universität wien

MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Die Full- und die Split-Methode: zwei auf On-Line Setting basierende Herangehensweisen zur Konstruktion konformaler Konfidenzbereiche im Rahmen von Regressionsanalsysen“

verfasst von / submitted by

Mag. Mario Giuseppe Caputo, Bakk.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 951
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Hannes Leeb

Ich danke meiner ganzen Familie für ihre finanzielle und moralische Unterstützung, die sie mir während meiner Studienzeit gewährt hat, sowie den Kollegen aus dem Studium für eine produktive Zusammenarbeit. Insbesondere möchte ich aber meinem Betreuer Prof. Dr. Hannes Leeb meinen Dank aussprechen, ohne dessen Anregungen und Impulse diese Arbeit in dieser Form sicher nicht zustande gekommen wäre.

Inhaltsverzeichnis

1 Einleitung

Die Prozedur der *konformalen Inferenz* (*conformal prediction*) ist eine vor allem durch Gammerman et al. (2005) sowie durch Shafer und Vovk (2008) bekannt gemachte Methode zur Generierung von Konfidenzbereichen, welche sich für *Prognosezwecke* eignen und oftmals im Rahmen des sogenannten *On-Line Settings* konstruiert werden. Konformale Konfidenzbereiche basieren zumeist auf Algorithmen, die auf den Gebieten der *Klassifikation* und der *Regression* genutzt werden. Somit findet diese Technik in zwei für die *Inferenzstatistik* und das *maschinelle Lernen* wichtigen Bereichen eine entsprechende Anwendung.

Konzeptionell kann konformale Inferenz (so wie wir sie in den Fokus dieser Arbeit stellen) wie folgt umschrieben werden: wir betrachten zu einer gegebenen natürlichen Zahl n, die für den Umfang einer aus einer Grundgesamtheit gezogenen Stichprobe steht, die Werte $(x_1, y_1), ..., (x_n, y_n)$, welche als Realisationen von Zufallsvariablen betrachtet werden können und die in der Stichprobe vorkommenden Instanzen repräsentieren. Im Rahmen von Regressionsanalysen können wir die Werte $x_1, ..., x_n$ als Realisationen eines (möglicherweise auch mehrdimensionalen) *Regressors* (bzw. einer *erklärenden* Variable) auffassen, während wir $y_1, ..., y_n$ als Realisationen eines (zumeist eindimensionalen) Regressanden (bzw. einer *erklärten* Variable) betrachten können.[1] Wir ziehen nun aus der Grundgesamtheit eine neue Instanz mit bekanntem x_{n+1}, aber unbekanntem y_{n+1}. Die konformale Methode nutzt die verfügbaren Daten $(x_1, y_1), ..., (x_n, y_n)$, welche den Klassifikations- bzw. Regressionsalgorithmen zumindest zu einem Teil als *Trainingsdaten* dienen, sowie die Information x_{n+1} um einen Konfidenzebereich für den unbekannten Wert y_{n+1} zu erzeugen. Ein solcher Konfidenzbereich kann in diesem Kontext auch als Prognosebereich für y_{n+1} aufgefasst werden. Die „Regel“, die entscheidet welche Werte aus der Menge aller möglichen Werte des Regressanden in den Konfidenzbereich fallen, wird in der Literatur sehr oft durch ein sogenanntes *Nichtkonformitätsmaß* (*nonconformity measure*) beschrieben, welches die „Diskrepanz“ (bzw. „Unähnlichkeit“) zwischen den zum Beispiel in einer Punktprognose zusammengefassten Daten $(x_1, y_1), ..., (x_n, y_n), x_{n+1}$ und den möglichen Werten für y_{n+1} „misst“; vgl. Shafer und Vovk (2008, S.383-384). Je „ausgeprägter“ eine solche Diskrepanz in einem zu betrachteten Fall ist, um so größer die damit zusammenhängende Nichtkonformität. Nach dieser Regel fallen all jene Elemente aus dem Wertebereich des Regressanden in den konformalen Konfidenzbereich, bei denen die Diskrepanz zu den gegebenen Daten nicht zu ausgeprägt ist. In Regressionsanalysen erfüllen oft die Absolutbeträge der *Residuen* die Funktion

eines Nichtkonformitätsmaßes (siehe auch Kapitel 3).

