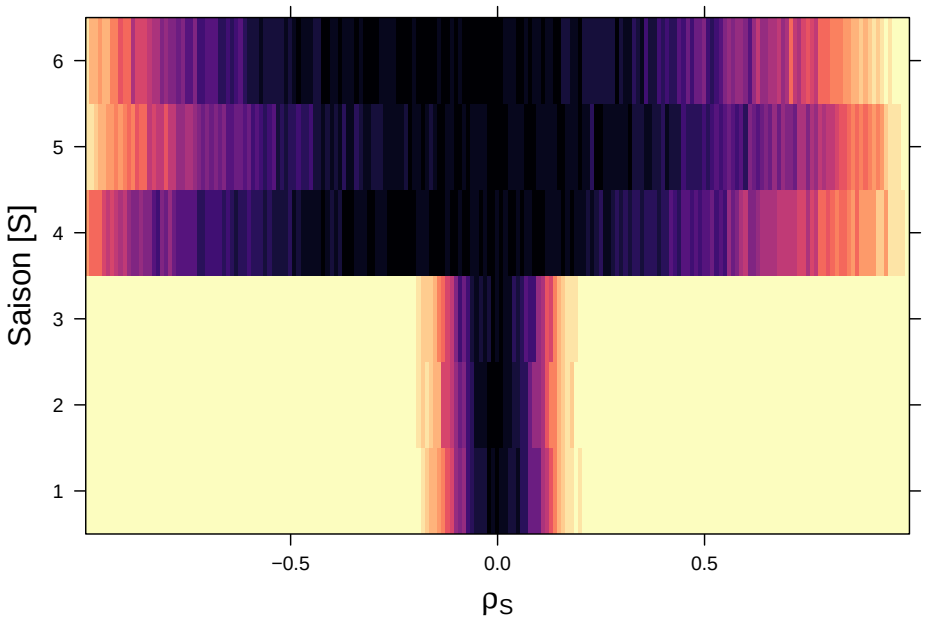
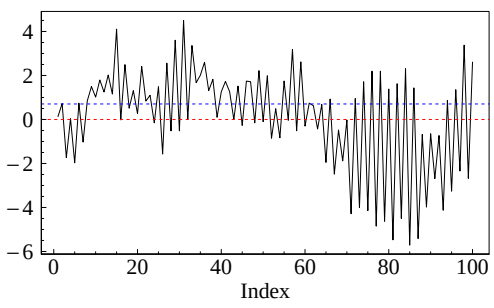


Abbildung 3.75: Simulierte Trennschärfen des LB-Tests mit $N = 500$ Beobachtungen zu unterschiedlichen Saisons, mit $H = 3$ in Abhängigkeit von ρ_S

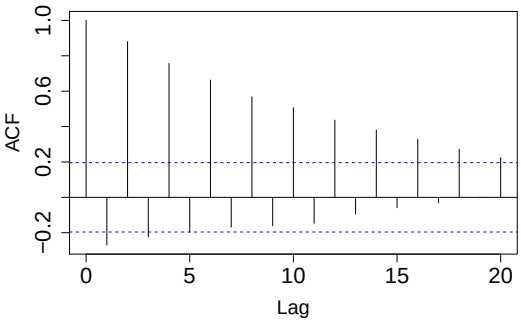
Daraus geht hervor, dass der Test lediglich gute Ergebnisse liefert, wenn die Saisonalität S in der Zeitreihe kleiner ist als der Parameter H. Ist dies nicht der Fall, so kann der LB-Test lediglich sehr extreme Korrelationen erkennen und seine Trennschärfe ist dabei insbesondere schlechter als die der K-VZ-Tests.

Um nachzuvollziehen, warum die verschiedenen Verfahren so unterschiedlich auf das Vorliegen von saisonalen Autokorrelationen reagieren, wurden zwei Zeitreihen mit starken negativen und positiven Korrelationen zur Saisonalität $S = 2$ mit den dazugehörigen Korrelogrammen in Abbildung 3.76 dargestellt.

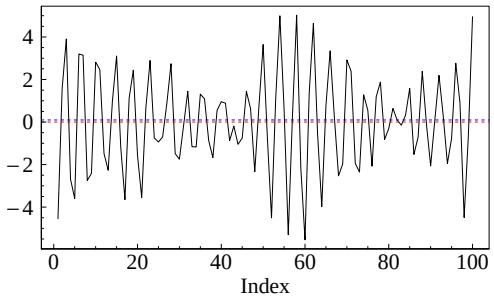
Anhand der abgebildeten Korrelogramme wird deutlich, weshalb es dem LB-Test im Gegensatz zum DW-Test gelingt, die Nullhypothese in solchen Situationen zu verwerfen. So ist jeder zweite empirische Autokorrelationskoeffizient stark signifikant erhöht, wohingegen $\hat{\rho}_1$ in beiden Fällen viel kleiner ausfällt. Außerdem wird deutlich, weshalb der DW-Test beim Vorliegen von negativen Autokorrelationen 2. Ordnung vergleichsweise mehr Schwierigkeiten bei der Verwerfung der Nullhypothese hat. So fällt $\hat{\rho}_1$ hier betragsmäßig deutlich geringer aus als im Fall von positiven Korrelationen, bei denen der Parameter sogar als signifikant eingestuft wird.



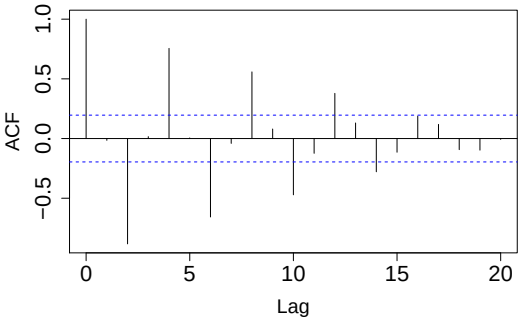
(a) ρ_2= 0.9



(b) ρ_2= 0.9



(c) ρ_2= -0.9



(d) ρ_2= -0.9

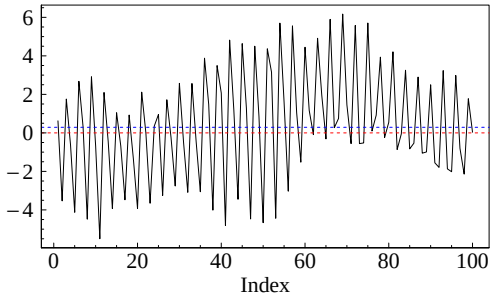
Abbildung 3.76: Zeitreihen aus saisonalen AR-Prozessen bei $S = 2$ mit $N = 100$ Beobachtungen, mit ρ_2= 0.9 (a) und ρ_2= -0.9 (c), zugehörigen Korrelogrammen (b) und (d), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

Diese Beobachtung ist damit zu begründen, dass positive Autokorrelationen zum Lag 2 das Niveau der Zeitreihe deutlich stärker schwanken lassen, als es bei negativen Korrelationen der Fall ist (vgl. Abb. 3.76). Ein solches Verhalten führt dazu, dass aufeinanderfolgende Beobachtungen dieser Zeitreihen − trotz des im Prinzip alternierenden Verhaltens, das aus Autokorrelationen 2. Ordnung folgt − Werte einer verhältnismäßig ähnlicheren Größenordnung annehmen. Als Konsequenz findet eine Überschätzung der empirischen Autokorrelationskoeffizienten 1. Ordnung statt. Bei stark negativen Korrelationen wird deutlich, dass das Niveau der Zeitreihe weitestgehend konstant bleibt und benachbarte Werte dementsprechend annähernd unabhängige Werte ausweisen, wodurch eine realistischere Schätzung von ρ_1 möglich ist als es bei positiven Korrelationen von einer gleichen Intensität der Fall ist.

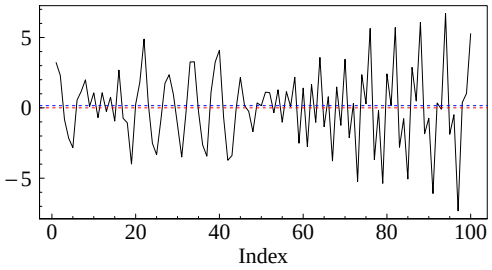
In Hinblick auf den VNRR- und den Runs-Test, sowie die K-VZ-Tests begünstigt dieses Verhalten ebenfalls eine Verwerfung der Nullhypothese bei positiven Korrelationen. So führt es sowohl zu weniger Runs als auch zu einer geringeren Anzahl an Vorzeichenblöcken. Außerdem hat es zur Konsequenz, dass benachbarte Beobachtungen ähnlichere Ränge aufweisen, als es unter der Nullhypothese zu erwarten wäre. In allen Fällen nehmen die verschiedenen Teststatistiken beim Vorliegen positiver Korrelationen zum Lag 2 Werte an, die eigentlich bei einer Abhängigkeitsstruktur 1. Ordnung zu erwarten wären. Bei negativen Korrelationen hingegen nehmen sie viel eher Werte an, die unter der Unabhängigkeit vorliegen.

Für weitere Untersuchungen zu den Asymmetrien der Trennschärfen sind exemplarisch Zeitreihen mit höheren saisonalen Lags von $S = 3$ und $S = 5$ in Abbildung 3.77 dargestellt. Dabei zeigt sich, dass die extremen Niveauänderungen bei stark positiven Werten von ρ_S mit größer werdendem S deutlich abgeschwächt werden. Das oben beschriebene Verhalten ist in diesen Fällen also deutlich schwächer ausgeprägt als bei kleineren Werten von S und die Annahmebereiche werden symmetrischer. Offensichtlich führen aber extreme positive und negative Autokorrelationen höherer Ordnung ebenfalls dazu, dass die diskutierten Teststatistiken die Korrelationsstruktur nicht erkennen können.

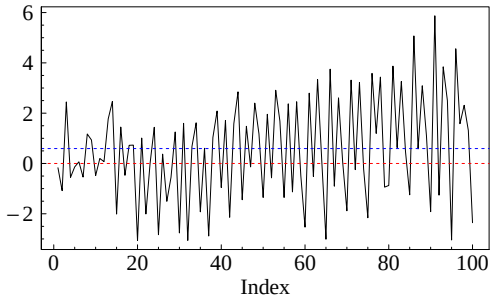
Die Ursache dafür, dass der TP-Test Autokorrelationen zum Lag 1 und 2 erkennen kann, wurde bereits in Kapitel 3.2 erläutert. So konnte dort gezeigt werden, dass Autokorrelationen zum Lag 2 die Wahrscheinlichkeit des Auftretens eines Turning-Points beeinflussen. Eine Erklärung dafür, dass eine Autokorrelation zum Lag 3 nicht mehr erkannt werden kann, liefert die Beschaffenheit der Teststatistik des TP-Tests. So sind jeweils 3 aufeinanderfolgende Beobachtungen an der Entstehung eines Turning-Points beteiligt und die Wahrscheinlichkeit, dass ein Turning-Points entsteht, beruht auf der Abhängigkeitsstruktur dieser 3 Beobachtungen. Bei saisonalen autoregressiven Modellen einer höheren Ordnung als 2, sind 3 aufeinanderfolgende Beobachtungen jedoch immer unabhängig, sodass hier die Wahrscheinlichkeit des Auftretens eines Turning-Points nicht von ρ_S abhängt. Offenbar führen sehr stark positive oder negative Korrelationen aufgrund der bereits erörterten leichten Niveauänderungen jedoch ebenfalls dazu, dass eine Verwerfung der Nullhypothese durch daraus resultierende Effekte in einigen Fällen



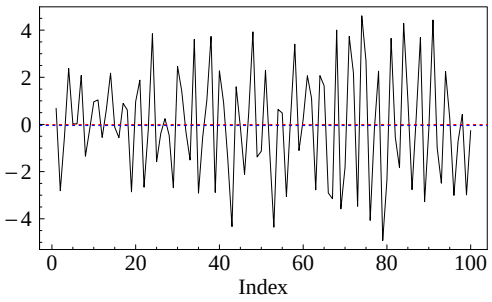
(a) ρ_S= 0.9 und $S = 3$



(b) ρ_S= -0.9 und $S = 3$



(c) ρ_S= 0.9 und $S = 5$



(d) ρ_S= -0.9 und $S = 5$

Abbildung 3.77: Saisonale Zeitreihen zum Lag $S = 3$ (a) und (b) sowie $S = 5$ (c) und (d) mit $N = 100$ Beobachtungen, mit ρ_S= 0.9 (a) und (c) und ρ_S= -0.9 (b) und (d), Nulllinien (rot) und empirischem Median (blau)

ermöglicht wird. Einen Sonderfall stellen positive Korrelationen zum Lag 3 dar. In diesem Fall ist die Verwerfung der Nullhypothese für den TP-Test aufgrund der Korrelationsstruktur nicht möglich. So wird beim Betrachten von Abbildung 3.77 (a) deutlich, dass in der Regel zu jeder 3. Beobachtung ein Hochpunkt vorliegt. Dies bedeutet, dass für 3 aufeinanderfolgende Werte jeweils 2 Turning-Points zu erwarten wären, wodurch die Wahrscheinlichkeit, bei einer beliebigen Beobachtung einen Turning-Point vorzufinden genau $2/3$ − also der Wahrscheinlichkeit unter der Nullhypothese − entspricht.

Insgesamt wird in diesem Kapitel deutlich, dass das Erkennen von saisonalen Abhängigkeiten vielen der gängigen Tests zur Überprüfung von Unabhängigkeitsannahmen einer Zeitreihe Schwierigkeiten bereitet. Mit Abstand am geeignetsten dafür ist der LB-Test, der zu sämtlichen hier betrachteten Lags Abhängigkeitsstrukturen mit ähnlich guter Trennschärfe wie im AR(1)-Fall erkennen kann. Dabei gilt, dass der Parameters H, der die von diesem Test betrachteten empirischen Autokorrelationskoeffizienten spezifiziert, die Saisonalität in der Zeitreihe S überschreiten sollte, damit er in der Lage ist, diese Abhängigkeiten zu erfassen. Auch mit dem TP-Test können zumindest saisonale Korrelationen bis zur 2. Ordnung detektiert werden, was ihn

von den übrigen Testverfahren abhebt. Besonders aber stechen die K-VZ-Tests in diesem Szenario heraus. Sie sind in der Lage, diverse, stark ausgeprägte saisonale Abhängigkeitsstrukturen zu erfassen, ohne das − wie beim LB-Test − eine geeignete Spezifikation des Testverfahrens erfolgen muss. Dies deutet darauf hin, dass die K-VZ-Tests in diesem Szenario eine breitere Alternative als die übrigen Tests umfassen und eventuell geeignet sind, um z. B. nach der Anpassung von ARMA(p,q)-Modellen, verbleibende Strukturen in den Residuen zu identifizieren. Sie profitieren jedoch − wie auch schon im AR(1)-Fall − kaum von einem wachsenden Stichprobenumfang, sodass ihre Trennschärfe allgemein eher schwach ausfällt.

3.4 MA(1)-Prozesse

Im Folgenden sollen die unterschiedlichen Testverfahren auf eine Auswahl von Moving-Average-Prozessen 1. Ordnung (MA(1)-Prozesse) angewendet werden (s. Kap. 2.2). Anders als bei autoregressiven Prozessen, wie sie in den vorherigen Kapiteln betrachtet wurden, stehen dabei nicht die Beobachtungen miteinander in Beziehung, sondern vielmehr die nicht beobachtbaren Zufallsgrößen. Diese Art von Prozessen findet Anwendung in Situationen, in denen sich Innovationen zu vergangenen Zeitpunkten immer im gleichem Maß auf zukünftige Beobachtungen auswirken. Konkret sind in diesem Abschnitt mathematisch Zeitreihen der Form

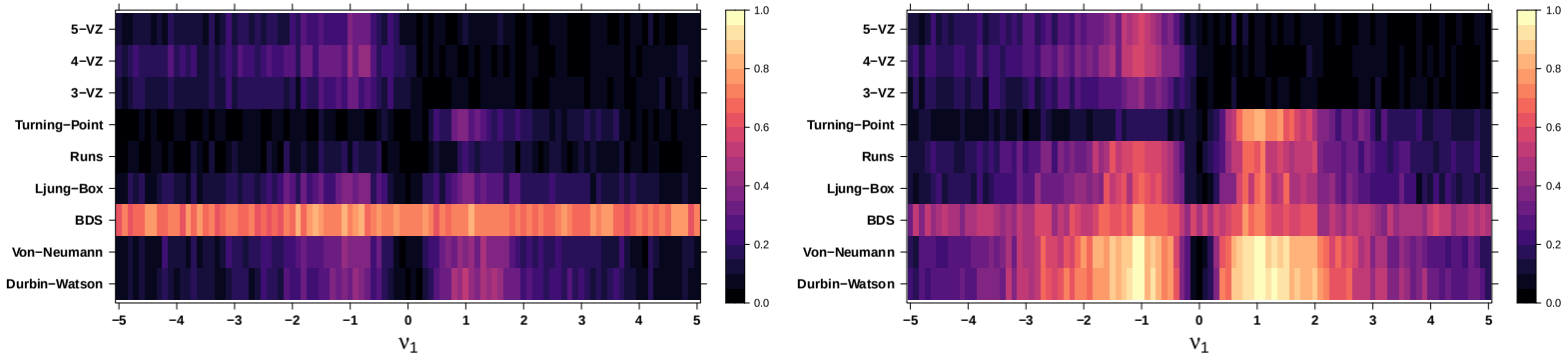
$$x_t = \mu + \nu_1 w_{t-1} + w_t, \quad w_t \sim WN(0, \sigma^2_{WN})$$

für $t \in \{2, ..., N\}$ mit $\mu = 0$ von Interesse. Es kann gezeigt werden, dass MA(1)-Prozesse für jeden Wert von ν_1 stationär sind (Shumway und Stoffer, 2017, S. 82). Die Unabhängigkeit eines solchen Prozesses ist dabei genau dann gegeben, wenn $\nu_1 = 0$ gilt. In dieser Arbeit werden aufgrund der wenigen neuen Systematiken bei betragsmäßig größeren Werten des Moving-Average-Parameters lediglich Alternativen von $\nu_1 \in$ [-5, 5] bei einer Gitterfeinheit von 0.01 betrachtet und die Trennschärfe der verschiedenen Testverfahren wurde für solche Alternativen untersucht. Die Ergebnisse für verschiedene Stichprobenumfänge sind dabei in Abbildung 3.78 dargestellt.

Beim Betrachten der Abbildungen fällt vor allem auf, dass sämtliche Verfahren bei Stichprobenumfängen, die kleiner als 100 sind, generelle Probleme bei der Verwerfung der Nullhypothese haben. So gelingt es lediglich dem DW- und dem VNRR-Test bei einem Wert des Parameters ν_1 von ca. 1 bzw. -1 in 95 % der Simulationen zu verwerfen. Auch bei einer Beobachtungszahl von $N = 100$ ist keines der Verfahren in der Lage, die Unabhängigkeit für betragsmäßig große Werte von ν_1 abzulehnen. Bei moderaten Werten von bis zu |2| können jedoch zumindest der VNRR- sowie der DW-Test eine Abhängigkeitsstruktur erkennen und sie schneiden damit bei sämtlichen Stichprobenumfängen mit Abstand am besten ab. Speziell bei $N = 500$ gelingt es diesen Verfahren bereits minimale Abweichungen des Parameters ν_1 von ca. 0.1 sehr zuverlässig erkennen. Auch erreichen sie eine Ablehnung der Nullhypothese bei allen übrigen Werten im Intervall [-5, 5].

Etwas schlechter erscheinen hier der LB- und der Runs-Test, die die Unabhängigkeit bei einer Beobachtungszahl von $N = 100$ lediglich für betragsmäßige Werte des Parameters ν_1 von bis zu ca. 1.5 verwerfen können. Im Bereich kleinerer Werte weisen sie dabei eine ähnliche Trennschärfe auf wie der DW- und der VNRR-Test. Bei einem Stichprobenumfang von $N = 500$ haben sie für Werte von $|\nu_1| \geq 4$ immer noch Probleme bei der Ablehnung der Nullhypothese.

Beim BDS-Test fällt auf, dass es ihm − wie schon im AR(1)- und AR(2)-Fall − nicht gelingt, das Niveau bei einem Stichprobenumfang von unter $N = 500$ einzuhalten. Auch bei einem für ihn geeigneten Stichprobenumfang hat er deutlich mehr Probleme, die Nullhypothese bei großen Werten von ν_1 zu verwerfen, als die oben genannten Verfahren.



(a) Stichprobenumfang $N = 20$

(b) Stichprobenumfang $N = 50$

(c) Stichprobenumfang $N = 100$

(d) Stichprobenumfang $N = 500$

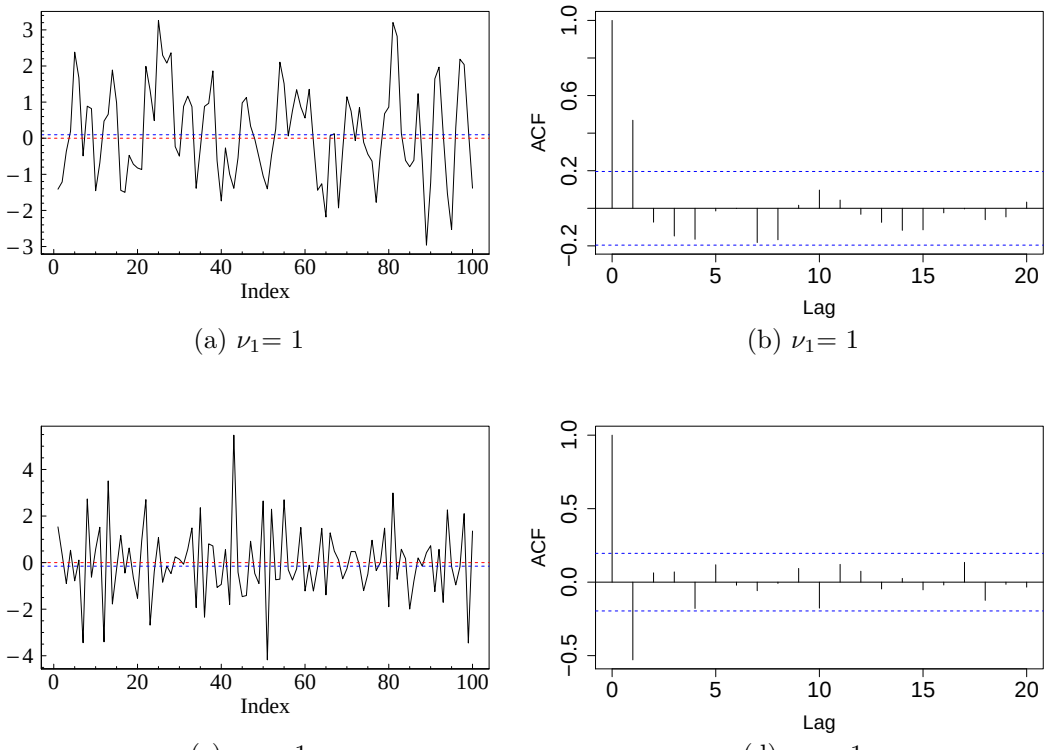
Abbildung 3.78: Simulierte Trennschärfe der Testverfahren bei ausgewählten, stationären MA(1)-Alternativen für unterschiedliche Stichprobenumfänge

Der TP-Test weist hier erneut eine asymmetrische Trennschärfe auf, wobei es ihm zu jedem Stichprobenumfang deutlich schwerer fällt, die Nullhypothese bei negativen Werten des Moving-Average-Koeffizienten zu verwerfen. Für positive Werte entspricht seine Trennschärfe in etwa der des LB- und des Runs-Tests.

Eine Ausnahme bilden hier die K-VZ-Tests, die erneut die schlechteste Trennschärfe unter allen Testverfahren aufweisen und insbesondere Alternativen mit positiven ν_1 unabhängig vom Stichprobenumfang nicht erkennen können. Aber auch für negative Werte des Parameters ist die Trennschärfe sehr schlecht, sodass die Nullhypothese für keinen Wert in mehr als 50 % der Fälle abgelehnt werden kann.

Auch wenn dieser Bereich hier nicht mehr dargestellt ist, fällt auf, dass die Testverfahren unabhängig von Stichprobenumfang bei hinreichend großem ν_1 erneut Probleme bei der Ablehnung der Nullhypothese bekommen.

Um nachvollziehen zu können, weshalb die unterschiedlichen Testverfahren in dieser Art und Weise auf die verschiedenen Alternativen reagieren, werden im Folgenden verschiedene MA(1)-Prozesse mit zugehörigen Korrelogrammen betrachtet. Für 2 exemplarische Zeitreihen mit $\nu_1 = 1$ und $\nu_1 = -1$ und einer Beobachtungszahl von $N = 100$ sind diese in Abbildung 3.79 dargestellt.



(a) ν_1= 1 (b) ν_1= 1

(c) ν_1= -1 (d) ν_1= -1

Abbildung 3.79: Zeitreihen von MA(1)-Prozessen mit $N = 100$ Beobachtungen, mit ν_1= 1 (a) und ν_1= -1 (c) zugehörigen Korrelogrammen (b) und (d), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

Dabei handelt es sich um Zeitreihen aus einem Bereich des Parameters ν_1, in dem die meisten Verfahren am ehesten eine Ablehnung der Nullhypothese erreichen.

