Technische Universität Dortmund

Fakultät für Statistik

Lehrstuhl für Statistik mit Anwendungen im Bereich der Ingenieurwissenschaften

Masterarbeit

Asymptotische Verteilungen von vollen Datentiefen

verfasst von Dennis Malcherczyk

Betreuerin:
Prof. Dr. Christine Müller

Gutachter:
Prof. Dr. Ivan Veselic

9. November 2018

Inhaltsverzeichnis

Vorwort

Ohne die Unterstützung folgender nennenswerter Menschen wäre die Erstellung meiner Masterarbeit in ihrer heutigen Form nicht realisierbar gewesen und daher möchte ich ihnen in diesem Vorwort meinen Dank aussprechen.
Vor allem bedanke ich mich bei meiner Betreuerin Prof. Dr. C.H. Müller für die Möglichkeit mich mit dem Thema dieser Arbeit zu beschäftigen. Das Fundament dieser Arbeit fußt auf den Forschungsergebnissen ihrer Arbeitsgruppe. In vielen Gesprächen hat sie mir Anregungen und Ideen mitgegeben, die der Schlüssel zur Findung einiger Resultate des vierten Kapitels dieser Arbeit gewesen sind. Für die Beantwortungen meiner Fragen ist sie stets offen und bereichernd. In ihrem Oberseminar konnte ich Resultate vorstellen und mit ihrer Arbeitsgruppe diskutieren.
Einen besonderen Dank möchte ich Dr. K. Leckey aussprechen. Durch seine Mitarbeit an dem Thema konnte die Vereinfachung von Beweisen und Verbesserung einiger Resultate aus Kustosz et al. (2016a) realisiert werden. Außerdem hat er die Idee gehabt, dass sich die volle Dreier-Tiefe exakt in linearer statt kubischer Laufzeit bestimmen lässt. Ich habe die Stellen in meiner Arbeit, die durch ihn geprägt worden sind, im Text kenntlich gemacht. Insbesondere möchte ich mich bei ihm dafür bedanken, dass er einen schwer versteckten Fehler in meinem Beweis in Kapitel 4 fand, mir einige mathematisch-technische Hinweise beim Lesen meiner Beweise mitgab und ich durch Diskussionen mit ihm mein Wissen erweitern durfte.
Ferner danke ich Prof. Dr. I. Veselic für sein Mitinteresse am Thema und seine Unterstützung als Gutachter aus der Fakultät Mathematik und Prof. Dr. H. Blum für die Genehmigung dieser Arbeit.
Meinen Eltern Marzanna und Andreas Malcherczyk sowie meiner Schwester Aleksandra Malcherczyk verdanke ich die Mitrealisierung und Unterstützung meines Studiums und in anderen Lebensbereichen, ohne die ich niemals das erreicht hätte, was ich heute in meinen Händen halte.
Zum Abschluss danke ich vier meiner Kommilitonen Xenia Kerkhoff, Lars Schroeder, Ulviyya Ibrahimli und vor allem Karl Burak Strebel für die vielen inspirierenden mathematischen Diskussionen und Anregungen während meiner Studienzeit!

1 Einleitung

In der mathematischen Statistik wird der Zusammenhang verschiedener Einflussgrößen auf eine Zielgröße durch Regressionsmodelle untersucht. Aus einer vorgegeben Klasse $\Theta \subseteq \mathbb{R}^p$ von Regressionsmodellen, indiziert durch p-dimensionale Parametervektoren $\vartheta \in \Theta$, soll sich anhand von Daten für ein passendes Regressionsmodell entschieden werden. Die Daten interpretieren wir als Realisationen eines Zufallsexperiments von Zufallsvariablen $Y_1, ..., Y_N$ mit der Gestalt:

$$Y_n = g_{\vartheta^*}(X_n) + E_n, \text{ für } n = 1, ..., N,$$

wobei zu jedem Parametervektor $\vartheta \in \Theta$ eine Regressionsfunktion g_ϑ definiert wird und $\vartheta^* \in \Theta$ der wahre Parameter sei. Die Zufallsvariablen $E_1, ..., E_N$ wirken als additive Fehler auf die Regressionsfunktion g_ϑ. Die Regressoren $X_1, ..., X_N$ können zufällig oder deterministisch sein. Für die Realisationen $y_1, ..., y_N$ bzw. $e_1, ..., e_n$ bzw. $x_1, ..., x_N$ der Zufallsvariablen $Y_1, ..., Y_N$ bzw. $E_1, ..., E_n$ bzw. $X_1, ..., X_N$ verwenden wir Kleinbuchstaben. Um aus gegebenen Daten die Qualität eines vorgeschlagenen Parameters ϑ zu beurteilen, benötigen wir einen Anpassungsbegriff, aus dem sich eine Möglichkeit zur Schätzung von ϑ ergibt. Ein klassischer Ansatz ist die Methode der Kleinste-Fehler-Quadrate. Dabei fällt die Entscheidung auf das Regressionsmodell aus der Klasse Θ, bei dem die quadratischen Abweichungen zwischen den beobachteten Daten und dem vermuteten Modell zu einem $\vartheta \in \Theta$ minimal sind, d.h. der Kleinste-Fehler-Quadrate Schätzer (KQ-Schätzer) besitzt die Gestalt (Toutenburg (2003), S. 89ff.):

$$\widehat{\vartheta}_{KQ} := \underset{\vartheta \in \Theta}{\operatorname{argmin}} \left(\sum_{n=1}^{N} \left(y_n - g_\vartheta(x_n) \right)^2 \right).$$

In Anwendungsfeldern mit extremen Werten beruhend auf starken Tails wie in der Hochwasserstatistik (siehe z.B. Fischer et al. (2015)) oder mit kontaminierten Daten, wo Ausreißer in Form von Messfehlern beobachtet werden (Büning (1991), S. 5ff.), sollten robuste Verfahren bevorzugt werden. Durch das nachfolgende Beispiel wird in der Abbildung 1 die Empfindlichkeit des KQ-Schätzers $\widehat{\vartheta}_{KQ}$ gegenüber Ausreißern

illustriert. Die schwarz markierten Punkte in der Abbildung 1 entsprechen 20 Realisationen des folgenden Modells von einer Geradengleichung für die deterministischen Regressoren $x_1 = 0.1, x_2 = 0.2, ..., x_{19} = 1.9, x_{20} = 2$:

$$y_n = x_n + e_n, \text{ für } n = 1, ..., 20. \tag{1.1}$$

Die angegebene Modellgleichung (1.1) wird dabei als spezielles Modell folgender Klasse von Regressionsmodellen von Geradengleichungen mit Regressionsfunktion $g_\vartheta(x) = \vartheta_0 + \vartheta_1 x$ für $x \in \mathbb{R}$ für den Parameter $\vartheta = (\vartheta_0, \vartheta_1)^\top = (0, 1)^\top$ aufgefasst:

$$y_n = \vartheta_0 + \vartheta_1 x_n + e_n, \text{ für } n = 1, ..., 20. \tag{1.2}$$

Wir gehen bei der Schätzung von $\widehat{\vartheta}_{KQ}$ davon aus, dass das Regressionsmodell in (1.2) den Daten zugrunde liegt und müssen aus den Daten ein geeignetes ϑ innerhalb dieser Klasse schätzen. Für $n = 1, ..., 20$ entsprechen e_n Realisationen von unabhängigen $\mathcal{N}\left(0, \frac{1}{2}\right)$-verteilten Zufallsvariablen. Die Anpassung der in schwarz

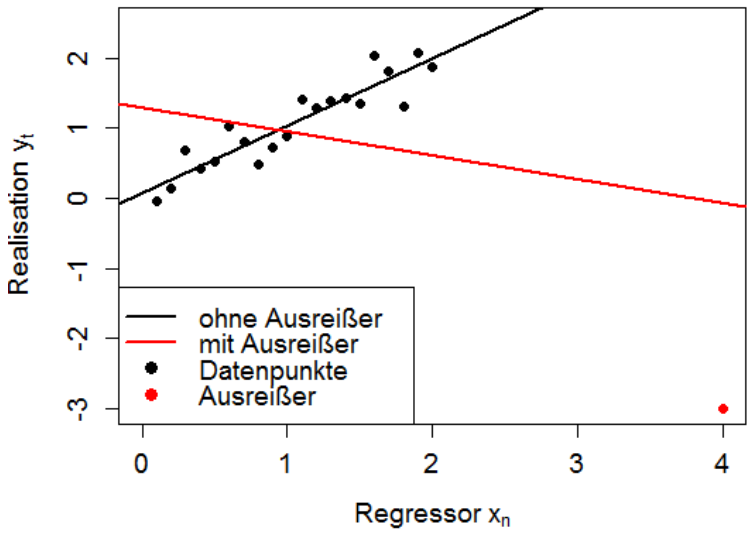


Abbildung 1: Zwei KQ-Schätzungen: in schwarz ohne und in rot mit Ausreißerpunkt

dargestellten Regressionsgeraden durch die KQ-Schätzung $\widehat{\vartheta}_{KQ} = (0.078, 0.960)^\top$ für die schwarzen Datenpunkte ist zufriedenstellend. Kontaminiert man den Da-

tensatz mit einem Ausreißer $(x_{21}, y_{21})^\top = (4, -3)^\top$, dargestellt als roter Punkt in Abbildung 1, so liefert die in rot dargestellte Regressionsgerade, bei der der Ausreißer zur Schätzung von ϑ mit verwendet wird, kein zufriedenstellendes Ergebnis.
Wir können die Regressionsgerade insbesondere beliebig durch einen neu hinzugefügten Punkt verändern und sprechen daher von einem **unrobusten Verfahren** (siehe Maronna et al. (2006), S.87ff, für eine ausführliche Diskussion).
Für einen robusten Ansatz wird der Anpassungsbegriff des Modells verändert:
Damit einzelne große Abweichungen keinen zu starken Einfluss besitzen, verwenden wir Anpassungsbegriffe beruhend auf Vorzeichen. Dazu fordern wir für den Median der Fehler: $\text{med}(E_n) = 0$ und $P(E_n \neq 0) = 1$ für $n = 1, ..., N$. Dadurch ergeben sich P-fast sicher Zufallsgrößen mit bestimmbaren Vorzeichen jeweils mit Wahrscheinlichkeit $\frac{1}{2}$. Diese Bedingung ist keine große Einschränkung, da in vielen Regressionsmodellen symmetrische Modellfehler verwendet werden. Ergeben sich für einen vorgeschlagenen Parametervekor $\vartheta \in \Theta$ gleich viele positive und negative Vorzeichen, so soll dies für eine gute Anpassung sprechen. Wir halten dies in einem ersten Fazit fest:

1. *Der KQ-Schätzer reagiert sensibel auf einzelne extreme Werte. Ein robuster Ansatz wird durch die Betrachtung von Vorzeichen gewährleistet.*

Die hier präsentierte Anwendung von Vorzeichen besitzt allerdings noch einige Schwächen, wie in einem weiteren Beispiel in der Abbildung 2 illustriert wird. In rot sehen wir dort die wahre Modellgleichung (1.3)

$$y_n = x_n^2 + e_n \text{ für } n = 1, ..., 11, \tag{1.3}$$

das für $\vartheta = (\vartheta_0, \vartheta_1, \vartheta_2)^\top = (0, 0, 1)^\top$ einem konkreten Modell der folgenden Modellklasse (1.4) entspricht:

$$y_n = \vartheta_0 + \vartheta_1 x_n + \vartheta_2 x_n^2 + e_n \text{ für } n = 1, ..., 11, \tag{1.4}$$

wobei $x_1 = -1, x_2 = -0.8, ..., x_{10} = 0.8, x_{11} = 1$ deterministische Regressoren sind und $g_\vartheta(x) = \vartheta_0 + \vartheta_1 x + \vartheta_2 x^2$ die Regressionsfunktion für ein $x \in \mathbb{R}$ ist. Die Vorzeichen

der Abweichungen von den Daten werden in der Abbildung 2 unten dargestellt. Wir erhalten für das wahre Modell eine gute Anpassung, da die Vorzeichen ein Verhältnis von 5 : 6 besitzen. Es gibt aber viele unpassende Modelle mit einer genauso guten

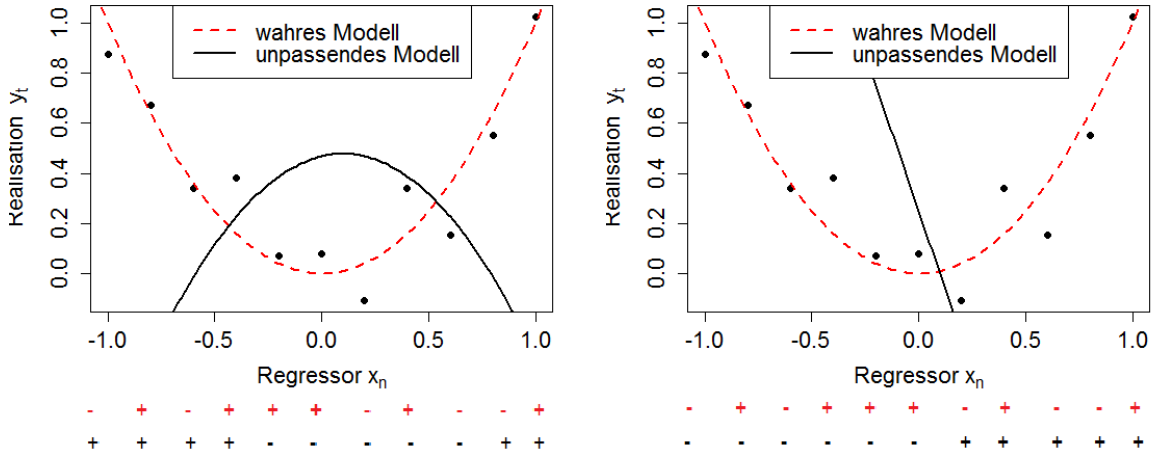


Abbildung 2: Vergleich: wahres Modell und unpassendes Modell und die Vorzeichen der elf Residuen pro Modell

Anpassung. Jeweils in schwarz links mit $(\vartheta_0, \vartheta_1, \vartheta_2)^\top = (0.4675, 0.2, -1)^\top$ und rechts mit $(\vartheta_0, \vartheta_1, \vartheta_2)^\top = (0.25, -2.5, 0)^\top$ wird ein unpassendes Modell angegeben, das aber ebenso ein Vorzeichen-Verhältnis von 5 : 6 besitzt. Wir halten fest:

2. Die reine Betrachtung der Vorzeichen-Verhältnisse ist zwar robust aber unzulänglich, da viele unpassende Modelle nicht aussortiert werden.

Analysiert man den obigen Ansatz, so können wir kritisieren, dass die Reihenfolge der auftretenden Vorzeichen nicht berücksichtigt wird. In den Gegenbeispielen kommen lange Läufe von gleichartigen Vorzeichen vor. Solche langen Läufe sind unter der Unabhängigkeit der Fehler $E_1, ..., E_N$ unwahrscheinlich. Die Verhältnisse unterschiedlicher Vorzeichen genügt nicht, um ein passendes Modell zu finden. Wir sollten ebenso Vorzeichenwechsel berücksichtigen.

3. Werden Vorzeichenwechsel berücksichtigt, so sind die Anforderungen an ein Modell-Kandidat höher. Wir erwarten so Verbesserungen des robusten Verfahrens.

In dieser Arbeit werden volle K-Datentiefen betrachtet. Das sind statistische Maßzahlen, die die relative Anzahl aller geordneten Tupel der Länge K aus den Ab-

weichungen vom vorgeschlagenen Modell-Kandidat und Daten mit $K-1$ enthaltenen Vorzeichenwechsel zählen. Ziel ist es, die asymptotische Verteilung der vollen K-Datentiefen herzuleiten, um robuste statistische Testverfahren beruhend auf der vollen Tiefe mit Hypothesenpaaren folgender Bauart zu konstruieren:

$$H_0 : \vartheta \in \Theta_0 \text{ vs. } H_1 : \vartheta \in \Theta_1,$$

wobei $\Theta_0 \uplus \Theta_1 = \Theta$ gilt. So können wir testen, ob vorgegebene Parameterbereiche Θ_0 innerhalb eines Regressionsmodells als unpassend deklariert werden können.
In Kapitel 2 wird die volle Zweier-Tiefe als einfachster Fall vorgestellt, bei dem die Vorzeichenwechsel aller Paare gezählt werden und bereits in Müller (2005) als Spezialfall betrachtet worden ist. Dieses Kapitel soll dazu dienen, die Konzepte und Methoden im simpelsten Fall vorzustellen. In Kapitel 3 werden mit der vollen Dreier-Tiefe Dreier-Tupel gezählt, bei denen zwei Vorzeichenwechsel vorliegen. Die Methoden zur Untersuchung der Asymptotik fußen auf den Resultaten des Kapitels 2 und besitzen eine analoge, aber ebenso aufwändigere Struktur. Die Grundideen stammen aus Kustosz et al. (2016a), wobei wir in dieser Masterarbeit viele Aussagen verschärfen und Beweise vereinfachen werden. Die Vereinfachungen liefern zum einen eine Möglichkeit zur Implementierung der vollen Dreier-Tiefe in linearer statt kubischer Laufzeit. Andererseits können durch Adaption der Vereinfachungen der Beweise aus Kustosz et al. (2016a) in Kapitel 4 höhere Tiefen betrachtet und eine Herleitung der asymptotischen Verteilung der Vierer-Tiefe bestimmt werden, deren asymptotische Verteilung bisher noch unbekannt gewesen ist. Insbesondere werden die Beweismittel und Ideen offen gelegt, um für die nächsten höheren Tiefen die asymptotische Verteilung zu bestimmen. Im Kapitel 5 wird ein Überblick zu einigen Anwendungsfeldern der bisherigen Resultate gegeben, sowie Anregungen für weitere Untersuchungen.
Statt der Verwendung von robusten Verfahren könnte man vorschlagen, Ausreißer zu entfernen. Hier sollen einige Argumente gegen ein solches Vorgehen genannt werden:

(a) Durch frei wählbare Kriterien, wie z.B. festgelegte Schwellenwerte, klassifiziert man Daten zu Ausreißern. Dabei liegt ein willkürlicher Spielraum bei der Wahl solcher Kriterien vor und die Gefahr von bewussten und unbewussten

Manipulationen besteht (Maronna et al. (2006), S. 3).

(b) In hochdimensionalen Datenstrukturen ist es sehr schwierig Ausreißer zu klassifizieren. Es werden gerade robuste Verfahren eingesetzt, um in solchen Datenstrukturen Ausreißer zu entdecken (Maronna et al. (2006), S. XVI).

(c) Das Entfernen von Daten verändert zugrundeliegende Verteilungsannahmen, die eigentlich bei den verwendeten Testverfahren mitberücksichtigt werden müssten (Maronna et al. (2006), S. 4).

Ferner existieren neben der Betrachtung von Vorzeichen andere robuste Methoden, wie das Trimmen von Datensätzen (Maronna et al. (2006), S. 31f.) oder z.B. die von Huber eingeführten verallgemeinerten Maximum-Likelihood-Schätzer (Maronna et al. (2006), S. 34f.). Allerdings arbeiten diese Verfahren mit frei zu wählenden Parametern (beim Trimmen) oder Score-Funktionen (bei M-Schätzern), wodurch sich ebenso im Gegensatz zu den Vorzeichen-Tiefen ein willkürlicher Spielraum ergibt.
Zum Abschluss der Einleitung soll erwähnt werden, dass die historische Entwicklung von Vorzeichen-Tiefen anders verlief als es in der obigen Darstellung motiviert wird. Die sogenannten Simplex-Tangent-Tiefen, eingeführt von Rousseeuw und Hubert (1999) und als Tangent-Tiefen von Mizera (2002) weitergeführt, lassen sich durch die Vorzeichen-Tiefen charakterisieren, sofern das Regressionsmodell Regularitäten erfüllt (siehe Kustosz et al. (2016b)). Für viele Modelle sind dort die Regularitätseigenschaften nachgeprüft worden. Mathematische Analysen der Simplex-Tangent-Tiefe mit ihrer ursprünglichen Definition erweisen sich als schwierig. Dadurch lässt sich das Studium von Vorzeichen-Tiefe motivieren, da in vielen Modellen die Begriffe äquivalent sind und die Untersuchungen von Vorzeichen-Tiefen oft die Untersuchungen von Simplex-Tangent-Tiefen einschließen.

2 Asymptotische Verteilung der vollen Zweier-Tiefe

Wir beginnen die Untersuchung der asymptotischen Verteilung von vollen Datentiefen mit der **vollen Zweier-Tiefe** als einfachsten Fall. Bevor wir in Kapitel 2.2 die asymptotische Verteilung der vollen Zweier-Tiefe herleiten, werden in Kapitel 2.1 das Regressionsmodell und dazugehörige Notationen im Detail dargestellt, welche global in der gesamten Arbeit gelten.

2.1 Grundbegriffe

Wir fassen das vorgestellte Regressionsmodell aus der Einleitung der Übersicht wegen in der nachfolgenden Definition zusammen (Kustosz et al. (2016a)):

Definition 2.1 (Globales Regressionsmodell der Arbeit).
Seien $E_1, ..., E_N$ unabhängig, identisch verteilte reellwertige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit $med(E_n) = 0$ und $P(E_n \neq 0) = 1$ für $n = 1, ..., N$. Seien $X_1, ..., X_N$ reellwertige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, die P-fast sicher folgende Ordnungsstruktur erfüllen

$$X_1 < X_2 < X_3 < ... < X_N.$$

Für jedes $\vartheta \in \Theta \subseteq \mathbb{R}^p$ definieren wir eine Regressionsfunktion $g_\vartheta : \mathbb{R} \to \mathbb{R}$. Wir definieren $Y_1, ..., Y_N$ durch folgenden Zusammenhang:

$$Y_n = g_{\vartheta^*}(X_n) + E_n \text{ für } n = 1, ..., N \text{ und ein festes } \vartheta^* \in \Theta.$$

*Seien $Z_n = (Y_n, X_n)$ für $n = 1, ..., N$ die zufälligen Datenpunkte des Regressionsmodells. Für jedes $\vartheta \in \Theta$ definieren wir die **Residuen**:*

$$res(\vartheta, Z_n) := Y_n - g_\vartheta(X_n) \text{ für } n = 1, ..., N.$$

Wir fordern für die Regressoren $X_1, ..., X_N$ keine stochastische Unabhängigkeit, um z.B. autoregressive Zeitreihenmodelle, wie in Kustosz et al. (2016a) zu verwenden. Deterministische Regressoren sind als Spezialfall von konstanten Zufallsvariablen ebenso erlaubt. Für $n = 1, ..., N$ liefert die Anforderung $P(E_n \neq 0) = 1$ an den Modellfehler ein eindeutiges Vorzeichen und $\text{med}(E_n) = 0$ wird benötigt, um die nachfolgenden Vorzeichen-Methoden zu verwenden. Die Residuen dienen dazu, die Anpassung eines Parameters ϑ zu bewerten. Zur Formulierung eines statistischen Entscheidungsproblems fassen wir $\text{res}(\vartheta, Z_1)$, ..., $\text{res}(\vartheta, Z_N)$ für jedes $\vartheta \in \Theta$ als Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P_\vartheta)$ auf. Dadurch können wir auf dem statistischen Raum $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta})$ arbeiten. Für die Zufallsvariablen $E_1, ..., E_N$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ entspricht hierbei $P = P_{\vartheta^*}$, wobei ϑ^* der wahre Parameter ist, da für ϑ^* sich die Residuen zum Modellfehler reduzieren:

$$\text{res}(\vartheta^*, Z_n) = E_n \text{ für } n = 1, ..., N \tag{2.1}$$

und für das Vorzeichen der Residuen gilt unter dem wahren Parameter:

$$P(\text{res}(\vartheta^*, Z_n) > 0) = P(\text{res}(\vartheta^*, Z_n) < 0) = \frac{1}{2} \text{ für } n = 1, ..., N. \tag{2.2}$$

Die folgende zusätzliche Anforderung an die Residuen für beliebige $\vartheta \in \Theta$ unter der wahren Verteilung P_{ϑ^*} ist für praktische Zwecke nützlich:

$$P(\text{res}(\vartheta, Z_n) > 0) + P(\text{res}(\vartheta, Z_n) < 0) = 1 \text{ für } n = 1, ..., N. \tag{2.3}$$

Unter dem wahren Parameter kennen wir das Verhalten der Vorzeichen der Residuen. Durch diese Modellierung können wir mit den Vorzeichen folgenden robusten Anpassungsbegriff verwenden (Kustosz et al. (2016a)):

Definition 2.2 (Die volle K-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 definieren wir für einen vorgegebenen Parameter $\vartheta \in \mathbb{R}^p$ und eine Realisation $z = (z_1, ..., z_N)$ von den Zufalls-

variablen $Z_1, ..., Z_N$ des Regressionsmodells die ***volle K-Tiefe****:*

$$d_S^K(\vartheta, z) = \frac{1}{\binom{N}{K}} \sum_{1 \leq n_1 < n_2 < ... < n_K \leq N} \left(\prod_{k=1}^{K} \mathbb{1}\{res(\vartheta, z_{n_k})(-1)^k > 0\} + \prod_{k=1}^{K} \mathbb{1}\{res(\vartheta, z_{n_k})(-1)^{k+1} > 0\} \right).$$

Die Indikatorfunktion $\mathbb{1}\{\text{Bedingung}(\vartheta, z_n)\}$ wird für ein gegebenes Paar (ϑ, z_n) folgendermaßen definiert:

$$\mathbb{1}\{\text{Bedingung}(\vartheta, z_n)\} := \begin{cases} 1, & \text{falls die Bedingung für } (\vartheta, z_n) \text{ erfüllt wird} \\ 0, & \text{falls die Bedingung für } (\vartheta, z_n) \text{ nicht erfüllt wird.} \end{cases}$$

Falls mehrere Bedingungen gleichzeitig erfüllt sein sollen, werden sie durch die Verwendung von Kommata aufgezählt. Ferner beschreibt $\sum_{1 \leq n_1 < n_2 < ... < n_K \leq N}$ die Summe über alle geordneten Teilmengen aus $\{1, ..., N\}$ indiziert durch $n_1, n_2, ..., n_K$ mit der Ordnungsstruktur $n_1 < n_2 < ... < n_K$. Insgesamt liegen $\binom{N}{K}$ Summanden vor, wodurch die volle K-Tiefe in Definition 2.2 Werte in $[0, 1]$ annimmt, da die Summanden $\{0, 1\}$-wertig sind. Die volle K-Tiefe beschreibt unter allen geordneten K-Tupeln den relativen Anteil mit $(K - 1)$ Vorzeichenwechsel. In Kustosz et al. (2016b) werden noch andere Tiefen und ihre Asymptotik vorgestellt, die nicht alle K-Tupel betrachten, um geringere Rechenzeiten zu ermöglichen. Da hier alle geordneten K-Tupel betrachtet werden, wird die Tiefe in Definition 2.2 *voll* genannt. Die Laufzeit der vollen K-Tiefe mit der Darstellung aus Definition 2.2 ist polynomiell vom Grad K. Zur asymptotischen Untersuchung der vollen K-Tiefe betrachten wir den Zufallsvektor $Z = (Z_1, ..., Z_N)$ aus dem Regressionsmodell in Definition 2.1 unter dem wahren Parameter ϑ^*. Wir müssen die Zufallsvariable $d_S^K(\vartheta^*, Z)$ zunächst zentrieren und bestimmen dazu den Erwartungswert unter der wahren Verteilung:

$$\mathbb{E}_{\vartheta^*}(d_S^K(\vartheta^*, Z)) = \frac{1}{\binom{N}{K}} \sum_{1 \leq n_1 < n_2 < ... < n_K \leq N} \mathbb{E}_{\vartheta^*} \left(\prod_{k=1}^{K} \mathbb{1}\{\text{res}(\vartheta^*, Z_{n_k})(-1)^k > 0\} + \prod_{k=1}^{K} \mathbb{1}\{\text{res}(\vartheta^*, Z_{n_k})(-1)^{k+1} > 0\} \right)$$

$$= \frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \Bigg(P\Big(\bigcap_{k=1}^{K} \{\mathrm{res}(\vartheta^*, Z_{n_k})(-1)^k > 0\} \Big) + P\Big(\bigcap_{k=1}^{K} \{\mathrm{res}(\vartheta^*, Z_{n_k})(-1)^{k+1} > 0\} \Big) \Bigg) \tag{2.4}$$

Da die Laufindizes in der geordneten Summe in (2.4) alle unterschiedlich sind, kann mit Anwendung von (2.1) auf die Residuen und mit (2.2) durch die Unabhängigkeit der Zufallsvariablen $E_1, ..., E_N$ die Formel (2.4) wie folgt dargestellt werden:

$$\begin{aligned}
&\frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left(\prod_{k=1}^{K} P(E_{n_k}(-1)^k > 0) + \prod_{k=1}^{K} P(E_{n_k}(-1)^{k+1} > 0) \right) \\
=&\frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left(\left(\frac{1}{2}\right)^K + \left(\frac{1}{2}\right)^K \right) \\
=&\frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left(\frac{1}{2}\right)^{K-1} = \left(\frac{1}{2}\right)^{K-1},
\end{aligned}$$

wobei die Annahme $\mathrm{med}(E_n) = 0$ für $n = 1, ..., N$ verwendet wird. Demnach wird $\mathbb{E}_{\vartheta^*}(d_S^K(\vartheta^*, Z)) = \left(\frac{1}{2}\right)^{K-1}$ zur Zentrierung der vollen K-Tiefe genutzt. Wir werden für den Fall $K = 2, 3, 4$ in den nachfolgenden Kapiteln sehen, dass der zusätzliche Vorfaktor N als Standardisierung für eine sinnvolle asymptotische Verteilung dient:

$$N \left(d_S^K(\vartheta^*, Z) - \left(\frac{1}{2}\right)^{K-1} \right).$$

Die Verwendung des Vorfaktors N kann mathematisch motiviert werden. Der Vorfaktor wird in den folgenden Rechnungen nach Umformungen den Vorfaktor $\frac{1}{\sqrt{N}}$ ergeben, was die Anwendung des Zentralen Grenzwertsatzes ermöglicht. Es wird vermutetet, dass für $K \ge 5$ der Vorfaktor N ebenso sinnvoll ist, was allerdings in dieser Arbeit nicht gezeigt wird.