Das Konzept des On-Line Settings kommt bei variabler Stichprobengröße zum Tragen und ist nach folgendem Schema aufgebaut: bei einer Stichprobengröße von 1 werden die Daten (x_1, y_1) und x_2 zur Generierung eines konformalen Konfidenzbereichs für y_2 genutzt, bei einer Stichprobengröße von 2 dienen die Daten (x_1, y_1) , (x_2, y_2) und x_3 zur Erzeugung eines konformalen Konfidenzbereichs für y_3, bei einer Stichprobengröße von n dienen $(x_1, y_1), ..., (x_n, y_n)$ und x_{n+1} zur Erzeugung eines konformalen Konfidenzbereichs für y_{n+1} und so weiter. Zu jeder Stichprobengröße werden also die Instanzen der Reihenfolge ihres Auftretens nach nummeriert und bei Erhöhung des Stichprobenumfangs wird die Menge der zur Bestimmung des konformalen Konfidenzbereichs herangezogenen Daten *sequentiell* um die neu hinzukommende Instanz erweitert. Dies steht in einem gewissen Gegensatz zum *Off-Line Setting*, bei dem die Daten beim Auftauchen neuer Instanzen nicht angepasst werden, d.h. heißt, dass bei Stichprobengröße n für alle neuen Instanzen die Produktion eines Konfidenzbereichs immer auf der *gleichen* Datenmenge $(x_1, y_1), ..., (x_n, y_n)$ beruht.[2]

In dieser Arbeit werden wir uns auf Regressionsprobleme beschränken und uns dabei mit zwei Arten, nach denen konformale Konfidenzbereiche innerhalb eines On-Line Settings konstruiert werden, auseinandersetzen. Das sind zum einen die *Full-Methode* und zum anderen die *Split-Methode*, welche beide in G'Sell et al. (2016) vorgestellt werden. Wir wollen zuerst zeigen, dass bereits unter *Vertauschbarkeit* jener Folge von Zufallsvariablen, deren Realisationen die Daten liefern, bei *fester* Stichprobengröße die nach diesen Methoden erzeugten Konfidenzbereiche *gültig* sind und zwar in dem Sinne, dass die *Überdeckungswahrscheinlichkeit* der Konfidenzbereiche (also die Wahrscheinlichkeit, dass y_{n+1} im betrachteten Konfidenzbereich enthalten ist) nicht kleiner als das gewünschte *nominelle Konfidenzniveau* ist und sich diesem mit *wachsender* Stichprobengröße sogar annähert. Dann werden wir uns mit einigen im Bezug auf wachsende Stichprobengröße *asymptotischen* Eigenschaften der beiden Methoden beschäftigen, die wir anschließend exemplarisch durch eine einfache *Simulation* eines On-Line Settings veranschaulichen. Dabei gehen wir von der stärkeren Annahme aus, dass die Daten von einer Folge *unabhängiger* und *identisch verteilter* Zufallsvariablen generiert werden. Die im Kontext der Asymptotik relevanten Größen werden die aus den Daten berechenbare *empirische Überdeckungsrate* sowie die *Breite* der Konfidenzbereiche sein.

Diese Arbeit ist in folgender Weise gegliedert. Im Kapitel 2 beschäftigen wir uns mit den für diese Arbeit wichtigen Fakten über vertauschbare Zufallsvariablen. Im Kapitel 3 werden wir die Full- und die Split-Methode einführen

und zeigen, dass bei endlicher Stichprobe die *Überdeckungswahrscheinlichkeit* der damit erzeugten Konfidenzbereiche nicht kleiner ist als das *nominellen Konfidenzniveau*. Außerdem beschäftigen wir uns dort kurz mit alternativen Techniken zur Bestimmung von Konfidenzbereichen. Kapitel 4 setzt sich mit der Asymptotik der beiden Methoden auseinander und Kapietl 5 beinhaltet schließlich die Simulation eines On-Line Settings zur Demonstration der beiden Methoden.

2 Die Grundlagen: vertauschbare Zufallsvariablen und Ränge

In diesem Kapitel wollen wir einige wichtige Resultate bezüglich *vertauschbarer* Zufallsvariablen erarbeiten. Diese werden dann in Kapitel 3 notwendig sein.