Anhand der Korrelogramme wird deutlich, warum die parametrischen Verfahren die Nullhypothese bei diesen Werten so zielsicher ablehnen können. So ist der empirische Autokorrelationskoeffizient 1. Ordnung in beiden Fällen signifikant erhöht. Insbesondere ist die theoretische Autokorrelation bei Werten von $\nu_1 = 1$ bzw. $\nu_1 =$ -1 mit $\rho_1 = 0.5$ maximal (vgl. Kap. 2.2). Typisch für MA(1)-Prozesse ist dabei weiterhin, dass die Korrelationskoeffizienten − anders als beim AR(1)-Prozess − nicht abklingen, sondern nach q Lags abreißen. Heuristisch lässt sich dies mit der Form des Prozesses erklären. So hängt ein Wert eines allgemeinen MA(q)-Prozesses lediglich von den q gewichteten Vergangenheitswerten des Weißen Rauschens ab und insbesondere sind x_t und x_{t+q+1} für alle $t \in \{1, ..., N - q - 1\}$ unabhängig. Dadurch erklärt sich auch die überlegene Trennschärfe des DW-Tests gegenüber der des LB-Tests, da letzterer 14 zusätzliche, nicht signifikant erhöhte, empirische Autokorrelationskoeffizienten betrachtet. Als Konsequenz erhöhen sich in diesem Fall lediglich die kritischen Werte, ab denen eine Ablehnung der Nullhypothese erfolgt, ohne dass der Test von deren Miteinbeziehung profitiert.

Mit früher angestellten Überlegungen wird auch die gute Trennschärfe des VNRR-Tests verständlich. So neigen aufeinanderfolgende Werte der Zeitreihe dazu, abhängig davon, ob ν_1 positiv oder negativ ist, ähnlichere bzw. unterschiedlichere Werte anzunehmen, als es unter der Unabhängigkeit der Fall wäre. Dies verdeutlichen die in Abbildung 3.79 dargestellten Korrelogramme. Auch das Verhalten des Runs- und BDS-Tests lässt sich durch die vorhandene empirische Korrelation in MA-Prozessen erklären, die in den betrachteten Fällen eine korrekte Ablehnung der Nullhypothese mit den gleichen Argumenten wie im AR(1)-Prozess ermöglicht.

Um die stark asymmetrische Trennschärfe des TP-Tests im MA(1)-Fall zu verstehen, macht es Sinn, die Auswirkungen von unterschiedlichen Werten des MA-Koeffizienten auf seine Teststatistik zu untersuchen. Dabei werden Situationen betrachtet, in denen eine Ablehnung der Nullhypothese heuristisch am wahrscheinlichsten ist, also in denen die empirische Autokorrelation maximal wird. So führen moderate positive Werte von ν_1 dazu, dass weniger Turning-Points auftreten, als es unter der Nullhypothese zu erwarten wäre. Dies ist exemplarisch für $\nu_1 = 1$ damit zu begründen, dass das Auftreten eines Hochpunkts (analog eines Tiefpunkts) im MA(1)-Prozess folgende Bedingungen erfordert:

$$\begin{aligned} & x_t > x_{t-1} && \wedge \quad x_t > x_{t+1} \\ \Leftrightarrow \quad & w_t + \nu_1 w_{t-1} > w_{t-1} + \nu_1 w_{t-2} && \wedge \quad w_t + \nu_1 w_{t-1} > w_{t+1} + \nu_1 w_t \\ \Leftrightarrow \quad & w_t > w_{t-2} && \wedge \quad w_{t-1} > w_{t+1}. \end{aligned}$$

Da die Zufallsgrößen einem Weißen Rauschen entstammen und die beiden erforderlichen Ereignisse unabhängig sind, ergibt sich hier die Wahrscheinlichkeit eines Hochpunkts von 0.25. Auf analoge Weise kann nachvollzogen werden, dass die Wahrscheinlichkeit eines Tiefpunktes ebenfalls 0.25 beträgt, sodass die Wahrscheinlichkeit eines Turning-Points genau ihrer Summe, also

0.5, entspricht. Damit ist das Auftreten eines Turning-Points in diesem Fall wesentlich weniger wahrscheinlich als unter der Unabhängigkeit, wo sie genau 2/3 entspricht (vgl. Kap. 2.6). Damit wird deutlich, warum hier eine Verwerfung der Nullhypothese stattfinden kann. Bei einem ähnlichen Vorgehen im Fall, dass ν_1 = -1 gilt − wobei eine größere Anzahl von Turning-Points zu erwarten wäre − ist die Wahrscheinlichkeit deutlich schwieriger zu berechnen. Das liegt daran, dass die Unabhängigkeit der beiden oben beschriebenen Ereignisse nicht ausgenutzt werden kann. Aus diesem Grund wurden 1000 Zeitreihen aus solchen MA(1)-Prozessen simuliert und die Anzahl der Turning-Points wurde für jede von ihnen berechnet, um eine Vorstellung von den Auswirkungen eines negativen MA-Parameters von -1 zu bekommen. Dabei ergab sich eine mittlere Anzahl von 71.44 Turning-Points, die nicht viel größer als die zu erwartende Anzahl von 66 ist. Insbesondere würde eine solche Anzahl von Turning-Points nicht dazu führen, dass eine Ablehnung der Nullhypothese stattfindet. Zur besseren Veranschaulichung wurde außerdem die mittlere Anzahl von Turning-Points für alle hier betrachteten Werte des Parameters ν_1 in Zeitreihen mit $N = 100$ Beobachtungen, gemeinsam mit den kritischen Werten, ab denen eine Ablehnung stattfindet, in Abbildung 3.80 dargestellt.

Dabei zeigt sich erneut die asymmetrische Auswirkung des MA-Parameters und insbesondere die Tatsache, dass negative Werte die Anzahl der Turning-Points weniger stark beeinflussen als positive. Insbesondere führen negative Korrelationen in diesem Fall nicht zu einer Verwerfung der Nullhypothese. Somit kann nachvollzogen werden, warum der TP-Test in diesem Szenario im Bereich negativer MA-Koeffizienten eine so schlechte Güte aufweist.

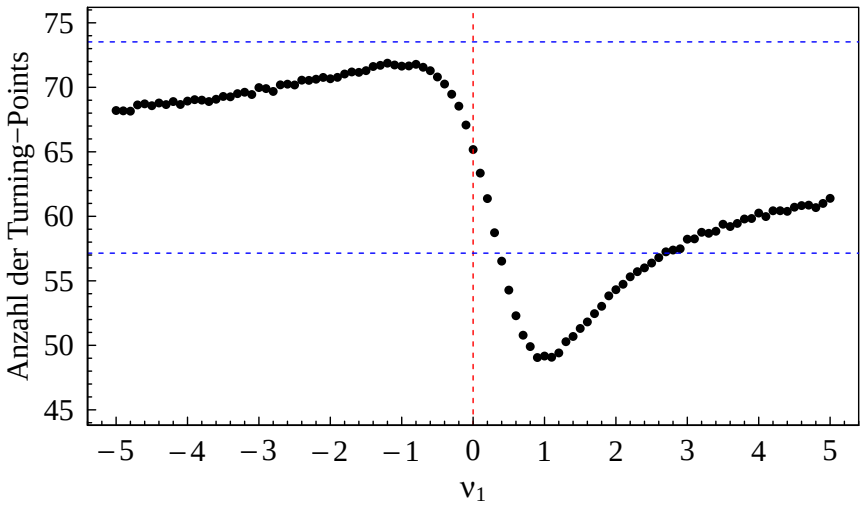
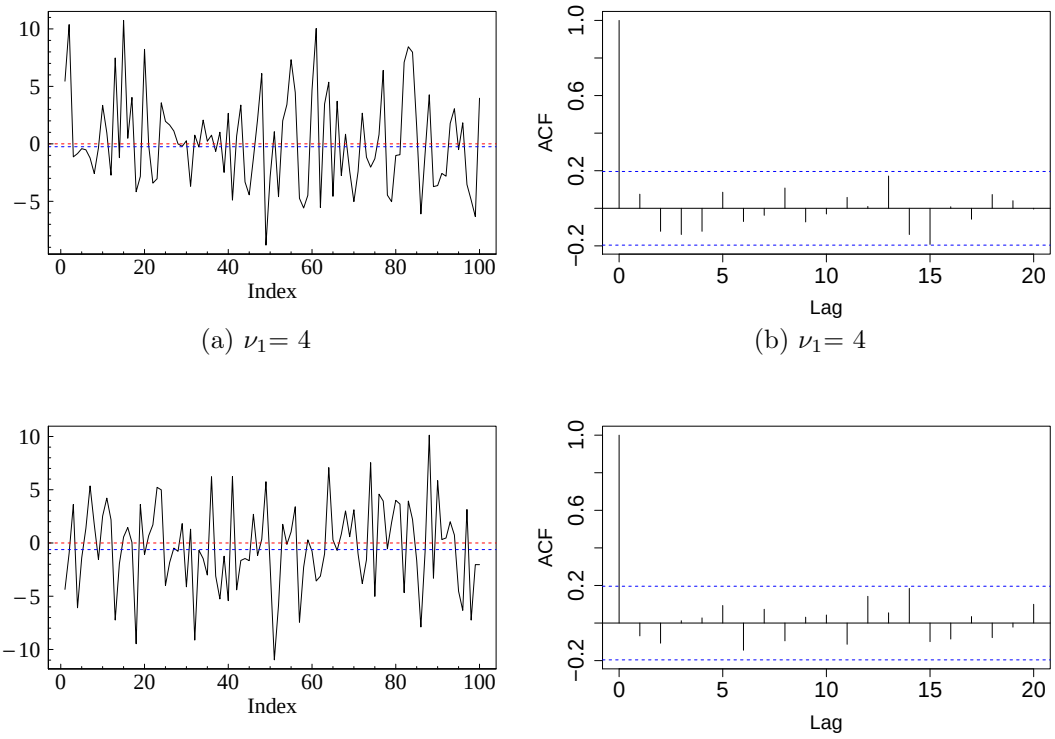


Abbildung 3.80: Mittlere Anzahl der Turning-Points in Abhängigkeit von ν_1, bei $N = 100$ Beobachtungen, mit Nullniveau (rot) und kritischen Werten (blau)

Weiter ist es von Interesse, weshalb die Verfahren ab einer bestimmten Größe von ν_1 nicht mehr in der Lage sind, die Abhängigkeitsstruktur der Zeitreihe zu erkennen. Um dies nachzuvollziehen, sind Zeitreihen mit einem Moving-Average-Koeffizienten von 4 bzw. -4 und bei einer Beobachtungszahl von $N = 100$ mit den zugehörigen Korrelogrammen in Abbildung 3.81 dargestellt. Damit stammen die Zeitreihen aus einem Bereich, in dem es keinem der betrachteten Testverfahren gelingt, die Nullhypothese der Unabhängigkeit zu verwerfen.

Anhand der Zeitreihen ist zunächst ersichtlich, dass die Beobachtungen des Prozesses aufgrund des betragsmäßig großen Wertes von ν_1 insgesamt größere Werte annehmen. Dabei sind die empirischen Korrelationskoeffizienten 1. Ordnung jedoch nicht mehr signifikant erhöht. Dies lässt sich darauf zurückführen, dass der theoretische Autokorrelationskoeffizient 1. Ordnung für $\nu_1 = 4$ deutlich kleiner ausfällt als z. B. im Fall $\nu_1 = 1$ (s. Kap. 2.2). Konkret strebt sein Wert für größere bzw. kleinere Werte als 1 bzw. -1, bei denen er sein Maximum annimmt, gegen 0. Aus diesem Grund haben die parametrischen Verfahren, deren Testentscheidung lediglich auf dem Wert von $\hat{\rho}_1$ beruht, auch Probleme die Nullhypothese abzulehnen. Mit zunehmender Beobachtungszahl wird jedoch der auf der Normalverteilung von den empirischen Autokorrela-



(a) ν_1= 4 (b) ν_1= 4

(c) ν_1= -4 (d) ν_1= -4

Abbildung 3.81: Zeitreihen eines MA(1)-Prozesses mit $N = 100$ Beobachtungen, mit ν_1= 4 (a) und ν_1= -4 (c), zugehörigen Korrelogrammen (b) und (d), Nulllinien (rot) und empirischem Median bzw. kritischen Werten (blau)

tionskoeffizienten basierende, kritische Wert immer kleiner, sodass auch betragsmäßig größere Werte des Parameters ν_1 zu signifikanten Autokorrelationen führen können. Zur Veranschaulichung ist die Funktion von $f(\nu_1) = \rho_1$ in Abhängigkeit von ν_1 in Abbildung 3.82 dargestellt. Dabei zeigt sich eine gewisse Ähnlichkeit mit der Anzahl der Turning-Points in einer Zeitreihe (s. Abb. 3.80), die offenbar maßgeblich von der Korrelation in der Zeitreihe abhängt.

Die Probleme der nichtparametrischen Verfahren sind durch ähnliche Überlegungen nachvollziehbar. So hängen die Ränge der Beobachtungen sowie die Anzahl der Vorzeichenwechsel oder der Runs im Wesentlichen mit der empirischen Autokorrelation zusammen. Ist diese zu klein, so entspricht die Anzahl der Mediandurchgänge, Nulldurchgänge sowie die sukzessiven Differenzen aufeinanderfolgender Beobachtungen eher einer, die unter der Nullhypothese zu erwarten wäre.

Insgesamt zeigt sich im Fall von MA(1)-Prozessen, dass die Trennschärfen der Testverfahren im Wesentlichen von den Werten des Autokorrelationsparameters ρ_1 abhängen. Da der maximale Wert dieses Parameters im Fall $|\nu_1| = 1$ erreicht wird, ist die Trennschärfe der Testverfahren hier am besten und eine Ablehnung der Nullhypothese erfolgt bei diesen Werten am ehesten. Für kleinere und größere Werte nimmt die Trennschärfe immer weiter ab. Am besten schneiden der DW- sowie der VNRR-Test in diesem Szenario ab, da sie auch im AR(1)-Fall unter Normalbedingung die beste Trennschärfe aufweisen. Da die empirischen Autokorrelationskoeffizienten im MA(1)-Fall nicht abklingen, ist die Trennschärfe des LB-Tests hier im Vergleich etwas schlechter als im AR(1)-Fall und entspricht in etwa der des Runs-Tests. Auffällig ist weiterhin, dass der TP-Test eine deutliche Asymmetrie entwickelt und die Nullhypothese bei positiven Werten von ν_1 häufiger verwerfen kann als bei negativen. Die K-VZ-Tests bilden im MA(1)-Fall das

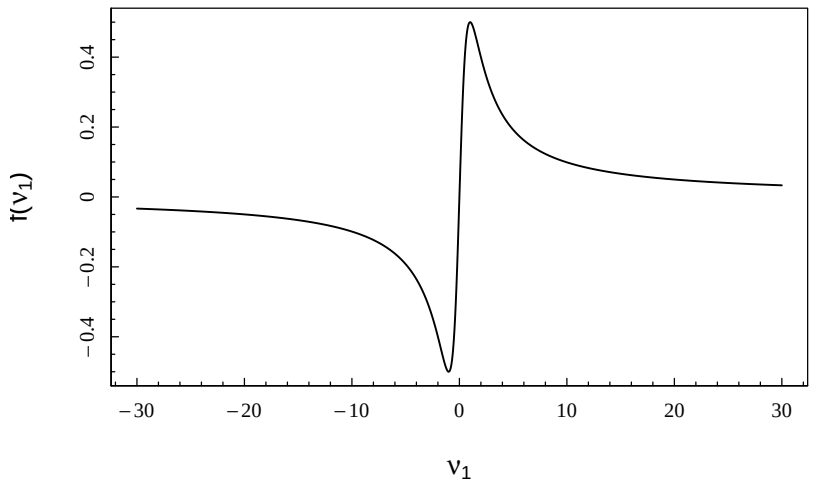


Abbildung 3.82: Funktion des theoretischen Autokorrelationskoeffizienten in MA(1)-Prozessen $f(\nu_1) = \rho_1$ in Abhängigkeit von ν_1

Schlusslicht und sind kaum in der Lage, Abweichungen von der Unabhängigkeit zu detektieren. Dabei ist besonders bemerkenswert, dass sie Alternativen mit positiven Werten des Parameters ν_1 überhaupt nicht erkennen können, während die Nullhypothese bei negativen Werten im Bereich starker Autokorrelationen zumindest bei 50 % der Zeitreihen abgelehnt werden kann. Offenbar stellen sie aber keine gute Wahl für die Erkennung von MA(1)-Prozessen dar, was damit zusammenhängen könnte, dass die Abhängigkeitsstruktur hier noch „lokaler“ ist als im AR(1)-Prozess. So wurde insbesondere erläutert, dass die Autokorrelation hier in jedem Fall nach bereits einem Lag abreißt, sodass die K-VZ-Tests anteilmäßig noch mehr uninformative Tupel zur Entscheidungsfindung heranziehen, als im AR(1)-Fall.

3.5 GARCH(1,1)-Prozesse

Eine weitere Abweichung von der Unabhängigkeitsannahme in einer Zeitreihe stellt eine Korrelation der Fehlerterme eines Prozesses dar. Anders als bisher sollen in diesem Kapitel also Zeitreihen aus Prozessen betrachtet werden, in denen die Varianzen der Innovationen miteinander korreliert sind. Dabei ist es von Interesse, ob die betrachteten Testverfahren in der Lage sind, Abhängigkeitsstrukturen dieser Art anhand von Zeitreihen zu detektieren.

Eine in der Praxis weit verbreitete Klasse von Modellen, in denen die Fehler als korreliert modelliert werden, stellen die GARCH(1,1)-Prozesse dar (s. Kap. 2.2). Besondere Anwendung finden sie bei der Modellierung von Finanzmarkt-Zeitreihen oder Aktienkursen, die in Kriesenperioden typischerweise deutlich stärker schwanken und damit sogenannte Volatillitätscluster aufweisen. Im Folgenden werden stationäre Prozesse der Form:

$$y_t = \epsilon_t, \quad \epsilon_t = \sigma_t w_t, \quad w_t \sim WN(0,1)$$
$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

mit $\omega = 0$ betrachtet. Es kann gezeigt werden, dass die Stationarität eines solchen Prozesses genau dann gegeben ist, wenn $\alpha + \beta < 1$ gilt (Verbeek, 2012, S. 299). Dabei implizieren Werte von $\alpha+\beta$ nahe 1 eine hohe Persistenz in der Volatilität. Eine Zeitreihe der Länge $N = 5000$ eines GARCH(1,1)-Prozesses mit typischen Parameterwerten von $\alpha = 0.09$ und $\beta = 0.9$ ist dabei in Abbildung 3.83 dargestellt.

Um die Trennschärfe der verschiedenen Testverfahren zu beurteilen, werden verschiedene Alternativen solcher Prozesse betrachtet. So wurden die mittleren Ablehnungsraten bei 100 simu-

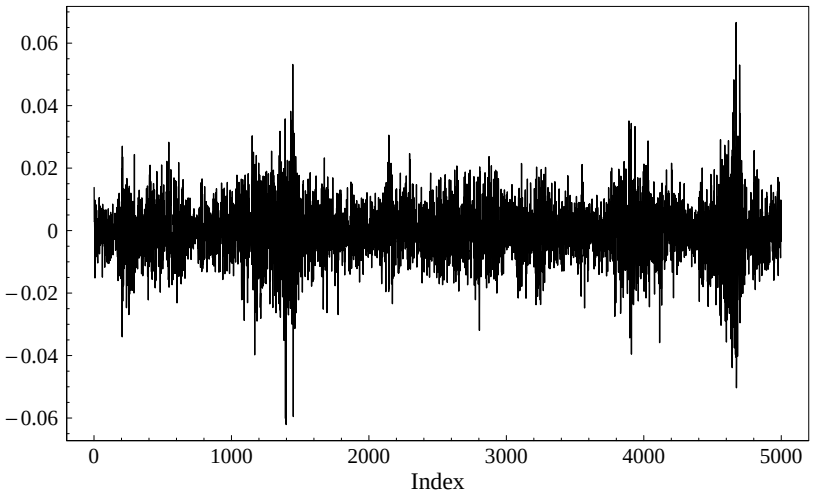


Abbildung 3.83: Zeitreihe der Länge $N = 5000$ von einem GARCH(1,1)-Prozess mit Parameterwerten von $\alpha = 0.09$ und $\beta = 0.9$

lierten Zeitreihen auf einem Gitter für Parameterwerte von $0.05 \leq \alpha \leq 0.1$ und $0.85 \leq \beta \leq 0.98$ mit einer Feinheit von 0.005 ermittelt. Dabei wurde sichergestellt, dass nur Kombinationen betrachtet wurden, für die $\alpha + \beta < 1$ gilt. Bei dieser Auswahl von Parameterwerten erfolgte eine Orientierung an den in der Praxis gängigen Kombinationen. Eine Unabhängigkeit des Prozesses ist dabei in keinem der betrachteten Fälle gegeben, sodass ein Verfahren die Nullhypothese im Idealfall auf dem gesamten Spektrum ablehnen sollte. Die entsprechenden Ergebnisse für Beobachtungszahlen $N = 500$ und $N = 5000$ sind in den Abbildungen 3.84 und 3.85 dargestellt.

Beim Betrachten dieser Grafiken wird deutlich, dass bis auf den Broock-Dechert-Schreinkman-Test keines der Verfahren in der Lage ist, die Abhängigkeitsstruktur eines GARCH(1,1)-Modells sicher als Abweichung von der Unabhängigkeit zu erkennen. Eine gewisse Ausnahme stellt der LB-Test dar, der die Nullhypothese bei $N = 5000$ zumindest bei Prozessen mit hoher Volatilitätspersistenz, also mit Werten, in denen $\alpha + \beta$ nahe 1 liegt, in ca. 50 % der Fälle zu erkennen scheint. Dem BDS-Test hingegen gelingt es, solche Alternativen mit höherer Sicherheit bereits bei $N = 500$ zu erkennen. Bei einer Beobachtungszahl von $N = 5000$ kann er die Nullhypothese auf dem gesamten betrachteten Spektrum zuverlässig ablehnen.

Mit den Ergebnissen aus Kapitel 3.1 und 3.2 sind diese Beobachtungen für den Fall, dass Varianzinhomogenitäten in den Zeitreihen auftreten, gut nachvollziehbar. So wurde dort deutlich, dass diejenigen Testverfahren, die lediglich auf dem sequenziellen Schema der Beobachtungen beruhen, nur im sehr geringen Maße und bei wenigen Beobachtungen von einer Varianzänderung beeinflusst werden. Dies wurde darauf zurückgeführt, dass die Varianz der Innovationen keinen Einfluss auf die Vorzeichen der Beobachtungen hat. Die Wahrscheinlichkeiten, dass Beobachtungen über bzw. unter dem empirischen Median liegen, werden ebenfalls nicht beeinflusst. Der Effekt auf die Ränge der Beobachtungen scheint dabei ebenfalls vernachlässigbar zu sein. Insbesondere sind der LB- und der BDS-Test die einzigen Verfahren, die überhaupt in der Lage sind, wachsende Varianzen als Abweichung von der Zufälligkeit der Zeitreihe zu detektieren.

Im nächsten Schritt soll untersucht werden, ob es dem BDS-Test gelingt, die Nullhypothese bei sehr geringen Volatilitätseffekten bzw. unter der Unabhängigkeit beizubehalten. Dabei liegt eine konstante Varianz − und damit die Unabhängigkeit der Innovationen − in der Zeitreihe genau dann vor, wenn $\alpha = 0$ und $\beta = 1$ sind. Die Simulationsergebnisse auf einem erweiterten Gitter, das diese Alternativen beinhaltet, ist in Abbildung 3.86 dargestellt.

Daraus geht hervor, dass es dem BDS-Test gelingt, das Niveau unter der Unabhängigkeit der Fehler einzuhalten. Offenbar hat er aber auch Schwierigkeiten bei der Ablehnung der Nullhypothese, falls der Parameter α sehr klein (< 0.04) ist. Dies lässt sich damit erklären, dass die Persistenz der Volatilitätscluster für solche Parameterwerte sehr gering ist, sodass die Clusterstruktur in der Zeitreihe schon rein optisch kaum erkennbar ist. Insbesondere nähert sich die Struktur für kleine α immer weiter der Unabhängigkeit an. Zur Veranschaulichung ist eine exemplarische Zeitreihe mit Parameterwerten von $\alpha = 0.01$ und $\beta = 0.98$, zu denen eine Ablehnung der Nullhypothese nicht erreicht werden kann, in Abbildung 3.87 dargestellt.