2.2 Asymptotik der vollen Zweier-Tiefe

In diesem Abschnitt beschäftigen wir uns mit der Asymptotik der vollen Zweier-Tiefe. Die volle Zweier-Tiefe entspricht der Definition 2.2 für $K = 2$:

$$d_S^2(\vartheta, z) = \frac{1}{\binom{N}{2}} \sum_{1 \le n_1 < n_2 \le N} \Big(\mathbb{1}\{\mathrm{res}(\vartheta, z_{n_1}) > 0, \mathrm{res}(\vartheta, z_{n_2}) < 0\} + \mathbb{1}\{\mathrm{res}(\vartheta, z_{n_1}) < 0, \mathrm{res}(\vartheta, z_{n_2}) > 0\} \Big).$$

Die Untersuchung der Asymptotik von vollen Datentiefen beruht auf der Verwendung von Zentralen Grenzwertsätzen. Allerdings können wir den Zentrale Grenzwertsatz nicht direkt anwenden, da die Summanden bei vollen Datentiefen nicht stochastisch unabhängig sind. Eine geschickte Umsortierung der Summe wird dieses Problem lösen. Dazu stellen wir die einzelnen Summanden als Produkt der Funktion $\Phi(x) := \mathbb{1}\{x < 0\} - \mathbb{1}\{x > 0\}$ dar und erweitern die geordnete Doppelsumme $\sum_{1 \le n_1 < n_2 \le N}$ zu einer Doppelsumme der Form $\sum_{n_1, n_2 = 1}^{N}$ durch Nulladditionen. Die Doppelsumme kann anschließend als eine einzelne Summe ins Quadrat dargestellt werden. Folgendes Lemma liefert für die volle Zweier-Tiefe summandenweise eine Produktdarstellung durch die Funktion Φ (Kustosz et al. (2016a)):

Lemma 2.3 (Φ-Darstellung der vollen Zweier-Tiefe).
Für Zufallsvariablen E_{n_1}, E_{n_2} mit $P(E_{n_i} \neq 0) = 1$ für $i = 1, 2$ gilt:

$$\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} - \frac{1}{2} = -\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}) \ \textit{P-fast sicher},$$

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

Tabelle 1: Fälle aller möglichen Realisationen bei zwei Zufallsvariablen

E_{n_1}	E_{n_2}	$\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\}$ $+\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} - \frac{1}{2}$	$-\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2})$
+	+	$0 + 0 - \frac{1}{2}$	$-\frac{1}{2} \cdot (-1) \cdot (-1)$
+	-	$1 + 0 - \frac{1}{2}$	$-\frac{1}{2} \cdot (-1) \cdot 1$
-	+	$0 + 1 - \frac{1}{2}$	$-\frac{1}{2} \cdot 1 \cdot (-1)$
-	-	$0 + 0 - \frac{1}{2}$	$-\frac{1}{2} \cdot 1 \cdot 1$

Beweis. Wir zeigen die Gleichheit in Tabelle 1, in dem wir alle $2^2 = 4$ Fälle für die Vorzeichen von E_{n_1}, E_{n_2} betrachten. Die Fälle $E_{n_1} = 0$ und $E_{n_2} = 0$ treten P-fast sicher nie ein und brauchen daher nicht betrachtet werden. □

Nun folgen zwei wichtige Hilfsmittel für die Untersuchung der Asymptotik. Der Zentrale Grenzwertsatz liefert asymptotische Normalität (Klenke (2006), S. 304ff.):

Satz 2.4 (Zentraler Grenzwertsatz).
Für eine Folge $(E_n)_{n\in\mathbb{N}}$ unabhängig, identisch verteilter reellwertiger Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit $Var(X_1) > 0$ gilt:

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \frac{E_n - I\!E(E_1)}{\sqrt{Var(X_1)}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

Die obige Notation $\xrightarrow[N\to\infty]{\mathcal{D}}$ stellt die Konvergenz in Verteilung von Zufallsvariablen dar (Klenke (2006), S. 243). Da uns quadrierte Summen von Zufallsvariablen begegnen werden, benötigen wir zudem ein Stetigkeitsargument in Form des Continuous-Mapping-Theorems (Klenke (2006), S. 245):

Satz 2.5 (Continuous-Mapping-Theorem).
Seien (S_i, d_i) metrische Räume versehen mit der Borel-σ-Algebra $\mathcal{B}(S_i)$ für $i = 1, 2$ und sei $\psi : S_1 \to S_2$ $\mathcal{B}(S_1) - \mathcal{B}(S_2)$-messbar. Dann gelten folgende Aussagen:

(i) Die Menge U_ψ der Unstetigkeitsstellen von ψ ist $\mathcal{B}(S_1)$-messbar.

(ii) Für eine Folge von S_1-wertiger Zufallsvariablen $(X_N)_{N\in\mathbb{N}}$ und eine S_1-wertige Zufallsvariable mit $P(X \in U_\psi) = 0$ und $X_N \xrightarrow[N\to\infty]{\mathcal{D}} X$ gilt,

$$\psi(X_N) \xrightarrow[N\to\infty]{\mathcal{D}} \psi(X).$$

Wir verwenden im nächsten Satz das Continuous-Mapping-Theorem für reellwertige Zufallsvariablen, die mit der Abbildung $\psi : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ zu einer neuen reellwertigen Zufallsvariable transformiert werden. Dazu versehen wir die Menge $\mathbb{R}$ kanonisch mit folgender Metrik

$$d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad d(x, y) := |x - y|$$

und gewährleisten die Stetigkeit von ψ. Das Continuous-Mapping-Theorem wird ferner in abstrakteren Räumen verwendet, wie wir im dritten und vierten Kapitel sehen werden. Mit dem Zentralen Grenzwertsatz und dem Continuous-Mapping-Theorem wird die asymptotische Verteilung der vollen Zweier-Tiefe im nächsten Satz hergeleitet.

Satz 2.6 (Asymptotische Verteilung der vollen Zweier-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 gilt

$$N\left(d_S^2(\vartheta^*, Z) - \frac{1}{2}\right) \xrightarrow[N\to\infty]{\mathcal{D}} \frac{1}{2}(1 - X),$$

wobei X eine χ_1^2-verteilte Zufallsvariable ist.

In Müller (2005) wird eine analoge Aussage für Simplex-Tangent-Tiefen mit Methoden für U-Statistiken gezeigt. Für die volle Zweier-Tiefe können die gleichen Methoden verwendet werden (Kustosz und Müller (2014)). Der folgende Beweis von Satz 2.6 ist vom Autor geführt und an der Beweisstruktur aus Kustosz et al. (2016a) zur Asymptotik der vollen Dreier-Tiefe angepasst worden. Durch den Beweis des Autors können die Analogien bei der Herleitung der asymptotischen Verteilung für den Fall $K = 2$ zum Fall $K = 3$ aufgezeigt werden.

Beweis. Wir verwenden die Φ-Darstellung der vollen Zweier-Tiefe aus Lemma 2.3 und erhalten P-fast sicher folgende Gleichheit:

$$\begin{aligned}
&N\left(d_S^2(\vartheta^*, Z) - \frac{1}{2}\right) \\
&= \frac{N}{\binom{N}{2}} \left(\sum_{1 \le n_1 < n_2 \le N} \left(\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} - \frac{1}{2} \right) \right) \\
&= \frac{N}{\binom{N}{2}} \sum_{1 \le n_1 < n_2 \le N} \left(-\frac{1}{2} \Phi(E_{n_1}) \Phi(E_{n_2}) \right). \qquad (2.5)
\end{aligned}$$

Das Ziel ist die Darstellung einer Doppelsumme, um den Zentralen Grenzwertsatz (Satz 2.4) und das Continuous-Mapping-Theorem (Satz 2.5) anzuwenden. Zunächst erweitern wir die geordnete Summe in (2.5) um Permutationen, sodass die Rollen von n_1, n_2 vertauscht werden können. Eine Korrektur mit dem Vorfaktor $\frac{1}{2}$ ergibt

folgende Darstellung von (2.5):

$$
\begin{aligned}
&-\frac{N}{4\binom{N}{2}} \sum_{1\le n_1\neq n_2\le N} \Phi(E_{n_1})\Phi(E_{n_2}) \\
=&-\frac{N}{2N(N-1)}\left(\sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\Phi(E_{n_1})\Phi(E_{n_2}) - \sum_{n=1}^{N}\Phi(E_n)^2\right) \qquad (2.6)\\
=&=-\frac{N}{2(N-1)}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(E_n)\right)^2 + \frac{1}{2(N-1)}\sum_{n=1}^{N}\Phi(E_n)^2 \\
=&-\frac{N}{2(N-1)}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(E_n)\right)^2 + \frac{N}{2(N-1)}. \qquad (2.7)
\end{aligned}
$$

In (2.6) werden Summanden mit $n_1 = n_2$ als Nulladdition hinzugefügt.
Mit $\Phi(E_n)^2 = 1$ P-fast sicher erhalten wir (2.7). Ferner gelten

$$\mathbb{E}(\Phi(E_n)) = 0 \text{ und } \mathrm{Var}(\Phi(E_{n_1})) = 1$$

und damit ist $X_N := \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(E_n)$ nach dem Zentralen Grenzwertsatz asymptotisch $\mathcal{N}(0,1)$-verteilt. Das Continuous-Mapping-Theorem liefert, dass X_N^2 asymptotisch χ_1^2-verteilt ist. Ferner konvergiert $\frac{N}{N-1}$ gegen 1 für $N \to \infty$ deterministisch und damit auch stochastisch. Mit dem Lemma von Slutzky (Bickel und Doksum (1997), S. 461) ergibt sich $\frac{1}{2}(1-X)$ als asymptotische Verteilung mit $X \sim \chi_1^2$. □

Der Beweis zeigt uns neben der Asymptotik eine alternative Darstellung für die volle Zweier-Tiefe, die in linearer statt quadratischer Laufzeit bestimmbar ist. Dabei sei zu beachten, dass die Darstellung in (2.7) für beliebige $\vartheta \in \Theta$ gilt, da auch in diesem Fall Lemma 2.3 angewendet werden kann, falls die Annahme (2.3) gilt:

$$N\left(d_S^2(\vartheta, Z) - \frac{1}{2}\right) = -\frac{N}{2(N-1)}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(\mathrm{res}(\vartheta, Z_n))\right)^2 + \frac{N}{2(N-1)}.$$

Die volle Zweier-Tiefe kann als Anteil von unterschiedlichen Vorzeichen mit passender Normierung verstanden werden, da in der Summation der $\Phi(\mathrm{res}(\vartheta, Z_n))$ sich unterschiedliche Vorzeichen gegenseitig aufheben. Die Art des häufiger vorkommenden Vorzeichens wird durch das Quadrat der Summe nicht berücksichtigt. Diese

Gleichheit wird uns im nächsten Unterkapitel zeigen, dass ein Test beruhend auf der voller Zweier-Tiefe asymptotisch einem bereits bekanntem Vorzeichen-Test entspricht. Ferner werden wir im dritten und vierten Kapitel uns solche analogen Darstellungen zunutze machen, um asymptotische Verteilungen zu bestimmen, sowie kürzere Laufzeiten zu gewinnen.

2.3 Testverfahren beruhend auf der vollen Zweier-Tiefe

Aus der asymptotischen Verteilung können wir ein Testverfahren aufstellen, das asymptotisch ein vorgegebenes Signifikanzniveau α einhält.

Korollar 2.7 (Testverfahren basierend auf voller Zweier-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 und das Hypothesenpaar $H_0 : \vartheta \in \Theta_0$ *vs.* $H_1 : \vartheta \in \Theta_1$ *hält das Testverfahren mit folgender Entscheidungsregel asymptotisch das Signifikanzniveau* α *ein:*

$$\textit{Man verwerfe } H_0, \textit{ falls } \sup_{\vartheta \in \Theta_0} \left(N \left(d_S^2(\vartheta, z) - \frac{1}{2} \right) \right) < q_\alpha^{(2)},$$

wobei $q_\alpha^{(2)}$ *das* α*-Quantil einer Zufallsvariablen* $\frac{1}{2}(1 - X)$ *mit* $X \sim \chi_1^2$ *entspricht.*

Beweis. Wir zeigen mit analoger Vorgehensweise wie in Müller (2005), dass der Fehler erster Art asymptotisch durch das Signifikanzniveau α beschränkt wird. Für jeden Parameter im Annahmebereich $\vartheta_0 \in \Theta_0$ gilt

$$\begin{aligned} & P_{\vartheta_0} \left(\sup_{\vartheta \in \Theta_0} \left(N \left(d_S^2(\vartheta, Z) - \frac{1}{2} \right) \right) < q_\alpha^{(2)} \right) \\ \leq\ & P_{\vartheta_0} \left(N \left(d_S^2(\vartheta_0, Z) - \frac{1}{2} \right) < q_\alpha^{(2)} \right) \xrightarrow[N \to \infty]{} \alpha \end{aligned}$$

nach Satz 2.6 und damit hält der Test asymptotisch das Signifikanzniveau ein. □

In der Φ-Darstellung der vollen Zweier-Tiefe erkennen wir, dass die Betrachtung der Vorzeichenwechsel von allen geordneten Paaren lediglich einer Zählung unterschiedlicher Vorzeichen entspricht. So ein Vorzeichen-Test ist bereits von Lehmann und D'Abrera (1975) für Zweistichproben-Fälle und von Huggins (1989) in einem allgemeineren Rahmen für stochastische Prozesse betrachtet worden. Dabei wird in den

angegebenen Quellen die Teststatistik ohne Quadrierung verwendet. Im nachfolgenden Satz 2.8 quadrieren wir die Teststatistik, um eine Vergleichbarkeit mit dem Test in Korollar 2.7 zu ermöglichen:

Satz 2.8 (Asymptotischer Vorzeichen-Test).
Für das gegebene Regressionsmodell in Definition 2.1 und das Hypothesenpaar $H_0 : \vartheta \in \Theta_0$ *vs.* $H_1 : \vartheta \in \Theta_1$ *hält das Testverfahren mit folgender Entscheidungsregel asymptotisch das Signifikanzniveau* α *ein:*

$$\textit{Man verwerfe } H_0, \textit{ falls } \inf_{\vartheta \in \Theta_0} T^N_{sign}(\vartheta, Z)^2 > w_{1-\alpha},$$

wobei $T^N_{sign}(\vartheta, Z) := \dfrac{1}{\sqrt{N}} \displaystyle\sum_{n=1}^{N} \left(\dfrac{\mathbb{1}\{res(\vartheta, Z) < 0\} - \frac{1}{2}}{\frac{1}{2}} \right)$ *und* w_α *das* α*-Quantil der* χ^2_1*-Verteilung sind.*

Die Tests in Korollar 2.7 und Satz 2.8 sind asymptotisch äquivalent, d.h. die Testentscheidungen sind für große N gleich. Das zeigen wir, indem wir die Entscheidungsregel des Tests aus Korollar 2.7 asymptotisch äquivalent als die Entscheidungsregel aus Satz 2.8 darstellen. Dabei wird verwendet, dass das α-Quantil der Zufallsvariablen $\frac{1}{2}(1-X)$ mit $X \sim \chi^2_1$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ die Gestalt $q^{(2)}_\alpha = \frac{1}{2}(1 - w_{1-\alpha})$ besitzt, denn

$$\alpha = P(X > w_{1-\alpha}) = P\left(\frac{1}{2}(1-X) < \underbrace{\frac{1}{2}(1 - w_{1-\alpha})}_{=q^{(2)}_\alpha} \right).$$

Wir verwenden (2.7) aus dem Beweis von Satz 2.6, um die die asymptotische Gleichwertigkeit der Tests zu zeigen:

$$\begin{aligned} & \sup_{\vartheta \in \Theta_0} \left\{ -\frac{N}{2(N-1)} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Phi(\mathrm{res}(\vartheta, Z)) \right)^2 \right\} + \frac{N}{2(N-1)} < q^{(2)}_\alpha \\ \Leftrightarrow \quad & \underbrace{\frac{N}{N-1}}_{\xrightarrow[N\to\infty]{} 1} \inf_{\vartheta \in \Theta_0} \left\{ \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \frac{\mathbb{1}\{\mathrm{res}(\vartheta, Z) < 0\} - \frac{1}{2}}{\frac{1}{2}} \right)^2 \right\} + \underbrace{1 - \frac{N}{N-1}}_{\xrightarrow[N\to\infty]{} 0} > w_{1-\alpha} \end{aligned} \tag{2.8}$$

Die asymptotische Gleichwertigkeit liegt vor, da sich Anteile in der umgeformten Entscheidungsregel asymptotisch vernachlässigen lassen, siehe Formel (2.8). Aus Simulationsergebnissen in Kustosz und Müller (2014) wird deutlich, dass die Tests nicht exakt gleichwertig sind, d.h. für kleine N können sich unterschiedliche Testentscheidungen ergeben. Es wäre wünschenswert, wenn die Testverfahren beruhend auf höheren vollen Tiefen neue Testverfahren liefern. Das lässt sich erwarten, da die Teststatistiken für $K \geq 3$ mehr Komplexität und andere Vorzeichenstrukturen ablesen. In der Tat können wir in Kapitel 3 und 4 neue Testverfahren herleiten.

Notwendiger Stichprobenumfang

Zum Abschluss werden wir einen notwendigen Stichprobenumfang N herleiten, für den das asymptotische Testverfahren überhaupt erst sinnvolle Ergebnisse liefern kann. Ist N zu niedrig, kann der kritische Wert $q_\alpha^{(2)}$ gar nicht unterschritten werden, sodass das Testverfahren nie die Nullhypothese ablehnen und signifikante Aussagen liefern kann. In so einem Fall sollte man mit den Quantilen der exakten Verteilung arbeiten, die sich z.B. durch Simulationen bestimmen lassen. Zur Bestimmung eines notwendigen Stichprobenumfangs betrachten wir den extremen Fall der Teststatistik mit einer minimalen Tiefe von 0. Die folgende Ungleichung gibt eine notwendige Bedingung an, die Nullhypothese abzulehnen und wird nach N aufgelöst:

$$-\frac{N}{2} \overset{!}{\leq} q_\alpha^{(2)} \quad \Leftrightarrow \quad N \overset{!}{\geq} -2q_\alpha^{(2)}.$$

Für häufig verwendete Signifikanzniveaus $\alpha \in \{0.1, 0.05, 0.01\}$ setzen wir das α-Quantil der asymptotischen Verteilung $\frac{1}{2}(1 - X)$ mit $X \sim \chi_1^2$ ein und bestimmen den notwendigen Stichprobenumfang aufgerundet auf die nächste natürliche Zahl (vgl. Tabelle 2). Die Quantile der Zufallsvariablen $\frac{1}{2}(1 - X)$ besitzen die Gestalt $q_\alpha^{(2)} = \frac{1}{2}(1 - w_{1-\alpha})$. Die Quantile der χ_1^2-Verteilung können dabei in `R` mit dem Befehl `qchisq()` berechnet werden. Es sei zu betonen, dass wir lediglich notwendi-

Tabelle 2: Notwendiger Stichprobenumfang N in Abhängigkeit von häufig verwendeten Signifikanzniveaus α für den Test aus Korollar 2.7

Signifikanzniveau α	0.1	0.05	0.01
notwendiges N	2	3	6

ge Bedingungen hergeleitet haben, sodass das asymptotische Testverfahren sinnvoll sein kann. Stichprobenumfänge, die nur leicht den notwendigen Wert überschreiten, könnten dennoch zu einer schlechten Güte führen, da die Asymptotik noch nicht wirkt, weswegen gegebenenfalls für kleine N das exakte Quantil besser sein kann.

3 Asymptotische Verteilung der vollen Dreier-Tiefe

Wir beschäftigen uns im dritten Kapitel mit der **vollen Dreier-Tiefe**. Die nachfolgenden Resultate bauen auf den Ergebnissen von Kustosz et al. (2016a) auf, ergänzen und präzisieren sie aber noch zusätzlich.
Durch Einsetzen von $K = 3$ in Definition 2.2 ergibt sich die volle Dreier-Tiefe:

$$d_S^3(\vartheta, z) = \frac{1}{\binom{N}{3}} \sum_{1 \le n_1 < n_2 < n_3 \le N} \Big(\mathbb{1}\{\mathrm{res}(\vartheta, z_{n_1}) > 0, \mathrm{res}(\vartheta, z_{n_2}) < 0, \mathrm{res}(\vartheta, z_{n_3}) > 0\} + \mathbb{1}\{\mathrm{res}(\vartheta, z_{n_1}) < 0, \mathrm{res}(\vartheta, z_{n_2}) > 0, \mathrm{res}(\vartheta, z_{n_3}) < 0\} \Big).$$

Eine Standardisierung erfolgt durch $N\left(d_S^3(\vartheta, z) - \frac{1}{4}\right)$. Ähnlich wie bei der vollen Zweier-Tiefe müssen wir die Abhängigkeitsstrukturen der geordneten Summe unter Kontrolle bekommen. Die folgende Liste beschreibt eine allgemeine Beweisstratgie, die wir bereits bei der vollen Zweier-Tiefe gesehen haben und auch im späteren Kapitel 4 bei höheren vollen Tiefen verwenden können:

1.Schritt: Wir leiten eine Darstellung mit der Funktion Φ her.

2.Schritt: Wir schreiben geordnete Summen in Quadrate von Summen um.

3.Schritt: Auf die Quadrate von Summen wenden wir eine Version des Zentralen Grenzwertsatzes und das Continuous-Mapping-Theorem an.

Die nächsten drei Unterabschnitte des Kapitels entsprechen jeweils einem Beweisschritt aus dieser Liste.

3.1 Φ-Darstellung der vollen Dreier-Tiefe

Analog wie bei der vollen Zweier-Tiefe besitzen die einzelnen Summanden bei der vollen Dreier-Tiefe stochastische Abhängigkeiten. Der erste Schritt ist die Gewinnung einer Darstellung, in der die stochastischen Abhängigkeiten voneinander getrennt werden. Das wird erneut durch die Funktion Φ erzielt.

Lemma 3.1 (Φ-Darstellung der vollen Dreier-Tiefe).
Für Zufallsvariablen $E_{n_1}, E_{n_2}, E_{n_3}$ *mit* $P(E_{n_i} \neq 0) = 1$ *für* $i = 1, 2, 3$ *gilt:*

$$\begin{aligned} &\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4} \\ = &-\frac{1}{4}\left(\Phi(E_{n_1})\Phi(E_{n_2}) - \Phi(E_{n_1})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_3})\right) \ P\text{-fast sicher}, \end{aligned}$$

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

In Kustosz et al. (2016a) wird eine analoge Aussage asymptotisch im L^2-Sinn gezeigt. Die Aussage in Lemma 3.1 wurde vom Autor dieser Arbeit als Analogon zu jener Aussage aus Kustosz et al. (2016a) mit der Funktion Φ aufgestellt und von ihm mit einer Beweisidee gezeigt, durch die sich eine allgemeinere Φ-Darstellung für höhere Tiefen herleiten lässt (vgl. Kapitel 4, Satz 4.1). Dr. K. Leckey erkannte zuvor bei jener analogen Aussage aus Kustosz et al. (2016a), dass sie exakt P-fast sicher gilt, durch einen ähnlichen Nachweis wie in Tabelle 3.

Beweis. Analog zu Lemma 2.3 können alle $2^3 = 8$ Kombinationen der Vorzeichen betrachtet und die Gleichheiten verifiziert werden, siehe Tabelle 3. Allerdings zeigt

Tabelle 3: Fälle aller möglichen Realisationen bei drei Zufallsvariablen

E_{n_1}	E_{n_2}	E_{n_3}	linke Seite	rechte Seite
$+$	$+$	$+$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left((-1)^2-(-1)^2+(-1)^2\right)$
$+$	$+$	$-$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left((-1)^2-(-1)\cdot 1+(-1)\cdot 1\right)$
$+$	$-$	$+$	$1+0-\frac{1}{4}$	$-\frac{1}{4}\left((-1)\cdot 1-(-1)^2+1\cdot(-1)\right)$
$+$	$-$	$-$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left((-1)\cdot 1-(-1)\cdot 1+(-1)^2\right)$
$-$	$+$	$+$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left(1\cdot(-1)-1\cdot(-1)+(-1)^2\right)$
$-$	$+$	$-$	$0+1-\frac{1}{4}$	$-\frac{1}{4}\left(1\cdot(-1)-1^2+1\cdot(-1)\right)$
$-$	$-$	$+$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left(1^2-1\cdot(-1)+1\cdot(-1)\right)$
$-$	$-$	$-$	$0+0-\frac{1}{4}$	$-\frac{1}{4}\left(1^2-1^2+1^2\right)$

dieser Beweis nicht, wie die Φ-Darstellung der vollen Dreier-Tiefe hergeleitet werden kann. Die Betrachtung der Herleitung lohnt sich, da wir aus ihr im Kapitel 4 einen allgemeineren Beweis ableiten können. Die Idee der folgenden Rechnung besteht dabei, durch Nulladditionen Ausdrücke der Form $\Phi(E_{n_i})\Phi(E_{n_j})$ zu erhalten:

$$\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4}$$

$$
\begin{aligned}
&= \left(\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} - \frac{1}{2} \right) \mathbb{1}\{E_{n_3} > 0\} \\
&\quad - \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} > 0\} + \frac{1}{2}\mathbb{1}\{E_{n_3} > 0\} \\
&\quad + \left(\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} - \frac{1}{2} \right) \mathbb{1}\{E_{n_3} < 0\} \\
&\quad - \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} < 0\} + \frac{1}{2}\mathbb{1}\{E_{n_3} < 0\} - \frac{1}{4}. \qquad (3.1)
\end{aligned}
$$

Da $P(E_{n_i} \neq 0) = 1$ für $i = 1, 2, 3$ ist, gilt $\mathbb{1}\{E_{n_3} > 0\} + \mathbb{1}\{E_{n_3} < 0\} = 1$ P-fast sicher. Ferner können wir Lemma 2.3 nutzen, um Φ-Darstellungen zu gewinnen und schreiben (3.1) wie folgt um:

$$
\begin{aligned}
&- \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2})\mathbb{1}\{E_{n_3} > 0\} - \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2})\mathbb{1}\{E_{n_3} < 0\} \\
&- \left(\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} < 0\} - \frac{1}{4} \right) \\
&= - \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}) - \Bigg(\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} > 0\} \\
&\qquad\qquad + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} < 0\} - \frac{1}{4} \Bigg). \qquad (3.2)
\end{aligned}
$$

In (3.2) gewinnen wir den zweiten Summanden, der sich analog zu obiger Überlegung folgendermaßen darstellen lässt:

$$
\begin{aligned}
&\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} < 0\} - \frac{1}{4} \\
&= - \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_3}) - \Bigg(\mathbb{1}\{E_{n_1} > 0, E_{n_2} > 0, E_{n_3} < 0\} \\
&\qquad\qquad + \mathbb{1}\{E_{n_1} < 0, E_{n_2} < 0, E_{n_3} > 0\} - \frac{1}{4} \Bigg). \qquad (3.3)
\end{aligned}
$$

Eine weitere Rechnung mit dem zweiten Summanden in (3.3) liefert analog:

$$
\mathbb{1}\{E_{n_1} > 0, E_{n_2} > 0, E_{n_3} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} < 0, E_{n_3} > 0\} - \frac{1}{4}
$$

$$= -\frac{1}{2}\Phi(E_{n_2})\Phi(E_{n_3}) - \Bigg(\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4}\Bigg).$$

Wir bemerken, dass durch mehrfache Durchführung dieser Rechnung sich der ursprünglich zu untersuchende Summand der vollen Dreier-Tiefe mit umgedrehten Vorzeichen ergibt. Nun setzen wir rekursiv die obigen Gleichheiten ein und erhalten:

$$\begin{aligned}
&\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4} \\
&= -\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}) \\
&- \left(\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} < 0\} - \frac{1}{4}\right) \\
&= -\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_3}) \\
&+ \left(\mathbb{1}\{E_{n_1} > 0, E_{n_2} > 0, E_{n_3} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} < 0, E_{n_3} > 0\} - \frac{1}{4}\right) \\
&= -\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_3}) - \frac{1}{2}\Phi(E_{n_2})\Phi(E_{n_3}) \\
&- \left(\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4}\right) \qquad (3.4)
\end{aligned}$$

Nach Addition des letzten Summanden und Division durch 2 in (3.4) erhalten wir:

$$\begin{aligned}
&\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0\} - \frac{1}{4} \\
&= -\frac{1}{4}\left(\Phi(E_{n_1})\Phi(E_{n_2}) - \Phi(E_{n_1})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_3})\right),
\end{aligned}$$

was der behaupteten Darstellung entspricht. Dabei sei zu beachten, dass nach der Verwendung der Voraussetzung $P(E_{n_i} \neq 0) = 0$ und des Lemmas 2.3 nur P-fast sichere Gleichheiten gelten. □

3.2 Umsortierung zur Doppelsumme mit separierten Summanden in Produktform

Bevor wir mit der Φ-Darstellung der vollen Dreier-Tiefe arbeiten, zeigen wir im Lemma 3.2 eine Integraldarstellung der Differenz von Maximum und Minimum, um im weiteren Verlauf eine symmetrische Darstellung der vollen Dreier-Tiefe zu gewinnen. Diese Integraldarstellung ist aus Kustosz et al. (2016a) entnommen, in der sie ebenso verwendet wird. Im Lemma 3.2 wird eine weitere Notation einer Indikatorfunktion $\mathbb{1}_A(x)$ für eine Menge $A \subseteq \mathbb{R}$ und $x \in \mathbb{R}$ verwendet:

$$\mathbb{1}_A(x) := \begin{cases} 1, & \text{für } x \in A \\ 0, & \text{für } x \notin A. \end{cases}$$

Lemma 3.2 (Differenz von Maximum und Minimum).
Für $n_1, n_2 \in \mathbb{R}$ und $N > 0$ gilt

$$\frac{|n_1 - n_2|}{N} = 1 - \int_{-\infty}^{\infty} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right)\, dt.$$

Falls zusätzlich $n_1, n_2 \in (0, N]$ sind, gilt

$$\int_{-\infty}^{\infty} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right)\, dt = \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right)\, dt.$$

Der Beweis von Lemma 3.2 ist vom Autor dieser Arbeit geführt worden.