Es sei ein Wahrscheinlichkeitsraum $(\Omega, \mathscr{A}, P)$ gegeben mit Ω als Grundgesamtheit, $\mathscr{A}$ als σ-Algebra auf Ω und P als Wahrscheinlichkeitsmaß auf $\mathscr{A}$. Sei außerdem $(U_i)_{i\in\mathbb{N}}$ eine Folge von Funktionen der Form

$$U_i : \Omega \to \mathbb{R}^d$$

für jedes $i \in \mathbb{N}$, wobei $d \in \mathbb{N}$ und $\mathbb{R}^d$ der d-dimensionale reelle euklidische Raum sei. Für beliebige $d \in \mathbb{N}$ stehe darüber hinaus der Ausdruck $\mathscr{B}_d$ für die Borelsche σ-Algebra auf $\mathbb{R}^d$. Wir nehmen für jedes $i \in \mathbb{N}$ die Funktion U_i als $\mathscr{A}$-$\mathscr{B}_d$-messbar an, womit wir $(U_i)_{i\in\mathbb{N}}$ als Folge von Zufallsvariablen auffassen können, welche auf den Wahrscheinlichkeitsraum $(\Omega, \mathscr{A}, P)$ definiert sind und Werte aus der Menge $\mathbb{R}^d$ annehmen. Für jede natürliche Zahl $n \in \mathbb{N}$ bezeichne außerdem S_n die Menge aller Permutationen von $\{1, ..., n\}$, also

$$S_n := \{\sigma : \{1, ..., n\} \to \{1, ..., n\} \,|\, \sigma \text{ ist bijektiv}\}.$$

Wir führen nun einen für diese Arbeit zentralen Begriff, nämlich jenen der *Vertauschbarkeit* von Zufallsvariablen, wie folgt ein:

Definition 2.1. *Für festes $n \in \mathbb{N}$ heißt eine endliche Folge von $\mathbb{R}^d$-wertigen Zufallsvariablen $U_1, ..., U_n$ bezüglich des Wahrscheinlichkeitsmaßes P vertauschbar, wenn für jede Permutation $\sigma \in S_n$ die Zufallsvektoren $(U_1, ..., U_n)$ und $(U_{\sigma(1)}, ..., U_{\sigma(n)})$ bezüglich P die gleiche Verteilung haben, d.h. wenn für beliebige Mengen $B_1, ..., B_n \in \mathscr{B}_d$ die Gleichheit*

$$P(\{U_1 \in B_1\} \cap ... \cap \{U_n \in B_n\}) = P(\{U_{\sigma(1)} \in B_1\} \cap ... \cap \{U_{\sigma(n)} \in B_n\})$$

gilt.

Wir nennen eine *unendliche* Folge von Zufallsvariablen $(U_i)_{i\in\mathbb{N}}$ vertauschbar, wenn für jedes $n \in \mathbb{N}$ die endliche Folge $U_1, ..., U_n$ vertauschbar im Sinne von Definition 2.1 ist. Für vertauschbare Zufallsvariablen gilt:

Lemma 2.1. *Für festes $n \in \mathbb{N}$ sei die Folge von $\mathbb{R}^d$-wertigen Zufallsvariablen $U_1, ..., U_n$ bezüglich des Wahrscheinlichkeitsmaßes P vertauschbar. Dann gilt, dass für jedes $i, j \in \{1, ..., n\}$ mit $i \neq j$ die Zufallsvariablen U_i und U_j bezüglich P die gleiche Verteilung haben.*

Beweis. Betrachte beliebige $i, j \in \{1, ..., n\}$ mit $i \neq j$ sowie eine Permutation $\sigma \in S_n$ mit der Eigenschaft $\sigma(i) = j$. Betrachte des Weiteren die Mengen $B_1, ..., B_n$, wobei $B_i \in \mathscr{B}_d$ beliebig gewählt sei und $B_k = \mathbb{R}^d$ für jedes $k \in \{1, ..., n\}$ mit $k \neq i$. Wegen der Vertauschbarkeit von $U_1, ..., U_n$ gilt dann

$$\begin{aligned} &P(U_i \in B_i) \\ &= P(\{U_1 \in B_1\} \cap ... \cap \{U_n \in B_n\}) = P(\{U_{\sigma(1)} \in B_1\} \cap ... \cap \{U_{\sigma(n)} \in B_n\}) \\ &= P(U_{\sigma(i)} \in B_i) = P(U_j \in B_i). \end{aligned}$$

□

Es gilt außerdem die folgende Implikation; vgl. auch Shafer und Vovk (2008, S.378):

Lemma 2.2. *Sei $(U_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}^d$-wertigen Zufallsvariablen, welche bezüglich P unabhängig und identisch verteilt sind. Dann ist $(U_i)_{i\in\mathbb{N}}$ vertauschbar.*