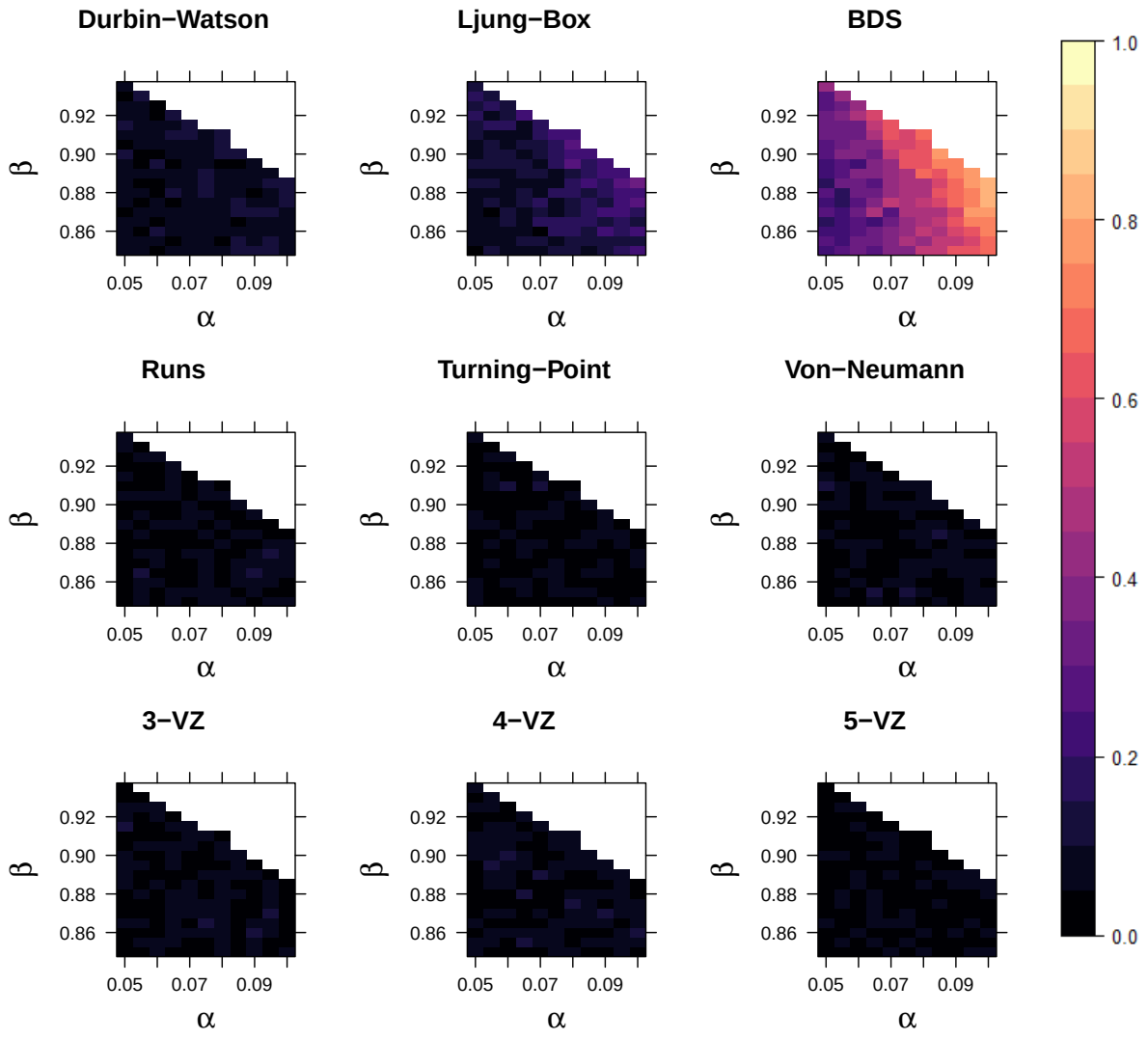


Abbildung 3.84: Simulierte Trennschärfen der Testverfahren bei praxisrelevanten, stationären GARCH(1,1)-Alternativen in Abhängigkeit von α und β bei einer Beobachtungszahl von $N = 500$

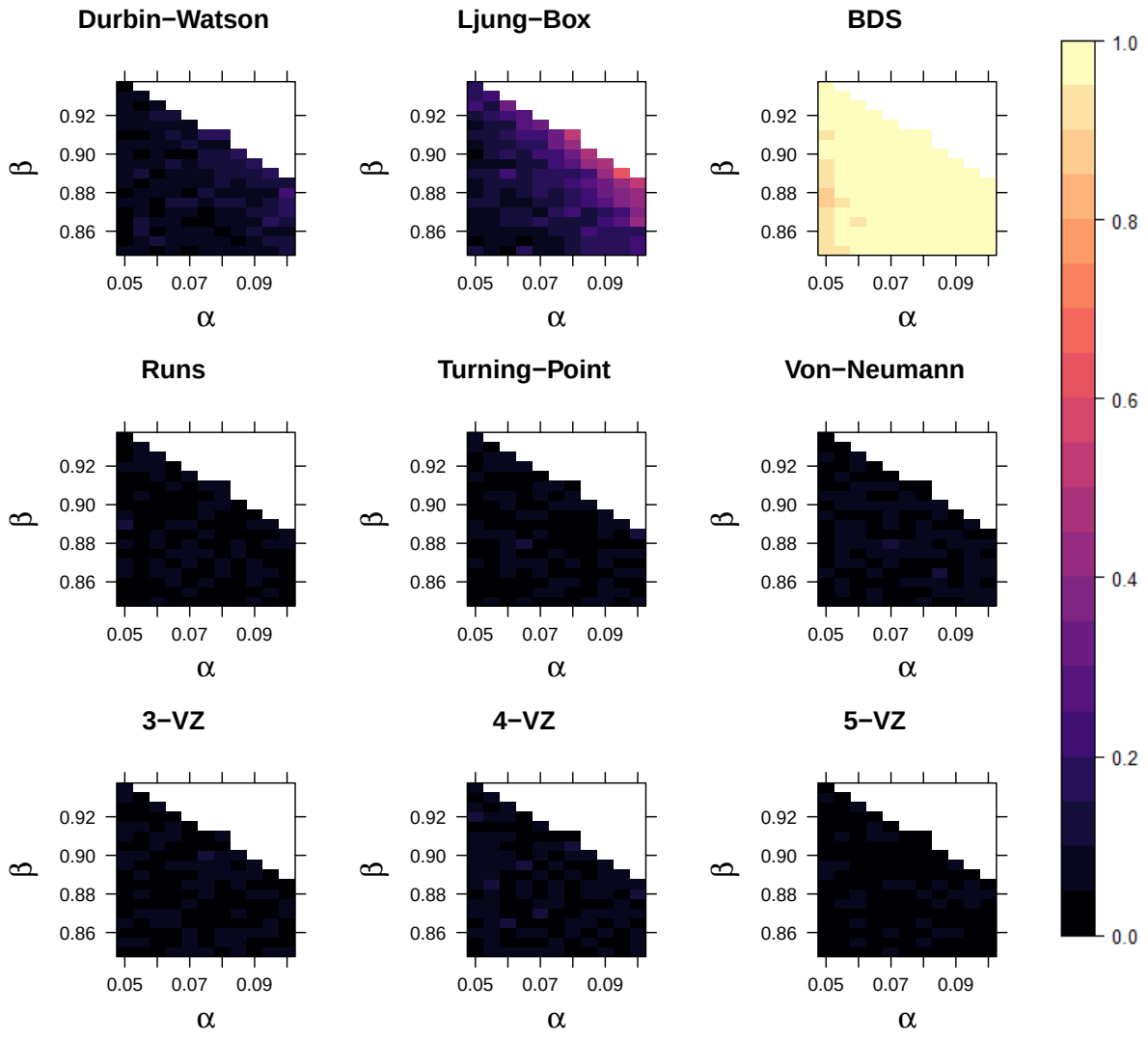


Abbildung 3.85: Simulierte Trennschärfen der Testverfahren bei praxisrelevanten, stationären GARCH(1,1)-Alternativen in Abhängigkeit von α und β bei einer Beobachtungszahl von $N = 5000$

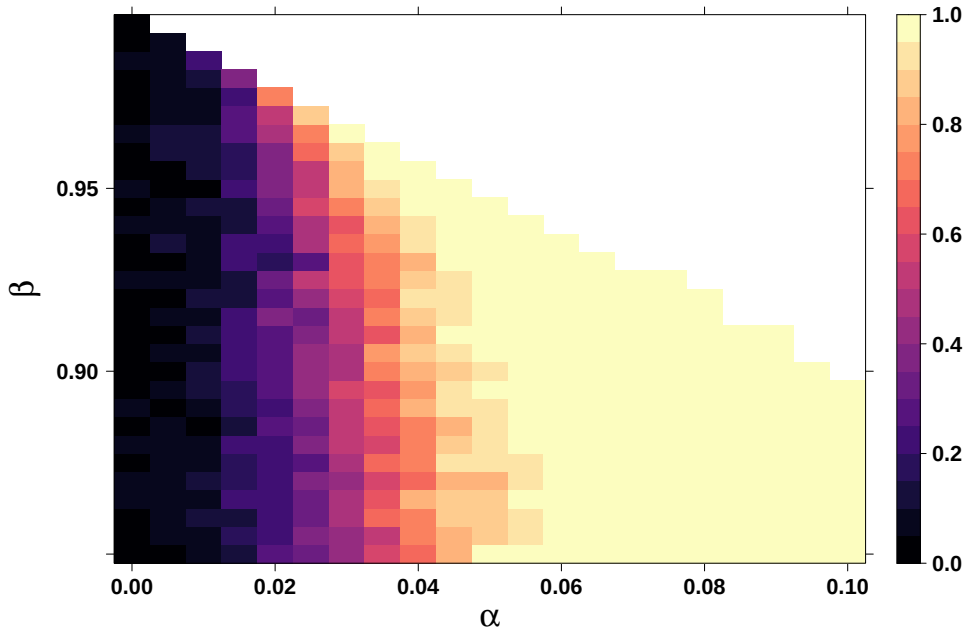


Abbildung 3.86: Simulationsergebnisse des BDS-Tests für ein breites Gitter von stationären GARCH(1,1)-Alternativen, das die Unabhängigkeit beinhaltet ($\alpha = 0$, $\beta = 1$)

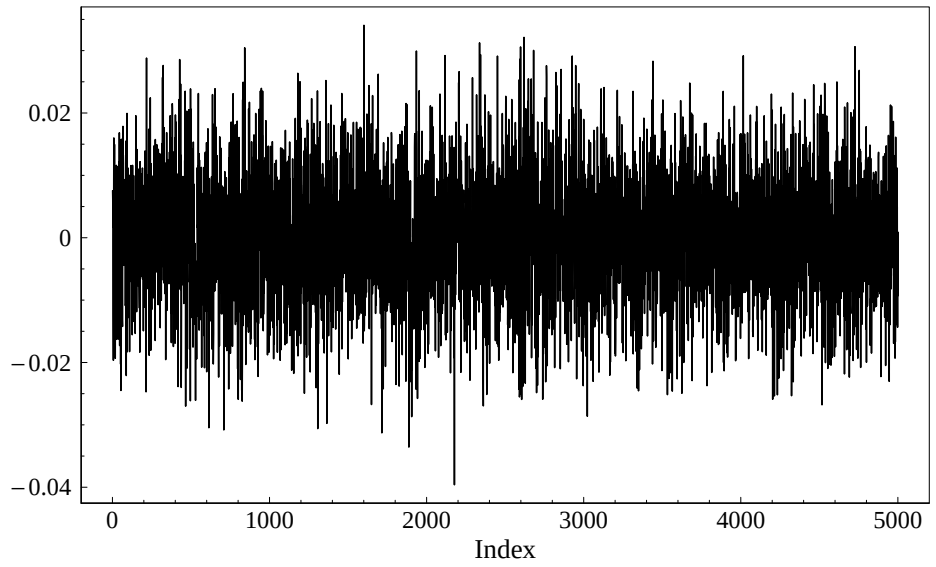


Abbildung 3.87: Zeitreihe eines stationären GARCH(1,1)-Prozesses mit einer Beobachtungszahl von $N = 5000$ und Parameterwerten von $\alpha = 0.01$ und $\beta = 0.98$

Da die Entscheidungsfindung des BDS-Tests maßgeblich von dem Abstand zwischen m-Historien abhängt, der in Volatilitätsclustern offensichtlich viel größer ist als in der restlichen Zeitreihe, wird somit auch nachvollziehbar, warum ihm eine Ablehnung der Nullhypothese in einem solchen Fall sehr schwer fällt.

Bei diesen Beobachtungen ist zu beachten, dass Parameterkombinationen, die in den Abbildung 3.84 und 3.85 untersucht wurden, praxisrelevante Prozesse umfassen. Somit spielen die Alternativen, in denen der BDS-Test keine Ablehnung der Nullhypothese erreichen kann (s. Abb. 3.86) in der Praxis keine Rolle und die gewonnenen Erkenntnisse sind lediglich von theoretischem Interesse.

Insgesamt zeigt sich in diesem Abschnitt, dass die Miteinbeziehung von den konkreten Größen der Beobachtungen bzw. von Abständen zwischen Beobachtungen essenziell für die Erkennung von Strukturen in Zeitreihen ist, in denen Varianzänderungen vorliegen. So gelingt es bei GARCH(1,1)-Prozessen lediglich dem BDS-Test befriedigende Ergebnisse zu liefern. Der LB-Test scheint zwar auch in der Lage zu sein, derartige Stukturen zu erkennen, seine Güte ist der des BDS-Tests jedoch selbst bei großen Beobachtungsumfängen weit unterlegen. Auch die Ergebnisse im Fall, dass wachsende Varianzen in der Zeitreihe vorhanden sind, legen nahe, dass der LB-Test keine gute Wahl zur Detektion von Varianzänderungen in der Zeitreihe ist.

3.6 Weitere Untersuchungen zu den K-Vorzeichentiefetests

In diesem Abschnitt sollen die K-Vorzeichentiefetests im Hinblick auf die in dieser Arbeit gewonnenen Erkenntnisse zu ihrer Trennschärfe weiter untersucht werden. So fiel einerseits auf, dass die Tests dazu neigen, eine Asymmetrie in ihrem Ablehnungsverhalten zu entwickeln und positive Korrelationen leichter zu erkennen als negative. Auf der anderen Seite wurde deutlich, dass ein wesentlicher Nachteil der Testverfahren darin besteht, dass sie nur wenig von einem wachsenden Stichprobenumfang profitieren können. Besonders auffällig ist dies, wenn sich die Abhängigkeitsstrukturen in der Zeitreihe wiederholen (vgl. Kap. 3.1.5). Deshalb ist es von Interesse, ob Modifikationen der Ablehnungsbereiche oder Teststatistiken von den K-VZ-Tests dazu führen können, dass sie eine bessere Trennschärfe aufweisen und somit für die praktische Anwendung an Relevanz gewinnen. Dabei beschränken sich die folgenden Untersuchungen auf den AR(1)-Fall unter Normalbedingungen, wie er in Kapitel 3.1 für die herkömmliche, zweiseitige Version der K-VZ-Tests behandelt wurde.

3.6.1 Asymmetrie

Zunächst steht die asymmetrische Trennschärfe der Testverfahren im Fokus der Untersuchungen. Diese ist vermutlich auf die asymmetrische Verteilung ihrer Teststatistiken zurückzuführen. So wurde bereits in Kapitel 2.9 erwähnt, dass eine symmetrische Wahl der kritischen Werte aufgrund der Form von den Dichteverteilungen nicht optimal erscheint. Zur Veranschaulichung sind die Dichten der K-Vorzeichentiefen für $K \in \{3, 4, 5\}$ in Abbildung 3.88 dargestellt.

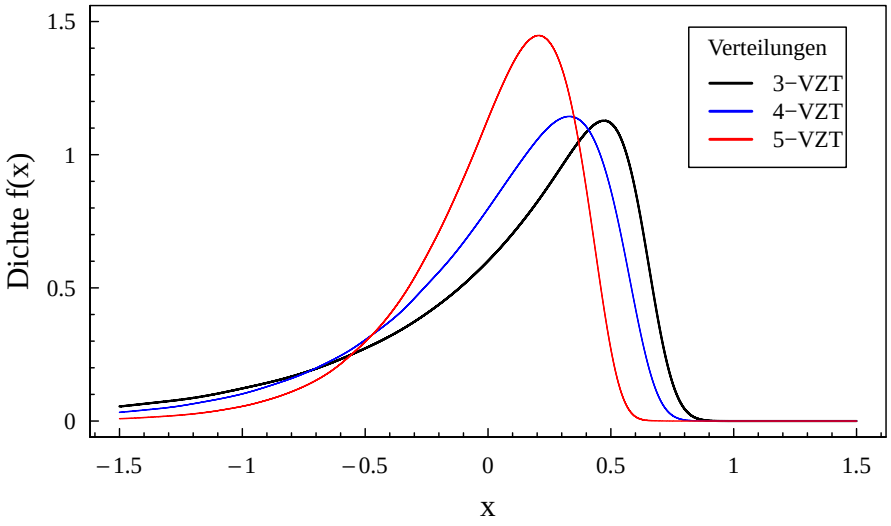


Abbildung 3.88: Vergleich der Dichtefunktionen von der 3-, 4-, und 5-Vorzeichentiefe

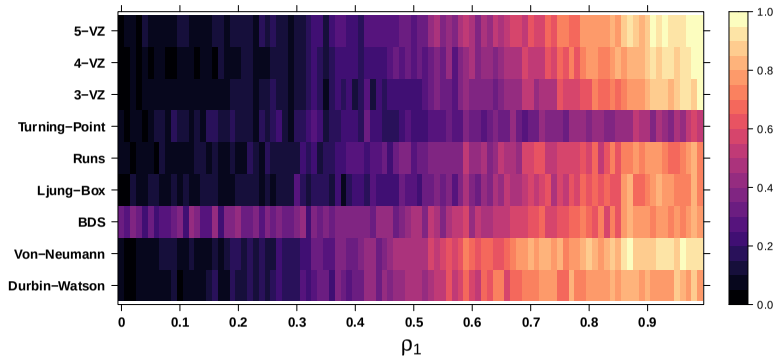
Diese Dichtefunktionen wurden für jedes K mithilfe der Funktion `ddepth` aus dem Paket `GSignTest` von Horn (2020) erzeugt. Dabei wurde die entsprechenden Werte von den Dichtefunktionen für jeden x-Wert in einem Intervall von -1.5 bis 1.5 auf einem gleichmäßigen Gitter der Feinheit 0.0001 ermittelt.

An dieser Stelle erscheint es sinnvoll, zu untersuchen, wie sich die Trennschärfen der verschiedenen Testverfahren in einer einseitigen Version des Unabhängigkeitstests verhalten, bei der die Asymmetrien ihrer Dichten keine Rolle spielen sollten. Dazu wird die Nullhypothese, dass keine positiven Korrelationen in der Zeitreihe vorliegen, also $H_0 : \rho_1 \leq 0$, gegen die Alternative $H_1 : \rho_1 > 0$ getestet. Die Ergebnisse dieser Simulation zeigt Abbildung 3.89.

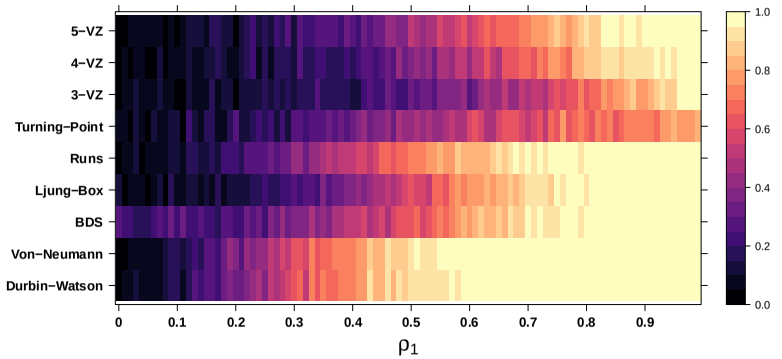
Hier wird noch einmal deutlich, dass die K-VZ-Tests den anderen Verfahren in kleinen Stichproben bezüglich ihrer Trennschärfen in nichts nachstehen. So ist ihre Trennschärfe in der einseitigen Version, zusammen mit der des VNRR-Tests, bei einem Stichprobenumfang von $N = 20$, unter allen betrachteten Tests am besten. Gleichzeitig gelingt es den Testverfahren, das Niveau unter der Unabhängigkeit einzuhalten. Bis zu einem Stichprobenumfang von $N = 50$ kann der 5-VZ-Test noch mit dem TP-Test mitzuhalten und erst für $N \geq 100$ schneiden die K-VZ-Tests am schlechtesten ab. Dieses verbesserte Verhalten gegenüber der zweiseitigen Version des Tests untermauert noch einmal die Vermutung, dass es Sinn macht, die Annahmebereiche der K-VZ-Tests entsprechend den asymmetrischen Verteilungen ihrer Teststatistiken anzupassen.

Ein Ansatz, um mit der Asymmetrie der K-Vorzeichentiefen umzugehen − falls eine zweiseitige Version des Unabhängigkeitstests von Interesse ist − stellt eine alternative Wahl der Ablehnungsbereiche von den K-VZ-Tests dar. Ein weitverbreitetes Vorgehen in solchen Fällen, wie z. B. bei der $\mathcal{X}^2$-Verteilung üblich, ist es, die Länge des Annahmebereichs zu minimieren. Dazu wurden verschiedene Alternativen für geeignete Quantile auf einem Gitter mit einer Feinheit von 0.00001 ausgewertet, wobei das obere und untere Quantil so variiert wurden, dass das Niveau $\alpha = 0.05$ stets eingehalten wird. Diejenige Konstellation, die zu dem Annahmebereich kürzester Länge führt, wurde dabei für jedes K separat ermittelt. Weiter sind die kritischen Werte zu den ermittelten unteren Quantilen sowie die Länge des kürzesten Annahmebereichs in Abhängigkeit von den verschiedenen Stichprobenumfängen N und dem Parameter K in der Tabelle 3.1 angegeben.

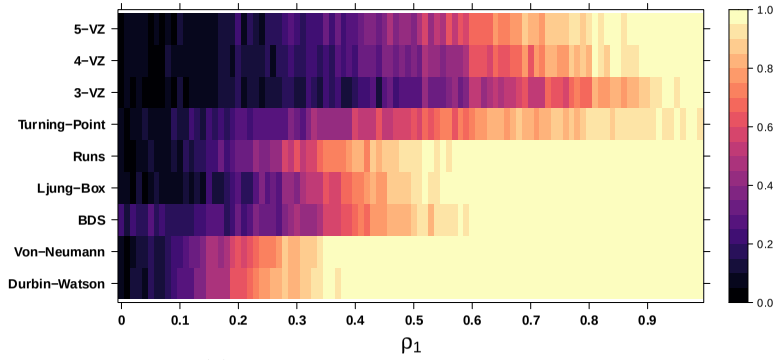
Dabei fällt vor allem auf, dass das untere Quantil zu dem kleinsten Annahmebereich im Fall $K = 3$ sehr klein ist (0.00095), mit steigendem K aber zunehmend etwas größer wird (0.00235 für $K = 4$ und 0.00355 für $K = 5$). Diese Systematik ist auf die Form der entsprechenden Dichten zurückzuführen, die in Abbildung 3.88 dargestellt sind. So fällt bei einer erneuten Betrachtung auf, dass sich die Schiefe der Verteilung mit zunehmenden K verringert und sich die Wahrscheinlichkeitsdichteverteilung der Symmetrie annähert. Somit wird die Systematik, dass sich der Annahmebereich mit minimaler Länge für größere Ks weiter nach rechts verschiebt, nachvollziehbar. Auch verändern sich die kritischen Werte − und damit die Länge der Annahmebereiche − mit zunehmendem Stichprobenumfang immer weniger.



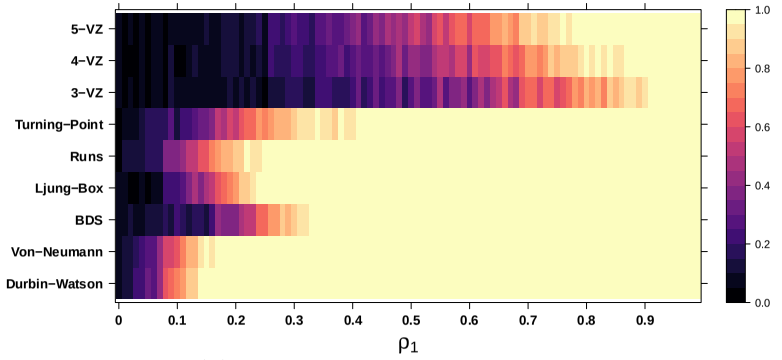
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



(d) Stichprobenumfang $N = 500$

Abbildung 3.89: Simulierte Trennschärfe der einseitigen Testverfahren bei stationären AR(1)-Alternativen unter Normalbedingungen für unterschiedliche Stichprobenumfänge

Tabelle 3.1: Untere kritische Werte (u. krit.) und Länge der Annahmeintervalle (Länge) zum breitesten Ablehnungsbereich der K-VZ-Tests in Abhängigkeit vom Stichprobenumfang N und dem Parameter K

	$N = 20$		$N = 50$		$N = 100$		$N = 500$	
	u. krit.	**Länge**	**u. krit.**	**Länge**	**u. krit.**	**Länge**	**u. krit.**	**Länge**
$K = 3$	-3.46	2.11	-3.54	2.27	-3.55	2.30	-3.55	2.30
$K = 4$	-2.00	0.96	-2.17	1.19	-2.22	1.25	-2.22	1.25
$K = 5$	-1.15	0.41	-1.33	0.63	-1.37	0.69	-1.37	0.69

Nun soll überprüft werden, inwieweit sich die auf diese Weise angepassten Ablehnungsbereiche auf die K-VZ-Tests auswirken. Dazu sind die so erhaltenen Trennschärfen den ursprünglichen Trennschärfen, die aus einer symmetrischen Wahl der Ablehnungsbereiche resultieren, in Abbildung 3.90 gegenübergestellt.