Beweis. Wir beginnen auf der rechten Seite. Dazu fassen wir die Indikatorfunktionen als Bedingungen auf, die den Integrationsbereich wie folgt einschränken:

$$\begin{aligned} & -\frac{1}{2} < \frac{n_1}{N} - t \leq \frac{1}{2} \quad \wedge \quad -\frac{1}{2} < \frac{n_2}{N} - t \leq \frac{1}{2} \\ \Leftrightarrow \quad & \frac{n_1}{N} - \frac{1}{2} \leq t < \frac{n_1}{N} + \frac{1}{2} \quad \wedge \quad \frac{n_2}{N} - \frac{1}{2} \leq t < \frac{n_2}{N} + \frac{1}{2}. \end{aligned}$$

Fassen wir die unteren und oberen Schranken von t zusammen, erhalten wir:

$$\Leftrightarrow \quad \frac{\max\{n_1, n_2\}}{N} - \frac{1}{2} \leq t < \frac{\min\{n_1, n_2\}}{N} + \frac{1}{2} \tag{3.5}$$

und können die Grenzen des Integrals auf der rechten Seite durch die berechneten Grenzen ersetzen. Ferner verwenden wir $\max\{n_1, n_2\} - \min\{n_1, n_2\} = |n_1 - n_2|$:

$$\begin{aligned} &1 - \int_{-\infty}^{\infty} \mathbb{1}_{(-\frac{1}{2}, \frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2}, \frac{1}{2}]}\left(\frac{n_2}{N} - t\right) dt = 1 - \int_{\frac{\max\{n_1,n_2\}}{N} - \frac{1}{2}}^{\frac{\min\{n_1,n_2\}}{N} + \frac{1}{2}} 1\, dt \\ =&1 - \left(-\frac{\max\{n_1, n_2\} - \min\{n_1, n_2\}}{N} + 1\right) = \frac{|n_1 - n_2|}{N}. \end{aligned}$$

Für die zweite Aussage des Lemmas geben wir für die Integrationsgrenzen eine untere bzw. obere Schranke an. Mit $0 < \frac{n_1}{N} \leq 1$, $0 < \frac{n_2}{N} \leq 1$ und den Ungleichungen in (3.5) folgen die Grenzen

$$\begin{aligned} t &\geq \frac{\max\{n_1, n_2\}}{N} - \frac{1}{2} > -\frac{1}{2}, \\ t &< \frac{\min\{n_1, n_2\}}{N} + \frac{1}{2} \leq \frac{3}{2}, \end{aligned}$$

die wir im Integral einsetzen können. □

Wir werden das Lemma im Beweis des nächsten Satzes 3.3 für den Fall, dass $n_1, n_2 \in \{1, ..., N\}$ Laufindizes einer Summe von 1 bis N sind, anwenden.

Satz 3.3 (Darstellung mit separierten Summanden in Produktform).
Für das gegebene Regressionsmodell in Definition 2.1 gilt

$$\begin{aligned} &N\left(d_S^3(\vartheta^*, Z) - \frac{1}{4}\right) = \frac{N^3}{8\binom{N}{3}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Phi(E_n)\right)^2 \\ &- \frac{N^3}{4\binom{N}{3}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2}, \frac{1}{2}]}\left(\frac{n}{N} - t\right) \Phi(E_n)\right)^2 dt + \frac{N^3}{8\binom{N}{3}} \; P\text{-fast sicher.} \end{aligned}$$

Der Satz 3.3 wird in Kustosz et al. (2016a) nur asymptotisch im stochastischen Sinne bewiesen, da einerseits dort eine analoge Variante von Lemma 3.1, die dort nur asymptotisch im L^2-Sinn gezeigt wird, verwendet wird und andererseits dort Summen von Zufallsvariablen mit dem Gesetz der großen Zahlen asymptotisch vernachlässigt werden. Der Autor dieser Masterarbeit erkannte, dass jene Summen von Zufallsvariablen P-fast sicher exakt 0 sind und hat Lemma 3.1 ebenso in einer P-fast sicher exakten Version präzisiert, wodurch Satz 3.3 P-fast sicher exakt gilt. Der

nachfolgende Beweis enthält beim Umgang mit geordneten Summen, sowie bei der Anwendung von Lemma 3.2 Ideen aus Kustosz et al. (2016a).

Beweis. Wir schreiben die Summanden der standardisierten vollen Dreier-Tiefe gemäß Lemma 3.1 in ihre P-fast sicher gültige Φ-Darstellungen um:

$$\begin{aligned}
&N\left(d_S^3(\vartheta^*, Z) - \frac{1}{4}\right)\\
=&\frac{N}{4\binom{N}{3}} \sum_{1\le n_1<n_2<n_3\le N} (-\Phi(E_{n_1})\Phi(E_{n_2}) + \Phi(E_{n_1})\Phi(E_{n_3}) - \Phi(E_{n_2})\Phi(E_{n_3}))\\
=&\frac{N}{4\binom{N}{3}}\Bigg(- \sum_{1\le n_1<n_2<n_3\le N} \Phi(E_{n_1})\Phi(E_{n_2}) + \sum_{1\le n_1<n_2<n_3\le N} \Phi(E_{n_1})\Phi(E_{n_3})\\
&- \sum_{1\le n_1<n_2<n_3\le N} \Phi(E_{n_2})\Phi(E_{n_3})\Bigg). \qquad (3.6)
\end{aligned}$$

In jeder Summe in (3.6) kommt ein Laufindex vor, von dem die Summanden nicht abhängen. Wir können den jeweiligen Laufindex in den Summen entfernen, wenn wir mit der Anzahl der Kombinationen ausgleichen und erhalten so folgende Darstellung von (3.6):

$$\begin{aligned}
&\frac{N}{4\binom{N}{3}}\Bigg(- \sum_{1\le n_1<n_2\le N} (N - n_2)\Phi(E_{n_1})\Phi(E_{n_2})\\
&\qquad + \sum_{1\le n_1<n_3\le N} (n_3 - n_1 - 1)\Phi(E_{n_1})\Phi(E_{n_3})\\
&\qquad - \sum_{1\le n_2<n_3\le N} (n_2 - 1)\Phi(E_{n_2})\Phi(E_{n_3})\Bigg)\\
=&\frac{N}{8\binom{N}{3}}\Bigg(- \sum_{1\le n_1\ne n_2\le N} (N - \max\{n_1, n_2\})\Phi(E_{n_1})\Phi(E_{n_2})\\
&\qquad + \sum_{1\le n_1\ne n_3\le N} (\max\{n_1, n_3\} - \min\{n_1, n_3\} - 1)\Phi(E_{n_1})\Phi(E_{n_3})\\
&\qquad - \sum_{1\le n_2\ne n_3\le N} (\max\{n_2, n_3\} - 1)\Phi(E_{n_2})\Phi(E_{n_3})\Bigg). \qquad (3.7)
\end{aligned}$$

Um (3.7) zu gewinnen, werden die Permutationen der Laufindizes hinzugefügt und gleichen mit dem Vorfaktor $\frac{1}{2!}$ aus.

Ferner benennen wir in (3.7) die Laufindizes in allen Summen einheitlich:

$$\begin{aligned}
&\frac{N}{8\binom{N}{3}}\Bigg(- \sum_{1\le n_1\neq n_2\le N} (N-\max\{n_1,n_2\})\Phi(E_{n_1})\Phi(E_{n_2}) \\
&+ \sum_{1\le n_1\neq n_2\le N} (\max\{n_1,n_2\}-\min\{n_1,n_2\}-1)\Phi(E_{n_1})\Phi(E_{n_2}) \\
&- \sum_{1\le n_1\neq n_2\le N} (\max\{n_1,n_2\}-1)\Phi(E_{n_1})\Phi(E_{n_2})\Bigg) \\
=&\frac{N}{8\binom{N}{3}} \sum_{1\le n_1\neq n_2\le N} (2|n_1-n_2|-N)\Phi(E_{n_1})\Phi(E_{n_2}), \qquad (3.8)
\end{aligned}$$

wobei sich (3.8) durch Ausmultiplizieren und Zusammenfassen der Summanden und mit Verwendung von $\max\{n_1,n_2\}-\min\{n_1,n_2\}=|n_1-n_2|$ ergibt. Wir fügen die Laufindizes mit $n_1=n_2$ als Nulladdition in (3.8) hinzu:

$$\begin{aligned}
&\frac{N}{8\binom{N}{3}} \sum_{n_1,n_2=1}^{N} (2|n_1-n_2|-N)\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{N^3}{8\binom{N}{3}} \\
=&-\frac{N^2}{8\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \Phi(E_{n_1})\Phi(E_{n_2}) \\
&+\frac{N^2}{4\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \frac{|n_1-n_2|}{N}\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{N^3}{8\binom{N}{3}}. \qquad (3.9)
\end{aligned}$$

Wir legen nun den Fokus auf die zweite Summe in (3.9) und verwenden Lemma 3.2:

$$\begin{aligned}
&\frac{N^2}{4\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \frac{|n_1-n_2|}{N}\Phi(E_{n_1})\Phi(E_{n_2}) \\
=&\frac{N^2}{4\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \left(1 - \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N}-t\right)\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N}-t\right)\,dt\right)\Phi(E_{n_1})\Phi(E_{n_2}).
\end{aligned}$$

Setzen wir diese Gleichheit in (3.9) ein, ergibt das insgesamt:

$$\begin{aligned}
&\frac{N^2}{8\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \Phi(E_{n_1})\Phi(E_{n_2}) \\
&-\frac{N^2}{4\binom{N}{3}} \sum_{n_1,n_2=1}^{N} \left(\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N}-t\right)\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N}-t\right)\,dt\right)\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{N^3}{8\binom{N}{3}}
\end{aligned}$$

$$
\begin{aligned}
=&\frac{N^3}{8\binom{N}{3}}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(E_n)\right)^2 \\
&-\frac{N^3}{4\binom{N}{3}}\int_{-\frac{1}{2}}^{\frac{3}{2}}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N}-t\right)\Phi(E_n)\right)^2 dt+\frac{N^3}{8\binom{N}{3}},
\end{aligned}
$$

wobei die Darstellung der Doppelsumme in eine Summe ins Quadrat die Behauptung von Satz 3.3 liefert. □

Der Satz 3.3 wird unter dem wahren Parameter ϑ^* mit Residuen $E_1, ..., E_N$ formuliert. Allerdings gilt er auch für $\text{res}(\vartheta, Z_1)$, ..., $\text{res}(\vartheta, Z_N)$, da auch für die Residuen die Φ-Darstellung aus Lemma 3.1 gilt. Aus den Resultaten in Kustosz et al. (2016a) kann die Laufzeit der vollen Dreier-Tiefe nur asymptotisch verkürzt werden, da die Darstellung nicht exakt gefunden wurde. Die Idee der Laufzeit-Reduktion der vollen Dreier-Tiefe ist erstmals von Dr. K. Leckey vorgeschlagen und in Kustosz et al. (2016a) nicht untersucht worden.

Bemerkung 3.4 (Verkürzung der Laufzeit der vollen Dreier-Tiefe).
Wir zeigen, dass die alternative Darstellung der vollen Dreier-Tiefe in Satz 3.3 in linearer statt kubischer Laufzeit bestimmbar ist und skizzieren, wie man mit einem Programmpaket, z.B. in `R` (R Core Team (2018)), die Berechnung implementieren sollte. Zunächst berechnen wir in linearer Laufzeit den Vektor $S \in \mathbb{R}^N$ der kumulierten Summen der $\Phi(\text{res}(\vartheta, Z_n))$ für $n = 1, ..., N$:

$$
S = (s_1, ..., s_N)^\top = \left(\sum_{n=1}^{i}\Phi(\text{res}(\vartheta, Z_n))\right)^\top_{1\le i\le N}
$$

So können wir den ersten Ausdruck aus Satz 3.3 in linearer Laufzeit bestimmen:

$$
\frac{N^3}{8\binom{N}{3}}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(\text{res}(\vartheta, Z_n))\right)^2 = \frac{3N}{4(N-1)(N-2)}s_N^2.
$$

Nun betrachten wir $\sum_{n=1}^{N}\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N}-t\right)\Phi(\text{res}(\vartheta, Z_n))$ und fassen den Bereich der Laufindizes der Summe mit der Bedingung der Indikatorfunktion zusammen. Zur

besseren Lesbarkeit schreiben wir statt $\max\{t,s\}$ bzw. $\min\{t,s\}$ nun $t \vee s$ bzw. $t \wedge s$:

$$\begin{aligned} -\frac{1}{2} < \frac{n}{N} - t \leq \frac{1}{2} \\ \Leftrightarrow \; N\left(t-\frac{1}{2}\right) < n \leq N\left(t+\frac{1}{2}\right). \end{aligned}$$

Da die Summe über $n \in \{1,...,N\}$ läuft, können wir die Schranken der Ungleichung durch Abrundung umschreiben:

$$\left\lfloor N\left(t-\frac{1}{2}\right)\right\rfloor + 1 \leq n \leq \left\lfloor N\left(t+\frac{1}{2}\right)\right\rfloor.$$

Die untere und obere Grenze der Summe kann also wie folgt dargestellt werden:

$$\begin{aligned} &\sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-t\right)\Phi(\mathrm{res}(\vartheta,Z_n)) = \sum_{n=\left(\left\lfloor N\left(t-\frac{1}{2}\right)\right\rfloor \vee 0\right)+1}^{\left\lfloor N\left(t+\frac{1}{2}\right)\right\rfloor \wedge N} \Phi(\mathrm{res}(\vartheta,Z_n)) \\ &= \sum_{n=1}^{\left\lfloor N\left(t+\frac{1}{2}\right)\right\rfloor \wedge N} \Phi(\mathrm{res}(\vartheta,Z_n)) - \sum_{n=1}^{\left\lfloor N\left(t-\frac{1}{2}\right)\right\rfloor \vee 0} \Phi(\mathrm{res}(\vartheta,Z_n)) \\ &= \begin{cases} \sum_{n=1}^{\left\lfloor N\left(t+\frac{1}{2}\right)\right\rfloor} \Phi(\mathrm{res}(\vartheta,Z_n)), & \text{für } -\frac{1}{2} \leq t < \frac{1}{2} \\ \sum_{n=1}^{N}\Phi(\mathrm{res}(\vartheta,Z_n)) - \sum_{n=1}^{\left\lfloor N\left(t-\frac{1}{2}\right)\right\rfloor} \Phi(\mathrm{res}(\vartheta,Z_n)), & \text{für } \frac{1}{2} \leq t \leq \frac{3}{2} \end{cases} \end{aligned} \tag{3.10}$$

Die Summe $\sum_{n=1}^{N}\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-t\right)\Phi(E_n)$ kann in Abhängigkeit von t als Treppenfunktion aufgefasst werden und die Höhe der Treppenstufen können wir nach der Fallunterscheidung in (3.10) durch folgenden Vektor darstellen:

$$D = \left(S, \underbrace{(s_N,...,s_N)^\top}_{\in\mathbb{R}^N} - S\right)^\top \in \mathbb{R}^{2N}.$$

Die Sprungstellen sind im Integrationsbereich $\left[-\frac{1}{2},\frac{3}{2}\right]$ im Abstand von $\frac{1}{N}$ voneinander entfernt, sodass wir den zweiten Ausdruck aus Satz 3.3 mit Integral in linearer

Laufzeit wie folgt über $D = (d_1, ..., d_{2N})^\top$ berechnen können:

$$
\begin{aligned}
&- \frac{N^3}{4\binom{N}{3}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n_1=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]} \left(\frac{n_1}{N} - t \right) \Phi(\mathrm{res}(\vartheta, Z_n)) \right)^2 dt \\
= &- \frac{3}{4(N-1)(N-2)} \sum_{j=1}^{2N} d_j^2
\end{aligned}
$$

Die Anzahl an benötigten arithmetischen Operationen, d.h. die Zählung aller Additionen und Multiplikationen (Korte und Vygen (2018), S. 6), bei der Bestimmung der vollen Dreier-Tiefe in obiger Darstellung beträgt $5N + 6$, wobei der Vorfaktor $\frac{3}{4(N-1)(N-2)}$ ausgeklammert und so nur einmal berechnet werden braucht. □

3.3 Anwendung des Invarianzprinzips von Donsker

Die Anwendung des Zentralen Grenzwertsatzes ist auf die Darstellung der vollen Dreier-Tiefe in Satz 3.3 nicht direkt möglich, da im Integralteil die Verteilung des Integranden von t abhängt, sodass die Grenzverteilung auch von t abhängen wird. Fassen wir den Integranden als stochastischen Prozess mit Zeitparameter t auf, so können wir eine verallgemeinerte Version des Zentralen Grenzwertsatzes für Irrfahrten, das Invarianzprinzip von Donsker, verwenden. Dieser funktionale Zentrale Grenzwertsatz liefert die schwache Konvergenz der Verteilung von zeitskalierten normierten Irrfahrten gegen das Wiener-Maß, das assoziiert als stochastischer Prozess der Brownschen Bewegung entspricht. Der Vollständigkeit wegen wird die mathematische Bedeutung der Brownschen Bewegung angegeben (siehe Klenke (2006), S. 436ff, S. 451ff.). Ein reellwertiger stochastischer Prozess $(B_t)_{t\in[0,1]}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **Brownsche Bewegung**, falls

(i) $B_0 = 0$,

(ii) B hat unabhängige, stationäre Zuwächse,

(iii) $B_t \sim \mathcal{N}(0, t)$ für alle $t \in (0, 1]$,

(iv) P-fast sicher gilt für jedes $\omega \in \Omega$, dass ein Pfad $t \mapsto B_t(\omega)$ stetig ist.

Da die Pfade der Brownschen Bewegung P-fast sicher stetig sind, können wir sie als Prozess auf $C[0,1] := \{f : [0,1] \to \mathbb{R}; f \text{ stetig}\}$ auffassen, um das **Wiener-Maß** als Wahrscheinlichkeitsmaß auf $C[0,1]$ einzuführen. Wir betrachten für $t \in [0,1]$ die Auswertungsfunktionale (auch kanonische Projektionen genannt)

$$\Pi_t : C[0,1] \to \mathbb{R}, \Pi_t(f) = f(t) \text{ für } t \in [0,1],$$

um eine σ-Algebra $\mathcal{C} := \sigma(\Pi_t, t \in [0,1])$ auf $C[0,1]$ für das Wiener-Maß zu definieren. Versehen wir $C[0,1]$ mit der uniformen Topologie, induziert durch folgende Norm:

$$\|f - g\|_\infty := \sup_{t \in [0,1]} |f(t) - g(t)| \text{ für } f, g \in C[0,1],$$

so erhalten wir mit einem $\frac{\varepsilon}{3}$-Argument und mit dem Approximationssatz von Weierstraß, dass $(C[0,1], \|.\|_\infty)$ ein separabler Banachraum ist (Werner (2018), S. 5f., S. 31f.). In diesem Fall sind $\mathcal{C}$ und die Borel-σ-Algebra identisch:

$$\mathcal{C} = \mathcal{B}\left(C[0,1], \|.\|_\infty\right).$$

Es existiert eine Menge $\tilde{\Omega} \in \mathcal{A}$ zur Brownschen Bewegung $(B_t)_{t \in [0,1]}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit der Eigenschaft:

$$P(\tilde{\Omega}) = 1 \text{ und } (t \mapsto B_t(\omega)) \in C[0,1] \text{ für alle } \omega \in \tilde{\Omega}$$

wegen der P-fast sicherer Stetigkeit der Pfade. Wir definieren $\tilde{\mathcal{A}} := \mathcal{A} \mid_{\tilde{\Omega}}$ als Spur-σ-Algebra von $\mathcal{A}$ bezüglich $\tilde{\Omega}$ und entsprechend $\tilde{P} := P \mid_{\tilde{\mathcal{A}}}$ als Wahrscheinlichkeitsmaß eingeschränkt auf $(\tilde{\Omega}, \tilde{\mathcal{A}})$ (Klenke (2006), S. 10). Der Operator B:

$$B : \tilde{\Omega} \to C[0,1], B(f) = f$$

ist $(\tilde{A}, \mathcal{C})$-messbar, wodurch wir das Bildmaß $W := \tilde{P}_B$ auf $(C[0,1], \mathcal{C})$ definieren können. W ist dann das Wiener-Maß und $(\Pi_t)_{t \in [0,1]}$ entspricht als kanonischer Prozess der Brownschen Bewegung.

Satz 3.5 (Existenz des Wiener-Maßes).
Es existiert ein Wahrscheinlichkeitsmaß W auf $(C[0,1], \mathcal{C})$ bezüglich dessen der kanonische Prozess eine Brownsche Bewegung ist.
*W heißt **Wiener-Maß** und $(C[0,1], \mathcal{C}, W)$ heißt **Wiener-Raum**.*

Das Wiener-Maß ist nach dem Satz von Donsker der Grenzwert bezüglich der schwachen Konvergenz von der Verteilung zeitskalierter Irrfahrten, die sich als Summe von zentrierten und standardisierten unabhängig, identisch verteilten Zufallsvariablen ergeben. Der Satz von Donsker wird für unsere Anwendungen auf dem Intervall $[0,1]$ formuliert, kann aber auch auf $[0,\infty)$ fortgesetzt werden (Klenke (2006), S. 456ff.):

Satz 3.6 (Funktionaler Zentraler Grenzwertsatz nach Donsker).
Sei $(E_N)_{N\in\mathbb{N}}$ eine Folge unabhängig, identisch verteilter reellwertiger Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dabei seien $\mathbb{E}(E_1) = \mu$ und $Var(E_1) = \sigma^2 > 0$. Für $t \in [0,1]$ und $N \in \mathbb{N}$ wird ein stochastischer Prozess $(\tilde{S}_t^N)_{t\in[0,1]}$ mit stetigen Pfaden durch folgende Abbildung

$$\tilde{S}^N(t) : [0,1] \to \mathbb{R}$$
$$mit\ \tilde{S}^N(t) := \tilde{S}_t^N := \frac{1}{\sqrt{N\sigma^2}} \sum_{i=1}^{\lfloor Nt \rfloor} (E_i - \mu) + \frac{Nt - \lfloor Nt \rfloor}{\sqrt{N\sigma^2}} (E_{\lfloor Nt \rfloor + 1} - \mu)$$

definiert. Im Sinne der schwachen Konvergenz von Wahrscheinlichkeitsmaßen bzw. der Konvergenz in Verteilung auf $(C[0,1], \mathcal{C})$ gilt (bzgl. der uniformen Topologie):

$$(\tilde{S}_t^N)_{t\in[0,1]} \xrightarrow[N\to\infty]{\mathcal{D}} W,$$

wobei $(\tilde{S}_t^N)_{t\in[0,1]}$ in $N \in \mathbb{N}$ eine Folge von stochastischen Prozessen auf $(C[0,1], \mathcal{C})$ ist und W dem Wiener-Maß auf $(C[0,1], \mathcal{C})$ im Zeitintervall $[0,1]$ entspricht.

Die stetige Interpolation der normierten Irrfahrt wird in der obigen Version des Invarianzprinzip von Donsker gefordert. In Satz 3.3 liegt allerdings nicht die Darstellung eines stetigen Prozesses vor. Als Ausweg können wir auf einer Obermenge, dem Raum der rechtsstetigen Funktionen mit linksseitig existierenden Grenzwerten,

kurz: **càdlàg-Funktionen**, arbeiten (Billingsley (1999), S. 121):

$D[0,1] := \{f : [0,1] \to \mathbb{R};$ für jedes $t \geq 0 : f(t) = \lim_{s \searrow t} f(s)$ und der linksseitige Grenzwert $\lim_{s \nearrow t} f(s)$ existiert für jedes $t > 0$ und ist endlich.$\}$

Der Satz von Donsker lässt sich derart erweitern, dass die Aussage auch für nicht notwendig stetig-interpolierte càdlàg-Prozesse gilt. Dabei wird der Raum $D[0,1]$ mit der sogenannten **Skorohod-Topologie** versehen, die durch die in Satz 3.7 vorgestellte Skorohod-Metrik d_D induziert wird. Wir nennen $(D[0,1], d_D)$ dann **Skorohod-Raum** (Billingsley (1999), S. 123ff., S. 127ff.):

Satz 3.7 (Skorohod-Raum).
Sei $\Lambda := \{\lambda : [0,1] \to [0,1]; \lambda$ *streng monoton wachsend, bijektiv*$\}$*, d.h. für alle* $\lambda \in \Lambda$ *gilt:* $\lambda(0) = 0, \lambda(1) = 1$*. Wir definieren für* $\lambda \in \Lambda$*:*

$$\|\lambda\|^\circ := \sup_{s \neq t \in [0,1]} \left| \log \left(\frac{\lambda(t) - \lambda(s)}{t - s} \right) \right|.$$

Die Skorohod-Metrik ist eine Abbildung der Form $D[0,1] \times D[0,1] \to \mathbb{R}$ *für* $(f,g) \mapsto d_D(f,g)$ *mit:*

$$d_D(f,g) := \inf_{\lambda \in \Lambda} \left\{ \|\lambda\|^\circ \vee \sup_{t \in [0,1]} |f(t) - (g \circ \lambda)(t)| \right\}.$$

$(D[0,1], d_D)$ *bildet einen separablen, vollständigen metrischen Raum.*

Die Skorohod-Metrik misst den Abstand zwischen zwei Funktion in der Supremumsnorm mit einer zusätzlichen hinreichend kleinen zeitlichen Parametrisierung durch λ. Anders formuliert folgt aus $d_D(f,g) \leq \varepsilon$:

$$\|\lambda\|^\circ \leq \varepsilon \text{ und } \sup_{t \in [0,1]} |f(t) - (g \circ \lambda)(t)| \leq \varepsilon,$$

wobei $\|\lambda\|^\circ \leq \varepsilon$ bedeutet, dass die Parametrisierung sich nicht zu stark von der Identität unterscheiden darf. Ferner wird der Einfluss der Parametrisierung im Falle der Konvergenz hinreichend klein, d.h. für eine Folge $(f_n)_{n \in \mathbb{N}} \subseteq D[0,1]$ und $f \in$

$D[0,1]$ gelten (Billingsley (1999), S. 124):

$$\begin{aligned} f_n \xrightarrow[n\to\infty]{d_D} f \quad \Leftrightarrow \text{ Es gibt eine Folge } (\lambda_n)_{n\in\mathbb{N}} \subseteq \Lambda, \text{ sodass} \\ f_n \circ \lambda_n \xrightarrow[n\to\infty]{\|.\|_\infty} f \text{ mit} \\ \lambda_n \xrightarrow[n\to\infty]{\|.\|_\infty} id_{[0,1]}, \end{aligned}$$

wobei $id_{[0,1]}$ die Identität auf $[0,1]$ beschreibt. Ferner führen wir mithilfe der Topologie die Borel-σ-Algebra auf $D[0,1]$ mit $\mathcal{D} = \mathcal{B}(D[0,1], d_D)$ ein. $\mathcal{D}$ ist dabei die kleinste erzeugte σ-Algebra aus dem System der offenen Mengen auf $D[0,1]$ bezüglich der Skorohod-Topologie. Die kanonischen Auswertungen Π_t für $t \in [0,1]$ auf $D[0,1]$ seien (nun sind nicht mehr die kanonischen Projektionen auf $C[0,1]$ gemeint)

$$\Pi_t : D[0,1] \to \mathbb{R}, \Pi_t(f) = f(t) \text{ für } t \in [0,1]$$

und sind $\mathcal{D} - \mathcal{B}(\mathbb{R})$-messbar. Ein anderer nützlicher Erzeuger von $\mathcal{D}$ ist das System der Urbilder dieser kanonischen Auswertungen (Billingsley (1999), S. 133ff.):

$$\mathcal{D} = \sigma(\Pi_t : t \in [0,1]). \tag{3.11}$$

Dieser Zusammenhang hilft beim Nachweis der Messbarkeit von Abbildungen mit Definitionsmenge $D[0,1]$. Um den Satz von Donsker für càdlàg-Prozesse nun zu formulieren, müssen wir klären, was formal das Wiener-Maß auf dem Raum $(D[0,1], \mathcal{D})$ ist. Wir betrachten die Identitätsabbildung J

$$J : C[0,1] \to D[0,1], J(f) = f,$$

welche insbesondere $\mathcal{C}-\mathcal{D}$-messbar ist. Für das Wiener-Maß W auf $(C[0,1], \mathcal{C})$ liefert das Bildmaß W_J auf $D[0,1]$ die gleichen endlich-dimensionalen Verteilungen und wir können es als das Wiener-Maß auf $(D[0,1], \mathcal{D})$ auffassen (Billingsley (1999), S.146f.):

Satz 3.8 (Satz von Donsker für càdlàg-Funktionen).
Sei $(E_N)_{N\in\mathbb{N}}$ eine Folge unabhängig, identisch verteilter reellwertiger Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dabei seien $\mathbb{E}(E_1) = \mu$ und

$Var(E_1) = \sigma^2 > 0$. Für $t \in [0,1]$ und $N \in \mathbb{N}$ wird ein stochastischer Prozess $(S_t^N)_{t\in[0,1]}$ mit càdlàg-Pfaden durch folgende Abbildung

$$S^N(t) : [0,1] \to \mathbb{R},$$
$$\text{mit } S^N(t) := S_t^N := \frac{1}{\sqrt{N\sigma^2}} \sum_{i=1}^{\lfloor Nt \rfloor} (E_i - \mu)$$

definiert. Im Sinne der schwachen Konvergenz von Wahrscheinlichkeitsmaßen bzw. der Konvergenz in Verteilung auf $(D[0,1], \mathcal{D})$ gilt (bzgl. der Skorohod-Topologie):

$$(S_t^N)_{t\in[0,1]} \xrightarrow[N\to\infty]{\mathcal{D}} W_J$$

wobei $(S_t^N)_{t\in[0,1]}$ in $N \in \mathbb{N}$ eine Folge von stochastischen Prozessen auf $(D[0,1], \mathcal{D})$ ist und W_J dem Wiener-Maß auf $(D[0,1], \mathcal{D})$ im Zeitintervall $[0,1]$ entspricht.