Beweis. Sei $n \in \mathbb{N}$ und betrachte beliebige $B_1, ..., B_n \in \mathscr{B}_d$. Dann gilt aufgrund der Unabhängigkeit von $U_1, ..., U_n$, dass

$$P(\{U_1 \in B_1\} \cap ... \cap \{U_n \in B_n\}) = \prod_{i=1}^{n} P(U_i \in B_i)$$

sowie

$$P(\{U_{\sigma(1)} \in B_1\} \cap ... \cap \{U_{\sigma(n)} \in B_n\}) = \prod_{i=1}^{n} P(U_{\sigma(i)} \in B_i)$$

für jedes $\sigma \in S_n$. Für jedes $i \in \{1, ..., n\}$ gilt gleichzeitig

$$P(U_i \in B_i) = P(U_{\sigma(i)} \in B_i),$$

da U_i und $U_{\sigma(i)}$ nach Voraussetzung die gleiche Verteilung besitzen. Somit erhalten wir

$$P(\{U_1 \in B_1\} \cap ... \cap \{U_n \in B_n\}) = P(\{U_{\sigma(1)} \in B_1\} \cap ... \cap \{U_{\sigma(n)} \in B_n\}).$$

□

Für den Rest dieses Kapitels betrachten wir $(U_i)_{i\in\mathbb{N}}$ als eine Folge reellwertiger Zufallsvariablen, d.h. wir setzen $d = 1$. Wir können dann bei beliebigem $n \in \mathbb{N}$ für die Zufallsvariablen $U_1, ..., U_{n+1}$ den Rang von U_i mit $i \in \{1, ..., n+1\}$ als

$$rang(U_i) := \sum_{j=1}^{n+1} \mathbb{1}_{\{U_j \leq U_i\}}$$

definieren, wobei für jedes $i, j \in \{1, ..., n+1\}$ der Ausdruck $\mathbb{1}_{\{U_j \leq U_i\}}$ für die Indikatorfunktion des Ereignisses $\{U_j \leq U_i\}$ stehe. Der dazu korrespondierende Rangvektor sei

$$R_{n+1} := (rang(U_1), ..., rang(U_{n+1})),$$

welcher Werte aus der Menge des $(n + 1)$-fachen kartesischen Produkts von $\{1, ..., n + 1\}$, abgekürzt durch den Ausdruck $\{1, ..., n + 1\}^{n+1}$, annimmt. Wir formulieren das folgenden Hilfsresultat; vgl. auch G'Sell et al. (2016, S.7):

Lemma 2.3. *Für $n \in \mathbb{N}$ sei die endliche Folge von $\mathbb{R}$-wertigen Zufallsvariablen $U_1, ..., U_{n+1}$ vertauschbar, wobei unter diesen bezüglich P fast sicher keine Bindungen vorkommen, d.h. für jedes $i, j \in \{1, ..., n + 1\}$ mit $i \neq j$ gelte*

$$P(U_i \neq U_j) = 1.$$

Dann ist die Zufallsvariable $rang(U_{n+1})$ diskret gleichverteilt auf der Menge $\{1, ..., n + 1\}$, d.h. für jedes $k \in \{1, ..., n + 1\}$ gilt

$$P(rang(U_{n+1}) = k) = \frac{1}{n+1}.$$

Beweis. Zunächst zeigen wir, dass der Rangvektor R_{n+1} diskret gleichverteilt ist auf der Menge aller Vektoren aus $\{1, ..., n+1\}^{n+1}$ mit ungleichen Einträgen.

Für jeden Vektor $(r_1, ..., r_{n+1}) \in \{1, ..., n + 1\}^{n+1}$ mit der Eigenschaft, dass $r_i = r_j$ für mindestens ein Paar $i, j \in \{1, ..., n + 1\}$ mit $i \neq j$, gilt

$$P(R_{n+1} = (r_1, ..., r_{n+1})) \leq P(U_i = U_j) = 0,$$

und somit

$$P(R_{n+1} = (r_1, ..., r_{n+1})) = 0.$$

Wir betrachten nun Vektoren $(r_1, ..., r_{n+1}) \in \{1, ..., n + 1\}^{n+1}$ mit der Eigenschaft $r_i \neq r_j$ für jedes $i, j \in \{1, ..., n + 1\}$ mit $i \neq j$. Dann gibt es genau eine Permutation $\sigma \in S_{n+1}$, sodass $\sigma(i) = r_i$ für jedes $i \in \{1, ..., n + 1\}$. Daraus folgt

$$\{R_{n+1} = (r_1, ..., r_{n+1})\} = \{rang(U_1) = \sigma(1), ..., rang(U_{n+1}) = \sigma(n + 1)\}.$$