Aus dieser Gegenüberstellung geht hervor, dass die Trennschärfen aller K-VZ-Tests − zumindest für kleine Stichprobenumfänge von $N = 20$ und $N = 50$ − mit den angepassten kritischen Werten deutlich symmetrischer erscheinen. Für größere Beobachtungszahlen wird aber offensichtlich, dass die Tests positive Korrelationen trotz der Modifikationen noch besser erkennen können als negative. Eine leichte Verbesserung der Symmetrie der Ablehnungsbereiche gegenüber der bei einer Wahl der herkömmlichen Quantilen ist dort jedoch weiterhin erkennbar. Ein wesentlicher negativer Gesichtspunkt bei der alternativen Wahl der Ablehnungsbereiche scheint jedoch zu sein, dass der Bereich, in dem die Nullhypothese in weniger als ca. 5 % der Fälle stattfindet, deutlich in den Bereich positiver Korrelationen verschoben wird. Während dieser bei symmetrischen Quantilen noch relativ zentral um 0 liegt, befindet er sich bei der Modifikation für alle betrachteten Stichprobenumfänge und alle Ks deutlich weiter rechts (s. Abb. 3.90). Damit verbessert sich zwar die Fähigkeit der Tests, negative Korrelationen zu detektieren, allerdings werden positive Korrelationen von bis zu $\rho_1 = 0.5$ bei $K = 3$ bzw. $\rho_1 = 0.35$ bei $K = 5$ nicht mehr als Abweichungen von der Unabhängigkeit erkannt. So wird die Nullhypothese bei den oben genannten Werten von ρ_1 mit der Modifikation in weniger als 5 % der Simulationen abgelehnt, während dies mit den ursprünglichen kritischen Werten zumindest in ca. 50 % der Fälle gelungen ist.

Aufgrund dieser nicht wünschenswerten Eigenschaft scheint eine alternative Wahl des Ablehnungsbereichs nach dem hier beschriebenen Prinzip keine deutliche Verbesserung der Trennschärfen zu bewirken. Insbesondere führt sie dazu, dass die Fähigkeit der K-VZ-Tests, positive Korrelationen zu detektieren, deutlich abnimmt. Die vorgestellte Modifikation ist allerdings eine sinnvollere Vorgehensweise als die, symmetrische Quantile zu betrachten. Für weitere Untersuchungen der K-VZ-Tests sollte also ein anderer Ansatz gewählt werden.

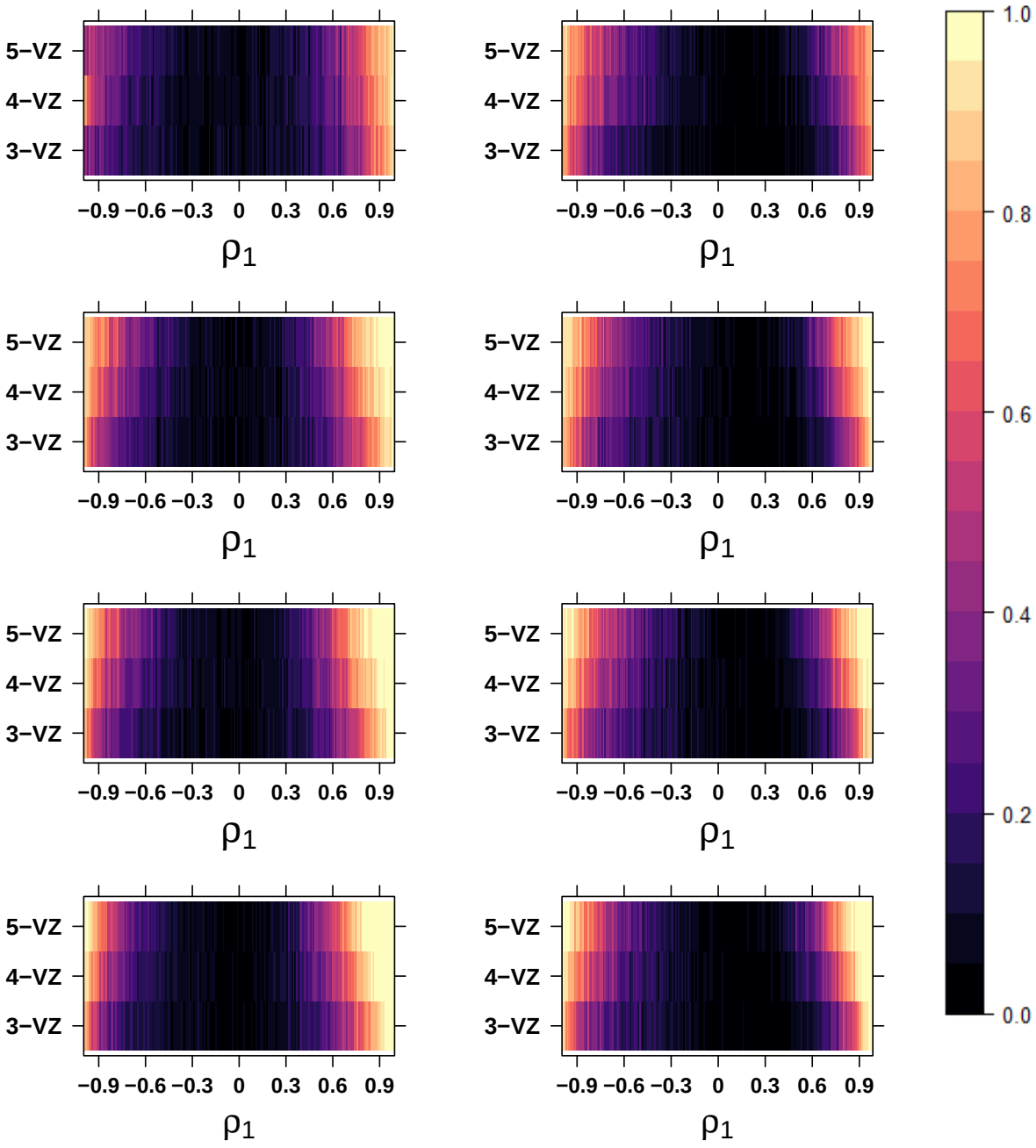


Abbildung 3.90: Simulierte Trennschärfen der K-VZ-Tests für die ursprünglichen (links) und modifizierten Quantile (rechts) in Abhängigkeit von ρ_1, für Stichprobenumfänge von $N = 20, 50, 100, 500$ (von oben nach unten)

3.6.2 Tests basierend auf der vereinfachten K-Vorzeichentiefe

Eine Erklärung dafür, dass die K-VZ-Tests nicht von einem wachsenden Stichprobenumfang profitieren, ist, dass mit wachsender Beobachtungszahl immer mehr Tupel betrachtet werden, die wenig Informationen über die Abhängigkeit des Prozesses beinhalten. So machen sich die Abhängigkeitsstrukturen der in dieser Arbeit betrachteten Prozesse hauptsächlich lokal bemerkbar, also bei kurz aufeinanderfolgenden Beobachtungen. Die Autokorrelationen in den AR(1)-Prozessen klingen mit zunehmendem Abstand exponentiell ab (vgl. Kap. 3.1, 3.2 u. 3.3) und bei MA(q)-Prozessen reißt die ACF sogar nach q Lags ab (vgl. Kap. 3.4). Dementsprechend scheint es zur Aufdeckung dieser Strukturen wenig zielführend, Vorzeichenwechsel von Tupeln aus weit auseinanderliegenden Beobachtungen zur Entscheidungsfindung heranzuziehen. Zumal der Anteil dieser nicht-informativen Tupel mit wachsendem Stichprobenumfang zunimmt.

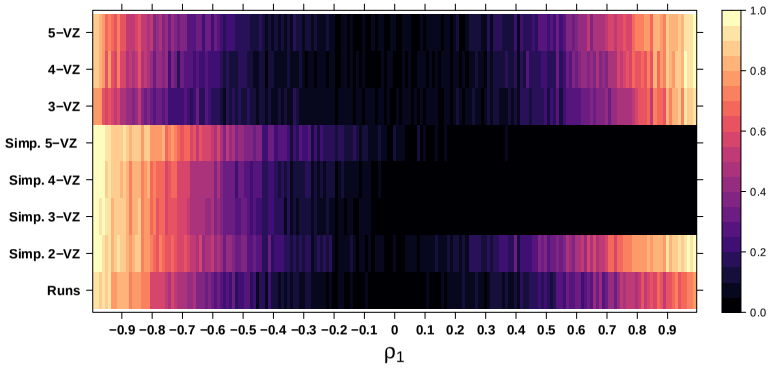
Es erscheint deshalb sinnvoll, die vereinfachten Versionen der K-VZ-Tests − wie sie bereits in Kapitel 2.9 vorgestellt wurden − zu betrachten und ihre Trennschärfen zu untersuchen. Bei diesen Tests werden nicht alle K-Tupel der Zeitreihe auf alternierende Vorzeichen untersucht, sondern lediglich aufeinanderfolgende Beobachtungen, wobei Überlappungen der betrachteten Tupel erlaubt sind. So basieren diese Tests, abhängig von dem gewählten K, auf dem relativen Anteil von K aufeinanderfolgenden Beobachtungen, deren Vorzeichen alternieren. Im Folgenden werden diejenigen K-Tupel, auf die das zutrifft, als K-Vorzeichenwechsel (K-VZW) bezeichnet.

3.6.2.1 Vergleich mit den vollständigen K-Vorzeichentiefetests

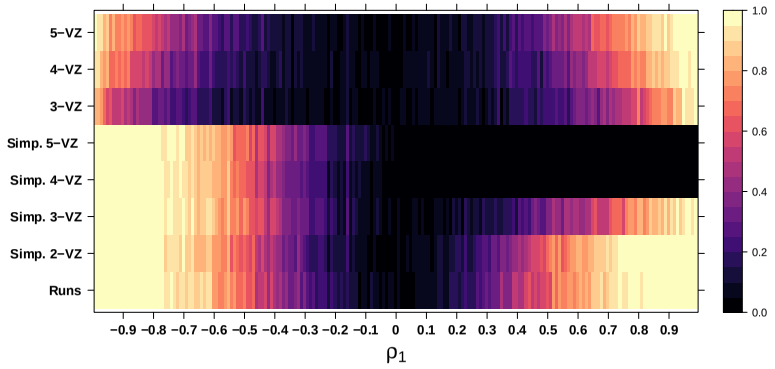
Diejenige Testgröße, auf der der vereinfachte 2-VZ-Test basiert, entspricht dabei der Anzahl der Vorzeichenwechsel − und damit der Runs − und die beiden Testverfahren sollten deshalb zumindest ähnliche Ergebnisse liefern (vgl. Kap. 2.9). Um die Trennschärfen der vereinfachten Versionen der K-VZ-Tests beurteilen zu können, wurden sie zunächst denen der vollständigen 3-, 4- und 5-VZ-Tests sowie der des Runs-Tests in Abbildung 3.91 für unterschiedliche Stichprobenumfänge gegenübergestellt.

Beim Betrachten der Ergebnisse fällt direkt ins Auge, dass die vereinfachten Versionen der K-VZ-Tests für geringe Beobachtungszahlen nicht in der Lage sind, Alternativen mit positiven Korrelationen abzulehnen. So gelingt es bei einem Stichprobenumfang von $N = 20$ lediglich dem vereinfachten 2-VZ-Test, positive Korrelationen zu detektieren. Ab einer Beobachtungszahl von $N = 50$ ist auch der vereinfachte 3-VZ-Test in der Lage, die Nullhypothese für solche Alternativen zu verwerfen. Damit der vereinfachte 4-VZ-Test eine Ablehnung erreichen kann, sind bereits $N = 100$ Beobachtungen nötig. Für den 5-VZ-Test muss hier schon ein Stichprobenumfang von $N = 500$ Beobachtungen gewählt werden.

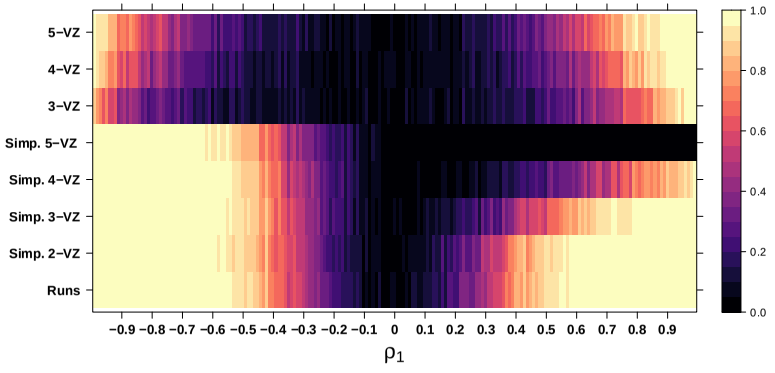
Weiter fällt auf, dass die Trennschärfen der vereinfachten K-VZ-Tests auch in Bereichen mit hinreichend großem Stichprobenumfang deutliche Asymmetrien aufweisen, die mit wachsendem K immer stärker werden. Dabei fällt ihnen die Erkennung von Abhängigkeitsstrukturen mit $\rho_1 < 0$ leichter.



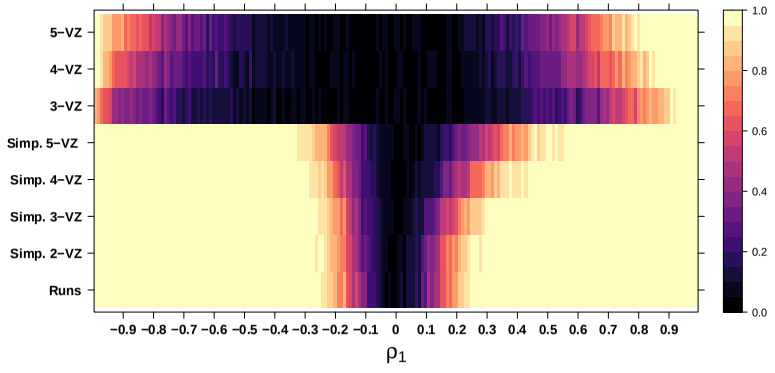
(a) Stichprobenumfang $N = 20$



(b) Stichprobenumfang $N = 50$



(c) Stichprobenumfang $N = 100$



(d) Stichprobenumfang $N = 500$

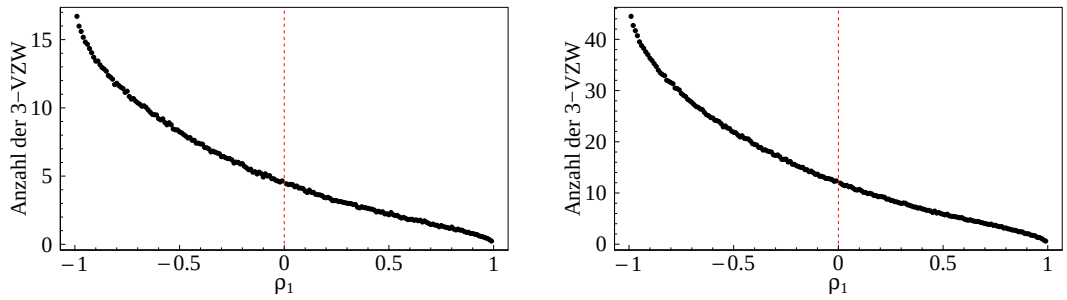
Abbildung 3.91: Simulierte Trennschärfen der vereinfachten K-VZ-Tests (Simp. K-VZ) im Vergleich mit ihren vollständigen Gegenstücken und dem Runs-Test, bei stationären AR(1)-Alternativen für unterschiedliche Stichprobenumfänge

In Hinblick auf den vereinfachten 2-VZ-Test wird die Ähnlichkeit zur Trennschärfe des Runs-Tests anhand der betrachteten Ergebnisse offensichtlich. Bereits ab einem Stichprobenumfang von $N = 50$ erscheinen ihre Trennschärfen weitestgehend identisch und die beiden Verfahren erzielen unter allen abgebildeten Tests die besten Ergebnisse. Bei einem kleinen Stichprobenumfang von $N = 20$ scheint der 2-VZ-Test dem Runs-Test sogar überlegen zu sein.

Eine weitere auffällige Systematik in Abbildung 3.91 bezieht sich auf den Einfluss des Wertes von K auf die Trennschärfe der K-VZ-Tests. Während die regulären K-VZ-Tests mit größerem K bessere Ergebnisse liefern, ist bei den vereinfachten Versionen ein gegenteiliger Effekt erkennbar. Neben der Tatsache, dass hier größere Ks einen größeren Stichprobenumfang benötigen, um überhaupt eine Ablehnung der Nullhypothese bei positiven Korrelationen erreichen zu können, nimmt die Trennschärfe im Fall von vielen Beobachtungen ($N = 500$) vor allem im Bereich positiver Korrelationen deutlich sichtbar ab.

Um die Beobachtungen zu der Asymmetrie und der Unfähigkeit, in kleinen Stichproben positive Korrelationen zu detektieren, zu verstehen, wurde exemplarisch für $K = 3$ untersucht, wie sich die Anzahl der 3-Vorzeichenwechsel (3-VZW) in Zeitreihen in Abhängigkeit von ρ_1 verhält. Zunächst ist dafür die mittlere Anzahl der 3-VZW für den Fall $N = 20$ und $N = 50$ anhand von 1000 Simulationen für jeden Wert von ρ_1 auf einem Gitter der Feinheit 0.01 berechnet worden. Die Ergebnisse sind in Abbildung 3.92 dargestellt.

Aus diesen Grafiken geht hervor, dass sich die Anzahl der 3-VZW in Abhängigkeit von ρ_1 in ähnlicher Weise asymmetrisch verhält, wie es schon bei der Anzahl der Turning-Points in Kapitel 3.1 beobachtet wurde. So fällt ihre Anzahl bei $N = 20$ für positive Korrelationen linear bis auf einen minimalen Wert von 0 ab, während sie für negative Korrelationen zumindest polynomiell bis auf einen maximalen Wert von ca. 17 ansteigt. Der mittlere Wert der 3-VZW unter der Unabhängigkeit beträgt dabei ungefähr 4.5 und entspricht damit dem Erwartungswert der vereinfachten 3-Vorzeichentiefe von (20 - 2) $\cdot$ $(1/2)^2$. Diese Systematiken bleiben auch für



(a) 3-VZW, $N = 20$ (b) 3-VZW, $N = 50$

Abbildung 3.92: Darstellung der mittleren Anzahl von 3-VZW in Abhängigkeit von ρ_1, basierend auf 1000 Simulationen bei $N = 20$ (a) und $N = 50$ Beobachtungen (b)

$N = 50$ erhalten und lediglich die maximale Anzahl sowie die Anzahl von 3-VZW unter der Unabhängigkeit ist deutlich höher. Auf diese Weise lässt sich die Asymmetrie der Trennschärfe des Tests ähnlich wie im Kapitel 3.1 für den TP-Test erklären.

Um nun zu verstehen, warum für positive Korrelationen bei einer zu kleinen Beobachtungszahl keine Ablehnung der Nullhypothese erfolgen kann, wurde die Anzahl der 3-VZW, ab denen eine Verwerfung der Nullhypothese stattfindet, für ausgewählte Stichprobenumfänge berechnet. Dabei entspricht das 0.025-Quantil der zugehörigen asymptotischen Normalverteilung bei $N = 20$ ca. -0.148 und für $N = 50$ in etwa 4.4. Da eine negative Anzahl an 3-VZW jedoch offensichtlich nicht auftreten kann, ist auch verständlich, warum eine Ablehnung der Nullhypothese hier nicht erfolgt. Die Tatsache, dass sie im Fall $N = 50$ möglich ist und für hinreichend große Korrelationen auch stattfindet, wird ebenfalls ersichtlich. Diese Erkenntnisse lassen sich gleichermaßen für die anderen vereinfachten K-VZ-Tests nachvollziehen, bei denen eine Ablehnung der Nullhypothese für positive Korrelationen erst bei größeren Stichprobenumfängen möglich wird. Somit muss bei einer Anwendung dieser Testverfahren bedacht werden, dass der Beobachtungsumfang hinreichend groß sein muss, damit eine Ablehnung der Nullhypothese zum angestrebten Signifikanzniveau überhaupt möglich ist. Um dieses Problem zu umgehen, könnten bei einer praktischen Anwendung dieser Testverfahren − zumindest bei kleinen Stichproben − auch die exakten Verteilungen der vereinfachten K-Vorzeichentiefen ermittelt und zur Fällung einer Testentscheidung herangezogen werden.

Um nachzuvollziehen, warum die vereinfachten K-VZ-Tests mit größer werdendem K im Bereich positiver Korrelationen an Trennschärfe verlieren, wurden die mittleren Anzahlen an 3-, 4-, und 5-VZW in Abhängigkeit von ρ_1 bei einer Gitterfeinheit von 0.01 auf der Grundlage von 1000 simulierten Zeitreihen bei $N = 500$ in Abbildung 3.93 dargestellt.

Daraus geht hervor, dass die Differenz zwischen der Anzahl an K-VZW unter der Unabhängigkeit und der minimalen Anzahl von K-VZW für größere Ks zunehmend geringer wird. Dies ist nachvollziehbar, da auch die erwartete Anzahl an Vorzeichenwechseln genau $(N - K)(1/2)^K$ entspricht. Die maximale Anzahl an Vorzeichenwechsel, die durch Korrelationen verursacht werden kann, bleibt jedoch für alle Ks konstant 0. Damit nimmt die Anzahl der K-VZW für positive Korrelationen mit größer werdendem K langsamer ab, als sie für negative Korrelationen zunimmt. Dieses Verhalten ist damit zu erklären, dass ein negatives ρ_1 ein alternierendes Verhalten der Zeitreihe maßgeblich beeinflusst. Das führt dazu, dass die Anzahl der K-VZW mit kleineren Werten des Parameters zunimmt. Auf der anderen Seite ist der Einfluss einer positiven Korrelation auf die K-VZW deutlich schwächer und ein Entstehen solcher Alternationen wird nicht im gleichen Maße verhindert, wie es durch negative Korrelationen verstärkt wird. Dieser Effekt wird offensichtlich mit größer werdendem K immer stärker. Da die kritischen Werte aus einer symmetrischen Verteilung ermittelt werden, können die Verfahren negative Korrelationen leichter detektieren als positive. Insbesondere nimmt damit die Fähigkeit letztere zu erkennen − und damit die Trennschärfe in diesem Bereich − für größere K deutlich ab.

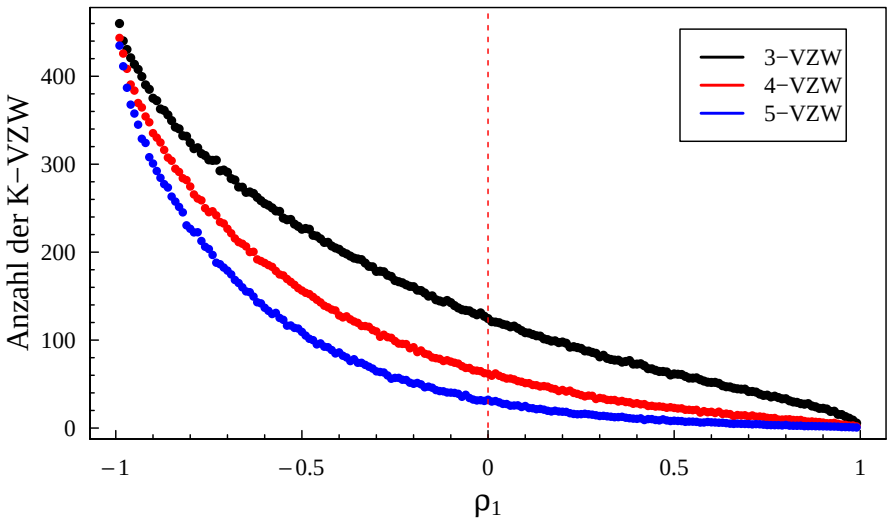


Abbildung 3.93: Simulierte mittlere Anzahl an 3-, 4- und 5-VZW in Abhängigkeit vom Parameter ρ_1 bei einem Stichprobenumfang von $N = 500$

3.6.2.2 Parallelen zu anderen Tests

Im Folgenden sollen die Parallelen zwischen dem vereinfachten 2-VZ-Test und dem Runs-Test weiter untersucht werden. An dieser Stelle ist zu bedenken, dass bei der Testentscheidung des Runs-Tests auf die Anzahl der Beobachtungen, die über dem Median liegen, bedingt wird. Dies ist bei dem vereinfachten 2-VZ-Test nicht der Fall. Somit ist auch von geringen Abweichungen zwischen Annahme- und Ablehnungsbereichen der Testverfahren − wie sie bereits beobachtet wurden − auszugehen. Eine zusätzliche Bedingung auf die Anzahl der positiven Beobachtungen beim vereinfachten 2-VZ-Test erscheint demnach naheliegend und sinnvoll. Dabei wäre davon auszugehen, dass diese Version des Testverfahrens Ergebnisse liefert, die denen des Runs-Tests noch ähnlicher sind als die des unbedingten Tests. Die kritischen Werte für die bedingte Version wurden dabei auf der Grundlage von 10 000 unabhängigen, simulierten Zeitreihen mit entsprechender Anzahl positiver Beobachtungen empirisch ermittelt. Die Trennschärfen des Runs-, des vereinfachten 2-VZ- und des bedingten vereinfachten 2-VZ-Tests sind einander für unterschiedliche Stichprobenumfänge in Abbildung 3.94 gegenübergestellt.