Wir werden im Beweis bei der Asymptotik der vollen Dreier-Tiefe in Satz 3.10 Donskers Invarianzprinzip für càdlàg-Prozesse mit dem Continous-Mapping-Theorem anwenden. Aus der Konstruktion des Bildmaßes W_J gilt, dass der Grenzprozess im Satz von Donsker in 3.8 in folgendem Sinne mit Wahrscheinlichkeit 1 stetige Pfade hat:

$$\text{Für alle } A \in \mathcal{D} \text{ mit } C[0,1] \subseteq A \text{ gilt } W_J(A) = 1. \tag{3.12}$$

Diese zunächst umständlich wirkende Formulierung ist wichtig, da nicht klar ist, ob $C[0,1]$ eine $\mathcal{D}$-messbare Menge ist und im Maß W_J ausgewertet werden kann. Es genügt für die Anwendung des Continuous-Mapping-Theorems die Stetigkeit eines Funktionals $\Psi : D[0,1] \to \mathbb{R}$ auf der Teilmenge $C[0,1] \subseteq D[0,1]$ nachzurechnen, falls W_J die Verteilung des Grenzwertes ist. Denn nach dem Continuous-Mapping-Theorem ist die Menge der Unstetigkeitsstellen U_Ψ eine $\mathcal{D}$-messbare Menge, da $\mathcal{D}$ die Borel-σ-Algebra auf $(D[0,1], d_D)$ ist. Es genügt zu zeigen, dass $W_J(U_\Psi) = 0$ ist. Wenn die Menge der Stetigkeitsstellen $U_\Psi^\complement$ die Menge $C[0,1]$ enthält, folgt:

$$W_J(U_\Psi) = 1 - W_J(U_\Psi^\complement) \overset{(3.12)}{=} 0.$$

Die Idee zur Anwendung des Continuous-Mapping-Theorems trotz eventueller Messbarkeitsproblematik stammt vom Autor dieser Arbeit. Für stetige Funktionen als Grenzwert sind die Skorohod-Topologie und die uniformen Topologie äquivalent, sodass im Beweis von Satz 3.10 bei der Anwendung des Continuous-Mapping-Theorems im Satz von Donsker auf die uniforme Topologie zurückgegriffen werden kann:

Lemma 3.9 (Stetige Funktion als Grenzwert).
Sei $(f_n)_{n\in\mathbb{N}} \subseteq D[0,1]$ eine Funktionenfolge von càdlàg-Funktionen und $f \in C[0,1]$ eine stetige Funktion. Dann gilt:

$$f_n \xrightarrow[n\to\infty]{d_D} f \quad \Leftrightarrow \quad f_n \xrightarrow[n\to\infty]{\|\cdot\|_\infty} f$$

Beweis. Die folgende Rechnung ist an Billingsley (1999), S. 124 angelehnt. Die Beweisrichtung „$\Leftarrow$“ gilt auch für $f \in D[0,1]$. Nach der Voraussetzung gilt:

$$\sup_{t\in[0,1]} |f_n(t) - f(t)| \xrightarrow[n\to\infty]{} 0. \tag{3.13}$$

Für $\lambda_0 = id_{[0,1]} \in \Lambda$ gilt einerseits $\|\lambda_0\|^\circ = 0$ und andererseits:

$$\begin{aligned}\inf_{\lambda\in\Lambda}\left\{\|\lambda\|^\circ \vee \sup_{t\in[0,1]} |f_n(t) - f(\lambda(t))|\right\} &\leq \sup_{t\in[0,1]} |f_n(t) - f(\lambda_0(t))| \\ &= \sup_{t\in[0,1]} |f_n(t) - f(t)|. \end{aligned}\tag{3.14}$$

Setzt man die Formeln (3.13) und (3.14) zusammen, ergibt sich:

$$\inf_{\lambda\in\Lambda}\left\{\|\lambda\|^\circ \vee \sup_{t\in[0,1]} |f_n(t) - f(\lambda(t))|\right\} \xrightarrow[n\to\infty]{} 0$$

d.h. $f_n \xrightarrow[n\to\infty]{d_D} f$. Um die Beweisrichtung „$\Rightarrow$“ zu zeigen, benötigen wir, dass der Grenzwert f gleichmäßig stetig ist, was als stetige Funktion auf einer kompakten Definitionsbereich erfüllt ist. Da $f_n \xrightarrow[n\to\infty]{d_D} f$ gilt, existiert eine Folge $(\lambda_n)_{n\in\mathbb{N}} \subseteq \Lambda$

mit $\lambda_n \xrightarrow[n\to\infty]{\|.\|_\infty} id_{[0,1]}$. Ferner gilt mit der Dreiecks-Ungleichung:

$$\sup_{t\in[0,1]} |f_n(t) - f(t)| \leq \sup_{t\in[0,1]} |f_n(t) - f(\lambda_n(t))| + \sup_{t\in[0,1]} |f(\lambda_n(t)) - f(t)|.$$

Einerseits gilt $\sup_{t\in[0,1]} |f_n(t) - f(\lambda_n(t))| \xrightarrow[n\to\infty]{} 0$, wegen $f_n \xrightarrow[n\to\infty]{d_D} f$ und andererseits gilt $\sup_{t\in[0,1]} |f(\lambda_n(t)) - f(t)| \xrightarrow[n\to\infty]{} 0$, da f gleichmäßig stetig ist und $\sup_{t\in[0,1]} |\lambda_n(t) - t| \xrightarrow[n\to\infty]{} 0$ vorliegt. □

Das Lemma 3.9 zeigt, dass die uniforme Topologie verwendet werden darf, wenn der Grenzwert eine stetige Funktion ist. Das wird sich im Beweis von Satz 3.10 zunutze gemacht, da wir in Satz 3.8 die Brownsche Bewegung als Grenzwert von (nicht notwendig stetigen) càdlàg-Prozessen erhalten.

Satz 3.10 (Asymptotische Verteilung der vollen Dreier-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 gelten

$$(i)\quad \left(\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Phi(E_n)\right)^2, \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N}-t\right)\Phi(E_n)\right)^2 dt\right)$$
$$\xrightarrow[N\to\infty]{\mathcal{D}} \left(\psi_1(B_\bullet)^2, \int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt\right),$$

$$(ii)\quad N\left(d_S^3(\vartheta^*, Z) - \frac{1}{4}\right) \xrightarrow[N\to\infty]{\mathcal{D}} \frac{3}{4}B_1^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt + \frac{3}{4},$$

wobei $(B_t)_{t\in[0,1]}$ die Brownsche Bewegung sei und $\psi : \mathcal{D}[0,1] \to \mathbb{R}\times D\left[-\frac{1}{2},\frac{3}{2}\right]$ ein Operator sei, der wie folgt koordinatenweise definiert wird:

$$\psi_1(f) := f(1),$$
$$\psi_2(f)(t) := f\left(\left(t+\frac{1}{2}\right)\wedge 1\right) - f\left(\left(t-\frac{1}{2}\right)\vee 0\right) \text{ für } t\in\left[-\frac{1}{2},\frac{3}{2}\right].$$

Die Schreibweisen $\psi_1(B_\bullet)$ bzw. $\psi_2(B_\bullet)_t$ mit dem dunklen Punkt sollen betonen, dass eine gesamte Funktion auf $\mathbb{R}$ bzw. auf $D\left[-\frac{1}{2},\frac{3}{2}\right]$ abgebildet wird. Analoge Schreibweisen sind für $\psi_1(S_\bullet^N)$ bzw. $\psi_2(S_\bullet^N)_t$ zu verstehen. Der Raum $D\left[-\frac{1}{2},\frac{3}{2}\right]$ entspricht dem Raum der càdlàg-Funktionen auf $\left[-\frac{1}{2},\frac{3}{2}\right]$. Im nachfolgenden Beweis

brauchen wir keine Topologie auf diesem Raum, da die Koordinaten in der Aussage (i) direkt auf reelle Zahlen abgebildet werden und wir diesen Raum überspringen. In Kustosz et al. (2016a) wird Satz 3.10 in einer äquivalenten Form aufwändiger über die endlich-dimensionalen Verteilungen des Prozesses und einem Straffheitsargument gezeigt. Der hier vorgestellte Beweis ist vom Autor dieser Arbeit und von Dr. K. Leckey gefunden worden und beruht auf die Anwendung des Satzes von Donsker in 3.8 (siehe auch Bemerkung 3.12 für einen detaillierteren Vergleich).

Beweis. Wir zeigen zunächst die $\mathcal{D} - \mathcal{B}(\mathbb{R}^2)$-Messbarkeit des folgenden Funktionals

$$\Psi : D[0,1] \to \mathbb{R}^2, \Psi(f) := \left(\psi_1(f)^2, \int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(f)^2(t)\, dt \right),$$

wobei wir dazu koordinatenweise die $\mathcal{D} - \mathcal{B}(\mathbb{R})$-Messbarkeit zeigen. Die erste Koordinate lässt sich über die kanonische Auswertung $\psi_1(f)^2 = \Pi_1(f)^2$ darstellen, welche $\mathcal{D} - \mathcal{B}(\mathbb{R})$-messbar ist, siehe (3.11). Ferner liefert auch das Produkt von zwei solcher messbaren Funktionalen erneut ein messbares Funktional. Die zweite Koordinate von Ψ schreiben wir durch Verschiebungen auf den Integrationsbereich $[0,1]$ um:

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(f)^2(t)\, dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(f\left(t+\tfrac{1}{2}\right) - f(0) \right)^2 dt + \int_{\frac{1}{2}}^{\frac{3}{2}} \left(f(1) - f\left(t-\tfrac{1}{2}\right) \right)^2 dt \\
=& \int_{-\frac{1}{2}}^{\frac{1}{2}} f\left(t+\tfrac{1}{2}\right)^2 dt - 2f(0) \int_{-\frac{1}{2}}^{\frac{3}{2}} f\left(t+\tfrac{1}{2}\right) dt + f(0)^2 \\
&+ f(1)^2 - 2f(1) \int_{\frac{1}{2}}^{\frac{3}{2}} f\left(t-\tfrac{1}{2}\right) dt + \int_{\frac{1}{2}}^{\frac{3}{2}} f\left(t-\tfrac{1}{2}\right)^2 dt \\
=& 2\int_0^1 f(t)^2\, dt + f(0)^2 + f(1)^2 - 2(f(0)+f(1)) \int_0^1 f(t)\, dt
\end{aligned} \tag{3.15}$$

Wir zeigen die $\mathcal{D} - \mathcal{B}(\mathbb{R})$-Messbarkeit für das erste Integral. Für die anderen Summanden lässt sich die Messbarkeit analog zeigen. Wir schreiben das Integral als unendliche Summe und repräsentieren die Funktion durch die kanonischen Auswertungen Π_t, um die Messbarkeit zu zeigen:

$$\int_0^1 f(t)^2\, dt = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} f\left(\frac{k}{n}\right)^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \Pi_{\frac{k}{n}}(f)^2.$$

Die abzählbare Summe der kanonischen Auswertungen ist eine $\mathcal{D} - \mathcal{B}(\mathbb{R})$-messbare Abbildung, da die Konvergenz punktweise gilt wegen höchstens abzählbar vieler Sprungstellen (Kaballo (2000), S. 131ff.) und punktweise Limiten messbarer Abbildungen messbar sind (Bauer (1992), S.59f.), womit die Messbarkeit des Integrals gezeigt ist. Die zweite Koordinate entspricht einer Summe von mehreren messbaren Abbildungen, womit wir die $\mathcal{D} - \mathcal{B}(\mathbb{R})$-Messbarkeit der zweiten Koordinate von Ψ gezeigt haben. Damit ist Ψ $\mathcal{D} - \mathcal{B}(\mathbb{R}^2)$-messbar. Für die Aussage (i) stellen wir beiden Koordinate der linken Seite als Auswertung mit dem Operator ψ in $S^N_\bullet$ dar. Die Auswertung von $S^N_\bullet$ in der ersten Koordinate ψ_1 liefert:

$$\psi_1(S^N_\bullet) = \frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor N \cdot 1 \rfloor} \Phi(E_n)$$

Zur Untersuchung der zweiten Koordinate fassen wir den Bereich der Laufindizes der Summe $\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2}, \frac{1}{2}\right]} \left(\frac{n}{N} - t\right) \Phi(E_n)$ wie in Bemerkung 3.4 mit der Indikatorfunktion zusammen:

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2}, \frac{1}{2}\right]} \left(\frac{n}{N} - t\right) \Phi(E_n) = \frac{1}{\sqrt{N}} \sum_{n=\lfloor N\left(\left(t-\frac{1}{2}\right)\vee 0\right)\rfloor + 1}^{\lfloor N\left(\left(t+\frac{1}{2}\right)\wedge 1\right)\rfloor} \Phi(E_n) \\
&= \frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor N\left(\left(t+\frac{1}{2}\right)\wedge 1\right)\rfloor} \Phi(E_n) - \frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor N\left(\left(t-\frac{1}{2}\right)\vee 0\right)\rfloor} \Phi(E_n) \\
&= S^N_{\left(t+\frac{1}{2}\right)\wedge 1} - S^N_{\left(t-\frac{1}{2}\right)\vee 0}.
\end{aligned}$$

Wir können also die zweite Koordinate über die Abbildung ψ_2 identifizieren:

$$\psi_2(S^N_\bullet)_t = S^N_{\left(t+\frac{1}{2}\right)\wedge 1} - S^N_{\left(t-\frac{1}{2}\right)\vee 0}$$

Nach dem Satz von Donsker in 3.8 gilt wegen $\mathbb{E}(\Phi(E_n)) = 0$ und $\mathrm{Var}(\Phi(E_n)) = 1$ für $n \in \mathbb{N}$ für den vorliegenden càdlàg-Prozess $(S^N_t)_{t\in[0,1]}$:

$$\left(\frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor Nt \rfloor} \Phi(E_n) \right)_{t\in[0,1]} \xrightarrow[N\to\infty]{\mathcal{D}} (B_t)_{t\in[0,1]}.$$

Wenn gezeigt wird, dass auf der Menge $C[0,1]$ die Abbildung Ψ $(D[0,1], d_D) - (\mathbb{R}^2, \|.\|_2))$-stetig ist, wobei $\|.\|_2$ die euklidische Norm auf $\mathbb{R}^2$ beschreibt, so folgt mit dem Continuous-Mapping-Theorem:

$$\Psi(S_\bullet^N) \xrightarrow[N\to\infty]{\mathcal{D}} \Psi(B_\bullet),$$

da zum stochastischen Prozess $(B_t)_{t\in[0,1]}$ das assoziierte Maß W_J auf $(D[0,1], \mathcal{D})$ mit Wahrscheinlichkeit 1 stetige Pfade im Sinne von (3.12) hat. Es genügt also die Stetigkeit für $(f_n)_{n\in\mathbb{N}} \subseteq D[0,1]$ mit $f_n \xrightarrow[n\to\infty]{d_D} f$ für $f \in C[0,1]$ zu zeigen. Nach Lemma 3.9 können wir äquivalent $f_n \xrightarrow[n\to\infty]{\|.\|_\infty} f$ betrachten. Wir zeigen die Stetigkeit von Ψ koordinatenweise bezüglich dem Betragsabstand, da dies äquivalent zur Konvergenz in $\|.\|_2$ auf $\mathbb{R}^2$ ist. Für die erste Koordinate von Ψ gilt:

$$\begin{aligned}
&\left|\psi_1(f)^2 - \psi_1(f_n)^2\right| = \left|f(1)^2 - f_n(1)^2\right| \\
\leq &\sup_{t\in[0,1]} \left|f(t)^2 - f_n(t)^2\right| \leq \sup_{t\in[0,1]} |f(t) - f_n(t)| \cdot \sup_{t\in[0,1]} |f(t) + f_n(t)| \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

wobei $\sup_{t\in[0,1]} |f(t) + f_n(t)|$ sich mit der Dreiecks-Ungleichung beschränken lässt, da $(f_n)_{n\in\mathbb{N}}$ als konvergente Folge und f als stetige Funktion auf einem kompakten Definitionsbereich beschränkt sind. Für die zweite Koordinate von Ψ werden die Integrale wie in (3.15) auf das Intervall $[0,1]$ verschoben:

$$\begin{aligned}
&\left|\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(f)(t)^2\,dt - \int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(f_n)(t)^2\,dt\right| \\
=&\left|\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(f_n\left(t+\tfrac{1}{2}\right) - f_n(0)\right)^2\,dt + \int_{\frac{1}{2}}^{\frac{3}{2}} \left(f_n(1) - f_n\left(t-\tfrac{1}{2}\right)\right)^2\,dt\right. \\
&\left. - \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(f\left(t+\tfrac{1}{2}\right) - f(0)\right)^2\,dt - \int_{\frac{1}{2}}^{\frac{3}{2}} \left(f(1) - f\left(t-\tfrac{1}{2}\right)\right)^2\,dt\right| \\
\leq &\, 2\int_0^1 |f_n(t)^2 - f(t)^2|\,dt + |f_n(0)^2 - f(0)^2| + 2\int_0^1 |f_n(0)f_n(t) - f(0)f(t)|\,dt \\
&+ |f_n(1)^2 - f(1)^2| + 2\int_0^1 |f_n(1)f_n(t) - f(1)f(t)|\,dt
\end{aligned}$$

$$= 2 \int_0^1 |f_n(t)^2 - f(t)^2| \, dt + |f_n(0)^2 - f(0)^2| + |f_n(1)^2 - f(1)^2| \tag{3.16}$$

$$+2 \int_0^1 |f_n(0)f_n(t) - f_n(0)f(t) + f_n(0)f(t) - f(0)f(t)| \, dt \tag{3.17}$$

$$+2 \int_0^1 |f_n(1)f_n(t) - f_n(1)f(t) + f_n(1)f(t) - f(1)f(t)| \, dt \tag{3.18}$$

Die Summanden in (3.16) konvergieren für $n \to \infty$ gegen 0, analog wie es durch Bildung des Supremums bei der ersten Koordinate gezeigt wird. Da die Funktion auf dem gesamten kompakten Integrationsbereich $[0,1]$ definiert ist, ist die Bildung des Supremums nicht problematisch. (3.17) bzw. (3.18) ergibt sich durch Nulladdition mit $f_n(0)f(t)$ bzw. $f_n(1)f(t)$ und geht nach folgender Rechnung ebenso für $n \to \infty$ gegen 0 (exemplarisch für (3.17)):

$$\begin{aligned} &\int_0^1 |f_n(0)f_n(t) - f_n(0)f(t) + f_n(0)f(t) - f(0)f(t)| \, dt \\ &\leq \sup_{t\in[0,1]} |f_n(t)| \sup_{t\in[0,1]} |f_n(t) - f(t)| + \sup_{t\in[0,1]} |f(t)| \sup_{t\in[0,1]} |f_n(t) - f(t)| \end{aligned} \tag{3.19}$$

In (3.19) haben sich erneut Ausdrücke ergeben, die analog wie bei der ersten Koordinate durch Bildung der Suprema, für $n \to \infty$ gegen 0 konvergieren, wobei die entsprechende Vorfaktoren beschränkt sind. (3.18) kann vollständig analog untersucht werden. Damit ist (i) gezeigt. Um (ii) zu zeigen, betrachten wir folgende $(\mathbb{R}^2 - \mathbb{R})$-stetige Abbildung A:

$$A : \mathbb{R}^2 \to \mathbb{R}, A(x_1, x_2) := \frac{3}{4}x_1 - \frac{3}{2}x_2 + \frac{3}{4}.$$

Insbesondere ist die Komposition $A \circ \Psi$ eine $(D[0,1] - \mathbb{R})$-stetige Abbildung auf $C[0,1]$. Somit lässt sich das Continuous-Mapping-Theorem ebenso für $A \circ \Psi$ anwenden und wir erhalten mit Satz 3.3:

$$\begin{aligned} N\left(d_S^3(\vartheta^*, Z) - \frac{1}{4}\right) =& \frac{N^2}{(N-1)(N-2)} \left(\frac{3}{4}\psi_1(S_\bullet^N)^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(S_\bullet^N)_t^2 \, dt + \frac{3}{4}\right) \\ =& \frac{N^2}{(N-1)(N-2)} A(\Psi(S_\bullet^N)) \xrightarrow[N\to\infty]{\mathcal{D}} A(\Psi(B_\bullet)) \end{aligned}$$

$$=\frac{3}{4}\psi_1(B_\bullet)^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt + \frac{3}{4},$$

wobei das Lemma von Slutzky mit $\frac{N^2}{(N-1)(N-2)} \xrightarrow[N\to\infty]{} 1$ (Bickel und Doksum (1997), S. 461) verwendet wird. □

Wir betten wie zu Beginn in Kapitel 2.1 das Regressionsmodell in einen statistischen Raum ein, um folgendes Testproblem zu beschreiben:

Korollar 3.11 (Testverfahren basierend auf voller Dreier-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 und das Hypothesenpaar $H_0 : \vartheta \in \Theta_0$ *vs.* $H_1 : \vartheta \in \Theta_1$ *hält das Testverfahren mit folgender Entscheidungsregel asymptotisch das Signifikanzniveau* α *ein:*

$$\textit{Man verwerfe } H_0, \textit{ falls } \sup_{\vartheta\in\Theta_0}\left(N\left(d_S^3(\vartheta, z) - \frac{1}{4}\right)\right) < q_\alpha^{(3)},$$

wobei $q_\alpha^{(3)}$ *das* α*-Quantil einer Zufallsvariable mit* $\frac{3}{4} + \frac{3}{4}B_1^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2 dt$ *ist mit* $(B_t)_{t\in[0,1]}$ *als Brownsche Bewegung und* ψ_2 *als Teilkoordinate des in Satz 3.10 angegebenen Operators.*

Beweis. Analog zum Korollar 2.7 mit Verwendung von Satz 3.10 □

In verschiedenen Arbeiten, wie z.B. in Kustosz et al. (2016a) sind Güteeigenschaften dieses Tests beruhend auf der vollen Dreier-Tiefe untersucht worden. In den bisherigen Untersuchungen ist noch kein äquivalentes Testverfahren, das anders motiviert wird, gefunden worden, sodass wir anders als bei der vollen Zweier-Tiefe ein neues Testverfahren entwickelt haben.
Zum Abschluss dieses Unterkapitels soll auf die Grenzverteilung bzw. der Zwischenprozess, über den integriert wird, eingegangen werden. Analog wie im Beweis von Satz 3.10(i) können wir einen bivariate Grenzprozess $(\psi_1(B_\bullet), \psi_2(B_\bullet)_t)_{t\in\left[-\frac{1}{2},\frac{3}{2}\right]}$ durch

folgende Konvergenz formulieren für $t \in \left[-\frac{1}{2}, \frac{3}{2}\right]$:

$$\left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Phi(E_n), \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2}, \frac{1}{2}\right]} \left(\frac{n_1}{N} - t \right) \Phi(E_n) \right)_t \right) \xrightarrow[N \to \infty]{\mathcal{D}} (B_1, \psi_2(B_\bullet)_t). \tag{3.20}$$

Dabei muss ein Operator der Bauart $D[0,1] \to \mathbb{R} \times D\left[-\frac{1}{2}, \frac{3}{2}\right]$ konstruiert werden, wobei der Wertebereich mit der kanonischen Produkttopologie versehen wird (Werner (2018), S. 37). Aus Gründen der Übersicht ist in Satz 3.10(i) direkt auf $\mathbb{R}^2$ abgebildet worden, allerdings kann das Continuous-Mapping-Theorem mit gleicher Argumentation verwendet werden, um (3.20) nachzurechnen. Die erste Koordinate des Prozesses ist der Wert eines Pfades der Brownschen Bewegung $(B_t)_{t \in [0,1]}$ zum Zeitpunkt $t = 1$. Die zweite Koordinate entspricht im Intervall $\left[-\frac{1}{2}, \frac{1}{2}\right]$ dem gleichen Pfad der Brownschen Bewegung mit Zeitverschiebung um $t = \frac{1}{2}$. Im Intervall $\left[\frac{1}{2}, \frac{3}{2}\right]$ wird vom Wert B_1 der gesamte gleiche Pfad abgezogen. Eine alternative Darstellung wird durch eine Fallunterscheidung in (3.21) illustriert:

$$\psi_2(B_\bullet)_t = \begin{cases} B_{t+\frac{1}{2}}, & \text{für } -\frac{1}{2} \leq t < \frac{1}{2} \\ B_1 - B_{t-\frac{1}{2}}, & \text{für } \frac{1}{2} \leq t \leq \frac{3}{2} \end{cases} \tag{3.21}$$

Dabei sei zu bemerken, dass ein Pfad des Prozesses $(\psi_2(B_\bullet)_t)_{t \in \left[-\frac{1}{2}, \frac{3}{2}\right]}$ bereits im Intervall $\left[-\frac{1}{2}, \frac{1}{2}\right]$ bestimmt wird. In Abbildung 3 wird ein simulierter Pfad dieses Prozesses dargestellt. In Kustosz et al. (2016a) wird diese Darstellung des Zwischenprozesses nicht gefunden, da nicht mit dem Satz von Donsker gearbeitet wird, sondern durch einen bivariaten Gauß-Prozess charakterisiert (siehe Bemerkung 3.12 und Satz 3.14). Durch die alternative Beweisführung von Satz 3.10 konnte der Autor dieser Arbeit die Charakterisierung des Zwischenprozesses durch Brownsche Bewegungen finden.

Bemerkung 3.12 (Vergleich mit dem Beweis in Kustosz et. al (2016)).
In dieser Arbeit wird in Satz 3.3 eine exakte Darstellung der vollen Dreier-Tiefe mit der Funktion Φ hergeleitet. In Kustosz et. al (2016) wird eine analoge Aussage nur asymptotisch gezeigt, da dort die Problemstellung aus einer anderen Perspektive

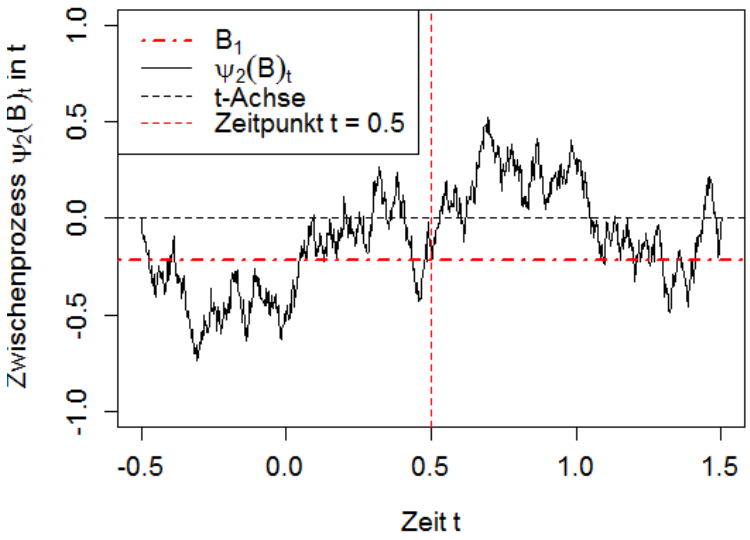


Abbildung 3: Simulierter bivariater Grenzprozess in Satz 3.10(i), für die Vorgehensweise zur Simulation, siehe Kapitel 3.4

mit Methoden zur Untersuchung von U-Statistiken bearbeitet wird. Gilt der Satz 3.3 nur asymptotisch, lässt sich dennoch das Resultat in Satz 3.10 zeigen. Allerdings kann keine exakte Darstellung in linearer Laufzeit wie in Bemerkung 3.4 angegeben werden.

Ferner wird die schwache Konvergenz von Verteilungen im Satz 3.10 in Kustosz et al. (2016a) mit einem Straffheitsargument gezeigt. Zunächst wird der Grenzwert in endlich-dimensionalen Verteilungen mit einer multivariaten Version des Zentralen Grenzwertsatzes von Lindeberg als zentrierter Gauß-Prozess mit einer speziellen Kovarianzstruktur (siehe Satz 3.14) identifiziert. Mit dem Eindeutigkeitssatz von Kolmogorov ist der stochastische Prozess durch die berechneten endlich-dimensionalen Verteilungen eindeutig bestimmt. Der Nachweis der Straffheit liefert dann die schwache Konvergenz gegen den Gauß-Prozess. Anschließend kann analog wie in Satz 3.10 mit dem Continuous-Mapping-Thoerem argumentiert werden, wobei man dabei die P-fast sichere Stetigkeit der Pfade des Gauß-Prozesses braucht, um die Stetigkeit des Funktionals für das Continuous-Mapping-Thoerem zu rechtfertigen. Dazu wird mit dem Satz von Kolmogorov-Chentsov gezeigt, dass eine Modifikation des Prozesses mit P-fast sicheren Pfaden existiert. Auf die konkrete Anwendung des Continuous-

Mapping-Theorems wird in Kustosz et al. (2016a) nicht eingegangen und ist in dieser Arbeit selbstständig vom Autor überlegt worden.
Der Aufbau dieses Beweises erinnert in einem Spezialfall an den Beweis vom Satz von Donsker; insbesondere auch deswegen, da wir den Satz von Donsker auf das Problem anwenden können. Der Vorteil bei der Verwendung des Satzes von Donsker ist, dass wir den Zwischenprozess als Brownsche Bewegung darstellen können und nicht eine abstraktere Darstellung als zentrierten Gauß-Prozess mit gegebener Kovarianzstruktur erhalten. Vor allem für praktische Anwendungen ist das Repertoire von Methoden zur Brownschen Bewegung umfangreicher. Beispielsweise sind die Berechnungen der Quantile (vgl. Kapitel 3.4) dadurch effizienter im Vergleich zur Berechnung in Kustosz et al. (2016a), da dort der Gauß-Prozess im Intervall $\left[\frac{1}{2}, \frac{3}{2}\right]$ zusätzlich simuliert wird. Ein weiterer Vorteil ist die Vereinfachung des Beweises, durch die die Suche der asymptotischen Verteilungen für höhere Tiefen (wie z.B. der vollen Vierer-Tiefe, siehe Kapitel 4.2) erleichtert wird. □

Im letzten Teil dieses Unterkapitels soll auf den in Kustosz et al. (2016a) hergeleiteten Gauß-Prozess eingegangen werden. Ein stochastischer Prozess $(X_t)_{t\in\mathbb{R}}$ heißt Gauß-Prozess, falls für beliebige Zeitpunkte $\{t_1, ..., t_n\}$ mit beliebigem $n \in \mathbb{N}$ die endlichen dimensionalen Verteilugen $P_{X_{t_1},...,X_{t_n}}$ multivariaten Normalverteilungen folgen. Der nächste Satz stellt eine eindeutige Charakterisierung von Gauß-Prozessen bis auf Äquivalenz von stochastischen Prozessen dar (Bauer (2002), S. 380f.):

Satz 3.13 (Charakterisierung von Gauß-Prozessen).
Ein Gauß-Prozess $(X_t)_{t\in\mathbb{R}}$ ist durch folgende Abbildungen:

$$\textit{Erwartungswertfunktion: } m : \mathbb{R} \to \mathbb{R}, m(t) := I\!E(X_t)$$
$$\textit{Kovarianzfunktion: } \Gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \Gamma(t, s) := Cov(X_t, X_s)$$

bis Äquivalenz eindeutig festgelegt. Zwei stochastische Prozesse $(X_t)_{t\in\mathbb{R}}$ bzw. $(\Pi_t)_{t\in\mathbb{R}}$ auf den Wahrscheinlichkeitsräumen $(\Omega, \mathcal{A}, P)$ bzw. $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ heißen dabei äquivalent, falls für beliebige Zeitpunkte $\{t_1, ..., t_n\}$ für $k \in \mathbb{N}$ ihre endlich-dimensionalen Verteilung $P_{(X_{t_1},...,t_n)}, \tilde{P}_{(Y_{t_1},...,t_n)}$ identisch sind (Bauer (2002), S. 331).

In Kustosz et al. (2016a) wird der Zwischenprozess im Grenzwert von Satz 3.10(i) als folgender Gauß-Prozess charakterisiert:

Satz 3.14 (Kovarianzstruktur des Zwischenprozesses).
Für das gegebene Regressionsmodell in Definition 2.1 gilt:

$$\left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Phi(E_n), \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]} \left(\frac{n_1}{N} - t \right) \Phi(E_n) \right) \xrightarrow[N\to\infty]{\mathcal{D}} (G_t)_{t\in\mathbb{R}} = (G_{1,t}, G_{2,t})_{t\in\mathbb{R}},$$

wobei $(G_t)_{t\in\mathbb{R}} = (G_{1,t}, G_{2,t})_{t\in\mathbb{R}}$ *ein zentrierter bivariater Gauß-Prozess mit folgender Kovarianzstruktur unter den Koordinaten ist:*

$$Cov(G_t, G_s) = \begin{pmatrix} 1 & \int_0^1 \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}(x-s)\,dx \\ \int_0^1 \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}(x-t)\,dx & \int_0^1 \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}(x-s)\mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}(x-t)\,dx \end{pmatrix}$$

Beweis. In Kustosz et al. (2016a) wird die Aussage, wie in Bemerkung 3.12 bereits skizziert, gezeigt.