Anhand der Grafik lässt sich erkennen, dass die 3 Varianten für alle betrachteten Stichprobenumfänge allgemein sehr ähnliche Ergebnisse liefern. Mit wachsendem Beobachtungsumfang werden sich ihre Trennschärfen dabei immer ähnlicher, sodass bei $N = 500$ kaum noch Unterschiede zwischen den Tests erkennbar sind. Für $N = 20$ hingegen scheint die unbedingte Version des vereinfachten 2-VZ-Tests die besten Ergebnisse zu liefern. Bei Stichprobenumfängen

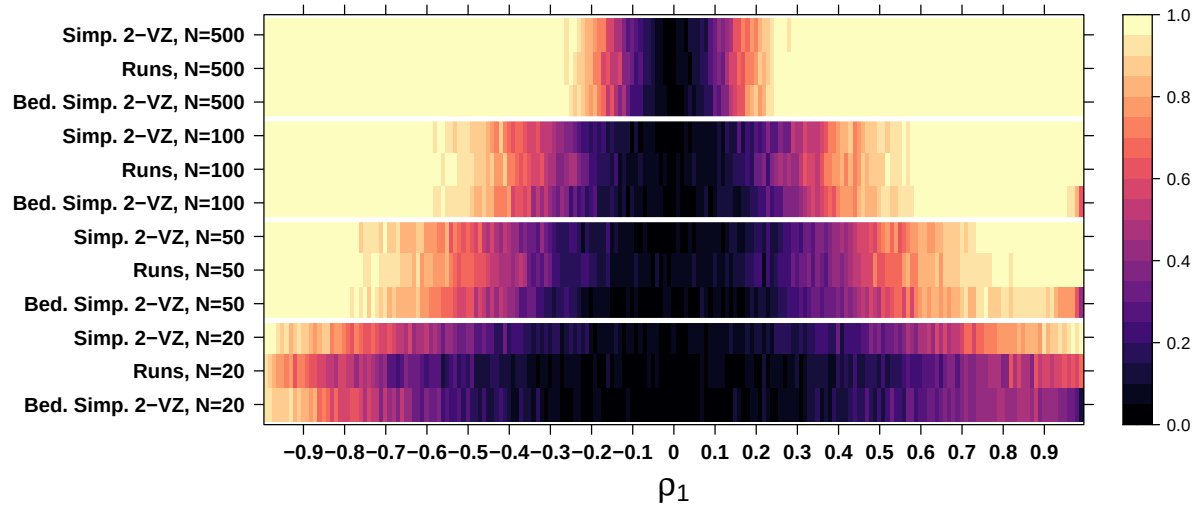


Abbildung 3.94: Vergleich der Trennschärfen des bedingten (Bed. Simp.) und unbedingten, vereinfachten (Simp.) 2-VZ-Tests sowie des Runs-Tests für unterschiedliche Stichprobenumfänge in Abhängigkeit von ρ_1

von $N = 50$ und $N = 100$ wirkt es, als wäre der Runs-Test den anderen Testverfahren etwas überlegen.

Eine weitere Auffälligkeit im Bezug auf die bedingte Version des vereinfachten 2-VZ-Test ist, dass die Nullhypothese bei Stichprobenumfängen von bis zu $N = 100$ im Bereich extrem positiver Korrelationen nicht verworfen werden kann. Dies ist damit zu begründen, dass Beobachtungen aus Prozessen mit derartig starken Korrelationsstrukturen dazu neigen, sehr ähnliche Werte anzunehmen. So kommt es hier häufig vor, dass kein einziger Vorzeichenwechsel auftritt. In einem solchen Fall sind demzufolge entweder alle oder aber keine der Beobachtung positiv. Deshalb können durch das oben erläuterte Vorgehen zur Bestimmung der kritischen Werte auch keine gültigen Quantile ermittelt werden. Die Nullhypothese der Unabhängigkeit kann in diesen Fällen also nicht abgelehnt werden. Die Wahrscheinlichkeit, dass kein Vorzeichenwechsel stattfindet, wird dabei jedoch mit wachsendem Stichprobenumfang immer geringer, sodass diese Effekte bei $N = 500$ keine Rolle mehr zu spielen scheinen und eine Ablehnung auf dem gesamten Spektrum erfolgen kann.

Die Tatsache, dass die bedingte Version des vereinfachten 2-VZ-Tests nicht äquivalent zu dem Runs-Test ist, liegt vermutlich daran, dass die Anzahl der Beobachtungen über dem Median in der Stichprobe durch die Zentrierung beim Runs-Test stets entweder $N/2$, $\lfloor N/2 \rfloor$ oder $\lceil N/2 \rceil$ be-

trägt. In der bedingten Version des vereinfachten 2-VZ-Tests kann diese Anzahl jedoch deutlich andere Werte annehmen. Aufgrund der Voraussetzung, dass die Verteilung der Beobachtungen bei den K-VZ-Tests einen Median von 0 aufweist, wird es dabei immer wahrscheinlicher, dass ähnlich viele positive wie negative Werte beobachtet werden. Aus diesem Grund nähern sich die Trennschärfen der Testverfahren auch immer weiter an. Damit die Teststatistiken zu jeglichen Stichprobenumfängen äquivalent sind, müsste vor dem Bedingen eine Zentrierung der Datenpunkte durch den empirischen Median vorgenommen werden.

Die bereits erwähnten Parallelen zwischen der Anzahl der 3-VZW und den Turning-Points in einer Zeitreihe können durch eine genauere Betrachtung dieser Testgrößen nachvollzogen werden. So ist zunächst offensichtlich, dass ein 3-VZW in einer Zeitreihe impliziert, dass bei den 3 betrachteten Beobachtungen ein Turning-Point auftritt, denn es gilt für ein $i \in \{1, ..., N \text{ - } 2\}$:

$$x_i < 0 \wedge x_{i+1} > 0 \wedge x_{i+2} < 0 \Rightarrow x_i < x_{i+1} \wedge x_{i+1} > x_{i+2} \quad \text{sowie}$$
$$x_i > 0 \wedge x_{i+1} < 0 \wedge x_{i+2} > 0 \Rightarrow x_i > x_{i+1} \wedge x_{i+1} < x_{i+2}.$$

Eine Umkehrung dieser Implikation gilt dabei nicht, da Turning-Points auch bei 3 positiven bzw. negativen Beobachtungen auftreten können. Betrachtet man allerdings die differenzierte Zeitreihe $\Delta(x_1, ..., x_N) = (x_2 \text{ - } x_1, ..., x_N \text{ - } x_{N-1}) = (\tilde{x_1}, ..., x_{\tilde{N-1}})$, so wird klar, dass das Auftreten eines 2-VZWs in dieser Zeitreihe äquivalent mit dem eines Turning-Points in der ursprünglichen Zeitreihe ist, denn es gilt für ein beliebiges $i \in \{1, ..., N \text{ - } 3\}$:

$$\begin{aligned} & \tilde{x_i} > 0 \wedge \tilde{x_{i+1}} < 0 \\ \Leftrightarrow \quad & x_{i+1} \text{ - } x_i > 0 \wedge x_{i+2} \text{ - } x_{i+1} < 0 \\ \Leftrightarrow \quad & x_i < x_{i+1} \wedge x_{i+1} > x_{i+2} \end{aligned}$$

bzw.

$$\begin{aligned} & \tilde{x_i} < 0 \wedge \tilde{x_{i+1}} > 0 \\ \Leftrightarrow \quad & x_{i+1} \text{ - } x_i < 0 \wedge x_{i+2} \text{ - } x_{i+1} > 0 \\ \Leftrightarrow \quad & x_i > x_{i+1} \wedge x_{i+1} < x_{i+2}\,. \end{aligned}$$

Die Trennschärfen des vereinfachten 2-VZ-Tests angewendet auf differenzierte Zeitreihen und des TP-Tests auf die originalen Zeitreihen sind jedoch aufgrund der unterschiedlichen asymptotischen kritischen Werten ziemlich unterschiedlich. Dies rührt vermutlich daher, dass die Beobachtungen einer differenzierten, unabhängigen Zeitreihe nicht mehr unabhängig sind, sondern zwei aufeinanderfolgende Beobachtungen insbesondere die Kovarianz $\text{-}\sigma^2$ aufweisen, da gilt:

$$\begin{aligned} & Cov(\tilde{x_i},\, \tilde{x_{i+1}}) = Cov(x_{i+1} \text{ - } x_i,\, x_{i+2} \text{ - } x_{i+1}) = Cov(x_{i+1}, x_{i+2}) \text{ - } Cov(x_{i+1}, x_{i+1}) \\ & \text{ - } Cov(x_i, x_{i+2}) + Cov(x_i, x_{i+1}) = \text{ -}Var(x_{i+1}) = \text{ -}\sigma^2. \end{aligned}$$

Das bedeutet, dass die differenzierten Zeitreihen von unabhängigen Zeitreihen stets negativ korreliert sind. Eine Verwerfung der Nullhypothese findet anhand der differenzierten Zeitreihe in diesem Fall nur für extreme positive Korrelationen in der Ausgangszeitreihe statt, da insbesondere im Fall $\rho_1 = 1$ gilt:

$$Cov(\tilde{x_i},\, \tilde{x_{i+1}}) = Cov(x_{i+1} \text{ - } x_i,\, x_{i+2} \text{ - } x_{i+1}) = Cov(x_i + w_{i+1} \text{ - } x_i,\, x_{i+1} + w_{i+1} \text{ - } x_{i+1})$$
$$= Cov(w_{i+1},\, w_{i+2}) = 0.$$

Aus diesem Grund wurden der vereinfachte 2-VZ-Test auf differenzierte Zeitreihen angewendet, wobei aber eine Ablehnung der Nullhypothese nicht bei einer Überschreitung der üblichen kritischen Werte der Teststatistik stattfindet. Vielmehr wurden die Quantile der exakten Verteilung in der differenzierten Reihe für den jeweiligen Stichprobenumfang auf der Grundlage von 10 000 unabhängigen Ausgangszeitreihen empirisch ermittelt. Damit sollten die Abhängigkeiten in der differenzierten Zeitreihe nicht zu deutlichen Verzerrungen der Testentscheidungen führen, wie in dem Fall, in dem der herkömmliche vereinfachte 2-VZ-Test auf die differenzierte Zeitreihe angewendet wird. Dabei wäre davon auszugehen, dass es deutlichere Parallelen zu dem TP-Test gibt, als es zwischen dem vereinfachten 3-VZ-Test und dem TP-Test in der ursprünglichen Zeitreihe der Fall ist. Zur Veranschaulichung wurden die Trennschärfen der 3 Testverfahren einander in Abbildung 3.95 gegenübergestellt.

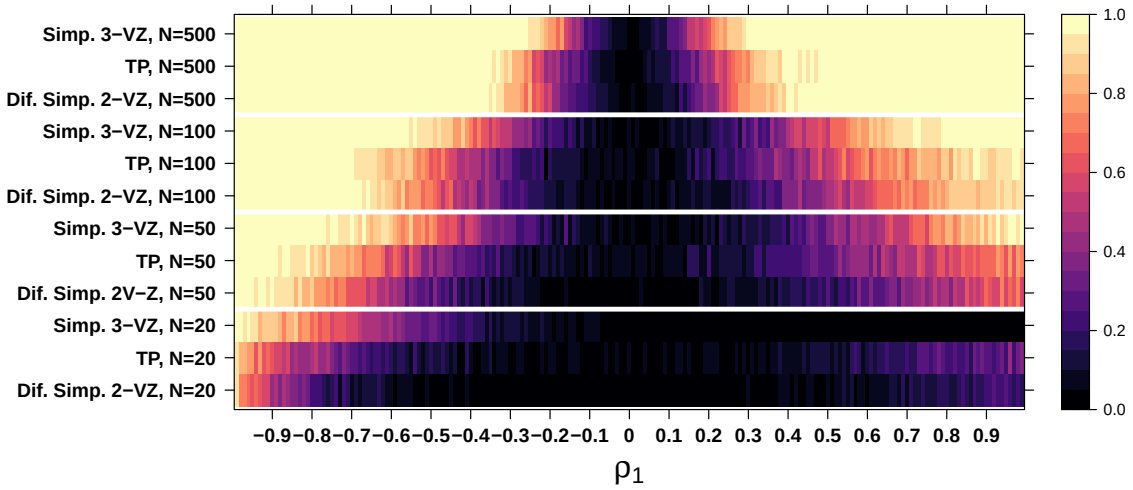


Abbildung 3.95: Trennschärfen des TP-Tests, des vereinfachten (Simp.) 3-VZ-Tests sowie des modifizierten (Dif. Simp.) 2-VZ-Tests anhand der differenzierten Zeitreihe in Abhängigkeit von ρ_1 zu unterschiedlichen Stichprobenumfängen

Hier sind die Parallelen zwischen den Trennschärfen der 3 Testverfahren deutlich zu erkennen. So gewinnen sie mit wachsendem Stichprobenumfang im ähnlichen Maße an Trennschärfe hinzu und weisen eine ähnliche Asymmetrie auf. Die Übereinstimmungen zwischen dem modifizierten, vereinfachten 2-VZ-Test auf der differenzierten Zeitreihe und dem TP-Test sind dabei erwartungsgemäß stärker als mit dem vereinfachten 3-VZ-Test. Insbesondere sind ihre Trennschärfen bei einer Beobachtungszahl von $N = 500$ nahezu identisch. Insgesamt schneidet der vereinfachte 3-VZ-Test bei sämtlichen Stichprobenumfängen, ab denen eine Verwerfung für positive Korrelationen möglich ist ($N \geq 50$), am besten ab. Dabei werden die Unterschiede zwischen den Verfahren mit zunehmender Beobachtungszahl immer geringer.

3.6.2.3 Einordnung der Trennschärfen

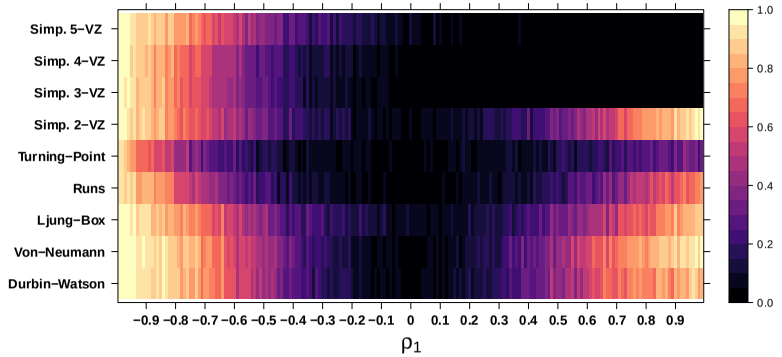
Um eine Einordnung der Trennschärfe von den vereinfachten K-VZ-Tests zu erreichen, werden die Trennschärfen dieser Testverfahren im Folgenden kurz denen von ausgewählten, in dieser Arbeit betrachteten Verfahren gegenübergestellt. Besonders interessant sind dabei die Fälle, dass es sich um einen AR(2)-Prozess (s. Kap. 3.2), bzw. um saisonale Prozesse (s. Kap. 3.3) handelt, da hier schon Überlegenheiten der vollständigen K-VZ-Tests gegenüber vielen der anderen Verfahren festgestellt wurden. Aufgrund ihrer insgesamt schlechten Trennschärfe wäre eine Anwendung dieser Testverfahren in der Praxis aber nicht sinnvoll. Die vereinfachten Versionen scheinen aufgrund ihrer stärkeren Konsistenzeigenschaften besser geeignet.

Die Trennschärfen für den Fall eines AR(1)-Prozesses sind in Abbildung 3.96, die Ergebnisse für den AR(2)-Fall in den Abbildungen 3.97 und 3.98, saisonale Prozesse in den Abbildungen 3.99 und 3.100 und ihre Trennschärfen bei MA(1)-Alternativen in Abbildung 3.101 dargestellt.

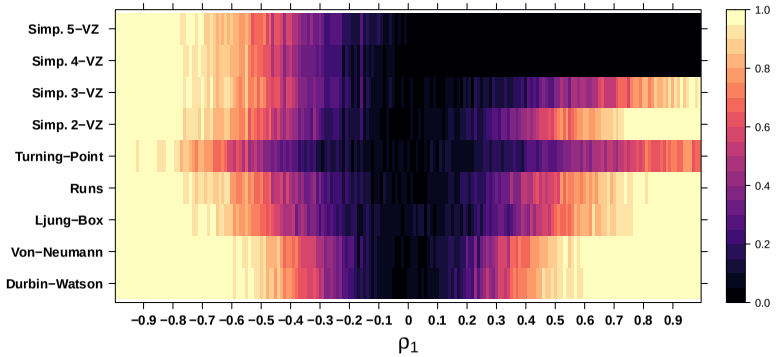
Dabei zeigt sich für den AR(1)-Fall noch einmal, dass die vereinfachten K-VZ-Tests deutlich bessere Trennschärfen aufweisen, als diejenigen, die auf der vollständigen Vorzeichentiefe basieren. Ein wesentlicher Vorteil besteht dabei darin, dass sie von einem wachsenden Stichprobenumfang in ähnlichem Maße profitieren wie die anderen Testverfahren. Am geeignetsten erscheint dabei der 2-VZ-Test, der bereits bei einem Stichprobenumfang von $N = 20$ eine gute Trennschärfe aufweist und hier zusammen mit dem VNRR-Test die besten Ergebnisse liefert. Auch bei einem wachsendem Stichprobenumfang erscheint seine Trennschärfe ähnlich gut wie die des Runs-Tests (s. Kap. 3.6.2.2) bzw. des LB-Tests.

Wie bereits erörtert, benötigen Tests mit einem größeren K mehr Beobachtungen um Alternativen mit positiven Korrelationen zu erkennen als solche mit kleinen Ks und ihre Trennschärfe wird im AR(1)-Fall mit wachsendem K zunehmend schlechter. Dies lässt sich damit erklären, dass die stärkste Abhängigkeitsstruktur bei AR(1)-Prozessen zwischen zwei aufeinanderfolgenden Beobachtungen vorliegt und die Betrachtung von noch mehr Beobachtungen bei den VZW die Ergebnisse eher verzerrt, ohne dabei deutlich mehr Informationen zu liefern.

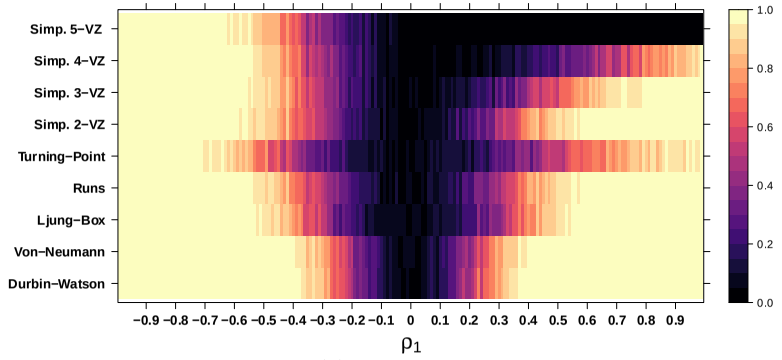
Auch wenn es sich um AR(2)-Prozesse handelt, zeigt sich die Überlegenheit der vereinfachten K-VZ-Tests zu ihren vollständigen Gegenstücken. Dabei entspricht die Trennschärfe des



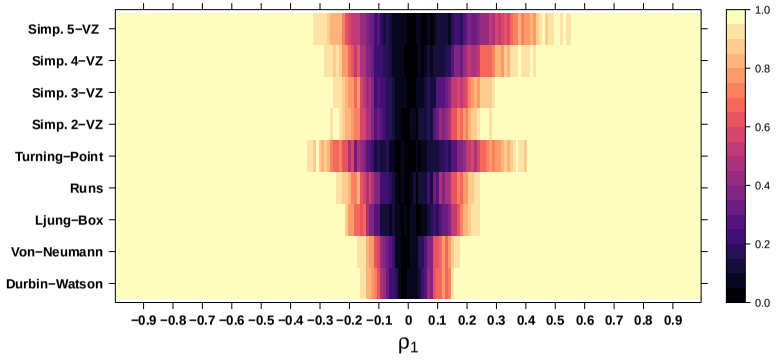
(a) $N = 20$



(b) $N = 50$



(c) $N = 100$



(d) $N = 500$

Abbildung 3.96: Vergleich der simulierten Trennschärfen der vereinfachten K-VZ-Tests mit anderen Testverfahren bei stationären AR(1)-Alternativen in Abhängigkeit von ρ_1 für unterschiedliche Stichprobenumfänge

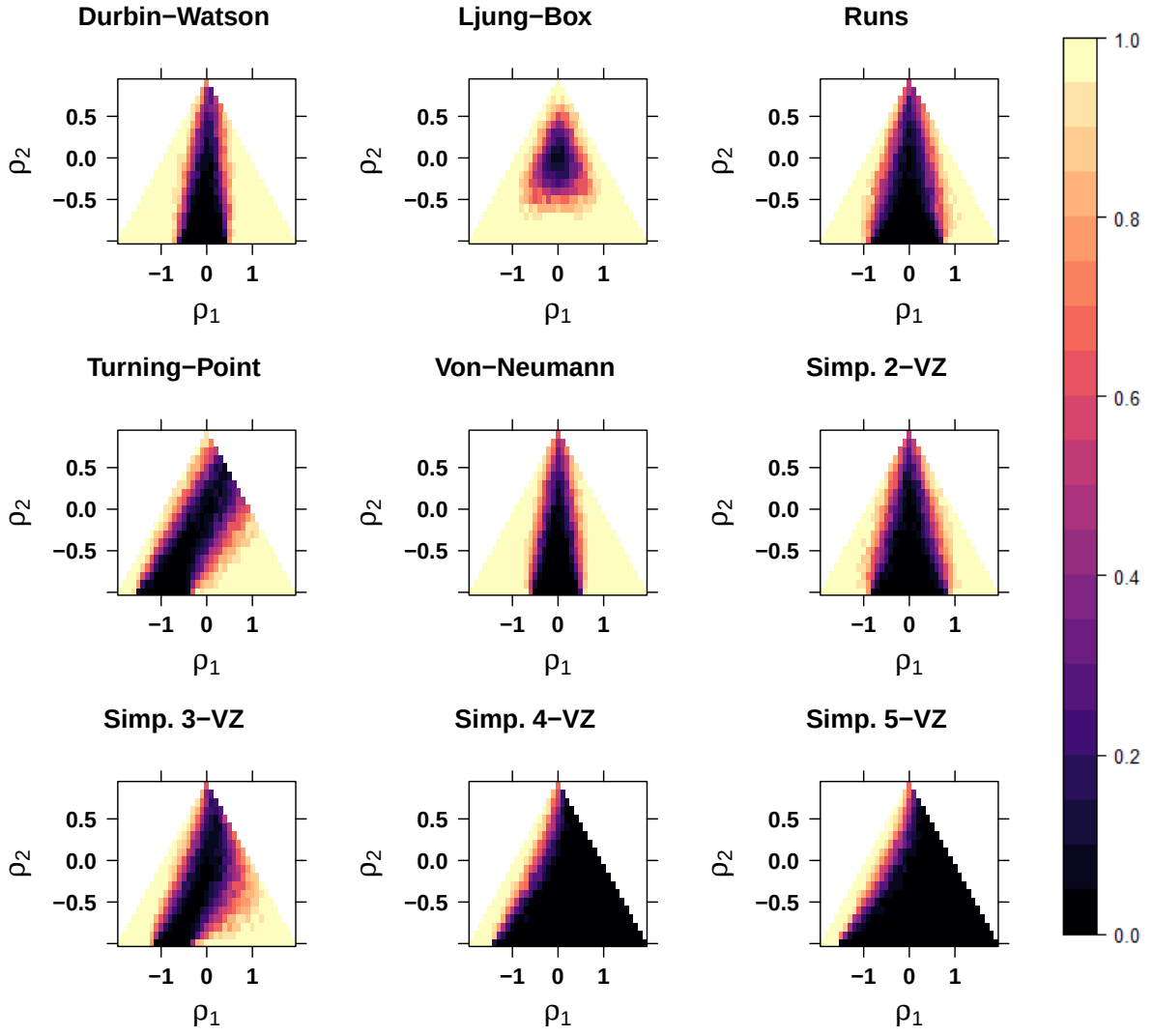


Abbildung 3.97: Simulierte Trennschärfen der vereinfachten K-VZ-Tests bei stationären AR(2)-Alternativen im Vergleich mit anderen Testverfahren in Abhängigkeit von ρ_1 und ρ_2 für $N = 50$ Beobachtungen

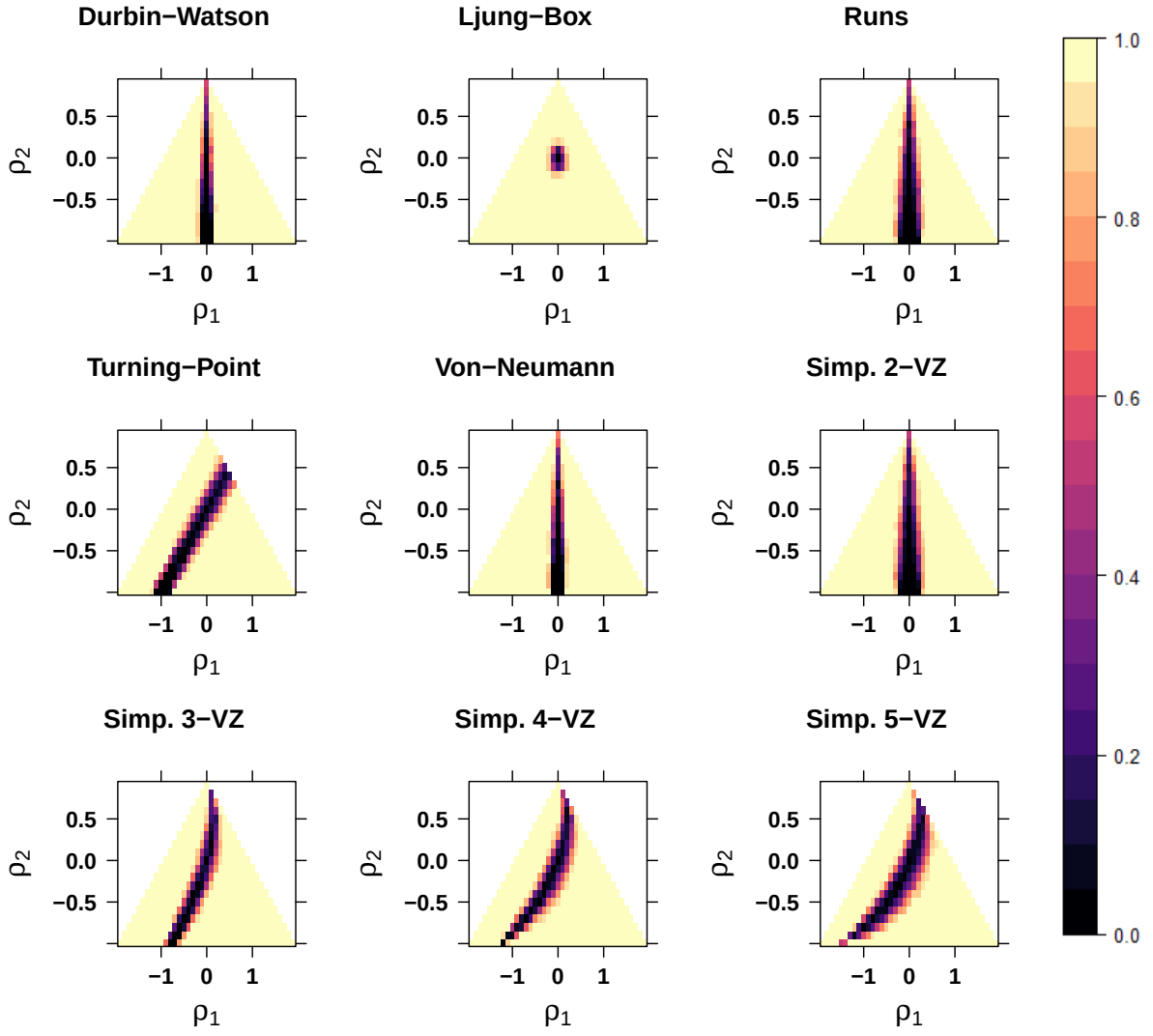


Abbildung 3.98: Simulierte Trennschärfen der vereinfachten K-VZ-Tests bei stationären AR(2)-Alternativen im Vergleich mit anderen Testverfahren in Abhängigkeit von ρ_1 und ρ_2 für $N = 500$ Beobachtungen

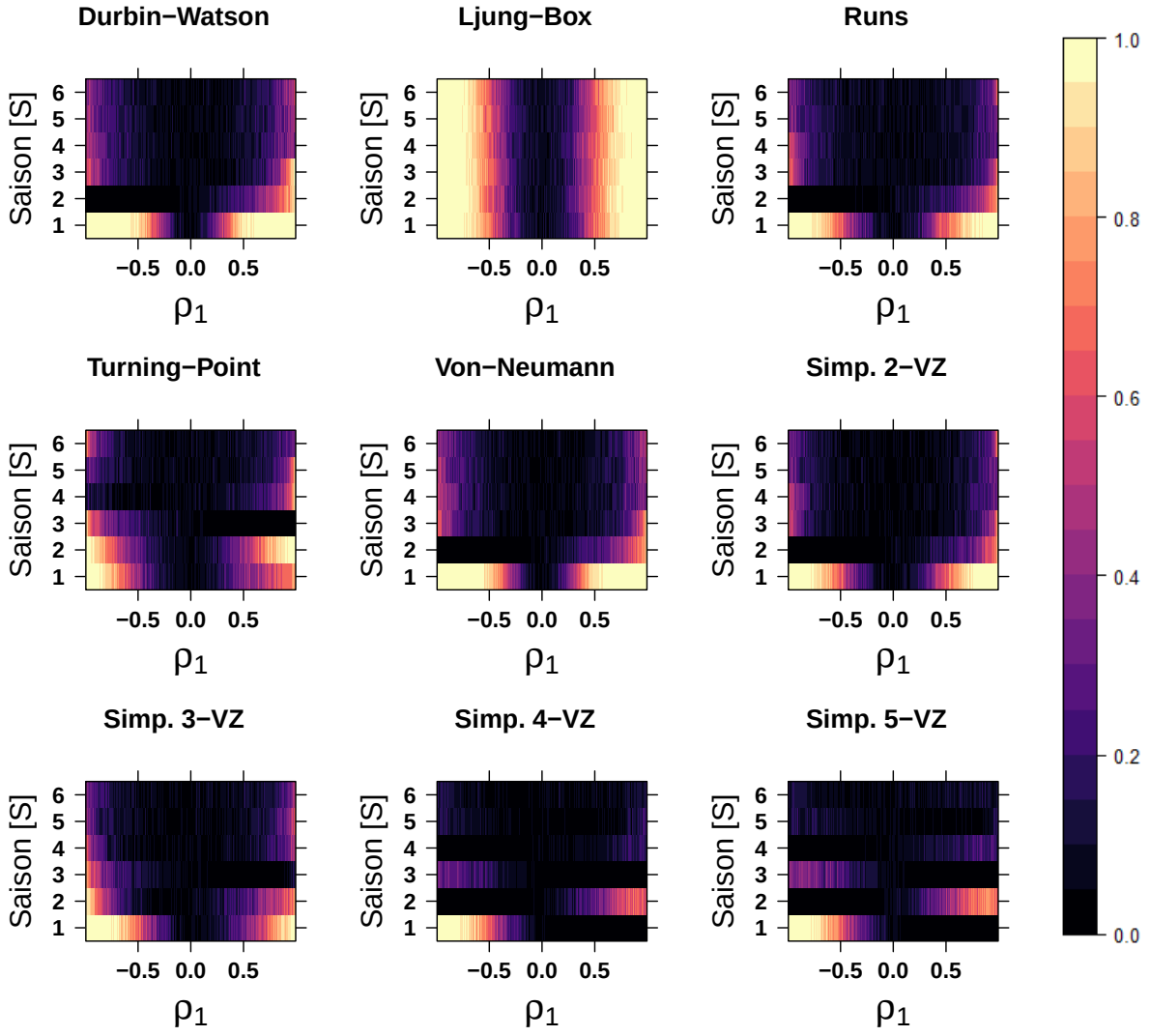


Abbildung 3.99: Simulierte Trennschärfen der vereinfachten K-VZ-Tests bei stationären, saisonalen autoregressiven Prozessen 1. Ordnung im Vergleich mit anderen Testverfahren in Abhängigkeit von ρ_S für $N = 50$ Beobachtungen

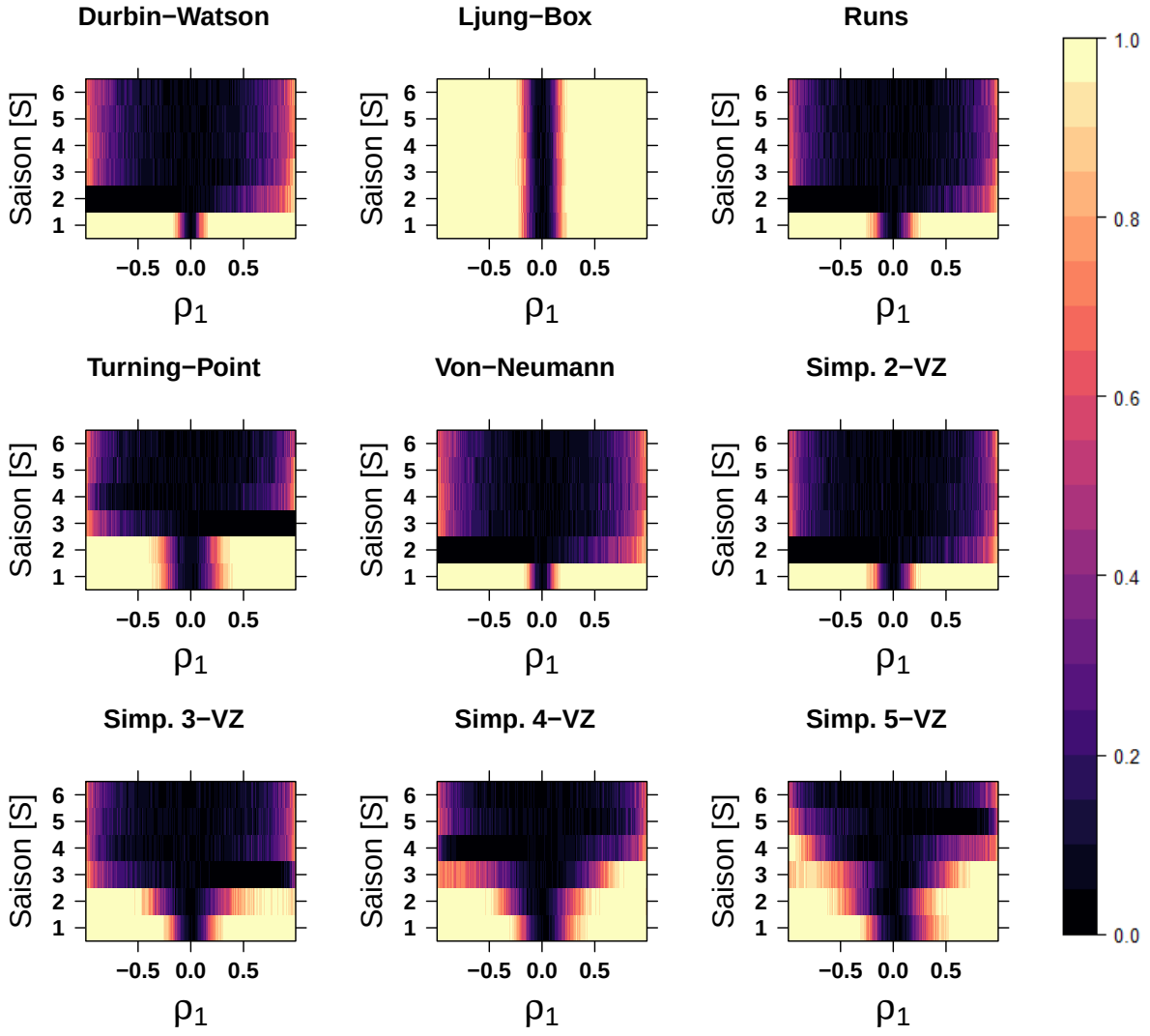
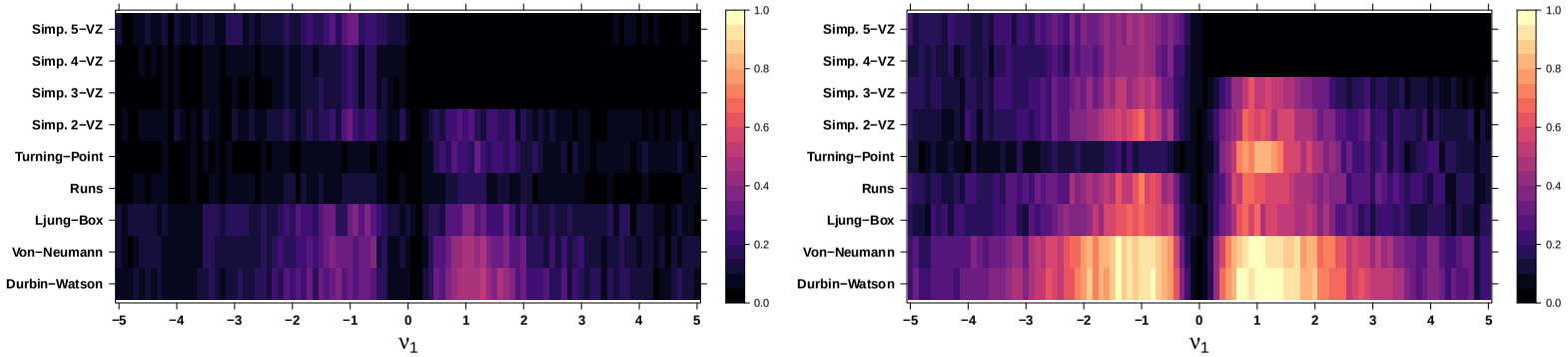
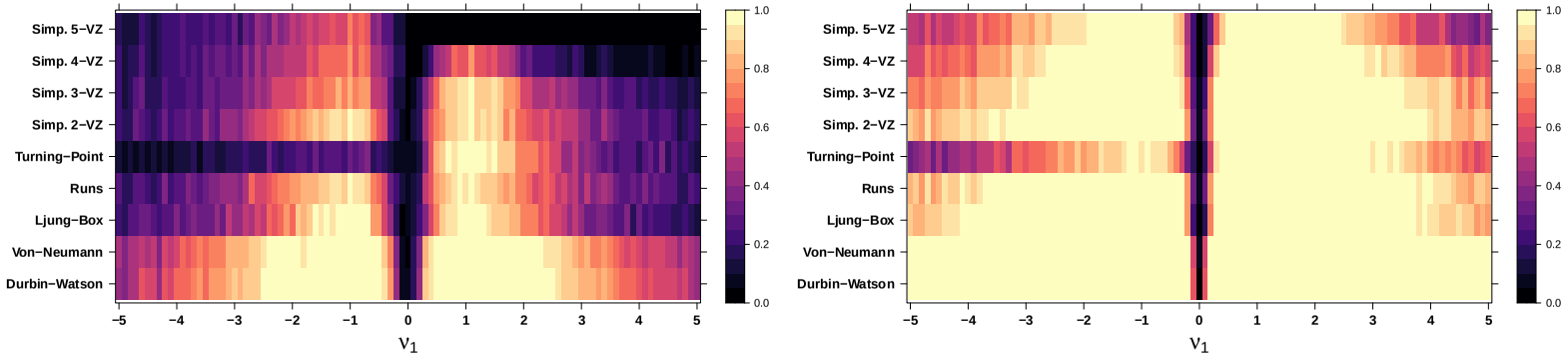


Abbildung 3.100: Simulierte Trennschärfen der vereinfachten K-VZ-Tests bei stationären, saisonalen autoregressiven Prozessen 1. Ordnung im Vergleich mit anderen Testverfahren in Abhängigkeit von ρ_S für $N = 500$ Beobachtungen



(a) $N = 20$

(b) $N = 50$



(c) $N = 100$

(d) $N = 500$

Abbildung 3.101: Vergleich der simulierten Trennschärfen von den vereinfachten K-VZ-Tests mit anderen Testverfahren bei stationären MA(1)-Alternativen in Abhängigkeit von ν_1 für unterschiedliche Stichprobenumfänge

vereinfachten 2-VZ-Tests hier in etwa der des Runs-Tests und damit gelingt es dem Verfahren lediglich, Abweichungen des Parameters ρ_1 zu detektieren. Für die vereinfachten Vorzeichentiefen mit größerem K ist ein auffällig anderes Verhalten erkennbar. So treten ähnliche Phänomene auf, wie sie beim TP-Test beobachtet wurden (vgl. Kap. 3.2) und gewisse Kombinationen der Parameter ρ_1 und ρ_2 führen dazu, dass eine Ablehnung der Nullhypothese nicht stattfinden kann. Hier wird die Trennschärfe für $K \geq 3$ auch immer schlechter, das heißt, es existieren mehr Alternativen, in denen die Nullhypothese nicht abgelehnt werden kann. Auch verändert sich die Form des Ablehnungsbereichs. So können bei $K = 3$, ähnlich wie beim TP-Test, Alternativen, in denen annähernd $\rho_1 = \rho_2$ gilt, nicht abgelehnt werden, während der Parameter ρ_1 mit wachsendem K im Verhältnis zu ρ_2 immer größer werden muss, damit keine Verwerfung stattfindet. Dieses Schema lässt sich auf die mit wachsendem K zunehmende Asymmetrie der Trennschärfen von den vereinfachten K-VZ-Tests zurückführen. Insgesamt scheinen die vereinfachten K-VZ-Tests in diesem Szenario nach dem LB-Test am besten abzuschneiden, da der Effekt des Parameters ρ_2 nicht ignoriert wird und verhältnismäßig wenige Alternativen zur Unabhängigkeit beibehalten werden. Im AR(2)-Prozess scheint also die Wahl eines höheren Ks als 2 sinnvoll, weil dadurch auch Abhängigkeiten von mehr als 2 aufeinanderfolgender Beobachtungen erfasst werden können. Damit wirkt es, als würden größere Ks dazu führen, dass komplexere Abhängigkeitsstrukturen erfasst werden.

Diese Beobachtungen und Vermutungen bestätigen sich bei der Betrachtung von den Trennschärfen der vereinfachten K-VZ-Tests, wenn sie auf saisonale Prozesse 1. Ordnung angewendet werden. Hier wird deutlich, dass die Wahl des Parameters K die Fähigkeit der Verfahren, saisonale Abhängigkeiten zu detektieren, maßgeblich beeinflusst. Während der vereinfachte 2-VZ-Test hier lediglich Abweichungen von ρ_1 zu erkennen vermag, gelingt es dem 3-VZ-Test − ähnlich wie dem TP-Test − zusätzlich Abweichungen des Parameters ρ_2 von 0 zu detektieren. Für wachsendes K scheint sich diese Tendenz fortzusetzen, sodass der vereinfachte K-VZ-Test allgemein in der Lage zu sein scheint, Abweichungen des Parameters ρ_{K-1} zu detektieren. Allerdings nimmt die Trennschärfe hier mit größer werdender Saisonalität drastisch ab, sodass der 5-VZ-Test nur noch extreme Abweichungen des Parameters ρ_4 erkennen kann. Auch treten hier für größere Saisonalitäten und Ks zunehmende Asymmetrien bezüglich der Trennschärfen auf. Allgemein scheint es, als würden für gerade K keine negativen Korrelationen für die Saisonalität $S = K$ detektiert werden können, während für ungerade Ks bei positiven Korrelationen keine Verwerfung zur gleichen Saisonalität stattfinden kann. Ein ähnliches Verhalten wurde für den TP-Test in Kapitel 3.3 festgestellt und kann hier mit ähnlichen Überlegungen nachvollzogen werden.

Im Fall von MA(1)-Prozessen werden erneut Überlegenheiten der vereinfachten K-VZ-Tests im Vergleich mit den vollständigen Testverfahren deutlich (vgl. Kap. 2.2). Während die vollständigen K-VZ-Tests hier sehr schlechte Trennschärfen aufgewiesen haben und positive Werte des Parameters ν_1 nicht als Abweichungen von der Unabhängigkeit erkennen konnten, gelingt dies den vereinfachten Versionen für größere Stichproben vergleichsweise gut. So entspricht die

Trennschärfe des vereinfachten 2-VZ-Tests hier wieder in etwa der des Runs-Tests oder des LB-Tests und ist somit besser als die des TP-Tests. Weiter fällt auf, dass die Trennschärfe dieser Testverfahren mit größerem K zunehmend schlechter wird. Bei einem $K \geq 4$ sind dabei für Stichprobenumfänge von $N \leq 50$ ähnliche Muster wie bei den vollständigen K-VZ-Tests erkennbar. Diese Systematiken lassen sich erneut auf die sehr lokale Abhängigkeitsstruktur in MA(1)-Prozessen und die zu kleine Beobachtungszahl zurückführen. So sollten hier lediglich Korrelationen zwischen 2 aufeinanderfolgenden Beobachtungen auftreten, sodass durch eine Wahl von Ks ≥ 2 keine zusätzliche Struktur im Prozess erfasst werden kann.

Insgesamt lässt sich feststellen, dass die vereinfachten Versionen im Bezug auf alle betrachteten Prozesse eine deutliche Verbesserung gegenüber den vollständigen K-VZ-Tests darstellen. Dabei scheint vor allem ihre Fähigkeit, von wachsenden Stichprobenumfängen zu profitieren, eine zentrale Rolle zu spielen. Auffällig ist, dass mit wachsendem K breitere Alternativen von Abhängigkeiten (z. B. längere Saisons) erfasst werden können, wodurch ihre Trennschärfe bei einfacheren Abhängigkeitsstrukturen im Abtausch deutlich abnimmt. Durch diese Erkenntnisse gewinnen die Verfahren in der praktischen Anwendung deutlich an Relevanz und stellen eine bemerkenswerte Alternative zu den anderen in dieser Arbeit betrachteten Tests dar.

Im Folgenden soll kurz erläutert werden, wie sich die vereinfachten Tiefen unter anderen, in dieser Arbeit bereits betrachteten Szenarien, verhalten. Dabei gilt, dass die Verfahren lediglich auf dem sequenziellen Schema der Beobachtungen beruhen und damit robuste Eigenschaften aufweisen. So legen Simulationen nahe, dass sie robust gegenüber Kontaminationen, innovativen Ausreißern, anderen Verteilungen der Innovationen sowie wachsenden Varianzen in der Zeitreihe sind. Weiterhin wird untersucht, wie die Verfahren auf Änderungen des Niveaus in der Zeitreihe reagieren. So wurde bei den vollständigen K-VZ-Tests beobachtet, dass sie Sprünge, Trends und eine Oszillation in der Zeitreihe als Abhängigkeitsstruktur erkennen können und somit auch als Testverfahren auf das Vorhandensein derartiger Strukturen herangezogen werden könnten. Für einen Stichprobenumfang von $N = 500$ ist die Trennschärfe der vereinfachten K-VZ-Tests im Fall von Trends in Abbildung 3.102 dargestellt. Das Verhalten bei einem Sprung lässt sich anhand von Abbildung 3.103 nachvollziehen und ihre Reaktionen auf eine wachsende bzw. feste Anzahl von Oszillationen in Abbildungen 3.104 bzw. 3.105.