Alternativ kann man den Zwischenprozess im Grenzwert von Satz 3.10(i) betrachten und mit Satz 3.13 zeigen, dass er genau diesem Gauß-Prozess entspricht. Dazu schreiben wir $\left(\tilde{G}_{1,t}, \tilde{G}_{2,t}\right)_{t\in\left[-\frac{1}{2},\frac{3}{2}\right]} = (B_1, \psi_2(B_\bullet)_t)_{t\in\left[-\frac{1}{2},\frac{3}{2}\right]}$. Wir sehen unmittelbar ein, dass die endlich-dimensionalen Verteilungen koordinatenweise multivariaten Normalverteilungen entsprechen. Nach Satz 3.13 können wir durch den Gauß-Prozess durch die Erwartungswert- und Kovarianzfunktion bis auf Äquivalenz eindeutig charakterisieren. Für den Erwartungswert gilt für $t \in \left[-\frac{1}{2}, \frac{3}{2}\right]$:

$$\begin{aligned} \mathbb{E}(\tilde{G}_{1,t}) &= \mathbb{E}(B_1) = 0 \\ \mathbb{E}(\tilde{G}_{2,t}) &= \mathbb{E}(\psi_2(B_\bullet)_t) = \mathbb{E}\left(B_{(t+\frac{1}{2})\wedge 1}\right) - \mathbb{E}\left(B_{(t-\frac{1}{2})\vee 0}\right) = 0, \end{aligned}$$

da $B_t \sim \mathcal{N}(0,t)$ für $t \in [0,1]$. Für die nachfolgenden Rechnungen verwenden wir $\mathbb{E}(B_t B_s) = t \wedge s$ für $t, s \in [0,1]$ (Klenke (2006), S. 437), um die Kovarianzfunktion

zu bestimmen. Für den Matrix-Eintrag an der Stelle $(1,1)$ gilt:

$$\mathrm{Cov}(\tilde{G}_{1,t}, \tilde{G}_{1,s}) = \mathbb{E}(B_1^2) = 1.$$

Für den Matrix-Eintrag an der Stelle $(1,2)$ gilt:

$$\begin{aligned}\mathrm{Cov}(\tilde{G}_{1,t}, \tilde{G}_{2,s}) &= \mathbb{E}\left(B_1\left(B_{(s+\frac{1}{2})\wedge 1} - B_{(s-\frac{1}{2})\vee 0}\right)\right) = \left(s+\frac{1}{2}\right)\wedge 1 - \left(s-\frac{1}{2}\right)\vee 0 \\ &= \int_{(s-\frac{1}{2})\vee 0}^{(s+\frac{1}{2})\wedge 1} 1\,dx = \int_0^1 \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}(x-s)\,dx.\end{aligned}$$

wobei $s \leq \frac{3}{2}$ verwendet wird. Die letzte Gleichheit ergibt sich analog wie im Beweis von Lemma 3.2 zur Bestimmung von Formel (3.5). Der Eintrag an der Stelle $(2,1)$ ist vollständig analog berechenbar. Die Berechnung der Matrix-Stelle $(2,2)$ ist etwas umfangreicher, da zwei Zeitparameter t, s berücksichtigt werden:

$$\begin{aligned}&\mathrm{Cov}(\tilde{G}_{2,t}, \tilde{G}_{2,s}) = \mathbb{E}\left(\left(B_{(t+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee 0}\right)\left(B_{(s+\frac{1}{2})\wedge 1} - B_{(s-\frac{1}{2})\vee 0}\right)\right) \\ &= \left(t+\tfrac{1}{2}\right)\wedge\left(s+\tfrac{1}{2}\right)\wedge 1 - \left(\left(t+\tfrac{1}{2}\right)\wedge 1\right)\wedge\left(\left(s-\tfrac{1}{2}\right)\vee 0\right) \\ &- \left(\left(t-\tfrac{1}{2}\right)\vee 0\right)\wedge\left(\left(s+\tfrac{1}{2}\right)\wedge 1\right) + \left(\left(t-\tfrac{1}{2}\right)\vee 0\right)\wedge\left(\left(s-\tfrac{1}{2}\right)\vee 0\right) \\ &= \begin{cases} \left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right), & \text{für } (t,s)\in\left[-\frac{1}{2},\frac{1}{2}\right)^2 \\ \left(t+\frac{1}{2}\right) - \left(t+\frac{1}{2}\right)\wedge\left(s-\frac{1}{2}\right), & \text{für } (t,s)\in\left[-\frac{1}{2},\frac{1}{2}\right)\times\left[\frac{1}{2},\frac{3}{2}\right] \\ \left(s+\frac{1}{2}\right) - \left(t-\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right), & \text{für } (t,s)\in\left[\frac{1}{2},\frac{3}{2}\right]\times\left[-\frac{1}{2},\frac{1}{2}\right) \\ 1 - \left(t-\frac{1}{2}\right) - \left(s-\frac{1}{2}\right) + \left(t-\frac{1}{2}\right)\wedge\left(s-\frac{1}{2}\right), & \text{für } (t,s)\in\left[\frac{1}{2},\frac{3}{2}\right]^2 \end{cases}\end{aligned} \tag{3.22}$$

Der zweite und dritte Fall in (3.22) ist 0, falls $|t-s| \geq 1$. Ferner kann dies im ersten und vierten Fall nie eintreten. Wir können daher alle Fälle mit der Indikatorfunktion $\mathbb{1}\{|t-s| \leq 1\}$ multiplizieren. Zudem können wir den vierten Fall durch die

Darstellung mit dem Maximum verkürzt aufschreiben. (3.22) ist also identisch mit:

$$
\begin{cases}
\left(\left(t+\frac{1}{2}\right) \wedge \left(s+\frac{1}{2}\right)\right) \mathbb{1}\{|t-s| \leq 1\}, & \text{für } (t,s) \in \left[-\frac{1}{2}, \frac{1}{2}\right)^2 \\
\left(\left(t+\frac{1}{2}\right) - \left(s-\frac{1}{2}\right)\right) \mathbb{1}\{|t-s| \leq 1\}, & \text{für } (t,s) \in \left[-\frac{1}{2}, \frac{1}{2}\right) \times \left[\frac{1}{2}, \frac{3}{2}\right] \\
\left(\left(s+\frac{1}{2}\right) - \left(t-\frac{1}{2}\right)\right) \mathbb{1}\{|t-s| \leq 1\}, & \text{für } (t,s) \in \left[\frac{1}{2}, \frac{3}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right) \\
\left(1 - \left(t-\frac{1}{2}\right) \vee \left(s-\frac{1}{2}\right)\right) \mathbb{1}\{|t-s| \leq 1\}, & \text{für } (t,s) \in \left[\frac{1}{2}, \frac{3}{2}\right]^2
\end{cases}
$$

$$
= \left(\left(t+\tfrac{1}{2}\right) \wedge \left(s+\tfrac{1}{2}\right) \wedge 1 - \left(t-\tfrac{1}{2}\right) \vee \left(s-\tfrac{1}{2}\right) \vee 0\right) \mathbb{1}\{|t-s| \leq 1\}
$$

$$
= \int_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 1}^{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 0} \mathbb{1}\{|t-s| \leq 1\}\, dx = \int_0^1 \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}(x-t) \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}(x-s)\, dx.
$$

Die letzte Gleichheit soll nun erklärt werden. Dazu betrachten wir die Bedingungen der Indikatorfunktionen auf der rechten Seite der Gleichheit:

$$
\begin{aligned}
t - \frac{1}{2} < x \leq t + \frac{1}{2}, \\
s - \frac{1}{2} < x \leq s + \frac{1}{2}, \\
0 < x \leq 1.
\end{aligned}
$$

Wir können die Ungleichungen zusammenfassen, wenn wir das Maximum der unteren Schranken und das Minimum der oberen Schranken betrachtet. Dabei muss beachtet werden, dass die unteren Grenzen einer Ungleichung nicht größer als die obere Grenze einer anderen Ungleichung sein darf. Dies lässt sich mit der Bedingung $|t-s| \leq 1$ als zusätzliche Bedingungen zusammenfassen, wodurch sich die Indikatorfunktion auf der linken Seite im Integranden ergibt:

$$
\begin{aligned}
\left(t-\tfrac{1}{2}\right) \vee \left(s-\tfrac{1}{2}\right) \vee 0 < x \leq \left(t+\tfrac{1}{2}\right) \wedge \left(s+\tfrac{1}{2}\right) \wedge 1, \\
|t-s| \leq 1.
\end{aligned}
$$

Damit ist gezeigt, dass die Erwartungswert- und Kovarianzfunktion mit dem Gauß-Prozess aus Satz 3.14 übereinstimmen, womit nach Satz 3.13 die Äquivalenz der Prozesse gezeigt ist.

3.4 Berechnung der Quantile der asymptotischen Verteilung der vollen Dreier-Tiefe

Aus den Resultaten von Satz 3.10 können wir numerisch durch eine Simulation die Quantile der asymptotischen Verteilung der normierten vollen Dreier-Tiefe bestimmen. Die Simulation entspricht einer Anwendung des Satzes von Glivenko-Cantelli. Dazu führen wir die Notation für das theoretische bzw. empirische Quantil einer beliebigen reellwertigen Zufallsvariable X mit Verteilungsfunktion F_X bzw. einer Stichprobe $X_1(\omega), ..., X_N(\omega)$ als Realisation von reellwertigen Zufallsvariablen mit empirischer Verteilungsfunktion F_N und einem $\alpha \in (0,1)$ ein:

- $q_\alpha^X := \inf\{x \in \mathbb{R}; F_X(x) \geq \alpha\}$ heißt theoretisches α-Quantil von X.
- $q_\alpha^N := q_{X_1(\omega),...,X_N(\omega)} = \inf\{x \in \mathbb{R}; F_N(x) \geq \alpha\}$ heißt empirisches α-Quantil von $X_1(\omega), ..., X_N(\omega)$.

Nach dem Satz von Glivenko-Cantelli (Klenke (2006), S. 111) lassen sich die theoretischen Quantile durch die empirischen Quantile (P-fast sicher in der Supremumsnorm) approximieren. Für eine Folge $(X_n)_{n\in\mathbb{N}}$ von unabhängig, identisch verteilten Zufallsvariablen mit Verteilungsfunktion F_X gilt:

$$q_\alpha^N \xrightarrow[N\to\infty]{} q_\alpha^X \ P\text{-fast sicher.} \tag{3.23}$$

Wir simulieren $S = 80.000$ unabhängige identische Wiederholungen von Pfaden der Brownschen Bewegung $(B_t^1)_{t\in[0,1]}, ..., (B_t^S)_{t\in[0,1]}$ im Intervall $[0,1]$ mit dem `R`-Paket `somebm` und dem Befehl `bm()`. Wir identifizieren dabei im Folgenden $(B_t^s)_{t\in[0,1]}$ für $s = 1, ..., S$ bereits als eine Realisation eines Pfades der Brownschen Bewegung und schreiben nicht $(B_t^s(\omega))_{t\in[0,1]}$. Ein Pfad der Brownschen Bewegung wird wie folgt approximativ gebildet. Seien dazu $X_k^1, ..., X_k^T$ unabhängig, identisch verteilte $\mathcal{N}(0,1)$-Zufallsvariablen für $k = 1, ..., T$, wobei hier $T = 10.000$ gesetzt wird. Wir diskretisieren das Intervall $[0,1]$ äquidistant in $\frac{1}{T}$-Schritten und beginnen zum Zeitpunkt $t = 0$ mit dem Startwert $X_0^i = 0$ für den i-ten Pfad einer Brownschen Bewegung. Wir addieren sukzessive die normalverteilten Zufallsvariablen X_k^i als un-

abhängige Zuwächse des i-ten Pfades im Intervall $\left[\frac{k}{T}, \frac{k+1}{T}\right]$ auf

$$B_t^i = \frac{1}{\sqrt{T}} \sum_{k=0}^{t} X_k^i.$$

Für hohes T und linearer Interpolation der Punkte ergibt sich nach dem Satz von Donsker approximativ ein Pfad der Brownschen Bewegung, da wir $(X_k^i)_{k\in\{1,\dots,T\}}$ als Zuwächse einer zentrierten Irrfahrt mit Varianz 1 auffassen können. Wir werten jedes der 80.000 Pfade einer Brownschen Bewegung durch folgendes Funktional aus:

$$\Psi : \mathcal{D}[0,1] \to \mathbb{R},\ \Psi(B_\bullet) = \frac{3}{4}B_1^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt + \frac{3}{4},$$

wobei ψ_2 die zweite Koordinate des in Satz 3.10 definierten Operators ist. Das Integral wird dabei zunächst vereinfacht:

$$\begin{aligned}
\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt &= \int_{-\frac{1}{2}}^{\frac{1}{2}} B_{t+\frac{1}{2}}^2\, dt + \int_{\frac{1}{2}}^{\frac{3}{2}} (B_1 - B_{t-\frac{1}{2}})^2\, dt \\
&= \int_0^1 B_t^2\, dt + \int_0^1 (B_1^2 - 2B_1B_t + B_t^2)\, dt \\
&= 2\int_0^1 B_t^2\, dt + B_1^2 - 2B_1 \int_0^1 B_t\, dt. \qquad (3.24)
\end{aligned}$$

Diese Integrale werden numerisch mit der Trapezregel bestimmt, wodurch für großes N eine Approximation der Integrale gewährleistet ist (Oevel (1995), S. 429ff.):

$$\begin{aligned}
\int_0^1 B_t\, dt &= \lim_{N\to\infty} \frac{1}{2N} \sum_{n=1}^{N} \left(B_{\frac{n}{N}} + B_{\frac{n-1}{N}}\right) \\
\int_0^1 B_t^2\, dt &= \lim_{N\to\infty} \frac{1}{2N} \sum_{n=1}^{N} \left(B_{\frac{n}{N}}^2 + B_{\frac{n-1}{N}}^2\right)
\end{aligned}$$

Hierbei sollte einerseits der Diskretisierungsfehler bei der Trapezregel und andererseits bei der Konstruktion eines Pfades einer Brownschen Bewegung erwähnt bleiben. Bekannte Referenzwerte, wie das arithmetische Mittel aus den Daten, das nach dem Gesetz der großen Zahlen den Erwartungswert approximiert, können als Referenzen für die Höhe des Diskretisierungsfehlers dienen. Der Erwartungswert des Grenzwert

beträgt 0, denn mit der Darstellung des Integrals in (3.24) und mit dem Satz von Fubuni (Klenke (2006), S. 265) folgt

$$\begin{aligned}
&\mathbb{E}\left(\frac{3}{4}B_1^2 - \frac{3}{2}\int_{-\frac{1}{2}}^{\frac{3}{2}} \psi_2(B_\bullet)_t^2\, dt + \frac{3}{4}\right) \\
=&\frac{3}{4} - \frac{3}{2}\mathbb{E}\left(2\int_0^1 B_t^2\, dt + B_1^2 - 2\int_0^1 B_1 B_t\, dt\right) + \frac{3}{4} \\
=&-3\int_0^1 \mathbb{E}(B_t^2)\, dt + 3\int_0^1 \mathbb{E}(B_1 B_t)\, dt = -\frac{3}{2} + \frac{3}{2} = 0.
\end{aligned}$$

So erhalten wir 80.000 reelle Zahlen, aus denen wir nach (3.23) simulativ die theoretischen Quantile bestimmen können. Ein Vergleich mit den Quantilen aus dem R-Paket `rexpar` von Szugat und Kustosz (2016) ergibt geringfügige Unterschiede. Das Stichprobenmittel der simulierten Daten beträgt -0.001, was ein Indikator für eine wirkende Approximation sein kann. Einige häufig verwendete Quantile aus den Berechnungen werden in Tabelle 4 angegeben. Im R-Paket `rexpar` können mit dem

Tabelle 4: Quantile für die asymptotische Verteilung der vollen Dreier-Tiefe

α	0.1	0.05	0.01	0.001
$q_\alpha^{(3)}$	-0.797	-1.227	-2.246	-3.654

Befehl `SimQuants()` die Quantile der vollen Dreier-Tiefe mit einer in Kustosz et al. (2016a) dargestellten Methode alternativ bestimmt werden.

Notwendiger Stichprobenumfang

Mit dem gleichen Ansatz wie im Ende von Kapitel 2 können wir aus den simulierten asymptotischen Quantilen einen notwendigen Stichprobenumfang berechnen, für den das asymptotische Testverfahren erst sinnvoll sein kann. Dazu betrachten den Fall, dass die volle Dreier-Tiefe den Wert 0 annimmt und lösen nach dem Stichprobenumfang N aufgerundet auf die nächste natürliche Zahl auf:

$$-\frac{N}{4} \overset{!}{\leq} q_\alpha^{(3)} \quad \Leftrightarrow \quad N \overset{!}{\geq} -4q_\alpha^{(3)}.$$

In der Tabelle 5 sind die berechneten Stichprobenumfänge N für typische Signifikanzniveaus angegeben. Falls ein vorliegender Stichprobenumfang unterhalb des

Tabelle 5: Notwendiger Stichprobenumfang N in Abhängigkeit von häufig verwendeten Signifikanzniveaus α für den Test aus Korollar 3.11

α	0.1	0.05	0.01	0.001
notwendiges N	4	5	9	15

notwendigen N liegt, sollten die exakten Quantile der vollen Dreier-Tiefe verwendet werden, da der asymptotische Test nicht ablehnen kann und so nie zu signifikanten Ergebnissen kommen wird. Die exakten Quantile lassen sich durch Simulationen bestimmen.

4 Asymptotische Verteilung höherer Tiefen

Im vierten Kapitel wird die asymptotische Verteilung der **vollen Vierer-Tiefe** hergeleitet. Die Beweisstruktur ist analog zur vollen Dreier-Tiefe. Zunächst wird die Φ-Darstellung für beliebige volle Datentiefen im Kapitel 4.1 hergeleitet, wodurch der erste Beweisschritt bereits für beliebige Datentiefen vollzogen wird. Im Kapitel 4.2 wird der Fokus auf die Asymptotik der vollen Vierer-Tiefe gelegt. Es werden an einigen Stellen zusätzliche Argumente im Vergleich zur vollen Dreier-Tiefe benötigt. Die Resultate dieses Kapitels beruhen vollständig auf der Arbeit des Autors und bauen auf der Vorarbeit des zweiten und dritten Kapitels auf.

4.1 Φ-Darstellung von höheren Tiefen

Die Φ-Darstellung für volle Datentiefen haben eine vorbereitende Funktion für die Anwendung eines Zentralen Grenzwertsatzes. Für höhere Datentiefen werden allerdings nicht nur Produkte aus zwei Faktoren von der Funktion Φ benötigt. Der nachfolgende Satz ist vom Autor der Arbeit selbst gefunden und bewiesen worden:

Satz 4.1 (Φ-Darstellung von allgemeinen vollen Datentiefen).
Für Zufallsvariablen $E_{n_1}, ..., E_{n_K}$ mit $P(E_{n_i} \neq 0) = 1$ für $i = 1, ..., K$ gilt für $K \in \mathbb{N}$:

$$\begin{aligned}&\prod_{i=1}^{K} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K-1}} \\ =&\frac{1}{2^{K-1}} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<...<m_{2L} \\ \subseteq\{n_1,...,n_K\}}} \prod_{j \in \mathcal{N}(m_1,...,m_{2L})} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \ P\text{-fast sicher.}\end{aligned}$$

wobei $\mathcal{N}(m_1, ..., m_{2L}) = \{j \in \{1, .., K\}; \exists i \in \{1, ..., 2L\} : m_i = n_j\}$ eine von den Laufindizes abhängige Menge ist und $\Phi(x) := \mathbb{1}\{x < 0\} - \mathbb{1}\{x > 0\}$ ist.

Die Summe $\sum_{\substack{m_1<...<m_{2L} \\ \subseteq\{n_1,...,n_K\}}}$ beschreibt hierbei die Summation über alle geordneten $2L$-elementigen Teilmengen aus $\{n_1, ..., n_K\}$. Ferner schreiben wir der Übersicht wegen im Folgenden statt $\mathcal{N}(m_1, ..., m_{2L}) = \mathcal{N}$. Bevor wir den Satz 4.1 beweisen, soll die vorgestellte Formel erklärt werden. Der in Kapitel 4.2 relevante Spezialfall für $K = 4$ wird als Korollar zur Veranschaulichung angegeben:

Korollar 4.2 (Φ-Darstellung der vollen Vierer-Tiefe).
Für Zufallsvariablen $E_{n_1}, E_{n_2}, E_{n_3}, E_{n_4}$ mit $P(E_{n_i} \neq 0) = 1$ für $i = 1, 2, 3, 4$ gilt:

$$\begin{aligned}
&\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0, E_{n_3} > 0, E_{n_4} < 0\} \\
&+\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0, E_{n_3} < 0, E_{n_4} > 0\} - \frac{1}{8} \\
&= \frac{1}{8}\Big(\prod_{i=1}^{4} \Phi(E_{n_i}) - \Phi(E_{n_1})\Phi(E_{n_2}) + \Phi(E_{n_1})\Phi(E_{n_3}) - \Phi(E_{n_1})\Phi(E_{n_4}) \\
&\qquad\qquad - \Phi(E_{n_2})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_4}) - \Phi(E_{n_3})\Phi(E_{n_4})\Big) \ \textit{P-fast sicher.}
\end{aligned}$$

Für höhere K-Tiefen kommen Produkte mit gerader Länge bis $\lfloor \frac{K}{2} \rfloor$ vor, wie die erste Summe in Satz 4.1 ausgedrückt. Es werden alle Kombinationen von Tupeln gerader Längen betrachtet und als Produkte mit der Funktion Φ geschrieben. Das Vorzeichen des Produkts ergibt sich aus der Position der Indizes im Tupel. Das Produkt wird negativ gewichtet, falls die Summe der betrachteten Positionen ungerade ist.

Beweis. Wir zeigen die Aussage mit vollständiger Induktion, wobei jeweils induktiv über die geraden Zahlen mit $2K \to 2(K+1)$ und die ungeraden Zahlen mit $2K+1 \to 2(K+1)+1$ geschlossen wird. Es genügt ein Induktionsbeweis für $K \to K+2$ durchzuführen, wenn der Induktionsanfang für $K = 2$ und $K = 3$ gezeigt ist, was in Lemma 2.3 und Lemma 3.1 durchgeführt wird. Sei die Aussage in Satz 4.1 für ein K wahr und folgere, dass sie auch für $K+2$ wahr ist.

$$\begin{aligned}
&\prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K+1}} \\
&= \prod_{i=3}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} \left(\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} + \mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} - \frac{1}{2} \right) \\
&-\mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} \prod_{i=3}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \frac{1}{2} \prod_{i=3}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} \\
&+ \prod_{i=3}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} \left(\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} + \mathbb{1}\{E_{n_1} > 0, E_{n_2} < 0\} - \frac{1}{2} \right) \\
&-\mathbb{1}\{E_{n_1} < 0, E_{n_2} > 0\} \prod_{i=3}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} + \frac{1}{2} \prod_{i=3}^{K+1} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K+1}}.
\end{aligned} \tag{4.1}$$

Durch Ausklammern des Vorfaktors $-\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2})$ und Durchführung einer Nulladdition mit $-\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2})$ ist (4.1) gleich:

$$
\begin{aligned}
&-\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_2})\left(\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}+\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K-1}}\right)\\
&-\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2})+\frac{1}{2}\left(\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}+\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K-1}}\right)\\
&-\Bigg(\mathbb{1}\{E_{n_1}>0,E_{n_2}<0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}\\
&\quad+\mathbb{1}\{E_{n_1}<0,E_{n_2}>0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\Bigg).
\end{aligned}
\tag{4.2}
$$

Nach der Induktionsvoraussetzung können wir die Φ-Darstellung für die volle K-Tiefe nutzen und erhalten aus (4.2):

$$
\begin{aligned}
&-\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2})\sum_{L=1}^{\lfloor\frac{K}{2}\rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_3,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2})+\frac{1}{2^K}\sum_{L=1}^{\lfloor\frac{K}{2}\rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_3,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\Bigg(\mathbb{1}\{E_{n_1}>0,E_{n_2}<0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}\\
&\quad+\mathbb{1}\{E_{n_1}<0,E_{n_2}>0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\Bigg).
\end{aligned}
\tag{4.3}
$$

Der negative Summand in (4.3) wird nun untersucht. Eine analoge Rechnung wie oben liefert mit der Induktionsvoraussetzung durch Ausklammern von $-\frac{1}{2}\Phi(E_{n_1})\Phi(E_{n_3})$

$$
\begin{aligned}
&\Bigg(\mathbb{1}\{E_{n_1}>0,E_{n_2}<0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}\\
&\quad+\mathbb{1}\{E_{n_1}<0,E_{n_2}>0\}\prod_{i=3}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_3})\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_3})+\frac{1}{2^K}\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\Bigg(\mathbb{1}\{E_{n_1}<0,E_{n_2}<0,E_{n_3}>0\}\prod_{i=4}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}\\
&\quad+\mathbb{1}\{E_{n_1}>0,E_{n_2}>0,E_{n_3}<0\}\prod_{i=4}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\Bigg).
\end{aligned}
\tag{4.4}
$$

Erneut wird der letzte Summand in (4.4) analog zu oben mit Anwendung der Induktionsvoraussetzung und mit Ausklammern von $-\frac{1}{2}\Phi(E_{n_2})\Phi(E_{n_3})$ betrachtet:

$$
\begin{aligned}
&\Bigg(\mathbb{1}\{E_{n_1}<0,E_{n_2}<0,E_{n_3}>0\}\prod_{i=4}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}\\
&\quad+\mathbb{1}\{E_{n_1}>0,E_{n_2}>0,E_{n_3}<0\}\prod_{i=4}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\Bigg)\\
&= -\frac{1}{2^K}\Phi(E_{n_2})\Phi(E_{n_3})\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\frac{1}{2^K}\Phi(E_{n_2})\Phi(E_{n_3})+\frac{1}{2^K}\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor}\sum_{\substack{m_1<\ldots<m_{2L}\\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}}\prod_{j\in\mathcal{N}}(-1)^j\prod_{i=1}^{2L}\Phi(E_{m_i})\\
&-\left(\prod_{i=1}^{K+2}\mathbb{1}\{E_{n_i}(-1)^i>0\}+\prod_{i=1}^{K+2}\mathbb{1}\{E_{n_i}(-1)^{i+1}>0\}-\frac{1}{2^{K+1}}\right),
\end{aligned}
\tag{4.5}
$$

wodurch sich in der letzten Zeile von (4.5) ein Summand der vollen $(K+2)$-Tiefe mit negativen Vorzeichen ergibt. Wir werden die oben berechneten Darstellungen von (4.2) bis (4.4) rekursiv einsetzen, bis der Ausdruck in (4.5) übrig bleibt und

erhalten für einen Summanden der vollen $(K+2)$-Tiefe:

$$
\begin{aligned}
&\prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K+1}} \\
&= -\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,\ldots,n_K\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad -\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_2}) + \frac{1}{2^K} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad +\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad +\frac{1}{2^K}\Phi(E_{n_1})\Phi(E_{n_3}) - \frac{1}{2^K} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad -\frac{1}{2^K}\Phi(E_{n_2})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad -\frac{1}{2^K}\Phi(E_{n_2})\Phi(E_{n_3}) + \frac{1}{2^K} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&\quad -\left(\prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K+1}}\right)
\end{aligned}
$$

Die Addition mit (4.5) und die Division durch 2 liefern folgende Darstellung:

$$
\begin{aligned}
&\prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K+2} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K+1}} \\
=&\frac{1}{2^{K+1}}\Bigg(- \Phi(E_{n_1})\Phi(E_{n_2}) + \Phi(E_{n_1})\Phi(E_{n_3}) - \Phi(E_{n_2})\Phi(E_{n_3}) && (4.6)\\
&+ \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.7)\\
&- \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.8)\\
&+ \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.9)\\
&- \Phi(E_{n_1})\Phi(E_{n_2}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.10)\\
&+ \Phi(E_{n_1})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.11)\\
&- \Phi(E_{n_2})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i})\Bigg) && (4.12)
\end{aligned}
$$

Bei der Addition von (4.7) und (4.8) fallen alle Kombinationen, in denen weder n_2 noch n_3 auftreten, weg, da die Summen unterschiedliche Vorzeichen besitzen. In (4.9) sind jene weggefallenen Kombinationen dabei enthalten:

$$
\begin{aligned}
&\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&- \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
=&\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1=n_3<m_2<\ldots<m_{2L} \\ \subseteq\{n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.13)\\
&- \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1=n_2<m_2<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) && (4.14)
\end{aligned}
$$

Ferner lässt sich (4.9) in eine Summe mit Summanden zerlegen, die n_1 enthält bzw. nicht n_1 enthält:

$$\sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i})$$

$$= \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1=n_1<m_2<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \tag{4.15}$$

$$+ \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \tag{4.16}$$

In (4.13) bis (4.15) liegen alle Kombinationen der Länge $2, \ldots, 2\left\lfloor \frac{K}{2} \right\rfloor$ vor, die nur genau eines der Indizes $\{n_1, n_2, n_3\}$ enthalten. In Formel (4.16) liegen alle Kombinationen der Länge $2, \ldots, 2\left\lfloor \frac{K}{2} \right\rfloor$ vor, die genau keines der Indizes $\{n_1, n_2, n_3\}$ enthalten. Ferner betrachten wir die Summe der Formeln (4.10) und (4.11). Dabei fallen alle Kombinationen weg, bei denen n_3 in (4.10) bzw. n_2 in (4.11) vorkommt. Ferner nehmen wir die Vorfaktoren $\Phi(E_{n_i})\Phi(E_{n_j})$ für $(i,j) = (1,2)$ bzw. $(i,j) = (1,3)$ in die Summe auf und summieren so ab $L = 2$ bis $\left\lfloor \frac{K+2}{2} \right\rfloor$:

$$- \Phi(E_{n_1})\Phi(E_{n_2}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i})$$

$$+ \Phi(E_{n_1})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i})$$

$$= - \sum_{L=2}^{\lfloor \frac{K+2}{2} \rfloor} \sum_{\substack{m_1=n_1<m_2=n_2<m_3<\ldots<m_{2L} \\ \subseteq\{n_1,n_2,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \tag{4.17}$$

$$+ \sum_{L=2}^{\lfloor \frac{K+2}{2} \rfloor} \sum_{\substack{m_1=n_1<m_2=n_3<m_3<\ldots<m_{2L} \\ \subseteq\{n_1,n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \tag{4.18}$$

Wir zerlegen (4.12) in zwei Summen, sodass in der einen Summe n_1 nicht vorkommt

und in der anderen Summe n_1 auftritt:

$$
\begin{aligned}
&- \Phi(E_{n_2})\Phi(E_{n_3}) \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<\ldots<m_{2L} \\ \subseteq\{n_1,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \\
&= - \sum_{L=2}^{\lfloor \frac{K+2}{2} \rfloor} \sum_{\substack{m_1=n_2<m_2=n_3<m_3<\ldots<m_{2L} \\ \subseteq\{n_2,n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \qquad (4.19) \\
&\quad - \sum_{L=2}^{\lfloor \frac{K+2}{2} \rfloor} \sum_{\substack{m_1=n_1<m_2=n_2<m_3=n_3<m_4<\ldots<m_{2L} \\ \subseteq\{n_1,n_2,n_3,n_4,\ldots,n_{K+2}\}}} \prod_{j\in\mathcal{N}} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}) \qquad (4.20)
\end{aligned}
$$

In (4.17) bis (4.19) treten alle Kombinationsmöglichkeiten der Länge $4, \ldots, 2\left\lfloor \frac{K+2}{2} \right\rfloor$ auf, in denen jeweils genau zwei Indizes aus $\{n_1, n_2, n_3\}$ vorkommen, wobei der dritte komplementäre Index genau nicht vorkommt. Nimmt man die einzelnen Summanden aus (4.6) in (4.17) bis (4.19) auf, ergeben sich auch jene Kombinationsmöglichkeiten zusätzlich mit Länge 2. In (4.20) kommen alle Kombinationsmöglichkeiten der Länge $4, \ldots, 2\left\lfloor \frac{K+2}{2} \right\rfloor$ vor, die alle Indizes $\{n_1, n_2, n_3\}$ gleichzeitig enthalten. Die Kombinationsmöglichkeiten der Länge 2 können in diesem Fall nicht auftreten. Zusammenfassend erhalten wir mit (4.13) bis (4.20) jeweils einen Fall, in dem eine Möglichkeit unter $2^3 = 8$ Möglichkeiten vorkommt, bei der bis zu drei Elemente aus der Menge $\{n_1, n_2, n_3\}$ gezogen werden. Die einzelnen Summanden entsprechen einander disjunkten Fällen und ergeben nach Addition mit der kombinatorischen Überlegung die behauptete Formel für $K + 2$. □

Der Satz 4.1 gilt auch für $K = 1$, da dann auf beiden Seiten 0 steht. Man hätte auf den Beweis von Lemma 3.1 verzichten können und stattdessen als Induktionsvoraussetzung im ungeraden Fall mit $K = 1$ argumentieren und Satz 4.1 in Kapitel 3 vorziehen können. In Kapitel 3 wird der Beweis für einen Spezialfall vorgestellt, um den Fokus von der vollen Dreier-Tiefe nicht zu verlieren und damit besser der Beweisidee von Satz 4.1 gefolgt werden kann.

4.2 Asymptotische Verteilung der vollen Vierer-Tiefe

In der Φ-Darstellung der vollen Vierer-Tiefe tritt neben zweifachen Produkten der Form $\Phi(E_{n_i})\Phi(E_{n_j})$ für $i,j = 1,...,4$ mit $i \neq j$ ein vierfaches Produkt auf. Im nächsten Lemma sehen wir, dass wir den Anteil der Darstellung asymptotisch vernachlässigen können.

Lemma 4.3 (Asymptotische Vernachlässigbarkeit des Vierer-Produkts). *Für eine Folge $(E_N)_{N\in\mathbb{N}}$ von unabhängigen Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit $P(E_n \neq 0) = 1$ für $n \in \mathbb{N}$ gilt:*

$$\frac{N}{\binom{N}{4}} \sum_{1\leq n_1<n_2<n_3<n_4\leq N} \prod_{i=1}^{4} \Phi(E_{n_i}) \xrightarrow[N\to\infty]{L^2} 0.$$

Insbesondere folgt daraus die stochastische Konvergenz

$$\frac{N}{\binom{N}{4}} \sum_{1\leq n_1<n_2<n_3<n_4\leq N} \prod_{i=1}^{4} \Phi(E_{n_i}) \xrightarrow[N\to\infty]{P} 0.$$

Die asymptotische Vernachlässigbarkeit in Lemma 4.3 wurde vom Autor dieser Arbeit selbst gefunden und bewiesen.

Beweis. Wir berechnen das Integral des Quadrats der linken Seite:

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{N}{\binom{N}{4}} \sum_{1\leq n_1<n_2<n_3<n_4\leq N} \prod_{i=1}^{4} \Phi(E_{n_i})\right)^2\right) \\
&= \frac{N^2}{\binom{N}{4}^2} \sum_{1\leq n_1<n_2<n_3<n_4\leq N} \sum_{1\leq \overline{n_1}<\overline{n_2}<\overline{n_3}<\overline{n_4}\leq N} \mathbb{E}\left(\prod_{i=1}^{4} \Phi(E_{n_i})\Phi(E_{\overline{n_i}})\right). \qquad (4.21)
\end{aligned}$$

Der Erwartungswert in (4.21) ist genau dann ungleich 0, wenn für alle $n_i = \overline{n_i}$ für $i = 1,...,4$ gilt. Ansonsten gibt es mindestens ein Paar $n_i \neq \overline{n_j}$ für $i,j = 1,...,4$, sodass $\Phi(E_{n_i})$ und $\Phi(E_{n_j})$ stochastisch unabhängig sind und wegen $\mathbb{E}(\Phi(E_{n_i})) = 0$ der Erwartungswert gleich 0 ist. Es gibt insgesamt $\binom{N}{4}$ Kombinationen, bei denen der Erwartungswert wegen $\mathbb{E}\left(\Phi(E_{n_i})^2\right) = 1$ für beliebiges $i = 1,...,N$ gleich 1 ist.

Damit lässt sich (4.21) wie folgt weiter vereinfachen:

$$\frac{N^2}{\binom{N}{4}^2} \cdot \binom{N}{4} = \frac{24N}{(N-1)(N-2)(N-3)} \xrightarrow[N\to\infty]{} 0.$$

Mit der Tschebyscheff-Ungleichung (Klenke (2006), S. 104) folgt unmittelbar die stochastische Konvergenz:

$$P\left(\left|\frac{N}{\binom{N}{4}} \sum_{1\le n_1<n_2<n_3<n_4\le N} \prod_{i=1}^{4} \Phi(E_{n_i})\right| > \varepsilon\right) \le \frac{24N}{\varepsilon^2(N-1)(N-2)(N-3)} \xrightarrow[N\to\infty]{} 0,$$

für beliebiges $\varepsilon > 0$, da $\mathbb{E}\left(\frac{N}{\binom{N}{4}} \sum_{1\le n_1<n_2<n_3<n_4\le N} \prod_{i=1}^{4} \Phi(E_{n_i})\right) = 0$ ist. □

Der Nachweis der stochastischen Konvergenz durch die Tschebyscheff-Ungleichung zeigt, dass der Ausdruck mindestens mit Ordnung N^2 gegen 0 geht. In Kapitel 4.4 werden detailliertere quantitative Untersuchungen für diesen Fehler in Abhängigkeit eines gegebenen Stichprobenumfangs N durchgeführt. Nun können wir analog zum Satz 3.3 für die volle Dreier-Tiefe im nachfolgenden Satz 4.4 für die volle Vierer-Tiefe eine asymptotische Darstellung mit separierten Summanden in Produktform gewinnen. Anders als in Satz 3.3, wo eine exakte Darstellung vorliegt, verlieren wir durch die Anwendung von Lemma 4.3 die Exaktheit und erhalten einen additiven Term $o_P(1)$, der stochastisch gegen 0 konvergiert. Satz 4.4 ist vom Autor dieser Arbeit selbst formuliert und bewiesen worden.

Satz 4.4 (Darstellung mit separierten Summanden in Produktform).
Für das gegebene Regressionsmodell in Definition 2.1 gilt:

$$\begin{aligned}
N\left(d_S^4(\vartheta^*, Z) - \frac{1}{8}\right) &= \frac{N^4}{32\binom{N}{4}} - \frac{N^4}{32\binom{N}{4}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Phi(E_n)\right)^2 \\
&+ \frac{N^4}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N} - t\right) \Phi(E_n)\right)^2 dt \\
&- \frac{N^4}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N} - s\right) \Phi(E_n)\right)^2 dt\, ds + o_P(1).
\end{aligned}$$

Beweis. Wir verwenden summandenweise die Φ-Darstellung aus Korollar 4.2:

$$\begin{aligned}
&N\left(d_S^4(\vartheta^*, Z) - \frac{1}{8}\right) \\
=&\frac{N}{8\binom{N}{4}} \sum_{1\le n_1<n_2<n_3<n_4\le N} \Bigg(\prod_{i=1}^{4} \Phi(E_{n_i}) + \Phi(E_{n_1})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_4}) \\
&- \Phi(E_{n_1})\Phi(E_{n_2}) - \Phi(E_{n_1})\Phi(E_{n_4}) - \Phi(E_{n_2})\Phi(E_{n_3}) - \Phi(E_{n_3})\Phi(E_{n_4}) \Bigg) \qquad (4.22)
\end{aligned}$$

Nach Lemma 4.3 konvergiert $\frac{N}{8\binom{N}{4}} \sum_{1\le n_1<n_2<n_3<n_4\le N} \prod_{i=1}^{4} \Phi(E_{n_i})$ für $N \to \infty$ stochastisch gegen 0. Somit können wir diesen Ausdruck nach dem Lemma von Slutzky asymptotisch vernachlässigen (Bickel und Doksum (1997), S. 461) und können (4.22) umschreiben. Analog zu Satz 3.3 trennen wir dann die Summanden voneinander in separate Summen:

$$\begin{aligned}
&\frac{N}{8\binom{N}{4}} \sum_{1\le n_1<n_2<n_3<n_4\le N} \Bigg(\Phi(E_{n_1})\Phi(E_{n_3}) + \Phi(E_{n_2})\Phi(E_{n_4}) \\
&-\Phi(E_{n_1})\Phi(E_{n_2}) - \Phi(E_{n_1})\Phi(E_{n_4}) - \Phi(E_{n_2})\Phi(E_{n_3}) - \Phi(E_{n_3})\Phi(E_{n_4}) \Bigg) + o_P(1) \\
&= \frac{N}{8\binom{N}{4}} \Bigg(\sum_{1\le n_1<n_3\le N} (n_3 - n_1 - 1)(N - n_3)\Phi(E_{n_1})\Phi(E_{n_3}) \\
&\qquad + \sum_{1\le n_2<n_4\le N} (n_4 - n_2 - 1)(n_2 - 1)\Phi(E_{n_2})\Phi(E_{n_4}) \\
&\qquad - \frac{1}{2} \sum_{1\le n_1<n_2\le N} (N - n_2)(N - n_2 - 1)\Phi(E_{n_1})\Phi(E_{n_2}) \\
&\qquad - \sum_{1\le n_2<n_3\le N} (n_2 - 1)(N - n_3)\Phi(E_{n_2})\Phi(E_{n_3}) \\
&\qquad - \frac{1}{2} \sum_{1\le n_1<n_4\le N} (n_4 - n_1 - 1)(n_4 - n_1 - 2)\Phi(E_{n_1})\Phi(E_{n_4}) \\
&\qquad - \frac{1}{2} \sum_{1\le n_3<n_4\le N} (n_3 - 1)(n_3 - 2)\Phi(E_{n_3})\Phi(E_{n_4}) \Bigg) + o_P(1) \qquad (4.23)
\end{aligned}$$

Nun erlauben wir in allen sechs Summen in (4.23) die Permutation der Laufindizes

und gleichen mit dem Vorfaktor $\frac{1}{2!}$ aus:

$$\begin{aligned}
&\frac{N}{16\binom{N}{4}}\Bigg(\sum_{1\le n_1\neq n_3\le N}((n_1\vee n_3)-(n_1\wedge n_3)-1)(N-(n_1\vee n_3))\Phi(E_{n_1})\Phi(E_{n_3})\\
&+\sum_{1\le n_2\neq n_4\le N}((n_2\vee n_4)-(n_2\wedge n_4)-1)((n_2\wedge n_4)-1)\Phi(E_{n_2})\Phi(E_{n_4})\\
&-\frac{1}{2}\sum_{1\le n_1\neq n_2\le N}(N-(n_1\vee n_2))(N-(n_1\vee n_2)-1)\Phi(E_{n_1})\Phi(E_{n_2})\\
&-\sum_{1\le n_2\neq n_3\le N}((n_2\wedge n_3)-1)(N-(n_2\vee n_3))\Phi(E_{n_2})\Phi(E_{n_3})\\
&-\frac{1}{2}\sum_{1\le n_1\neq n_4\le N}((n_1\vee n_4)-(n_1\wedge n_4)-1)((n_1\vee n_4)-(n_1\wedge n_4)-2)\Phi(E_{n_1})\Phi(E_{n_4})\\
&-\frac{1}{2}\sum_{1\le n_3\neq n_4\le N}((n_3\wedge n_4)-1)((n-3\wedge n_4)-2)\Phi(E_{n_3})\Phi(E_{n_4})\Bigg)+o_P(1)
\end{aligned}\tag{4.24}$$

Die Laufindizes aller Summanden in (4.24) werden einheitlich benannt:

$$\begin{aligned}
&\frac{N}{16\binom{N}{4}}\Bigg(\sum_{1\le n_1\neq n_2\le N}((n_1\vee n_2)-(n_1\wedge n_2)-1)(N-(n_1\vee n_2))\Phi(E_{n_1})\Phi(E_{n_2})\\
&+\sum_{1\le n_1\neq n_2\le N}((n_1\vee n_2)-(n_1\wedge n_2)-1)((n_1\wedge n_2)-1)\Phi(E_{n_1})\Phi(E_{n_2})\\
&-\frac{1}{2}\sum_{1\le n_1\neq n_2\le N}(N-(n_1\vee n_2))(N-(n_1\vee n_2)-1)\Phi(E_{n_1})\Phi(E_{n_2})\\
&-\sum_{1\le n_1\neq n_2\le N}((n_1\wedge n_2)-1)(N-(n_1\vee n_2))\Phi(E_{n_1})\Phi(E_{n_2})\\
&-\frac{1}{2}\sum_{1\le n_1\neq n_2\le N}((n_1\vee n_2)-(n_1\wedge n_2)-1)((n_1\vee n_2)-(n_1\wedge n_2)-2)\Phi(E_{n_1})\Phi(E_{n_2})\\
&-\frac{1}{2}\sum_{1\le n_1\neq n_2\le N}((n_1\wedge n_2)-1)((n_1\wedge n_2)-2)\Phi(E_{n_1})\Phi(E_{n_2})\Bigg)+o_P(1)\\
&=\frac{N}{16\binom{N}{4}}\Bigg(\sum_{1\le n_1\neq n_2\le N}\Big(2N(n_1\vee n_2)-2N(n_1\wedge n_2)+\frac{1}{2}N-\frac{1}{2}N^2-2(n_1\vee n_2)^2\\
&-2(n_1\wedge n_2)^2+4(n_1\vee n_2)(n_1\wedge n_2)-1\Big)\Phi(E_{n_1})\Phi(E_{n_2})\Bigg)+o_P(1)
\end{aligned}\tag{4.25}$$

Dabei ergibt sich die Gleichheit in (4.25) durch Ausklammern mit $\Phi(E_{n_1})\Phi(E_{n_2})$ und Zusammenfassen aller Summanden. Mit der binomischen Formel kann (4.25)

wie folgt dargestellt werden:

$$\frac{N^3}{8\binom{N}{4}}\Bigg(\sum_{1\leq n_1\neq n_2\leq N}\Bigg(-\left(\frac{(n_1\vee n_2)-(n_1\wedge n_2)}{N}-\frac{1}{2}\right)^2 + \frac{1}{4N}-\frac{1}{2N^2}\Bigg)\Phi(E_{n_1})\Phi(E_{n_2})\Bigg)+o_P(1) \tag{4.26}$$

Nun werden in (4.26) die Laufindizes mit $n_1 = n_2$ als 0 addiert. Die Anteile der Summe mit den Summanden $\frac{1}{4N}-\frac{1}{2N^2}$ sind dabei von niedriger Ordnung und können asymptotisch vernachlässigt werden:

$$\begin{aligned}&\frac{N^3}{\binom{N}{4}}\sum_{1\leq n_1<n_2<n_3<n_4\leq N}\left(\frac{1}{4N}-\frac{1}{2N^2}\right)\Phi(E_{n_1})\Phi(E_{n_2})\\ =&\frac{N}{\binom{N}{4}}\left(\frac{1}{4N}-\frac{1}{2N^2}\right)\left(\frac{1}{N}\sum_{n=1}^{N}\Phi(E_n)\right)^2-\frac{N^4}{8\binom{N}{4}}\left(\frac{1}{4N}-\frac{2}{N^2}\right)\end{aligned} \tag{4.27}$$

In (4.27) geht der erste Summand mit der Anwendung des Gesetzes der großen Zahlen (Klenke (2006), S. 108), des Continuous-Mapping-Theorems und des Lemmas von Slutzky gegen 0, da $\mathbb{E}(\Phi(E_n)) = 0$ und $(E_n)_{n\in\mathbb{N}}$ unabhängig, identisch verteilte Zufallsvariablen sind. Der zweite Summand ist ebenso asymptotisch vernachlässigbar, wodurch mit dem Lemma von Slutzky die gesamte Formel stochastisch gegen 0 geht. In (4.26) werden demnach die Ausdrücke in (4.27) durch $o_P(1)$ ersetzt:

$$-\frac{N^3}{8\binom{N}{4}}\sum_{n_1,n_2=1}^{N}\left(\frac{|n_1-n_2|}{N}-\frac{1}{2}\right)^2\Phi(E_{n_1})\Phi(E_{n_2})+\frac{N^4}{4\cdot 8\binom{N}{4}}+o_P(1) \tag{4.28}$$

Mit $\frac{|n_1-n_2|}{N}-\frac{1}{2}=\frac{1}{2}-\int_{-\frac{1}{2}}^{\frac{3}{2}}\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n_1}{N}-t\right)\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n_2}{N}-t\right)dt$ nach Lemma 3.2 schreiben wir (4.28) folgendermaßen um:

$$\begin{aligned}&-\frac{N^3}{8\binom{N}{4}}\sum_{n_1,n_2=1}^{N}\left(\frac{1}{2}-\int_{-\frac{1}{2}}^{\frac{3}{2}}\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n_1}{N}-t\right)\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n_2}{N}-t\right)dt\right)^2\Phi(E_{n_1})\Phi(E_{n_2})\\ &+\frac{N^4}{4\cdot 8\binom{N}{4}}+o_P(1)\end{aligned} \tag{4.29}$$

Multiplizieren wir nun (4.29) mit der binomischen Formel aus und teilen die sich

ergebenen Summanden auf, so erhalten wir folgende Darstellung:

$$= -\frac{N^3}{8\binom{N}{4}} \sum_{n_1,n_2=1}^{N} \left(\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right) dt \right)^2 \Phi(E_{n_1})\Phi(E_{n_2}) \tag{4.30}$$

$$+ \frac{N^3}{8\binom{N}{4}} \sum_{n_1,n_2=1}^{N} \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right) \Phi(E_{n_1})\Phi(E_{n_2})\, dt \tag{4.31}$$

$$- \frac{N^3}{32\binom{N}{4}} \sum_{n_1,n_2=1}^{N} \Phi(E_{n_1})\Phi(E_{n_2}) + \frac{N^4}{4 \cdot 8\binom{N}{4}} + o_P(1) \tag{4.32}$$

Die Untersuchung der Summanden in den Zeilen (4.31) und (4.32) erfolgt vollständig analog zum Satz 3.3. Lediglich der erste Summand in (4.30) erfordert eine erweiterte Idee. Man schreibt das Produkt der Integrale als Doppelintegral durch verschiedene Integrationsvariablen und kann so den Ausdruck in Zeile (4.30) umformulieren:

$$\begin{aligned}
&- \frac{N^3}{8\binom{N}{4}} \sum_{n_1,n_2=1}^{N} \left(\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right) dt \right)^2 \Phi(E_{n_1})\Phi(E_{n_2}) \\
&= -\frac{N^4}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \frac{1}{N} \sum_{n_1,n_2=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - t\right) \\
&\qquad\qquad \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_1}{N} - s\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n_2}{N} - s\right) \Phi(E_{n_1})\Phi(E_{n_2})\, dt\, ds \\
&= -\frac{N^4}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N} - t\right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]}\left(\frac{n}{N} - s\right) \Phi(E_n) \right)^2 dt\, ds
\end{aligned}$$

Das Einsetzen dieser letzten Rechnung in (4.30) und die Darstellung von (4.31) und (4.32) durch Quadrate einer Summe liefern die Behauptung von Satz 4.4. □

Bemerkung 4.5 (Verkürzung der Laufzeit der vollen Vierer-Tiefe).
Ähnlich zur vollen Dreier-Tiefe hat der Autor dieser Arbeit eine Möglichkeit gefunden, wie sich die volle Vierer-Tiefe in linearer statt quadratischer Laufzeit bestimmen lässt, obwohl in der Darstellung ein Doppelintegral vorliegt. Die Darstellung mit verkürzter Laufzeit gilt allerdings nicht mehr exakt, da wir das Vierer-Produkt aus Lemma 4.3 für die Darstellung in Satz 4.4 asymptotisch vernachlässigt haben. Da die Berechnung dieses Vierer-Produkts in der Φ-Darstellung in Korollar 4.2 po-

lynomielle Laufzeit vom Grad 4 hat, würde ihre Berechnung die gesamte Ordnung verschlechtern. Im Kapitel 4.4 wird der Konvergenzfehler im Detail analysiert. Ferner fallen im Beweis von Satz 4.4 einige Nullfolgen weg, die wir in konstanter Laufzeit berechnen können und nehmen sie daher bei der alternativen Berechnung der vollen Vierer-Tiefe mit auf:

$$
\begin{aligned}
N\left(d_S^4(\vartheta, Z) - \frac{1}{8}\right) &= \frac{N^4}{32\binom{N}{4}} - \frac{N^3}{32\binom{N}{4}} \left(\sum_{n=1}^{N} \Phi(E_n)\right)^2 \\
&+ \frac{N^3}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N} - t\right) \Phi(E_n)\right)^2 dt \\
&- \frac{N^4}{8\binom{N}{4}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N} - t\right) \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N} - s\right) \Phi(E_n)\right)^2 dt\, ds \\
&+ \frac{N^2}{8\binom{N}{4}} \left(\frac{1}{4} - \frac{1}{2N}\right) \left(\left(\sum_{n=1}^{N} \Phi(E_n)\right)^2 - N\right) + o_P(1),
\end{aligned} \tag{4.33}
$$

wobei in der letzten Zeile die asymptotisch vernachlässigten Terme aus dem Beweis von Satz 4.4 in (4.27) stehen. Bis auf das Doppelintegral in der dritten Zeile der Formel (4.33) können die Ausdrücke wie in Bemerkung 3.4 in linearer Laufzeit durch die dort angegebenen Vektoren S und D berechnet werden. Doppelintegrale sind allgemein in quadratischer Laufzeit berechenbar, da zwei Integrationsvariablen vorliegen. Im vorliegenden Fall werden wir dennoch eine Möglichkeit herleiten können, in linearer Laufzeit das Doppelintegral zu berechnen. Analog zur Bemerkung 3.4 fassen wir die Summe im Doppelintegral mit den Indikatorfunktionen zusammen:

$$
\begin{aligned}
-\frac{1}{2} < \frac{n}{N} - t \leq \frac{1}{2}, \\
-\frac{1}{2} < \frac{n}{N} - s \leq \frac{1}{2}
\end{aligned}
$$

und erhalten eine untere und obere Schranke für den Laufindex n:

$$
\begin{aligned}
N\left(t - \frac{1}{2}\right) \leq n \leq N\left(t + \frac{1}{2}\right), \\
N\left(s - \frac{1}{2}\right) \leq n \leq N\left(s + \frac{1}{2}\right).
\end{aligned}
$$

Wir können eine gemeinsame untere und obere Schranke für n in Abhängigkeit von t und s aus beiden Ungleichungen zusammenfassen, müssen aber als Zusatzbedingung fordern, dass die untere Schranke jeweils einer Ungleichung kleiner als die obere Schranke der anderen Ungleichung ist:

$$N\left(\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\right)\leq n\leq N\left(\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\right),$$
$$N\left(t-\frac{1}{2}\right)\leq N\left(s+\frac{1}{2}\right),$$
$$N\left(s-\frac{1}{2}\right)\leq N\left(t+\frac{1}{2}\right).$$

Die letzten beiden Ungleichungen können wir als Bedingung zusammenfassen, dass $|t-s|\leq 1$ gilt. Die erste Bedingung kann nach Abrunden der unteren und oberen Schranke auf die Summe angewendet werden:

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-t\right)\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-s\right)\Phi(E_n)\right)^2 dt\,ds\\
&=\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}}\mathbb{1}\{|t-s|\leq 1\}\left(\frac{1}{\sqrt{N}}\sum_{n=\left\lfloor N\left(\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 0\right)\right\rfloor+1}^{\left\lfloor N\left(\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 1\right)\right\rfloor}\Phi(E_n)\right)^2 dt\,ds\\
&=\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}}\mathbb{1}\{|t-s|\leq 1\}\left(S^N_{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 1}-S^N_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 0}\right)^2 dt\,ds
\end{aligned}$$

Dabei ist $S^N_t=\frac{1}{\sqrt{N}}\sum_{n=1}^{\lfloor Nt\rfloor}\Phi(E_n)$ für $t\in[0,1]$. Den Integranden kann man besser durch eine Fallunterscheidung verstehen:

$$\left(S^N_{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 1}-S^N_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 0}\right)\mathbb{1}\{|t-s|\leq 1\} \tag{4.34}$$

$$=\begin{cases}
S^N_{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)}, & \text{für } (t,s)\in\left[-\frac{1}{2},\frac{1}{2}\right)^2\\
\left(S^N_{t+\frac{1}{2}}-S_{s-\frac{1}{2}}\right)\mathbb{1}\{|t-s|\leq 1\}, & \text{für } (t,s)\in\left[-\frac{1}{2},\frac{1}{2}\right)\times\left[\frac{1}{2},\frac{3}{2}\right]\\
\left(S^N_{s+\frac{1}{2}}-S_{t-\frac{1}{2}}\right)\mathbb{1}\{|t-s|\leq 1\}, & \text{für } (t,s)\in\left[\frac{1}{2},\frac{3}{2}\right]\times\left[-\frac{1}{2},\frac{1}{2}\right)\\
S^N_1-S_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)}, & \text{für } (t,s)\in\left[\frac{1}{2},\frac{3}{2}\right]^2
\end{cases} \tag{4.35}$$

Damit ergeben sich vier Fälle, über die wir vier verschiedene Integrale auf vier ver-

schiedenen Integrationsgebieten bestimmen können, um das gesamte Doppelintegral zu berechnen. Das Integral über eine zweidimensionalen Treppenfunktionen in (t, s) können wir durch eine Matrix in einer Doppelsumme visualisieren. Insbesondere lassen sich die Summanden dieser Doppelsumme derartig zusammenfassen, dass wir einfache Summen erhalten, wodurch das Integral in linearer Laufzeit berechenbar ist. Dazu definieren wir $s_k := S^N_{\frac{k}{N}}$ für $k = 1, ..., N$. Beginnen wir mit dem Fall für $(t, s) \in \left[-\frac{1}{2}, \frac{1}{2}\right)^2$ und definieren die Matrix D_1:

$$D_1 := (d_1(l,k)^2)_{\substack{1 \le l \le N \\ 1 \le k \le N}}$$
$$\text{mit } d_1(l,k) = s_{l \wedge k} \text{ für } (l,k) \in \{1, ..., N\}^2.$$

Wir können die Matrix D_1 wie folgt darstellen:

$$D_1 = \begin{pmatrix} s_1^2 & s_1^2 & \cdots & s_1^2 \\ s_1^2 & s_2^2 & \cdots & s_2^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^2 & s_2^2 & \cdots & s_N^2 \end{pmatrix}$$

Die Doppelsumme über alle Treppenstufen $\frac{1}{N^2} \sum_{l,k=1}^N d_1(l,k)^2$ zur Berechnung des Integrals kann als eine einfache Summe vereinfacht werden, da s^2_{N-j+1} genau $(2j-1)$-mal vorkommt. Somit können wir schreiben:

$$\begin{aligned} &\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(S^N_{\left(t+\frac{1}{2}\right) \wedge \left(s+\frac{1}{2}\right)} \right)^2 dt\, ds \\ =&\frac{1}{N^3} \sum_{l,k=1}^{N} d_1(l,k)^2 = \frac{1}{N^3} \sum_{j=1}^{N} (2j-1) s^2_{N-j+1} \end{aligned}$$

Für den Fall $(t, s) \in \left[\frac{1}{2}, \frac{3}{2}\right]^2$ definieren wir die Matrix D_2:

$$D_2 := (d_2(l,k)^2)_{\substack{1 \le l \le N \\ 1 \le k \le N}}$$
$$\text{mit } d_2(l,k) = (s_N - s_{l \vee k}) \text{ für } (l,k) \in \{1, ..., N\}^2.$$

Wir können die Matrix D_2 wie folgt darstellen:

$$D_2 = \begin{pmatrix} (s_N - s_1)^2 & \cdots & (s_N - s_{N-2})^2 & (s_N - s_{N-1})^2 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ (s_N - s_{N-2})^2 & \cdots & (s_N - s_{N-2})^2 & (s_N - s_{N-1})^2 & 0 \\ (s_N - s_{N-1})^2 & \cdots & (s_N - s_{N-1})^2 & (s_N - s_{N-1})^2 & 0 \\ 0 & \cdots & 0 & 0 & 0 \end{pmatrix}$$

Analog wie im ersten Fall können wir das Doppelintegral über die Treppenfunktion in (t, s) durch eine einfache Summe statt über folgende Doppelsumme $\frac{1}{N^2} \sum_{l,k=1}^{N} d_2(l,k)^2$ darstellen, wenn geschickt aufsummiert wird. Hier taucht der Summand $(s_N - s_j)^2$ insgesamt $(2j-1)$-mal auf für $j = 1, ..., N-1$, wodurch sich ergibt:

$$\begin{aligned} &\int_{\frac{1}{2}}^{\frac{3}{2}} \int_{\frac{1}{2}}^{\frac{3}{2}} \left(S_1^N - S^N_{\left(t-\frac{1}{2}\right) \vee \left(s-\frac{1}{2}\right)} \right)^2 dt\, ds \\ =&\frac{1}{N^3} \sum_{l,k=1}^{N} d_2(l,k)^2 = \frac{1}{N^3} \sum_{j=1}^{N} (2j-1)(s_N - s_j)^2 \end{aligned}$$

Für den Fall $(t, s) \in \left[-\frac{1}{2}, \frac{1}{2}\right) \times \left[\frac{1}{2}, \frac{3}{2}\right]$ definieren wir die Matrix D_3:

$$\begin{aligned} D_3 &:= (d_3(l,k)^2)_{\substack{1 \le l \le N \\ 1 \le k \le N}} \\ \text{mit } d_3(l,k) &= (s_l - s_k) \mathbb{1}\{l \ge k\} \text{ für } (l,k) \in \{1, ..., N\}^2. \end{aligned}$$