Daraus geht hervor, dass sich die vereinfachten K-VZ-Tests nicht besonders gut zur Detektion dieser Strukturen eignen. So verhält sich ihre Trennschärfe in den unterschiedlichen Szenarien eher wie die des Runs- oder des VNRR-Tests. So verschieben sich ihre Annahmebereiche bei zunehmendem Trend, zunehmender Sprunghöhe und zunehmender Amplitude der Oszillationen in den Bereich negativer Korrelationen und eignet sich nicht gut zu deren Erkennung. Zum Vergleich kann die Nullhypothese beim 5-VZ-Test und einer Sprunghöhe von 1 bzw. -1 nur bei extremen Korrelationen von -0.9 beibehalten werden, während sie bei den vereinfachten Versionen bei einem Parameter von $|\rho_1| \approx 0.15$ nicht verworfen werden kann. Auf ähnliche Weise regieren die vereinfachten Versionen auch viel weniger sensibel auf einen Trend und eine feste bzw. wach-

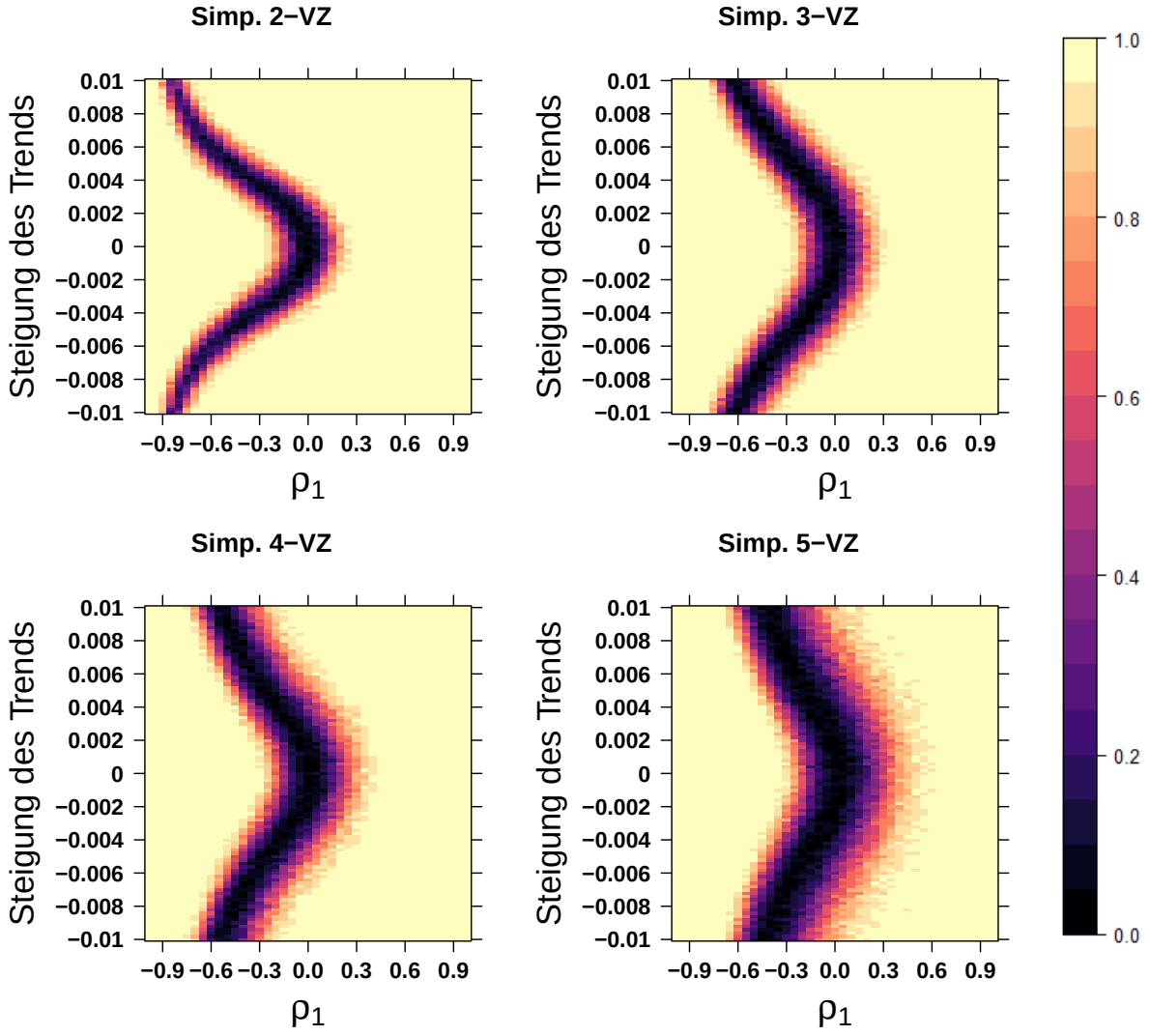


Abbildung 3.102: Simulierte Trennschärfen der vereinfachten (Simp.) K-VZ-Tests bei stationären AR(1)-Prozessen mit Trend, in Abhängigkeit von ρ_1 und der Steigung des Trends für $N = 500$ Beobachtungen

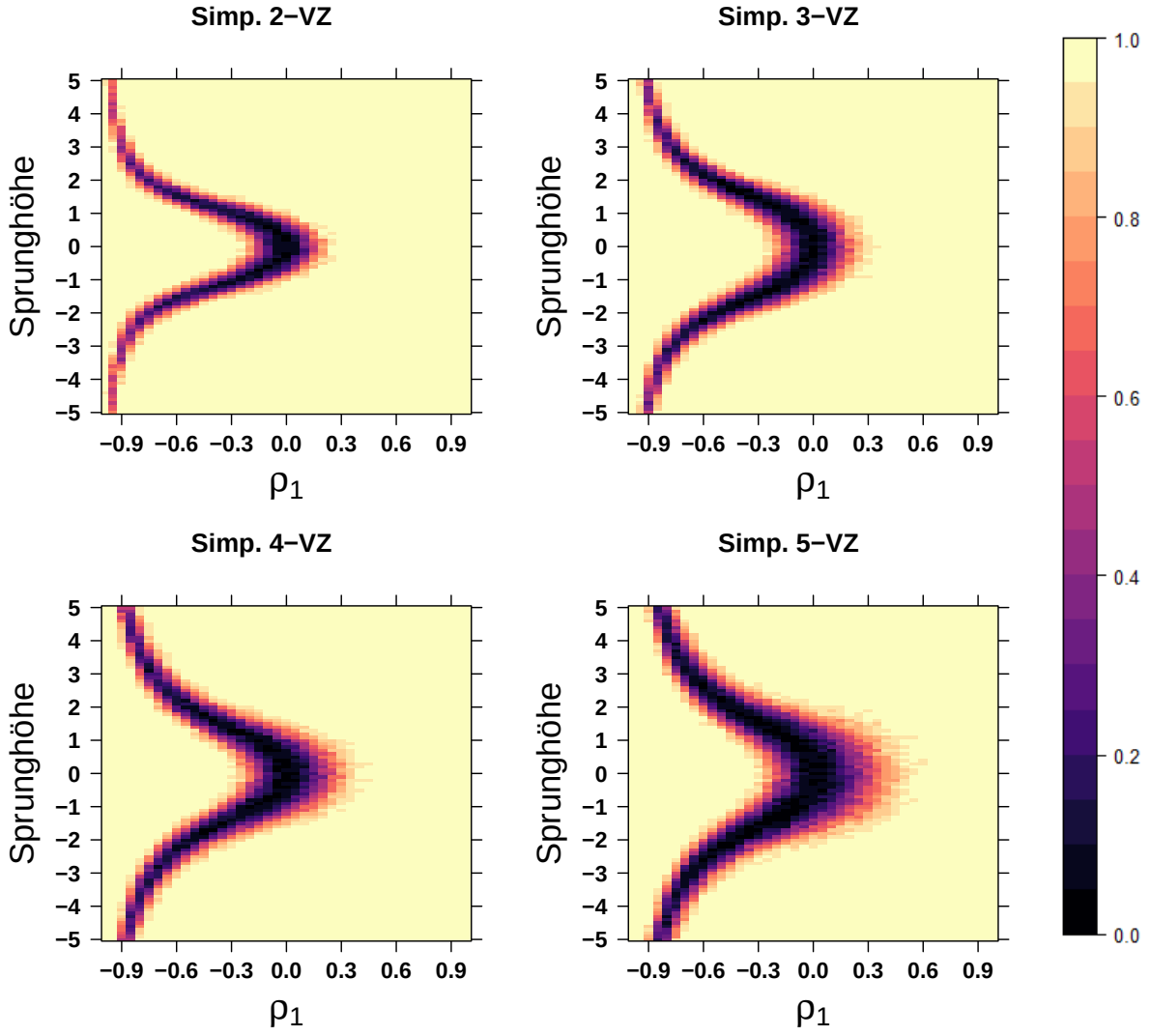


Abbildung 3.103: Simulierte Trennschärfen der vereinfachten (Simp.) K-VZ-Tests bei stationären AR(1)-Prozessen mit Sprung, in Abhängigkeit von ρ_1 und der Sprunghöhe für $N = 500$ Beobachtungen

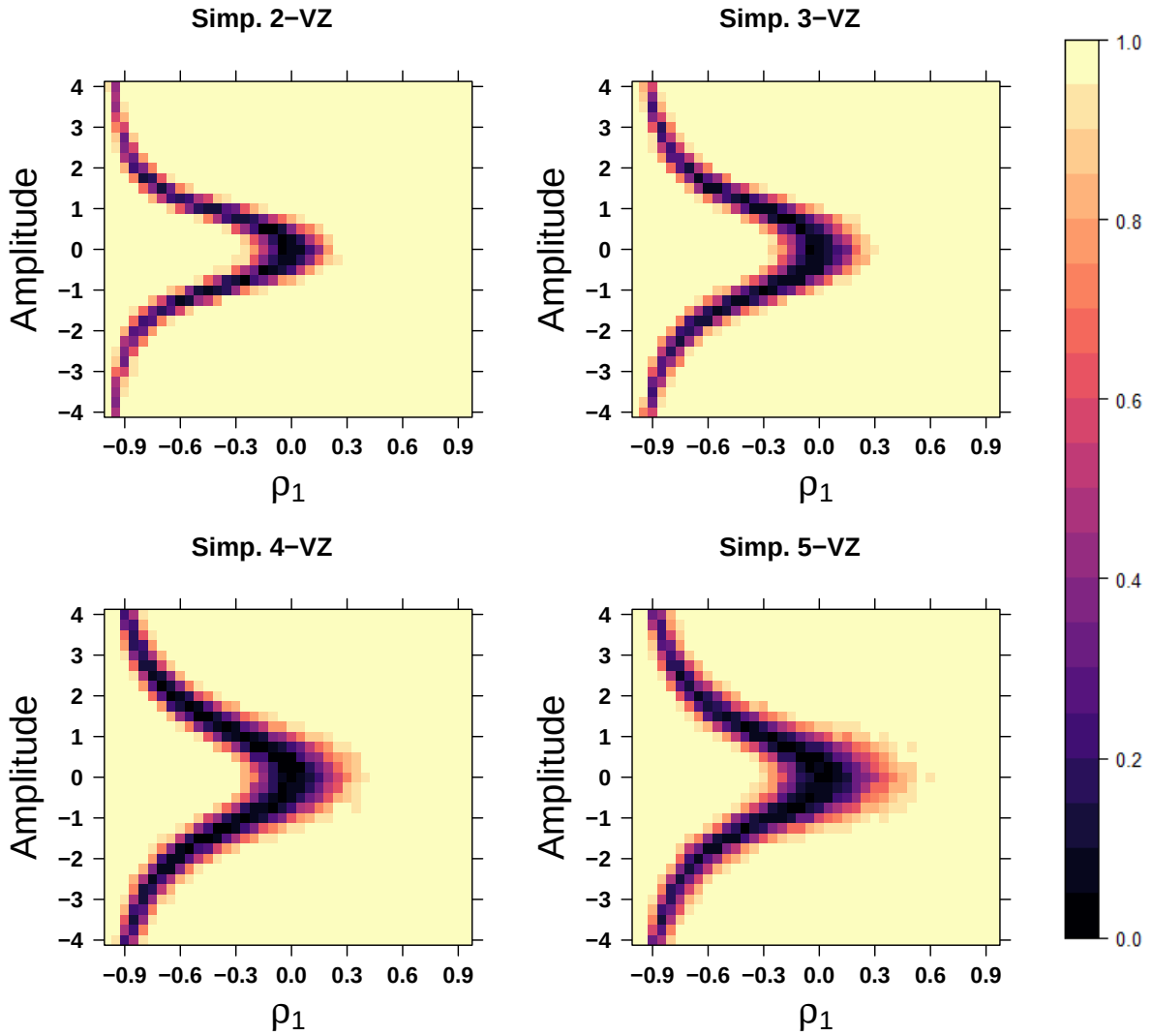


Abbildung 3.104: Simulierte Trennschärfen der vereinfachten (Simp.) K-VZ-Tests bei stationären AR(1)-Prozessen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 500$ Beobachtungen und einer wachsenden Anzahl von Oszillationen

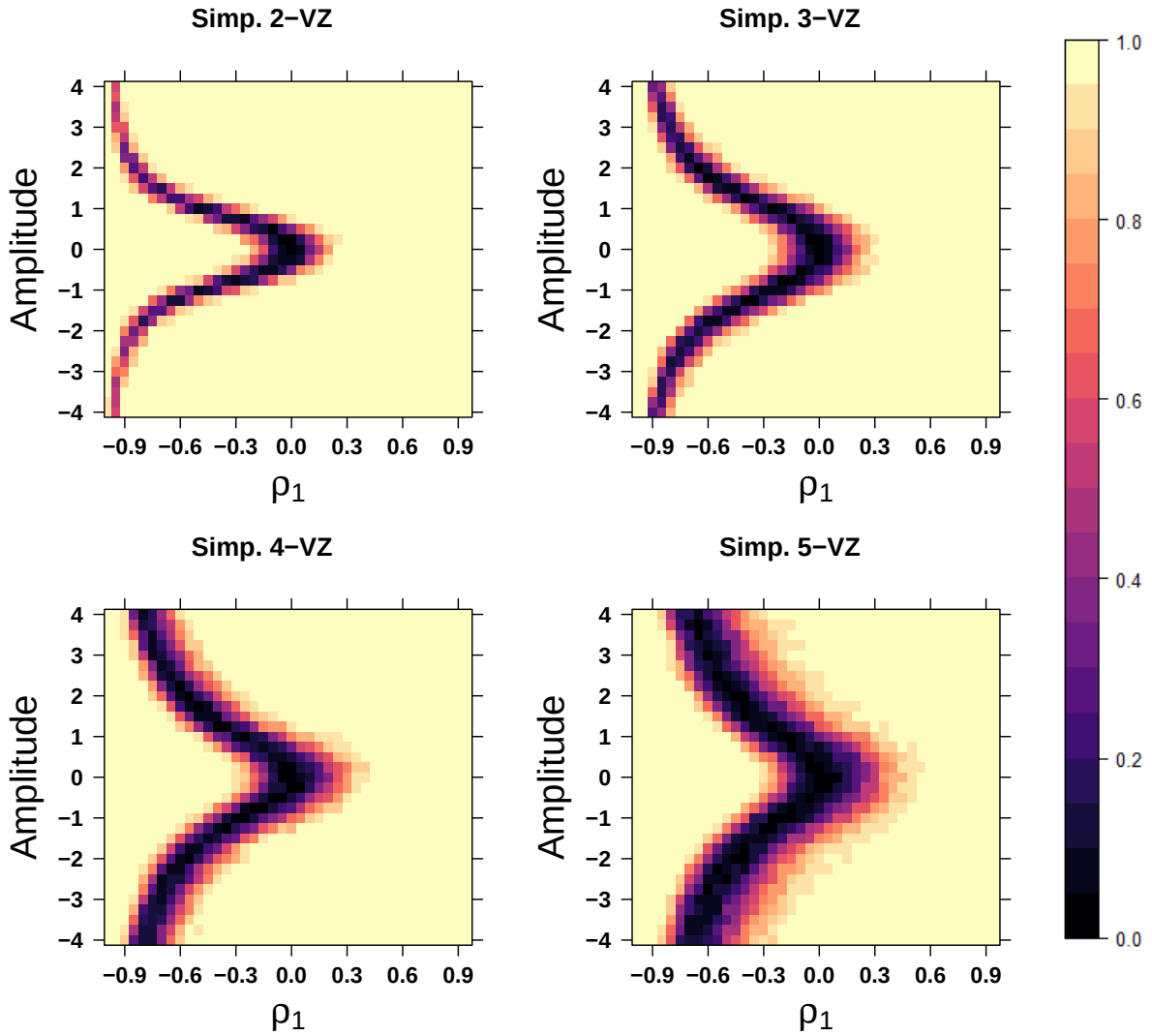


Abbildung 3.105: Simulierte Trennschärfen der vereinfachten (Simp.) K-VZ-Tests bei stationären AR(1)-Prozessen in Abhängigkeit von ρ_1 und der Amplitudenhöhe der Oszillationen, bei $N = 500$ Beobachtungen und einer festen Anzahl von Oszillationen

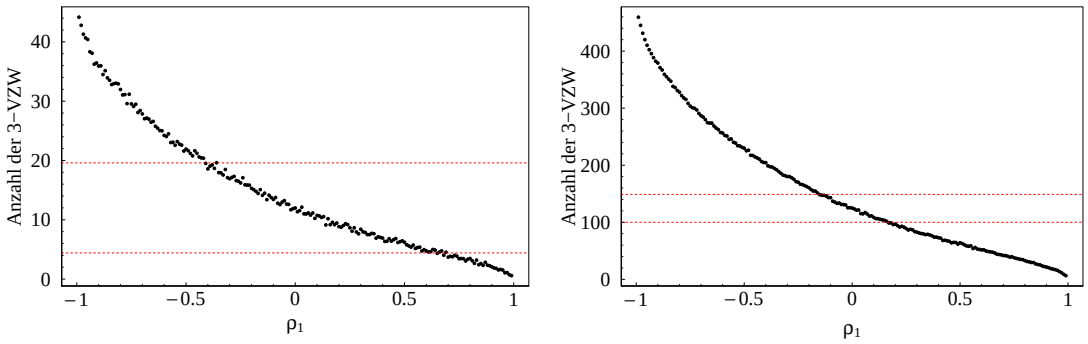
sende Anzahl von Oszillationen in der Zeitreihe. Auch unterscheiden sich die Verhaltensweisen für unterschiedliche Ks nicht auffällig stark. Allerdings scheint es, als würde die Verschiebung in den Bereich negativer Korrelationen im Verhältnis zur Amplituden- und Sprunghöhe sowie zur Steigung des Trends für größer werdende Ks zunehmend schwächer ausfallen. Auch wird die Trennschärfe etwas schlechter, was sich durch breitere Annahmebereiche bemerkbar macht.

Insgesamt zeigt sich, dass die vereinfachten K-VZ-Tests wesentlich besser als die vollständigen K-VZ-Tests zur Erkennung von kurzfristigen Abhängigkeitsstrukturen (AR-, MA-, SAR-Prozesse) in Zeitreihen geeignet sind. Auf der anderen Seite gelingt es ihnen weitaus weniger gut, langfristige Strukturen wie Niveauänderungen in der Zeitreihe zu detektieren. Sie eignen sich aus diesem Grund nicht zu deren Erkennung.

3.6.3 Konsistenzeigenschaften

Im diesem Abschnitt wird untersucht, warum die vereinfachten K-VZ-Tests deutlich stärker von einem größer werdenden Stichprobenumfang profitieren als die herkömmlichen Versionen. Dazu wird betrachtet, wie sich die Anzahl der vereinfachten K-Vorzeichenwechsel, ab der eine Verwerfung der Nullhypothese stattfindet, in Abhängigkeit von ρ_1 bei wachsendem Stichprobenumfang verhält. Dazu wurde die mittlere Anzahl an 3-VZW in Zeitreihen der Länge $N = 50$ bzw. $N = 500$ basierend auf 1000 Simulationen in Abhängigkeit von ρ_1 zusammen mit den aus der Normalverteilungs-Approximation resultierenden „kritischen“ Anzahlen in Abbildung 3.106 dargestellt.

Es ist ersichtlich, dass die Auswirkungen der Autokorrelationen auf die Anzahl der 3-VZW schematisch ähnlich bleiben, während die Spanne der kritischen Werte − zumindest relativ gesehen − kleiner wird. Dadurch kann eine Verwerfung auch schon bei geringeren Korrelationen erfolgen. Außerdem verdeutlicht diese Darstellung noch einmal die Asymmetrie der Testentscheidung. Das gleiche Vorgehen wurde für die Anzahl aller alternierenden 3er-Tupel durchgeführt,



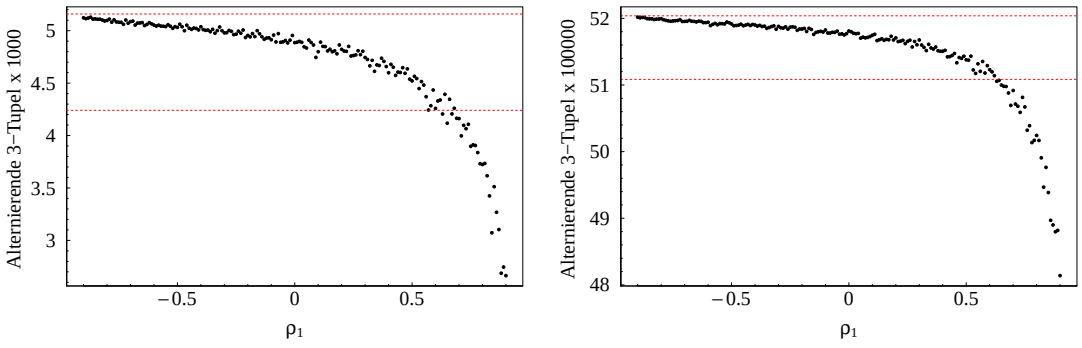
(a) $N = 50$ (b) $N = 500$

Abbildung 3.106: Darstellung der mittleren Anzahlen von 3-VZW in Abhängigkeit von ρ_1 zusammen mit den kritischen Werten (rote Linien)

was der Teststatistik des vollständigen 3-VZ-Tests entspricht. Die Ergebnisse sind für Werte von ρ_1 zwischen -0.99 und 0.99 mit einer Gitterfeinheit von 0.01 in Abbildung 3.107 dargestellt.

Daraus geht hervor, dass das beobachtete Verhalten für die Teststatistik der vereinfachten Version bei der Menge aller 3er-Tupel deutlich schwächer ausgeprägt ist. So gibt es zwischen den beiden betrachteten Stichprobenumfängen keine deutlichen Unterschiede bezüglich der Werte von ρ_1, ab denen ein Verwerfung stattfinden kann. Heuristisch sind diese Beobachtungen damit zu erklären, dass sich Autokorrelationen 1. Ordnung vor allem auf das sequenzielle Schema naheliegender Beobachtungen auswirken und die Abhängigkeitsstruktur über einen größeren Abstand immer geringer wird. Demnach ist die Aussagekraft eines K-Tupels von 3 Beobachtungen, die sehr weit auseinanderliegen, denkbar gering und liefert kaum Hinweise auf das Vorhandensein einer Autokorrelation. Die herkömmlichen K-VZ-Tests betrachten also mit zunehmender Beobachtungszahl immer mehr Tupel, die sehr wenig Informationen über die Abhängigkeitsstruktur der Zeitreihe enthalten, was einem Zugewinn an Informationen durch eine größere Beobachtungszahl entgegenzuwirken scheint. Da die vereinfachten Versionen lediglich direkt aufeinanderfolgende Beobachtungen betrachten, nimmt die Zahl an aussagekräftigen Tupeln mit steigender Beobachtungszahl immer weiter zu, was eine Verwerfung der Nullhypothese erleichtert.

Im Hinblick auf die Erkenntnisse aus dem vorherigen Abschnitt kann nachvollzogen werden, warum den vollständigen K-VZ-Tests eine Erkennung von Trends, Sprüngen und einer Oszillationen leichter fällt. So enthält ein viel größerer Anteil der betrachteten Tupel Informationen über diese Strukturen. Insbesondere liefert im Fall eines Sprunges jedes Tupel, dass Beobachtungen aus der ersten und zweiten Hälfte der betrachteten Zeitreihe enthält, Informationen und ein Trend macht sich bei weiter auseinanderliegenden Beobachtungen sogar stärker bemerkbar als bei aufeinanderfolgenden. Auch bei einer Oszillation ist ihr Effekt auf die Werte der Zeitreihe bei weit auseinanderliegenden Beobachtungen stärker ausgeprägt. Damit erklärt sich auch die viel geringere Sensitivität der vereinfachten K-VZ-Tests in solchen Situationen.



(a) $N = 50$ (b) $N = 500$

Abbildung 3.107: Darstellung der mittleren Anzahlen aller alternierenden 3-Tupel in Abhängigkeit von ρ_1 zusammen mit den kritischen Werten (rote Linien)

4 Fazit

4.1 Zusammenfassung der Ergebnisse

Das Ziel dieser Arbeit bestand darin, Verhalten und Trennschärfe unterschiedlicher statistischer Testverfahren zur Überprüfung von Unabhängigkeits- bzw. Zufälligkeitsannahmen in Zeitreihen unter verschiedenen stationären Prozessen zu untersuchen. Außerdem wurde betrachtet, wie die Verfahren reagieren, wenn andere Strukturen als Korrelationen in der Zeitreihe oder Abweichung von den Testvoraussetzungen vorliegen. Das Hauptaugenmerk lag bei diesen Untersuchungen auf den K-Vorzeichentiefetests und auf ihren Reaktionen bei derartigen Alternativen zur Zufälligkeit. Vor den Untersuchungen in dieser Arbeit gab es noch keine Erkenntnisse zu ihren Trennschärfen im Kontext von stationären Prozessen, obwohl Resultate von Leckey et al. (2020) nahelegen, dass sie als Testverfahren für derartige Szenarien geeignet sein könnten.