Wir können die Matrix D_3 wie folgt darstellen:

$$D_3 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ (s_2 - s_1)^2 & 0 & \cdots & 0 & 0 \\ (s_3 - s_1)^2 & (s_3 - s_2)^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ (s_N - s_1)^2 & (s_N - s_2)^2 & \cdots & (s_N - s_{N-1})^2 & 0 \end{pmatrix}$$

Erneut stellen wir das Doppelintegral über die Treppenfunktion in (t, s) durch eine einfache Summe statt über folgende Doppelsumme $\frac{1}{N^2} \sum_{l,k=1}^{N} d_3(l,k)^2$ dar. Multipli-

ziert man alle Einträge der Matrix mit $l \geq k$ aus, liefert das:

$$d_3(l,k)^2 = (s_l - s_k)^2 = s_l^2 + s_k^2 - 2s_l s_k.$$

Summiert man über alle diese Einträge auf, so ergibt sich

$$\begin{aligned}\sum_{l,k=1}^{N} d_3(l,k)^2 &= (N-1)\sum_{l=1}^{N} s_l^2 - 2\sum_{\substack{l,k=1\\ l>k}}^{N} s_l s_k \\ &= N\sum_{l=1}^{N} s_l^2 - 2\sum_{\substack{l,k=1\\ l>k}}^{N} s_l s_k - \sum_{l=1}^{N} s_l^2 \\ &= N\sum_{l=1}^{N} s_l^2 - \left(\sum_{l=1}^{N} s_l\right)^2. \end{aligned} \tag{4.36}$$

Aus (4.36) ergibt sich die Möglichkeit das Doppelintegral über die Treppenfunktion in linearer Laufzeit zu berechnen:

$$\begin{aligned}&\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\}\left(S^N_{t+\frac{1}{2}} - S^N_{s-\frac{1}{2}}\right)^2 dt\, ds \\ =&\frac{1}{N^3}\sum_{l,k=1}^{N} d_3(l,k)^2 = \frac{1}{N^3}\left(N\sum_{l=1}^{N} s_l^2 - \left(\sum_{l=1}^{N} s_l\right)^2\right)\end{aligned}$$

Ferner ist der Fall für $(t,s) \in \left[\frac{1}{2},\frac{3}{2}\right] \times \left[-\frac{1}{2},\frac{1}{2}\right)$ symmetrisch zum vorherigen Fall, da

$$(s_l - s_k)^2 = (s_k - s_l)^2 \text{ gilt,}$$

womit wir auch folgendes Integral bestimmen können:

$$\begin{aligned}&\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\}\left(S^N_{s+\frac{1}{2}} - S^N_{t-\frac{1}{2}}\right)^2 dt\, ds \\ =&\frac{1}{N^3}\left(N\sum_{l=1}^{N} s_l^2 - \left(\sum_{l=1}^{N} s_l\right)^2\right)\end{aligned}$$

Somit können wir numerisch das zu untersuchende Doppelintegral in einer linearen

Laufzeit berechnen:

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]} \left(\frac{n}{N} - t \right) \mathbb{1}_{(-\frac{1}{2},\frac{1}{2}]} \left(\frac{n}{N} - s \right) \Phi(E_n) \right)^2 dt\, ds \\
=&\frac{1}{N^3} \left(\sum_{j=1}^{N} (2j-1) s_{N-j}^2 + \sum_{j=1}^{N} (2j-1)(s_N - s_j)^2 + 2N \sum_{l=1}^{N} s_l^2 - 2 \left(\sum_{l=1}^{N} s_l \right)^2 \right)
\end{aligned}$$

Für die volle Vierer-Tiefe gilt folgende asymptotische Darstellung:

$$\begin{aligned}
&N \left(d_S^4(\vartheta, Z) - \frac{1}{8} \right) \\
=&\frac{3N^3}{(N-1)(N-2)(N-3)} \left(\frac{1}{4} - \frac{1}{4N} s_N^2 + \frac{1}{N^2} \sum_{j=1}^{2N} d_j^2 - \frac{1}{N^3} \sum_{j=1}^{N} (2j-1) s_{N-j}^2 \right. \\
&- \frac{1}{N^3} \sum_{j=1}^{N} (2j-1)(s_N - s_j)^2 - \frac{2}{N^2} \sum_{j=1}^{N} s_j^2 + \frac{2}{N^3} \left(\sum_{j=1}^{N} s_j \right)^2 \\
&\left. + \left(\frac{1}{4N^2} - \frac{1}{2N^3} \right) (s_N^2 - N) \right) + o_P(1).
\end{aligned}$$

$o_P(1)$ entspricht dabei der stochastischen Nullfolge aus Lemma 4.3. Hierbei kann mit der Annahme (2.3) ein beliebiges $\vartheta \in \Theta$ verwendet werden und in der obigen Rechnung E_n durch $\mathrm{res}(\vartheta, Z_n)$ ersetzt werden. □

Aus Satz 4.4 können wir die asymptotische Verteilung der vollen Vierer-Tiefe herleiten und gehen dabei analog zu Satz 3.10 vor. Insbesondere ergeben sich beim Nachweis der Stetigkeit des Funktionals im Satz 4.4 gleiche Rechnungen wie im Beweis von Satz 3.10, da die ersten beiden Koordinaten dieses Funktionals den Koordinaten des Funktionals im Beweis von Satz 3.10 entsprechen. Es ergibt sich im nachfolgenden Beweis von Satz 4.6 eine dritte Koordinate, die nur zusätzlich betrachtet werden braucht. Die Asymptotik der vollen Vierer-Tiefe wurde vom Autor dieser Arbeit selbst gefunden und bewiesen. Hierbei dankt der Autor für die Unterstützung von Dr. K. Leckey, der einen Fehler in einer vorherigen Version des Beweises von Satz 4.6 korrigieren konnte.

Satz 4.6 (Asymptotische Verteilung der vollen Vierer-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 gilt

$$N\left(d_S^4(\vartheta^*, Z) - \frac{1}{8}\right) \xrightarrow[N\to\infty]{\mathcal{D}} -\frac{3}{4}B_1^2 + 3\int_{-\frac{1}{2}}^{\frac{3}{2}} \left(B_{(t+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee 0}\right)^2 dt$$
$$- 3\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \le 1\} \left(B_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee(s-\frac{1}{2})\vee 0}\right)^2 dt\, ds + \frac{3}{4},$$

wobei $(B_t)_{t\in[0,1]}$ die Brownsche Bewegung sei.

Beweis. Wir gehen ähnlich wie in Satz 3.10 vor und stellen die linke Seite in der behaupteten Konvergenz als die Auswertung der zeitskalierte Irrfahrt $(S_t^N)_{t\in[0,1]}$ in einem passenden Funktional Ψ dar:

$$\Psi : D[0,1] \to \mathbb{R}^3, \Psi(f) = \Big(f(1)^2, \int_{-\frac{1}{2}}^{\frac{3}{2}} \left(f\left(\left(t+\tfrac{1}{2}\right)\wedge 1\right) - f\left(\left(t-\tfrac{1}{2}\right)\vee 0\right)\right)^2 dt,$$
$$\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\} \left(f\left(\left(t+\tfrac{1}{2}\right)\wedge\left(s+\tfrac{1}{2}\right)\wedge 1\right) - f\left(\left(t-\tfrac{1}{2}\right)\vee\left(s-\tfrac{1}{2}\right)\vee 0\right)\right)^2 dt\, ds\Big).$$

Die $\mathcal{D}-\mathcal{B}(\mathbb{R})$-Messbarkeit der ersten beiden Koordinaten von Ψ ist in Satz 3.10 gezeigt. Für $\mathcal{D}-\mathcal{B}(\mathbb{R})$-Messbarkeit in der dritten Koordinaten stellen wir das Doppelintegral zunächst anders dar, indem wir das Gebiet $\left[-\frac{1}{2}, \frac{3}{2}\right]^2$ in vier Teilgebiete zerlegen, siehe auch (4.35) in Bemerkung 4.5:

$$\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\} \left(f\left(\left(t+\tfrac{1}{2}\right)\wedge\left(s+\tfrac{1}{2}\right)\wedge 1\right) - f\left(\left(t-\tfrac{1}{2}\right)\vee\left(s-\tfrac{1}{2}\right)\vee 0\right)\right)^2 dt\, ds$$
$$= \int_0^1\int_0^1 \left(f(t\wedge s) - f(0)\right)^2 dt\, ds + \int_0^1\int_0^1 \left(f(1) - f(t\vee s)\right)^2 dt\, ds$$
$$+2\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\} \left(f\left(t-\tfrac{1}{2}\right) - f\left(s+\tfrac{1}{2}\right)\right)^2 dt\, ds$$

Wir können die Doppelintegrale durch Riemann-Doppelsummen darstellen und als in der kanonischen Projektion Π_t in $D[0,1]$ ausgewertete Summanden schreiben. Für das erste Integral gilt:

$$\int_0^1\int_0^1 \left(f(t\wedge s) - f(0)\right)^2 dt\, ds = \lim_{n\to\infty} \frac{1}{n^2}\sum_{s,t=1}^{n} \left(f\left(\tfrac{t}{N}\wedge\tfrac{s}{N}\right) - f(0)\right)^2$$

$$= \lim_{n\to\infty} \frac{1}{n^2} \sum_{s,t=1}^{n} \left(\Pi_{\frac{t}{N}\wedge\frac{s}{N}}(f) - \Pi_0(f) \right)^2 . \tag{4.37}$$

Damit erhalten wir eine abzählbare Summe von $\mathcal{D}-\mathcal{B}(\mathbb{R})$-messbaren Abbildungen mit punktweise existierendem Limes, womit das betrachtete Doppelintegral $\mathcal{D}-\mathcal{B}(\mathbb{R})$-messbar ist. Für das zweite Doppelintegral kann analog vorgegangen werden. Für das dritte Integral zeigen wir die Messbarkeit genauso und schränken die Summation durch die vorgegebene Indikatorfunktion ein:

$$\begin{aligned} &2 \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \le 1\} \left(f\left(t-\tfrac{1}{2}\right) - f\left(s+\tfrac{1}{2}\right) \right)^2 dt\, ds \\ &= \lim_{n\to\infty} \frac{1}{n^2} \sum_{\substack{s,t=1 \\ t\le s}}^{n} \left(f\left(\tfrac{t}{n}\right) - f\left(\tfrac{s}{n}\right) \right)^2 = \lim_{n\to\infty} \frac{1}{n^2} \sum_{\substack{s,t=1 \\ t\le s}}^{n} \left(\Pi_{\frac{t}{n}}(f) - \Pi_{\frac{s}{n}}(f) \right)^2 \end{aligned} \tag{4.38}$$

und erhalten erneut eine abzählbare Summe von $\mathcal{D}-\mathcal{B}(\mathbb{R})$-messbaren Abbildungen mit punktweise existierendem Limes. Damit ist Ψ eine $D[0,1]-\mathcal{B}(\mathbb{R}^3)$-messbare Abbildung und kann für das Continuous-Mapping-Theorem mit dem Satz von Donsker verwendet werden, wenn die Stetigkeit auf $C[0,1]$ gezeigt wird. Für die ersten beiden Koordinaten ist der Beweis der Stetigkeit auf $C[0,1]$ bereits in Satz 3.10 durchgeführt worden. Für die dritte Koordinate schreiben wir zunächst wie in Bemerkung 4.5 in (4.34), den Integranden des Doppelintegrals aus Satz 4.4 um:

$$\begin{aligned} &\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-t\right) \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}\left(\frac{n}{N}-s\right) \Phi(E_n) \\ &= \left(S^N_{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 1} - S^N_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 0} \right) \mathbb{1}\{|t-s| \le 1\} \end{aligned}$$

Schreiben wir zusätzlich die Integrale dazu, erhalten wir genau die dritte Koordinate Ψ_3 der Abbildung Ψ:

$$\Psi_3(S^N_\bullet) = \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \le 1\} \left(S^N_{\left(t+\frac{1}{2}\right)\wedge\left(s+\frac{1}{2}\right)\wedge 1} - S^N_{\left(t-\frac{1}{2}\right)\vee\left(s-\frac{1}{2}\right)\vee 0} \right)^2 dt\, ds$$

Wir können also das Doppelintegral bei der Darstellung der vollen Vierer-Tiefe in Satz 4.4 durch S^N_t ausgewertet in Ψ_3 identifizieren. Es bleibt die Stetigkeit von Ψ_3

auf $C[0,1]$ zu zeigen. Sei $f_n \xrightarrow[n\to\infty]{d_D} f$ für eine Funktionenfolge $(f_n)_{n\in\mathbb{N}}$ in $D[0,1]$ mit $f \in C[0,1]$. Hierbei können wir wegen Lemma 3.9 auf die uniforme Konvergenz zurückgreifen:

$$
\begin{aligned}
&|\Psi_3(f_n) - \Psi_3(f)| \\
\leq &\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} | \left(f_n\left(\left(t+\tfrac{1}{2}\right) \wedge \left(s+\tfrac{1}{2}\right)\right) - f_n(0)\right)^2 \\
&\qquad - \left(f\left(\left(t+\tfrac{1}{2}\right) \wedge \left(s+\tfrac{1}{2}\right)\right) - f(0)\right)^2 | \, dt\, ds \qquad (4.39) \\
+2&\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\} | \left(f_n\left(s+\tfrac{1}{2}\right) - f_n\left(t-\tfrac{1}{2}\right)\right)^2 \\
&\qquad - \left(f\left(s+\tfrac{1}{2}\right) - f\left(t-\tfrac{1}{2}\right)\right)^2 | \, dt\, ds \qquad (4.40) \\
+&\int_{\frac{1}{2}}^{\frac{3}{2}} \int_{\frac{1}{2}}^{\frac{3}{2}} | \left(f_n(1) - f_n\left(\left(t-\tfrac{1}{2}\right) \vee \left(s-\tfrac{1}{2}\right)\right)\right)^2 \\
&\qquad - \left(f(1) - f\left(\left(t-\tfrac{1}{2}\right) \vee \left(s-\tfrac{1}{2}\right)\right)\right)^2 | \, dt\, ds \qquad (4.41)
\end{aligned}
$$

Wir untersuchen nun (4.39)-(4.41) separat. Zunächst verschieben wir in (4.39) den Integrationsbereiche auf $[0,1]$ und trennen die Fälle für $t \leq s$ und $t > s$. Dabei beachte man die Verwendung der binomischen Formel und Dreiecks-Ungleichung:

$$
\begin{aligned}
&\int_0^1 \int_0^1 \left|(f_n(t \wedge s) - f_n(0))^2 - (f(t \wedge s) - f(0))^2\right| \, dt\, ds \\
\leq &\int_0^1 \int_s^1 |f_n(s)^2 - f(s)^2| \, dt\, ds + \int_0^1 \int_0^s |f_n(t)^2 - f(t)^2| \, dt\, ds + |f_n(0)^2 - f(0)^2| \qquad (4.42) \\
+2&\int_0^1 \int_s^1 |f_n(0)f_n(s) - f(0)f(s)| \, dt\, ds + 2\int_0^1 \int_0^s |f_n(0)f_n(t) - f(0)f(t)| \, dt\, ds \qquad (4.43)
\end{aligned}
$$

Wir betrachten nun (4.42) und (4.43) separat. (4.42) lässt sich wie folgt abschätzen, da über ein rechtwinkliges Dreieck mit einem Flächeninhalt von $\frac{1}{2}$ integriert wird:

$$
\begin{aligned}
&\int_0^1 \int_s^1 |f_n(s)^2 - f(s)^2| \, dt\, ds + \int_0^1 \int_0^s |f_n(t)^2 - f(t)^2| \, dt\, ds + |f_n(0)^2 - f(0)^2| \\
\leq &\frac{1}{2} \sup_{s\in[0,1]} |f_n(s)^2 - f(s)^2| + \frac{1}{2} \sup_{t\in[0,1]} |f_n(t)^2 - f(t)^2| + \sup_{t\in[0,1]} |f_n(t)^2 - f(t)^2|. \qquad (4.44)
\end{aligned}
$$

Alle Summanden in (4.44) konvergieren für $n \to \infty$ gegen 0, was wie im Beweis von Satz 3.10 gezeigt werden kann. Zu (4.43) betrachten wir das erste Integral. Wir führen eine Nulladdition mit $f_n(0)f(s)$ durch und erhalten:

$$\begin{aligned}
&2\int_0^1 \int_s^1 |f_n(0)f_n(s) - f_n(0)f(s) + f_n(0)f(s) - f(0)f(s)|\, dt\, ds \\
\leq &|f_n(0)| \sup_{s\in[0,1]} |f_n(s) - f(s)| + |f_n(0) - f(0)| \sup_{s\in[0,1]} |f(s)| \\
\leq &\sup_{s\in[0,1]} |f_n(s)| \sup_{s\in[0,1]} |f_n(s) - f(s)| + \sup_{s\in[0,1]} |f_n(s) - f(s)| \sup_{s\in[0,1]} |f(s)| \qquad (4.45)
\end{aligned}$$

Für die Abschätzung in (4.45) wird genutzt, dass über ein rechtwinkliges Dreieck mit Flächeninhalt $\frac{1}{2}$ integriert wird und die Funktionen auf diesem Integrationsgebiet über ihr Supremum abgeschätzt werden können. Die einzelnen Ausdrücke konvergieren für $n \to \infty$ gegen 0 nach Voraussetzung, da die entsprechenden Vorfaktoren $\sup_{s\in[0,1]} |f_n(s)|$ bzw. $\sup_{s\in[0,1]} |f(s)|$ beschränkt sind. Für das zweite Integral in (4.43) zeigt man analog mit einer Nulladdition von $f_n(0)f(t)$, dass es für $n \to \infty$ gegen 0 konvergiert. Damit ist gezeigt, dass die Terme in (4.39) für $n \to \infty$ gegen 0 gehen. Für die Untersuchung von (4.40) gehen wir analog zu oben vor und verwenden die Dreiecks-Ungleichung:

$$\begin{aligned}
&2\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\} \Big| \left(f_n\left(s+\tfrac{1}{2}\right) - f_n\left(t-\tfrac{1}{2}\right)\right)^2 \\
&\qquad\qquad - \left(f\left(s+\tfrac{1}{2}\right) - f\left(t-\tfrac{1}{2}\right)\right)^2 \Big|\, dt\, ds \\
\leq &2\int_0^1 \int_0^1 \mathbb{1}\{|t-s| \leq 1\} \left(|f_n(s)^2 - f(s)^2| + |f_n(t)^2 - f(t)^2|\right)\, dt\, ds \qquad (4.46) \\
&+4\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\} \Big| f_n\left(s+\tfrac{1}{2}\right) f_n\left(t-\tfrac{1}{2}\right) \\
&\qquad\qquad - f\left(s+\tfrac{1}{2}\right) f\left(t-\tfrac{1}{2}\right) \Big|\, dt\, ds. \qquad (4.47)
\end{aligned}$$

Analog wie bei der Abschätzung von (4.44) kann gezeigt werden, dass (4.46) für $n \to \infty$ gegen 0 konvergiert, da jeweils wegen der Indikatorfunktion über ein rechtseitiges Dreieck mit Flächeninhalt $\frac{1}{2}$ integriert wird. (4.47) hat nach einer Nulladdition

folgende Gestalt:

$$
\begin{aligned}
&4\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\leq 1\}|f_n\left(s+\tfrac{1}{2}\right) f_n\left(t-\tfrac{1}{2}\right) - f\left(s+\tfrac{1}{2}\right) f_n\left(t-\tfrac{1}{2}\right) \\
&+ f\left(s+\tfrac{1}{2}\right) f_n\left(t-\tfrac{1}{2}\right) - f\left(s+\tfrac{1}{2}\right) f\left(t-\tfrac{1}{2}\right)| \, dt\, ds \\
\leq& 2 \sup_{s\in[0,1]} |f_n(s)-f(s)| \sup_{t\in[0,1]} |f_n(t)| + 2 \sup_{s\in[0,1]} |f(s)| \sup_{t\in[0,1]} |f_n(t)-f(t)|. \qquad (4.48)
\end{aligned}
$$

Dabei ist wieder zu beachten, dass durch die Indikatorfunktion über ein rechtwinkliges Dreieck mit Flächeninhalt $\frac{1}{2}$ integriert wird. In (4.48) konvergieren die Ausdrücke mit gleicher Argumentation zu oben für $n \to \infty$ gegen 0. Damit ist gezeigt, dass (4.40) für $n \to \infty$ gegen 0 geht. Um zu zeigen, dass (4.41) für $n \to \infty$ gegen 0 konvergiert, gehen wir analog wie in (4.39) vor und trennen die Integrationsbereiche für die Fälle $t \leq s$ und $t > s$:

$$
\begin{aligned}
&\int_0^1\int_0^1 |\left(f_n(1)-f_n\left(t\vee s\right)\right)^2 - \left(f(1)-f\left(t\vee s\right)\right)^2| \, dt\, ds \\
\leq& \int_0^1\int_s^1 |f_n(t)^2-f(t)^2| \, dt\, ds + \int_0^1\int_0^s |f_n(s)^2-f(s)^2| \, dt\, ds + |f_n(1)^2-f(1)^2| \qquad (4.49) \\
&+ 2\int_0^1\int_s^1 |f_n(1)f_n(t)-f(1)f(t)| \, dt\, ds + 2\int_0^1\int_0^s |f_n(1)f_n(s)-f(1)f(s)| \, dt\, ds \qquad (4.50)
\end{aligned}
$$

Analog wie bei (4.42) kann man zeigen, dass die Ausdrücke in (4.49) für $n \to \infty$ gegen 0 konvergieren. In (4.50) kann mit einer Nulladdition durch $f_n(1)f_n(t)$ und $f_n(1)f_n(s)$ analog wie in (4.43) die Konvergenz gegen 0 für $n \to \infty$ gezeigt werden. Damit ist die Stetigkeit von Ψ auf $C[0,1]$ gezeigt.
Nun betrachten wir folgende $(\mathbb{R}^3-\mathbb{R})$-stetige Abbildung A:

$$
A : \mathbb{R}^3 \to \mathbb{R}, A(x_1,x_2,x_3) := -\frac{3}{4}x_1 + 3x_2 - 3x_3 + \frac{3}{4}.
$$

Damit ist $A \circ \Psi$ eine $D[0,1]-\mathbb{R}$-stetige Abbildung auf der Menge $C[0,1]$ und wir

können das Continuous-Mapping-Theorem auf $A \circ \Psi$ auf Satz 4.4 anwenden:

$$N\left(d_S^4(\vartheta^*, Z) - \frac{1}{8}\right) = \frac{N^3}{(N-1)(N-2)(N-3)} A(\Psi(S_\bullet^N)) + o_P(1)$$
$$\xrightarrow[N\to\infty]{\mathcal{D}} A(\Psi(B_\bullet)) = -\frac{3}{4}B_1^2 + 3\int_{-\frac{1}{2}}^{\frac{3}{2}} \left(B_{(t+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee 0}\right)^2 dt$$
$$- 3\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \le 1\} \left(B_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee(s-\frac{1}{2})\vee 0}\right)^2 dt\, ds + \frac{3}{4}, \quad (4.51)$$

wobei $\frac{N^3}{(N-1)(N-2)(N-3)} \xrightarrow[N\to\infty]{} 1$ mit dem Lemma von Slutzky (Bickel und Doksum (1997), S. 461) verwendet wird. □

Korollar 4.7 (Testverfahren beruhend auf der vollen Vierer-Tiefe).
Für das gegebene Regressionsmodell in Definition 2.1 und das Hypothesenpaar $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$ hält das Testverfahren mit folgender Entscheidungsregel asymptotisch das Signifikanzniveau α ein:

$$\textit{Man verwerfe } H_0, \textit{ falls } \sup_{\vartheta\in\Theta_0} \left(N\left(d_S^4(\vartheta, z) - \frac{1}{8}\right)\right) < q_\alpha^{(4)},$$

wobei $q_\alpha^{(4)}$ das α-Quantil einer Zufallsvariable der Form $A \circ \Psi(B_\bullet)$ ist mit einer Brownschen Bewegung $(B_t)_{t\in[0,1]}$ und $A \circ \Psi$ als das Funktional, das im Beweis von Satz 4.6 in (4.51) verwendet wird.

Beweis. Analog zum Korollar 2.7 mit Verwendung von Satz 4.6 □

4.3 Berechnung der Quantile der asymptotischen Verteilung der vollen Vierer-Tiefe

Die Berechnung der Quantile der asymptotischen Verteilung der vollen Vierer-Tiefe beruhen wie im Kapitel 3.4 auf die Verwendung des Satzes von Glivenko-Cantelli durch eine numerische Simulation der Quantile. Dabei gehen wir analog vor und bestimmen 20.000 unabhängig, identische Wiederholungen von Brownschen Bewegungen mit Feinheit $T = 100$. Bis auf den Term mit dem Doppelintegral können wir wie im Kapitel 3.4 die Ausdrücke bestimmen.

Wir vereinfachen zunächst das Doppelintegral:

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\}\left(B_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee(s-\frac{1}{2})\vee 0}\right)^2 dt\,ds\\
=&\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} B^2_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})}\,dt\,ds + \int_{\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{1}\{|t-s|\le 1\}(B_{t-\frac{1}{2}} - B_{s+\frac{1}{2}})^2\,dt\,ds\\
+&\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\}(B_{s-\frac{1}{2}} - B_{t+\frac{1}{2}})^2\,dt\,ds + \int_{\frac{1}{2}}^{\frac{3}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} (B_1 - B_{(t+\frac{1}{2})\vee(s+\frac{1}{2})})^2\,dt\,ds.
\end{aligned}$$

Wir vereinfachen die Integrale durch Verschiebung der Integrationsbereiche oder durch Zusammenfassung bestimmter Integrale. Das erste Integral lässt sich folgendermaßen vereinfachen:

$$\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} B^2_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})}\,dt\,ds = \int_0^1\int_0^1 B^2_{t\wedge s}\,dt\,ds.$$

Für das zweite und dritte Integral gilt aus Symmetriegründen:

$$\begin{aligned}
&\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{1}\{|t-s|\le 1\}(B_{t-\frac{1}{2}} - B_{s+\frac{1}{2}})^2\,dt\,ds\\
&+\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s|\le 1\}(B_{s-\frac{1}{2}} - B_{t+\frac{1}{2}})^2\,dt\,ds\\
=&2\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{1}\{|t-s|\le 1\}(B_{t-\frac{1}{2}} - B_{s+\frac{1}{2}})^2\,dt\,ds
\end{aligned}$$

Das vierte Integral wird auch auf den Bereich $[0,1]$ verschoben und zusammengefasst:

$$\begin{aligned}
&\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{\frac{1}{2}}^{\frac{3}{2}} (B_1 - B_{(t+\frac{1}{2})\vee(s+\frac{1}{2})})^2\,dt\,ds = \int_0^1\int_0^1 (B_1 - B_{t\vee s})^2\,dt\,ds\\
=&\int_0^1\int_0^1 (B_1^2 - 2B_1B_{t\vee s} + B^2_{t\vee s})\,dt\,ds\\
=&B_1^2 - 2B_1\int_0^1\int_0^1 B_{t\vee s}\,dt\,ds + \int_0^1\int_0^1 B^2_{t\vee s}\,dt\,ds.
\end{aligned}$$

Zusammenfassend erhalten wir:

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\} \left(B_{(t+\frac{1}{2})\wedge(s+\frac{1}{2})\wedge 1} - B_{(t-\frac{1}{2})\vee(s-\frac{1}{2})\vee 0} \right)^2 dt\, ds \\
&= \int_0^1 \int_0^1 B_{t\wedge s}^2 \, dt\, ds + 2 \int_{\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{1}\{|t-s| \leq 1\} (B_{t-\frac{1}{2}} - B_{s+\frac{1}{2}})^2 \, dt\, ds \\
&\quad + B_1^2 - 2B_1 \int_0^1 \int_0^1 B_{t\vee s} \, dt\, ds + \int_0^1 \int_0^1 B_{t\vee s}^2 \, dt\, ds.
\end{aligned}$$

Wir können wie im Beweis von Satz 4.6 in (4.37) und in (4.38) bei der numerischen Berechnung der Doppelintegrale vorgehen. Diese Vorgehensweise wird in Davis und Rabinowitz (1975), S. 267f. gerechtfertigt:

- $\displaystyle\int_0^1 \int_0^1 B_{t\wedge s}^2 \, dt\, ds = \lim_{N\to\infty} \frac{1}{N^2} \sum_{t,s=1}^{N} (B_{\frac{t}{N}\wedge\frac{s}{N}})^2$
- $\displaystyle 2 \int_{-\frac{1}{2}}^{\frac{3}{2}} \int_{-\frac{1}{2}}^{\frac{3}{2}} \mathbb{1}\{|t-s| \leq 1\} \left(B_{t-\frac{1}{2}} - B_{s+\frac{1}{2}} \right)^2 dt\, ds = \lim_{N\to\infty} \frac{1}{N^2} \sum_{\substack{s,t=1 \\ t\leq s}}^{N} \left(B_{\frac{t}{N}} - B_{\frac{s}{N}} \right)^2$
- $\displaystyle\int_0^1 \int_0^1 B_{t\vee s}^2 \, dt\, ds = \lim_{N\to\infty} \frac{1}{N^2} \sum_{t,s=1}^{N} (B_{\frac{t}{N}\vee\frac{s}{N}})^2$
- $\displaystyle\int_0^1 \int_0^1 B_{t\vee s} \, dt\, ds = \lim_{N\to\infty} \frac{1}{N^2} \sum_{t,s=1}^{N} B_{\frac{t}{N}\vee\frac{s}{N}}.$

Die Darstellungen der Doppelintegrale als doppelte Riemann-Summen können ferner vollständig analog zu Bemerkung 4.5 in einfache Riemann-Summen umgeschrieben werden, um die Laufzeiten zu verringern. Einige der berechneten Quantile für häufig verwendete Signifikanzniveaus sind in Tabelle 6 gelistet. Das arithmetische Mittel der Daten beträgt 0.017, was in Zusammenhang mit der kleinen Wahl von T dafür spricht, dass sich die Quantile noch besser approximieren lassen sollten.

Tabelle 6: Quantile für die asymptotische Verteilung der vollen Vierer-Tiefe

α	0.1	0.05	0.01	0.001
$q_\alpha^{(4)}$	-0.629	-0.949	-1.715	-2.998

Notwendiger Stichprobenumfang

Mit gleichem Ansatz wie im Ende von Kapitel 2 und 3 werden aus den simulierten asymptotischen Quantilen der vollen Vierer-Tiefe notwendige Stichprobenumfänge durch folgende Ungleichung berechnet:

$$-\frac{N}{8} \overset{!}{\leq} q_\alpha^{(4)} \quad \Leftrightarrow \quad N \overset{!}{\geq} -8q_\alpha^{(4)}.$$

Aus dieser Ungleichung ergeben sich die notwendigen Stichprobenumfänge, siehe Tabelle 7. Auffällig ist, dass die Quantile der asymptotischen Verteilung für höhe-

Tabelle 7: Notwendiger Stichprobenumfang N in Abhängigkeit von häufig verwendeten Signifikanzniveaus α für den Test aus Korollar 4.7

α	0.1	0.05	0.01	0.001
notwendiges N	6	8	14	24

res K kleiner werden. Dadurch werden die benötigten Stichprobenumfänge für eine sinnvolle Verwendung des asymptotischen Tests größer. Allerdings können in allen Fällen auf die exakten Quantile zurückgegriffen werden.

4.4 Konvergenzordnung der verschwindenden Ausdrücke

In diesem Teilkapitel beschäftigen wir uns mit dem Konvergenzfehler im Lemma 4.3. Die Kontrolle dieses Konvergenzfehlers ist nützlich, da wir dann die volle Vierer-Tiefe in linearer Laufzeit bestimmen können und die maximalen zusätzlichen Abweichungen durch den Konvergenzfehler kennen. Aus dem Beweis von Lemma 4.3 entnehmen wir die folgende Abschätzung mit der Tschebyscheff-Ungleichung:

$$P\left(\left|\frac{N}{\binom{N}{4}} \sum_{1\leq n_1<n_2<n_3<n_4\leq N} \prod_{i=1}^{4} \Phi(E_{n_i})\right| > \varepsilon\right) \leq \frac{24N}{\varepsilon^2(N-1)(N-2)(N-3)} \overset{!}{\leq} \alpha,$$

für ein gegebenes kleines $\varepsilon > 0$, das den Fehler beschreibt, und für ein gegebenes kleines $\alpha \in (0,1)$, mit welchem wir die Sicherheit der Aussage skalieren. Ziel ist es für die gegebenen Parameter ε, α den kleinsten Stichprobenumfang N zu finden, sodass der Konvergenzfehler höchstens mit Wahrscheinlichkeit α maximal ε ist. Ist der

Fehler, der durch Vernachlässigung der Zufallsvariable $\frac{N}{\binom{N}{4}} \sum_{1 \leq n_1 < n_2 < n_3 < n_4 \leq N} \prod_{i=1}^{4} \Phi(E_{n_i})$ gemacht wird, hinreichend gut kontrollierbar, können wir die asymptotische Berechnung in linearer Laufzeit durchführen. Folgende kubische Ungleichung ist zu lösen

$$N^3 - 6N^2 + \left(11 - \frac{24}{\alpha\varepsilon^2}\right) N - 6 \overset{!}{\geq} 0, \tag{4.52}$$

wobei wir am kleinsten $N \in \mathbb{N}$ interessiert sind, sodass diese Ungleichung wahr ist. Diese Ungleichung ist im Allgemeinen nur numerisch lösbar, sodass wir in `R` z.B. mit dem Newton-Verfahren (Oevel (1995), S. 74ff.) für gegebene ε, α die Nullstellen des Polynoms $N^3 - 6N^2 + \left(11 - \frac{24}{\alpha\varepsilon^2}\right) N - 6$ in N berechnen. In der Tabelle 8 sind

Tabelle 8: Kleinster Stichprobenumfang, sodass Ungleichung (4.52) erfüllt ist.

$\varepsilon \backslash \alpha$	0.1	0.05	0.01	0.001
0.75	24	33	69	210
0.5	34	47	101	313
0.25	65	91	199	623
0.1	158	223	493	1553
0.05	313	442	983	3101

die kleinsten Stichprobenumfänge N angegeben. Sollte die berechnete volle Vierer-Tiefe nach Bemerkung 4.5 hinreichend weit von einem kritischen Wert, z.B. beim Testverfahren in Korollar 4.7 das Quantil, entfernt sein, so kann Rechenzeit gespart werden, wenn selbst bei Addition mit der maximalen Abweichung ε der kritische Wert nicht über- oder unterschritten wird. Bei Unsicherheiten kann man im Zweifel nachträglich den obigen Ausdruck ohne Verlust von Rechenzeit in polynomieller Laufzeit vom Grad 4 ausrechnen. Ein Nachteil dieser Methode ist, dass für kleine ε der Stichprobenumfang unüblich hoch sein muss.

5 Fazit mit Überblick

Das letzte Kapitel der Arbeit fasst im Unterkapitel 5.1 die wichtigsten Resultate zusammen und gibt im Unterkapitel 5.2 verschiedene Anwendungen der Ergebnisse wieder. Anschließend werden in Unterkapitel 5.3 noch nicht geklärte Fragen, Vermutungen und weitere Untersuchungsziele in einem Ausblick gelistet.

5.1 Zusammenfassung der Resultate

Ziel dieser Arbeit ist die Bestimmung der asymptotischen Verteilung von vollen Datentiefen unter der Nullhypothese für statistische Tests zur Testung der Anpassung von Parametern im Regressionsmodell. Bisher ist die Theorie für die volle Zweier-Tiefe und volle Dreier-Tiefe bekannt gewesen. Für die volle Dreier-Tiefe werden in dieser Arbeit die Aussagen aus Kustosz et al. (2016a) geschärft, durch welche sich nun die volle Dreier-Tiefe exakt in linearer Laufzeit bestimmen lässt. Außerdem lassen sich die in dieser Arbeit entwickelten Beweisideen für Satz 3.10 auf höhere Datentiefen anwenden. Für die höheren Datentiefen wird im Kapitel 4.1 eine Φ-Darstellung gezeigt, die den ersten Schritt zur Herleitung der Asymptotik höherer Datentiefen bildet. Für den Fall $K = 4$ ist in dieser Arbeit die asymptotische Verteilung unter dem wahren Parameter hergeleitet worden. Der Beweisaufbau für die asymptotische Verteilung der vollen Vierer-Tiefe ist analog wie zur vollen Dreier-Tiefe. Es müssen lediglich zusätzlich Ausdrücke asymptotisch vernachlässigt werden. Weiterhin ergibt sich bei der vereinfachten Darstellung der vollen Vierer-Tiefe ein Doppelintegral. Die Laufzeit ist asymptotisch linear, wobei der Konvergenzfehler der asymptotisch vernachlässigbaren Ausdrücke in Kapitel 4.4 analysiert wird. Die Arbeit legt einen Grundbaustein für die Analyse der Asymptotik von höheren vollen Datentiefen. Erwähnenswert ist ferner, dass wir nichtparametrische Verfahren vorliegen haben, sodass die Asymptotik bis auf die geforderten Eigenschaften nicht von den Verteilungsannahmen abhängt. Dadurch ist die Modellierung flexibler als bei Verfahren, die z.B. eine Normalverteilungsannahme der Fehler fordern.

5.2 Anwendungsbeispiele

Verschiedene Anwendungsmöglichkeiten der vollen Datentiefen in der mathematischen Statistik werden in diesem Kapitel präsentiert. Vor allem in Anwendungsfeldern mit häufig extremen Werten wie in der Hochwasserstatistik oder mit vielen Messfehlern können die nachfolgenden Verfahren Verwendung finden.

AR(1)-Modell mit Explosion

Wir verwenden folgendes Zeitreihen-Modell für das Modell aus Definition 2.1:

$$Y_n = \vartheta_0 + Y_{n-1}\vartheta_1 + E_n \text{ für n} = 1, ..., \text{N},$$

wobei der Start $Y_0 = y_0$ fest ist. Ferner gilt $\vartheta_1 > 1$, wodurch wir von einer Zeitreihe mit Explosion sprechen. Das sogenannte AR(1)-Zeitreihenmodell mit Explosion ist eine stochastische Version der Paris-Erdogan-Gleichung, die z.B. zur Modellierung von Risswachstum von Brücken verwendet wird. Während es für stationäre Zeitreihen (d.h. $|\vartheta_1| \leq 1$) eine Bandbreite an robusten Testverfahren für Zeitreihen gibt, ist das Spektrum für Zeitreihen mit Explosion gering (Kustosz et al. (2016a)). Durch die Resultate in Kapitel 3 und 4 können wir Verfahren zur Testung von Parametern $\vartheta = (\vartheta_1, \vartheta_2) \in \Theta := [0, \infty) \times (1, \infty)$ mit folgendem Hypothesenpaar konstruieren:

$$H_0 : \vartheta \in \Theta_0 \text{ vs. } H_1 : \vartheta \in \Theta_1,$$

wobei $\Theta_0 \uplus \Theta_1 = \Theta$ sei. Wir verwerfen die Nullhypothese H_0 wie in Korollar 3.11 bzw. 4.7 beruhend auf der vollen Dreier- bzw. Vierer-Tiefe:

$$\begin{aligned} \text{Man verwerfe } H_0 \text{ falls, } \sup_{\vartheta \in \Theta_0} N\left(d_S^3(\vartheta, Z) - \frac{1}{4}\right) &< q_\alpha^{(3)}, \\ \text{bzw. } \sup_{\vartheta \in \Theta_0} N\left(d_S^4(\vartheta, Z) - \frac{1}{8}\right) &< q_\alpha^{(4)}. \end{aligned}$$

In Kustosz et al. (2016a) werden unter Zusatzannahmen an den Parameterraum Konsistenzeigenschaften des Testverfahrens beruhend auf der vollen Dreier-Tiefe bewiesen, d.h. der Fehler zweiter Art wird bei wachsendem Stichprobenumfang be-

liebig klein. Außerdem werden Simulationsstudien durchgeführt, in der das Testverfahren beruhend auf der vollen Dreier-Tiefe mit anderen Testverfahren, wie dem Vorzeichen-Test in Satz 2.8 oder einem Test beruhend auf den KQ-Schätzer, unter verschiedenen Verteilungsannahmen an die Fehler $E_1, ..., E_N$ verglichen wird. In allen Situationen liefert das Testverfahren beruhend auf der vollen Dreier-Tiefe stabile Ergebnisse. Insbesondere ist die Güte der vollen Dreier-Tiefe in vielen Fällen höher als bei den anderen standardmäßigen Testverfahren. Für die volle Vierer-Tiefe sind weder solche Untersuchungen durchgeführt worden, noch ist umfangreich untersucht, ob oder unter welchen Bedingungen die volle Dreier- oder die volle Vierer-Tiefe bessere Resultate erzielt.

Konfidenzbereiche und Prognosen

Die Dualität zwischen Testverfahren und Konfidenzbereich (Czado und Schmidt (2011)) ermöglicht die Konstruktion von asymptotischen Konfidenzbereichen zum Konfidenzniveau $1 - \alpha$ aus den Testverfahren in Korollar 3.11 bzw. 4.6, indem wir den Annahmebereich des Testverfahren betrachten (analoge Idee aus Kustosz und Müller (2014), wo es für die volle Zweier-Tiefe durchgeführt wird):

$$\left\{\vartheta \in \Theta; N\left(d_S^3(\vartheta, Z) - \frac{1}{4}\right) \geq q_\alpha^{(3)}\right\}$$
$$\text{bzw. } \left\{\vartheta \in \Theta; N\left(d_S^3(\vartheta, Z) - \frac{1}{8}\right) \geq q_\alpha^{(4)}\right\}.$$

Die Parameter lassen sich in der Darstellung des Konfidenzbereichs nicht isolieren, sodass wir die Mengen numerisch z.B. durch eine Gittersuche bestimmen. Aus Konfidenzbereichen können wir robuste Prognoseintervalle für statistische Fragestellungen in Regressionsmodellen bilden, wie z.B. mit der Plug-In-Methode, siehe Szugat et al. (2016).

Vergleich von zwei Stichproben

Das angegebene Regressionsmodell in Definition 2.1 kann als Zweistichproben-Problem dargestellt werden:

$$Y_n = \begin{cases} \vartheta_1 + E_n, & \text{für } n = 1, ..., M \\ \vartheta_2 + E_n, & \text{für } n = M+1, ..., N \end{cases}$$

wobei die erste Stichprobe aus M und die zweite Stichprobe aus $N - M$ Daten besteht. Die beiden Stichproben unterscheiden sich nur vom Lageparameter ϑ_1 oder ϑ_2 und sind ansonsten identisch verteilt. Das Hypothesenpaar

$$H_0 : |\vartheta_1 - \vartheta_2| \leq \delta \text{ vs. } H_1 : |\vartheta_1 - \vartheta_2| > \delta$$

untersucht für einen Relevanzparameter $\delta \geq 0$ die Lageparameter ϑ_1 und ϑ_2 auf einen relevanten Unterschied. Die Entscheidungsregel lautet nach Korollar 3.11 bzw. 4.7:

$$\begin{aligned} \text{Man verwerfe } H_0, \text{ falls: } & \sup_{\vartheta \in \Theta_0} N \left(d_S^3(\vartheta, z) - \frac{1}{4} \right) < q_\alpha^{(3)}, \\ \text{bzw. } & \sup_{\vartheta \in \Theta_0} N \left(d_S^4(\vartheta, Z) - \frac{1}{4} \right) < q_\alpha^{(4)}. \end{aligned}$$

In Malcherczyk (2018) wird der Relevanz-Zweistichproben-Test untersucht und mit dem Relevanz-t-Test für Zweistichproben unter verschiedenen Annahmen verglichen. Unter der Cauchyverteilung und Verteilungen mit Kontaminationen liefert der Test beruhend auf der vollen Dreier-Tiefe deutlich bessere Ergebnisse als der t-Test. Die Güte der vollen Vierer-Tiefe ist dabei noch nicht untersucht worden.

Parameterschätzung und Klassifikation von Ausreißern

Wir betrachten das erste Beispiel in Kapitel 1 aus Abbildung 1, wo aus Realisationen von einer Regressionsgeraden ein Parameter ϑ mit der Kleinste-Fehler-Quadrate-Methode geschätzt wird. Stattdessen schätzen wir den Parameter ϑ durch folgenden

Maximum-Likelihood-Ansatz (Kustosz und Müller (2014)):

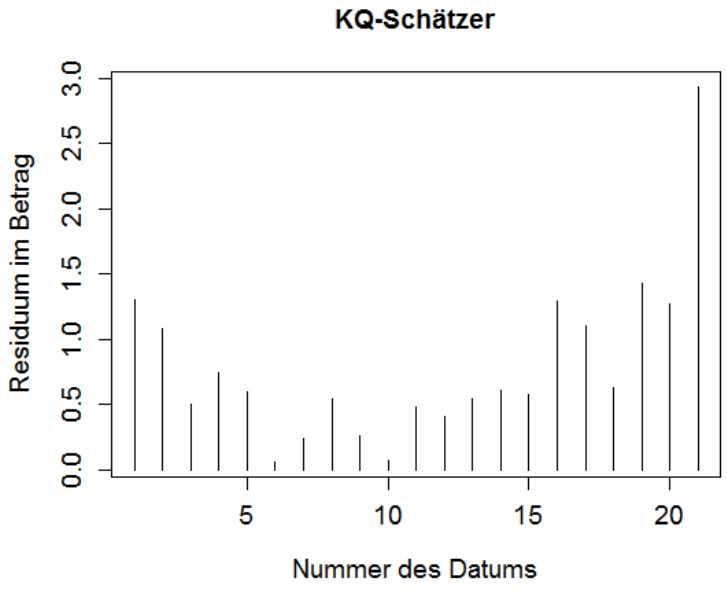
$$\widehat{\vartheta}_{(3)} := \underset{\vartheta \in \Theta}{\operatorname{argmax}}\, d_S^3(\vartheta, Z) \text{ oder } \widehat{\vartheta}_{(4)} := \underset{\vartheta \in \Theta}{\operatorname{argmax}}\, d_S^4(\vartheta, Z).$$

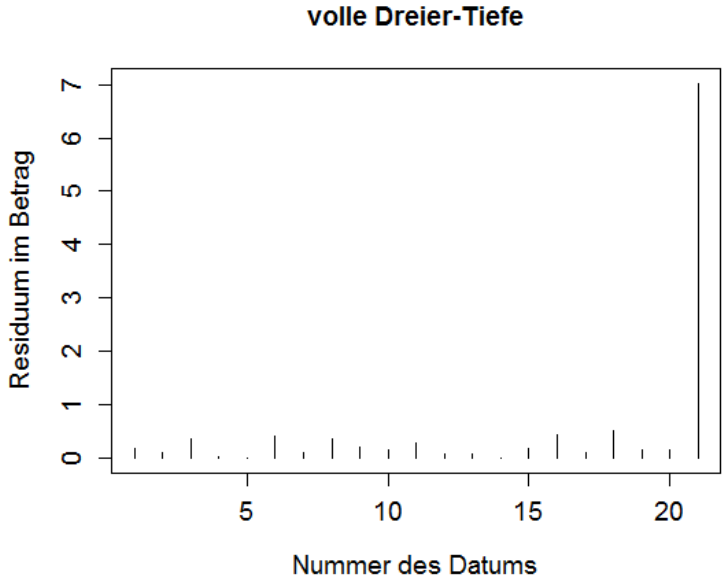
Das Maximum lässt sich numerisch durch eine Gittersuche auf einer diskreten, endlichen Teilmenge von Θ bestimmen. Wir führen die Gittersuche der Menge $\{(\vartheta_0, \vartheta_1) \in \mathbb{R}^2; \vartheta_0 \in \{-1, -0.999, ..., 0.999, 1\}, \vartheta_1 \in \{0, 0.001, ..., 1.999, 2\}$ durch. In

Tabelle 9: Schätzungen aus 20 Datenpunkten einer $\mathcal{N}\left(0, \frac{1}{2}\right)$-Verteilung und einem Ausreißer im Punkt $(4, -3)^\top$

	ϑ ohne Ausreißer	ϑ mit Ausreißer
KQ	$(0.078, 0.960)^\top$	$(1.297, -0.341)^\top$
d_S^3	$(-0.008, 1.078)^\top$	$(0.034, 0.996)^\top$
d_S^4	$(0.034, 0.996)^\top$	$(0.034, 0.996)^\top$

Tabelle 9 werden die Parameterschätzungen mit und ohne Ausreißer mit den Schätzungen der Kleinste-Fehler-Quadrate-Methode verglichen. Für den Datensatz ohne Ausreißer bilden alle drei Methoden zufriedenstellende Schätzungen Durch einen hinzugefügten Ausreißer verschlechtern sich die Schätzungen bei den Datentiefen im Gegensatz zum KQ-Schätzer nicht. Die Schätzung durch die volle Vierer-Tiefe verändert sich durch Hinzufügen des Ausreißers nicht. In Abbildung 4 werden die Residuen für die geschätzten Modelle mit Ausreißer angegeben. Die Idee der Betrachtung von Residuen aus einer robusten Parameterschätzung ist an Rosseeuw und Leroy (1987) angelehnt. Wir bemerken, dass durch die robuste Schätzungen mit den vollen Datentiefen der Ausreißer identifiziert wird. Beim KQ-Schätzer ist der Ausreißer auch auffällig und besitzt die höchste quadratische Abweichung der Residuen. Allerdings ist das Ergebnis weniger deutlich. Zwar ist in einem Punktwolken-Diagramm der Ausreißer noch per Augenmaß erkennbar, da die Regressoren eindimensional sind. In hochdimensionalen Strukturen ist diese Fragestellung jedoch weniger trivial (Rosseeuw und Leroy (1987), S. 7). Diese Anwendung zeigt, dass sich Datentiefen nicht nur für Qualitätsuntersuchungen durch statistische Testverfahren eignen, wie es in dieser Arbeit durchgeführt wird, sondern die Grundidee auch für robuste deskriptive Untersuchungen nützlich sind. Die hier dargestellten Punktschätzung-Methode sollte in der Praxis mit Konfidenzbereichen kombiniert werden, welche durch die





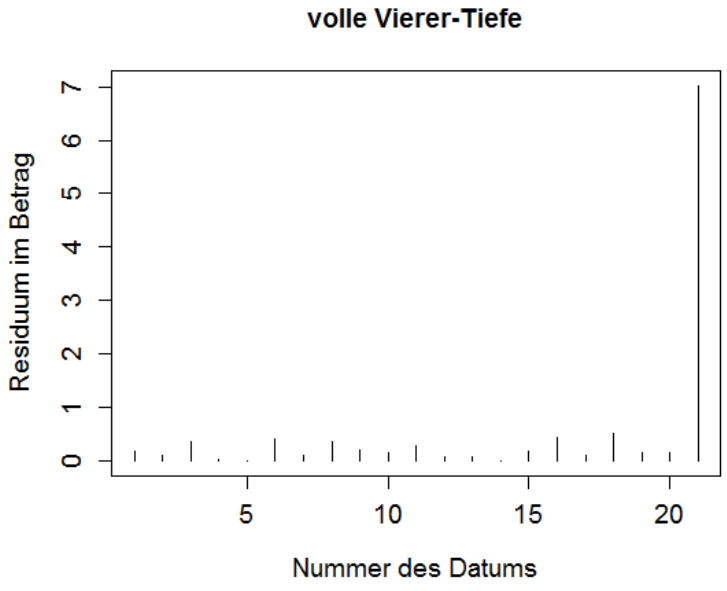


Abbildung 4: Vergleich: Residuen zur Ausreißererkennung (Datenpunkt 21 ist der Ausreißer)

Resultate dieser Arbeit bestimmbar sind, um die Qualität der Punktschätzungen zu beurteilen (z.B. anhand der Größe des Konfidenzbereichs).

5.3 Ausblick

Von hoher Interesse ist die Herleitung der asymptotischen Verteilungen der höheren vollen Datentiefen. Dazu ist die Betrachtung und der Vergleich der Herleitung der Asymptotik zur vollen Dreier-Tiefe und vollen Vierer-Tiefe besonders interessant. Auffällig ist die ähnliche Bauart von Korollar 3.11 und Korollar 4.7, welche Ideen geben, wie die Gestalt der asymptotische Verteilung der allgemeinen vollen K-Tiefen aussieht und sich herleiten lässt. Es ist naheliegend zu vermuten, dass ein Funktional Ψ wie im Beweis von Satz 4.6 bei der K-Tiefe insgesamt $K+1$ Koordinaten aufweisen wird, wobei die Anzahl an Zeitparametern pro Koordinate von 0 bis K zunimmt. Weiterhin wird vermutetet, dass die asymptotische Verteilung sich als eine Zufallsvariable mit $K+1$ Summanden darstellen lässt, wobei dabei eine konstante Zufallsvariable als skalierte Punktauswertung in B_1^2 und Integralterme mit bis zu $K-2$ Integralen über $\psi_L(B_\bullet)^2_{t_1,\dots,t_{L-2}}$ für Zeitparameter $t_1,\dots,t_{L-2}$ für $L=3,\dots,K$ mit passendem ψ_L und einer Konstanten vorliegen. Mit Satz 4.1 ist bereits durch vollständige Induktion die Φ-Darstellung der vollen Datentiefen gelungen. Der nächste Schritt ist die Gewinnung von im Produkt symmetrischen Darstellungen bei einer allgemeinen K-Tiefe. Dabei sind die kombinatorischen Argumente eine Herausforderung, wenn mit allgemeinem K gerechnet wird. Außerdem müssen die Anteile in der Φ-Darstellung gefunden werden, die asymptotisch vernachlässigbar sind. Ist eine symmetrische Darstellung wie in Satz 3.4 oder Satz 4.4 gefunden, kann mit dem Satz von Donsker und dem Continuous-Mapping-Theorem argumentiert werden. Beim Vergleich der Anwendung von $K=3$ und $K=4$ zeigt sich, dass die Probleme bei der Anwendung vom Satz von Donsker bereits im Fall $K=3$ gelöst sind und die Anwendung für $K=4$ analog ist. Daher wird vermutet, dass für allgemeines K die wahrschlichkeitstheoretischen Schwierigkeiten bereits überwunden sind. Weiterhin kann der Zwischenprozess in der Herleitung der asymptotischen Verteilung der vollen Vierer-Tiefe im Detail untersucht werden. Dieser entspricht einem Brownschen Blatt mit zeitlicher Skalierung in beiden Zeitparametern. Dazu müssen

mehrdimensionale Skorohod-Räume studiert werden, die z.B. in Łagodowski und Rychlik (1986) und Ferger (2010) diskutiert werden.

Ferner bleibt die Frage offen, inwieweit sich die höheren vollen Datentiefen für $K \geq 4$ in kürzerer Laufzeit bestimmen lassen. Für die volle Zweier-Tiefe und die volle Dreier-Tiefe ist eine exakte lineare Laufzeit möglich, während für die volle Vierer-Tiefe nur eine asymptotisch lineare Laufzeit bisher gewährleistet werden kann. Hier sind gerade für die Praxis weitere Analysen, insbesondere für die höheren Datentiefen, interessant. Eventuell gibt es z.B. Vereinfachungen bei den asymptotisch vernachlässigbaren Ausdrücken, wie dem in Kapitel 4.4 betrachteten Term durch kombinatorische Überlegungen.

Ferner können sich weitere Testverfahren überlegt werden. Es können z.B. Mehr-Stichproben-Vergleiche oder Testungen mit sehr vielen Parametern konstruiert werden. Heuristisch gesehen decken K-Tiefen mit höherem K intensiver Abhängigkeitsstrukturen und Wechselwirkungen zwischen den Parametern ab. Man kann daher vermuten, dass für Modelle mit vielen Parametern höhere K benötigt werden. Die Gütefunktionen solcher Testverfahren sollten in Rahmen von Simulationsstudien mit anderen robusten und nicht-robusten Testverfahren verglichen werden. Wünschenswert wären verbesserte Resultate gegenüber nicht-robusten Testverfahren bei zugrundeliegenden Fehlerverteilungen mit schweren Rändern, wie z.B. die Cauchy-Verteilung oder Doppel-Exponentialverteilung (Büning (1991), S. 5ff.) oder unter Kontaminationen (siehe Büning und Trenkler (1994), S. 24 und S.294ff.). Bei Vergleichen mit anderen robusten Verfahren erhofft man sich, dass höhere Anteile von Ausreißern geringeren Einfluss haben oder das die Gütefunktion unter bestimmten Annahmen gleichmäßig besser ist.

Neue Begriffe von Datentiefen können ebenso in zukünftigen Analysen relevant sein. Mit den Methoden dieser Arbeit können nur Daten mit univariaten Zielgrößen untersucht werden, während mehrdimensionale Zielgrößen nicht betrachtet werden können. In Kustosz et al. (2016b) werden vereinfachte Datentiefen untersucht, die nicht alle Kombinationen betrachten, um Laufzeiten zu verringern. Bisherige Untersuchungen zeigen schlechte Resultate, siehe Kustosz et al. (2016b) und Malcherczyk (2018). Allerdings können alternative Ansätze zu brauchbaren Maßzahlen führen.

Literatur

Bauer, H. (2002): *Wahrscheinlichkeitstheorie*, 5. Aufl., Walter de Gruyter, Erlangen.

Bauer, H. (1992): *Maß- und Integrationstheorie*, 2. Aufl., Walter de Gruyter, Erlangen.

Bickel, P.J. and Doksum, K.A. (1977): *Mathematical Statistics*, 1. Aufl., Holden-Day Inc., Berkeley.

Billingsley, P. (1999): *Convergence of Probability Measure*, 2. Aufl., Wiley series in probability and statistics, Chicago.

Büning H. und Trenkler G. (1994): *Nichtparametrische statistische Methoden*, 2. Aufl., Walter de Gruyter, Berlin/Hannover.

Büning H. (1991): *Robuste und adaptive Tests*, 1. Aufl., Walter de Gruyter, Berlin.

Czado, C. und Schmidt, T. (2011): *Mathematische Statistik*, 1. Aufl., Springer, München/Leipzig.

Davis P.J. and Rabinowitz P. (1975): *Methods of numerical integration*, 1. Aufl., Academic Press, New York.

Ferger (2010): Arginf-Sets of multivariate cadlag processes and their convergence in hyperspace topologies. *Theory of Stochastic Processes*, Vol. 20 (36), no. 2, 2015, pp. 13–41

Fischer, S., Fried R. and Schumann, A. H. (2015): Examination for robustness of parametric estimators for flood statistics in the context of extraordinary extreme events. *Hydrology and earth system sciences discussions*, 12(8), 8553-8576.

Huggins, R. (1989): The Sign Test for Stochastic Processes. *Australian and New Zealand Journal of Statistics*, 31, 153-165.

Junwen, H. (2013): *somebm: some Brownian motions simulation functions.* R package version 0.1. https://CRAN.R-project.org/package=somebm

Kaballo, W. (2000): *Einführung in die Analysis 1*, 2. Aufl., Spektrum, Dortmund.

Klenke, A. (2006): *Wahrscheinlichkeitstheorie*, 1. Aufl., Springer, Mainz.

Korte, B. und Vygen, J. (2018): *Kombinatorische Optimierung: Theorie und Algorithmen*, 3. Aufl., Springer, Bonn.

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

Kustosz, C. and Müller, C. (2014). Analysis of crack growth with robust, distributionfree estimators and tests for nonstationary autoregressive processes. *Statistical Papers* 55, 125-140.

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

Kustosz, C. and Szugat, S. (2016): *rexpar: Simplicial Depth for Explosive Autoregressive Processes*. R package version 1.1.

Łagodowski A. and Rychlik Z. (1986): Weak convergence of probability meaures on the function space $\mathcal{D}_d[0,\infty)$, *Bull. Polish Acad. Sci. Math.* 34 (1986), no. 5-6, 329–335.

Lehmann, E.L. und D'Abrera, H.J.M. (1975): *Statistical Methods based on Ranks*, 1. Aufl., Holden-Day Inc., Berkeley.

Malcherczyk, D.A. (2018): *Vergleich von Zwei-Stichproben-Relevanz-Tests basierend auf t-Tests und Datentiefen*. Bachelorarbeit, Technische Universität Dortmund.

Maronna, R.A., Martin, R.D. und Yohai, V.J. (2006): *Robust Statistics Theory and Methods*, 1. Aufl., John Wiley & Sons, Ltd.

Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* 30, 1681-1736.

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

Oevel W. (1995): *Einführung in die numerische Mathematik*, 1. Aufl., Spektrum, Paderborn.

Szugat, S., Heinrich, J., Maurer R. und Müller, C. (2016): Prediction intervals for the failure time of prestressed concrete beams. *Advances in Materials Science and Engineering.*

Toutenburg, H. (2003): *Lineare Modelle*, 2. Aufl., Physica-Verlag Heidelberg, München.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rousseeuw, P.J. and Hubert, M. (1999): Regression depth (with discussion). *J. Amer. Statist. Assoc.* 94, 388-433.

Rosseeuw, P.J. and Leroy, A.M. (1987): *Robust Regression and Outlier Detection*, 1. Aufl., John Wiley & Sons.

Werner, D. (2018): *Funktionalanalysis*, 8. Aufl., Springer, Berlin.

Eidesstattliche Versicherung

_____ _____

Name, Vorname Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/Masterarbeit* mit dem 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.

_____ _____

Ort, Datum Unterschrift

*Nichtzutreffendes bitte streichen

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 gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

_____ _____

Ort, Datum Unterschrift