Insgesamt zeigt die nun vorliegende Simulationsstudie, dass die K-Vozeichentiefe-Tests keine gute Wahl zur Detektion kurzfristiger Korrelationsstrukturen, wie sie hier betrachtet wurden, darstellen. So schneiden ihre Trennschärfen bei AR(1)-, AR(2)- und MA(1)-Prozessen unter allen betrachteten Verfahren am schlechtesten ab. Besonders auffällig ist, dass ihre Trennschärfe in kleinen Stichproben ähnlich gut ist wie die der anderen Tests. Sie scheinen jedoch kaum von einem wachsenden Stichprobenumfang zu profitieren. Allgemein konnte dabei festgestellt werden, dass größere Ks bei allen betrachteten Szenarien bessere Ergebnisse erzielten als kleinere. Bei den Simulationen zu saisonalen AR-Prozessen wurde jedoch deutlich, dass die K-VZ-Tests − im Gegensatz zu vielen anderen Verfahren − in der Lage sind, Korrelationen zu höheren Lags und damit komplexere Abhängigkeitsstrukturen zu erfassen.

Eine weitere auffallende Eigenschaft dieser Verfahren ist, dass sie anscheinend längerfristige Strukturen wie Sprünge, Trends oder eine feste Anzahl von Oszillationen als Abweichung von der Zufälligkeit der Zeitreihe erkennen können. Auch zeigen die Tests robuste Eigenschaften und können Abhängigkeitsstrukturen trotz anders verteilter Innovationen, geringen Anteilen von Kontaminationen, innovativen Ausreißern und einer wachsenden Varianz der Innovationen ähnlich gut erkennen wie unter Normalbedingungen.

Als eine Erklärung dafür, dass die vollständigen K-Vorzeichentests bei wachsendem Stichprobenumfang kaum an Trennschärfe gewinnen und damit für größere Stichproben ungeeignet sind, wurde erkannt, dass zu viele nicht-informative Tupel zur Testentscheidung herangezogen werden. So wächst die Anzahl von Tupeln mit weit entfernten Tupelelementen, bei denen in den hier betrachteten Prozessen kaum noch eine Autokorrelation vorhanden ist, mit steigender Be-

obachtungszahl stark an. Daraus lässt sich schließen, dass es für die Aufdeckung derartiger Abhängigkeitsstrukturen Sinn macht, nur zeitlich nah beieinanderliegende Beobachtungen in die Teststatistik einfließen zu lassen. Aus diesem Grund wurden weitere Untersuchungen mit einer vereinfachten Version der K-VZ-Tests angestellt, in der lediglich Vorzeichenwechsel aufeinanderfolgender Beobachtungen berücksichtigt werden. Im Zuge dessen wurden noch einmal die von Leckey et al. (2020) beschriebenen Parallelen zwischen dem Runs-Test und dem vereinfachten 2-VZ-Test aufgezeigt. Außerdem konnte ein Zusammenhang zwischen dem vereinfachten 2- bzw. 3-VZ-Test mit dem Turning-Point-Test herausgearbeitet werden.

Bei der Verwendung dieser vereinfachten Testverfahren konnte ein drastischer Zugewinn an Trennschärfe im Vergleich zu den vollständigen Version festgestellt werden. Besonders deutlich wurde dieser im Fall von AR(1)-, AR(2)- sowie MA(1)-Prozessen. Insbesondere profitieren diese Versionen deutlich stärker von einem wachsenden Stichprobenumfang. Ihre Trennschärfe nimmt − anders als bei ihren vollständigen Gegenstücken − mit größer werdendem K ab und es entwickelt sich dabei eine zunehmende Asymmetrie der Annahmebereiche. Im Hinblick auf saisonale autoregressive Prozesse wurde außerdem deutlich, dass die Wahl des Ks einen Einfluss auf die Fähigkeit der Testverfahren hat, saisonale Abhängigkeitsstrukturen zu detektieren. So zeigen die Simulationsergebnisse, dass ein vereinfachter K-VZ-Test Autokorrelationen bis zum Lag K-1 erkennen kann. Die Trennschärfe der Testverfahren nimmt jedoch mit größeren Lags zunehmend ab. Ein wesentlicher Nachteil gegenüber den herkömmlichen K-VZ-Tests ist, dass sie nicht in der Lage sind, Sprünge, Trends oder eine feste Anzahl von Oszillationen in der Zeitreihe zu detektieren.

Von den anderen betrachteten Testverfahren sticht der Ljung-Box-Test besonders hervor. Dieser weist sowohl bei einfachen Abhängigkeitsstrukturen (MA(1)- u. AR(1)-Prozesse) als auch bei komplexeren Strukturen (saisonale AR- und AR(2)-Prozesse) eine gute Trennschärfe auf. Weiterhin scheint er in sehr großen Stichproben zumindest ansatzweise in der Lage zu sein, Korrelationen der Fehler eines Prozesses, wie sie in GARCH(1,1)-Modellen existieren, zu erkennen. Auch eignet er sich − ähnlich wie die K-VZ-Tests − hervorragend zur Detektion von Sprüngen, Trends sowie einer festen und einer wachsenden Anzahl von Oszillationen. Da es sich jedoch um einen parametrischen Test handelt, werden seine Testentscheidungen stark von anders verteilten Innovationen, Kontaminationen sowie innovativen Ausreißern beeinflusst. Auch wachsende Varianzen bereiten diesem Test Probleme, sodass er weder in der Lage ist, diese eindeutig zu erkennen, noch robust zu reagieren. Des Weiteren muss ein Parameter H gewählt werden, der die Anzahl der vom Test betrachteten empirischen Autokorrelationskoeffizienten spezifiziert. Bei einer ungeeigneten Wahl ist es dem Test deshalb eventuell nicht möglich, die interessierenden Abhängigkeitstrukturen in der Zeitreihe zu erfassen.

Die beste Trennschärfe im Fall von AR(1)- und MA(1)-Prozessen unter Normalbedingung weist in dieser Arbeit der Durbin-Watson-Test auf. Seine Teststatistik ist dabei genau auf die Erkennung von Autokorrelationen 1. Ordnung ausgerichtet. Allerdings leidet der Test als pa-

rametrisches Verfahren deutlich unter Verletzungen seiner Voraussetzungen, Kontaminationen und innovativen Ausreißern. Auch kann dieser Test keine anderweitigen Abhängigkeitsstrukturen erkennen.

Auch den Von-Neumann-Ratio-Rang-Test zeichnet vor allem seine sehr gute Trennschärfe im AR(1)- und MA(1)-Fall aus, die nach der des Durbin-Watson-Tests die zweitbeste ist. Gleichzeitig zeigt er als rangbasiertes Testverfahren robuste Eigenschaften bei Kontaminationen, innovativen Ausreißern, anders verteilten Innovationen sowie einer wachsenden Varianz. Die verbesserte Trennschärfe gegenüber den anderen nicht-parametrischen Verfahren, wie dem Runs-Test, ist dabei auf die Mehrinformationen durch die Betrachtung der Ränge von den Beobachtungen zurückzuführen. Allerdings büßt der VNRR-Test dadurch geringfügig an Robustheit ein, was in dieser Arbeit z. B. im Fall von wachsenden Varianzen deutlich wird. Hier scheint der Runs-Test etwas geeigneter zu sein

Auch der Turning-Point-Test hebt sich von den anderen Testverfahren ab. So reagiert er in vielen Szenarien (Sprung, Trend, feste Anzahl von Oszillationen), wo andere Testverfahren deutliche Schwierigkeiten zeigten, äußerst robust und kann Autokorrelationen ähnlich gut wie unter Normalbedingungen erkennen. Weiterhin stellt die Einfachheit der Berechnung seiner Teststatistik einen wesentlichen Vorteil dieses Verfahrens dar. Seine Trennschärfe ist jedoch etwas schlechter als die der anderen nichtparametrischen Verfahren und die Annahmebereiche des Tests weisen Asymmetrien auf.

Eine Sonderstellung nimmt auch der Broock-Dechert-Schreinkman-Test ein. Er ist als einziger Test in der Lage, Fehlerkorrelationen im Rahmen von praxisrelevanten GARCH(1,1)-Prozessen zuverlässig zu erkennen. Auch zeigt er Ansätze, weitere komplexe Strukturen wie wachsende Varianzen, Oszillationen und Sprünge in der Zeitreihe detektieren zu können. Ein wesentlicher Nachteil dieses Tests ist jedoch, dass er einen großen Stichprobenumfang von ca. $N \geq 500$ benötigt, um zufriedenstellende Ergebnisse zu liefern. Damit ist er in kleinen Stichproben unbrauchbar. Auch wird er stark durch Kontaminationen in der Zeitreihe beeinflusst und weist in einigen Szenarien ein unberechenbares Verhalten auf.

Die gesamten Ergebnisse dieser Arbeit sind in Tabelle 4.1 zusammenfassend dargestellt. Dabei wurde für jedes betrachtete Verfahren beurteilt, wie gut es bei den unterschiedlichen zugrunde liegenden Prozessen und unter den verschiedenen Szenarien geeignet ist. Dabei ist eine grobe Einteilung in „gut geeignet“, „teilweise geeignet“ und „ungeeignet“ vorgenommen worden. Im Fall von Trends oder Sprüngen in der Zeitreihe war es weiterhin wichtig, zu unterscheiden, ob das jeweilige Verfahren robust regiert oder ob es die vorliegende Struktur erkennt.

Tabelle 4.1: Beurteilung der Eignung verschiedener Testverfahren in den betrachteten Prozessen und Szenarien (grün ≙ gut geeignet, gelb ≙ teilweise geeignet, rot ≙ ungeeignet; ✚ ≙ erkennt die Struktur, ✱ ≙ reagiert robust)

Test-verfahren	**Prozess**					**Szenario**							
	AR(1)	**AR(2)**	**SAR**	**MA(1)**	**Garch**	**a.ver.I.**	**Kont.**	**in.Au.**	**Sprung**	**Trend**	**w.Var.**	**f.Osz.**	**w.Osz.**
Durbin-Watson	✚	−	−	✚	−	✱	−	−	−	✱	✱	−	−
Ljung-Box	✚	✚	✚	✚	✚	✱	−	−	✚	✚	−	✚	✚
Von-Neumann	✚	−	−	✚	−	✱	✱	✱	−	−	✱	−	−
BDS	✚	−	−	✚	✚	✱	−	✱	✚	−	✚	✚	✚
Turning-Point	✚	✚	✚	−	−	✱	✱	✱	✱	✱	✱	✱	−
Runs	✚	−	−	✚	−	✱	✱	✱	−	−	✱	−	−
***K*-VZ**	✚	−	✚	−	−	✱	✱	✱	✚	✚	✱	✚	−
Simp. *K*-VZ	✚	✚	✚	✚	−	✱	✱	✱	−	−	✱	−	−

a.ver.I. ≙ anders verteilte Innovationen Kont. ≙ Kontaminationen in.Au. ≙ innovative Ausreißer w.Var. ≙ wachsende Varianzen
f.Osz. ≙ feste Anzahl Oszillationen w.Osz. ≙ wachsende Anzahl Oszillationen

4.2 Ausblick

Mit den gewonnenen Erkenntnissen über die Eigenschaften der K-Vorzeichentiefetests und ihrer vereinfachten Versionen kann ein breites Spektrum an Abweichungen von der Zufälligkeit getestet werden. So können lokale Abhängigkeiten durch nah beieinanderliegende K-Tupel erfasst werden, während weit auseinanderliegende K-Tupel Informationen über langfristige Strukturen enthalten. Die Wahl eines geeignetes Ks und der verwendeten Testversion trägt dabei massiv zur Trennschärfe der Verfahren in verschiedenen Szenarien bei.

Für zukünftige Untersuchungen erscheint es weiterhin sinnvoll, Kompromisse zwischen der vollständigen Version der K-VZ-Tests und ihren vereinfachten Versionen zu betrachten. Dabei wäre davon auszugehen, dass durch solche Modifikationen komplexere Strukturen erkannt werden können, während nicht zu viele nicht-informative Tupel betrachtet werden, die die Trennschärfe des Tests verschlechtern. Eine Möglichkeit besteht darin, K-Tupel aus allen möglichen Fenstern der Länge P mit $N \geq P \geq K$ in der Zeitreihe $(r_1, ..., r_N)$ zur Entscheidungsfindung heranzuziehen. Dabei wäre davon auszugehen, dass die gute Trennschärfe der vereinfachten Versionen des K-VZ-Tests mit kleinen Ks mit der breiteren Alternative der Testverfahren mit großen Ks kombiniert werden kann. Diese Teststatistik könnte als (K, P)-Vorzeichentiefe bezeichnet werden und ist dann definiert durch:

$$d_{K,P}(r_1, ..., r_N) = \frac{1}{(N-P+1) \cdot \binom{P}{K}} \sum_{i=1}^{N-P+1} \sum_{i \leq n_1 < ... < n_K \leq i+P-1} \left(\prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k r_{n_k} < 0\right\} \right).$$

Dabei entspricht diese Teststatistik bei der Wahl von $P = K$ derjenigen der vereinfachten Version der K-VZ-Tests $d_K^S(r_1, ..., r_N)$. Wird $P = N$ gewählt, so entspricht diese Größe genau der vollständigen K-Vorzeichentiefe $d_K(r_1, ..., r_N)$. Somit handelt es sich hier um eine logische Modifikation, die sich in die bereits vorhandenen Versionen des Tests einfügt.

Weiterhin konnte im Rahmen dieser Arbeit festgestellt werden, dass die K-VZ-Tests und ihre vereinfachten Versionen in vielen Situationen robuste Eigenschaften aufweisen. Dies hängt maßgeblich damit zusammen, dass lediglich die Vorzeichen der Beobachtungen zur Entscheidungsfindung herangezogen werden. Auf diese Weise gehen jedoch viele Informationen über die Zeitreihe verloren, die zu einer verbesserten Trennschärfe der Testverfahren führen könnten. So wurde in dieser Arbeit deutlich, dass die Trennschärfe des Von-Neumann-Ratio-Rang-Tests durch die Betrachtung der Ränge der Beobachtungen denjenigen Tests, die lediglich das sequenzielle Schema der Beobachtungen heranziehen, in vielen Situationen überlegen ist. Auch das Miteinbeziehen der Größen oder der Abstände zwischen Beobachtungen erlaubte es dem Ljung-Box- und dem Broock-Dechert-Schreinkmann-Test, Varianzänderungen in den Zeitreihen zu detektieren (s. Kap. 3.1, 3.2 u. 3.5).

Aus diesem Grund erscheint es sinnvoll, Modifikationen der vereinfachten K-VZ-Tests zu betrachten, die informativere Größen der Beobachtungen zur Entscheidungsfindung heranziehen. Dabei wäre davon auszugehen, dass derartige Modifikationen dazu führen, dass die Tests breitere Alternativen zur Zufälligkeit der Zeitreihe erfassen können und, dass ihre Trennschärfe bei Abhängigkeiten verbessert wird. Dafür wäre es denkbar, die Darstellung des wesentlichen Teils der K-VZ-Tests aus dem von Leckey et al. (2020) bewiesenen Lemma 1 auszunutzen (s. Kap. 2.9). Dieses Lemma besagt beispielhaft, dass die folgende Äquivalenz für den Fall der vereinfachten 2-VZ-Statistik gilt:

$$\sum_{n=1}^{N-1}\left(1\left\{r_n>0, r_{n+1}<0\right\}+1\left\{r_n<0, r_{n+1}>0\right\}-\frac{1}{2}\right)=\sum_{n=1}^{N-1}-\frac{1}{2}\psi\left(r_n\right)\psi\left(r_{n+1}\right),$$

wobei $\psi(x) = sign(x)$ hier der Vorzeichenfunktion entspricht, wie sie in Kapitel 2.9 definiert wurde.

Nun bietet es sich an, die Funktionen ψ anders zu definieren, sodass informativere Kenngrößen der Beobachtungen in die Teststatistik einfließen. Dabei ist zu beachten, dass der theoretische Median von $(\psi(R_1), ..., \psi(R_N))$ wieder 0 entsprechen sollte, damit die asymptotische Verteilung der Teststatistik nicht verändert wird. Dazu könnten die unveränderten Werte der Beobachtungen benutzt werden, das heißt, die Funktion könnte definiert werden als $\psi(x) = x$. Dies hätte jedoch zur Folge, dass der Test stark von Ausreißern beeinflusst würde, sodass er seine robusten Eigenschaften verlöre. Um einen Kompromiss zwischen Robustheit und Informationsausnutzung zu erzielen, könnten aber auch abgeschnittene Beobachtungen betrachtet werden. In diesem Fall könnte ψ für ein beliebigen Abschneidewert c definiert werden als:

$$\psi(x)=\begin{cases} x & \text{, falls } |x|\leq c \\ sign(x)\cdot c & \text{, falls } |x|>c. \end{cases}$$

Eine weitere Möglichkeit wäre es, ψ als Rangfunktion zu definieren, wodurch die asymptotische Verteilung der Teststatistik jedoch deutlich verändert werden würde.

Durch derartige Modifikation könnten die vollen und vereinfachten K-VZ-Tests insgesamt mehr Informationen aus den Zeitreihenwerten ziehen. Mit einer geeigneten Wahl von ψ könnten dabei gleichzeitig die robusten Eigenschaften der Testverfahren erhalten bleiben.

Literaturverzeichnis

R. L. Anderson. Distribution of the serial correlation coefficient. *The Annals of Mathematical Statistics*, 13(1):1–13, 1942.

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

R. Bender, A. Ziegler, und S. Lange. Multiple Regression - Artikel Nr. 13 der Statistik-Serie in der DMW. *DMW - Deutsche Medizinische Wochenschrift*, 127(Suppl. Statistik):8–10, 2002.

Bienaymé. Sur une question de probabilités. *Bulletin de la Societe Mathematique de France*, 2: 153–154, 1873.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

G. E. P. Box und D. A. Pierce. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

P. J. Brockwell und R. A. Davis. *Time Series: theory and methods*. Springer Series in Statistics. Springer, New York, NY, 2nd edition, reprint of the 1991 edition, 2006. ISBN 978-1-4419-0319-8.

P. J. Brockwell und R. A. Davis. *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer, New York, NY, 2nd edition, 2010. ISBN 0-387-95351-5.

W. A. Broock, D. A. Hsieh, und B. LeBaron. *Nonlinear dynamics, chaos, and instability: Statistical theory and economic evidence*. MIT Press, Cambridge, Mass., 3rd edition, 1993. ISBN 0-262-02329-6.

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

C. Brooks. Portmanteau model diagnostics and tests for nonlinearity: A comparative monte carlo study of two alternative methods. *Computational Economics*, 13(3):249–263, 1999.

F. Caeiro und A. Mateus. *randtests: Testing randomness in R*, 2014. URL `https://CRAN.R-project.org/package=randtests`. R package version 1.0.

G. M. Caporale, C. Ntantamis, T. Pantelidis, und N. Pittis. The bds test as a test for the adequacy of a garch(1,1) specification: a monte carlo study. *Economics and Finance Section, School of Social Sciences, Brunel University, Public Policy Discussion Papers*, 01 2004.

N. Davies und P. Newbold. Some Power Studies of a Portmanteau Test of Time Series Model Specification. *Biometrika*, 66(1):153, 1979.

N. Davies, C. M. Triggs, und P. Newbold. Significance Levels of the Box-Pierce Portmanteau Statistic in Finite Samples. *Biometrika*, 64(3):517–522, 1977.

J. Durbin und G. S. Watson. Testing for serial correlation in least squares regression. i. *Biometrika*, 37(3-4):409–428, 1950.

J. Durbin und G. S. Watson. Testing for serial correlation in least squares regression. iii. *Biometrika*, 58(1):1–19, 1971.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987, 1982.

R. W. Farebrother. Algorithm as 153: Pan's procedure for the tail probabilities of the durbin-watson statistic. *Applied Statistics*, 29(2):224, 1980.

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

G. D. Gupta und Z. Govindarajulu. Nonparametric tests of randomness against autocorrelated normal alternatives. *Biometrika*, 67(2):375–379, 1980.

M. J. Harrison und B. P. M. McCabe. Autocorrelation with heteroscedasticity: A note on the robustness of the durbin-watson, geary and henshaw tests. *Biometrika*, 62(1):214, 1975.

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

B. I. Hart und J. von Neumann. Tabulation of the Probabilities for the Ratio of the Mean Square Successive Difference to the Variance. *The Annals of Mathematical Statistics*, 13(2): 207–214, 1942.

M. Horn. *GSignTest: Robust Tests for Regression-Parameters via Sign Depth*, 2020. URL `https://github.com/melaniehorn/GSignTest`. R package version 1.0.5.

M. Horn und C. H. Müller. Tests based on sign depth for multiple regression. *SFB - discussion paper No. 823*, 2020. URL `https://eldorado.tu-dortmund.de/handle/2003/39065`.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48 (3-4):419–426, 1961.

T. Islam und E. Toor. Power comparison of autocorrelation tests in dynamic models. *International Econometric Review*, 11(2):58–69, 2019.

M. G. Kendall. *Time-series*. Griffin, London, 1973. ISBN 9780852642207.

H. S. Kim, D. S. Kang, und J. H. Kim. The BDS statistic and residual test. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 17(1-2):104–115, 2003.

W. Krämer. The power of the durbin-watson test for regressions without an intercept. *Journal of Econometrics*, 28(3):363–370, 1985.

C. P. Kustosz, A. Leucht, und C. H. Müller. Tests based on simplicial depth for ar(1) models with explosion. *Journal of Time Series Analysis*, 37(6):763–784, 2016a.

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

B. LeBaron. A fast algorithm for the bds statistic. *Studies in Nonlinear Dynamics & Econometrics*, 2(2), 1997.

K. Leckey, D. Malcherczyk, und 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`.

W. L. L'Esperance und D. Taylor. The power of four tests of autocorrelation in the linear regression model. *Journal of Econometrics*, 3(1):1–21, 1975.

R. Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18 (1):405–414, 1990.

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

A. Madansky. *Prescriptions for Working Statisticians*. Springer Texts in Statistics. Springer, New York, NY, 1988. ISBN 978-1-4612-3794-5.

D. Malcherczyk, K. Leckey, und C. H. Müller. K-sign depth: From asymptotics to efficient implementation. *SFB - discussion paper No. 823*, 2020. URL `https://eldorado.tu-dortmund.de/handle/2003/39100`.

A. Mateus und F. Caeiro. Comparing several tests of randomness based on the difference of observations. *AIP Conference Proceedings*, pages 809–812, 2013. doi: 10.1063/1.4825618.

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

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

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

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

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

N. E. Savin und K. J. White. The durbin-watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica*, 45(8):1989, 1977.

R. Schlittgen. *Zeitreihenanalyse*. Lehr- und Handbücher der Statistik. Oldenbourg Wissenschaftsverlag, s.l., 2001. ISBN 3-486-25725-0.

R. H. Shumway und D. S. Stoffer. *Time series analysis and its applications: With R examples*. Springer Texts in Statistics. Springer, Cham, 4th edition, 2017. ISBN 978-3-319-52451-1.

R. P. Stanley. A Survey of Alternating Permutations. *Department of Mathematics, M.I.T.*, 2009. URL `https://arxiv.org/pdf/0912.4240`.

A. Trapletti und K. Hornik. *tseries: Time Series Analysis and Computational Finance*, 2019. URL `https://CRAN.R-project.org/package=tseries`. R package version 0.10-47.

J. W. Tukey. Mathematics and the picturing of data. *Proc. int. Congr. Math., Vancouver 1974, Vol. 2, 523-531* , 1975.

M. Verbeek. *A guide to modern econometrics*. Wiley, Chichester, West Sussex, 4th edition, 2012.

V. A. Volkonskii und Y. A. Rozanov. Some limit theorems for random functions. i. *Theory of Probability & Its Applications*, 4(2):178–197, 1959.

J. von Neumann. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *The Annals of Mathematical Statistics*, 12(4):367–395, 1941.

A. Wald und J. Wolfowitz. On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

Y. Wang. Nonparametric tests for randomness. *Research report, UIUC*, 2003. URL `http://www.ifp.illinois.edu/~ywang11/paper/ECE461Proj.pdf`.

M. Wendler. *Empirical U-quantiles of dependent data*. Ruhr-Universität Bochum, Universitätsbibliothek, 2011.

L. C. Young. On randomness in ordered sequences. *The Annals of Mathematical Statistics*, 12 (3):293–300, 1941.

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

Eidesstattliche Versicherung
(Affidavit)

Dohme, Hendrik	209359
Name, Vorname (Last name, first name)	Matrikelnr. (Enrollment number)

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/Masterarbeit* mit dem folgenden Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present Bachelor's/Master's* thesis with the following title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution.

Titel der ~~Bachelor-~~/Masterarbeit*:
(Title of the ~~Bachelor's~~/ Master's* thesis):

Vergleich der Vorzeichentiefetests mit anderen Tests zur Überprüfung von Unabhängigkeitsannahmen in Zeitreihen

*Nichtzutreffendes bitte streichen
(Please choose the appropriate)

Dortmund, 13.01.21	H. Dohme
Ort, Datum (Place, date)	Unterschrift (Signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to €50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, section 63, subsection 5 of the North Rhine-Westphalia Higher Education Act (*Hochschulgesetz*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:**

Dortmund, 13.01.21	H. Dohme
Ort, Datum (Place, date)	Unterschrift (Signature)

**Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.