Betrachte nun ein beliebiges $l \in \{1, ..., n+1\}$. Da σ bijektiv ist, gibt es genau ein $k \in \{1, ..., n+1\}$, sodass $\sigma(k) = l$. Falls

$$rang(U_i) = \sigma(i)$$

für jedes $i \in \{1, ..., n+1\}$ gilt, dann ist $rang(U_k) = l$. Sei σ^{-1} die inverse Funktion von σ. Dann gilt $\sigma^{-1}(l) = k$, und somit

$$rang(U_{\sigma^{-1}(l)}) = l.$$

Wir erhalten also die Aussage

$$\{rang(U_1) = \sigma(1), ..., rang(U_{n+1}) = \sigma(n+1)\}$$

$$\subseteq \{rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1\}$$

Analog kann gezeigt werden, dass

$$\{rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1\}$$

$$\subseteq \{rang(U_1) = \sigma(1), ..., rang(U_{n+1}) = \sigma(n+1)\},$$

womit also

$$\{R_{n+1} = (r_1, ..., r_{n+1})\}$$

$$= \{rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1\}.$$

Da $\sigma^{-1} \in S_{n+1}$, gilt wegen der Vertauschbarkeit von $U_1, ..., U_{n+1}$ gleichzeitig

$$P(rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1))$$

$$= P(rang(U_1) = 1, ..., rang(U_{n+1}) = n+1).$$

Beachte, dass σ beliebig aus S_{n+1} gewählt wurde und in der rechten Seite der obigen Gleichung die Ränge der Zufallsvariablen $U_1, ..., U_{n+1}$ gemäß einer festen Permutation aus S_{n+1}, nämlich der Identitätsfunktion auf der Menge $\{1, ..., n+1\}$, angeordnet sind. Da außerdem die Menge S_{n+1} genau $(n+1)!$ Elemente hat und angenommen wird, dass unter den Zufallsvariablen $U_1, ..., U_{n+1}$ fast sicher keine Bindungen vorkommen, gilt

$$\sum_{\sigma \in S_{n+1}} P(rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1) = 1,$$

und somit

$$P(R_{n+1} = (r_1, ..., r_{n+1}))$$

$$= P(rang(U_{\sigma^{-1}(1)}) = 1, ..., rang(U_{\sigma^{-1}(n+1)}) = n+1) = \frac{1}{(n+1)!}.$$

Wir fahren fort, indem wir zu Hilfszwecken die Menge

$$\tilde{S}_k := \{\sigma \in S_{n+1} \,|\, \sigma(n+1) = k\}$$

mit $k \in \{1, ..., n+1\}$ einführen. Wir haben oben gezeigt, dass die Zufallsvariable R_{n+1} all jene $(n+1)!$ Vektoren aus der Menge $\{1, ..., n+1\}^{n+1}$, bei welchen die Einträge ungleich sind, mit Wahrscheinlichkeit $1/(n+1)!$ annimmt. Da von diesen Vektoren all jene, bei welchen der $(n+1)$-te Eintrag mit k ident ist, jeweils mit genau einem $\sigma \in \tilde{S}_k$ identifiziert werden können, und außerdem, wie oben gezeigt, all jene Vektoren mit mindestens einem Paar gleicher Einträge von R_{n+1} mit Wahrscheinlichkeit Null angenommen werden, erhalten wir unter Anwendung der σ-Additivität von Wahrscheinlichkeitsmaßen

$$P(rang(U_{n+1})) = k)$$

$$= P(\bigcup_{\sigma \in \tilde{S}_k} \{rang(U_1) = \sigma(1), ..., rang(U_n) = \sigma(n), rang(U_{n+1}) = k\})$$

$$= \sum_{\sigma \in \tilde{S}_k} P(R_{n+1} = (\sigma(1), ..., \sigma(n), k)).$$

Beachte, dass $\tilde{S}_k$ genau $n!$ Elemente besitzt. Wir erhalten also

$$\sum_{\sigma \in \tilde{S}_k} P(R_{n+1} = (\sigma(1), ..., \sigma(n), k))$$

$$= \sum_{\sigma \in \tilde{S}_k} \frac{1}{(n+1)!} = \frac{n!}{(n+1)!} = \frac{1}{n+1},$$

und somit

$$P(rang(U_{n+1}) = k) = \frac{1}{n+1}.$$

□

Zur Formulierung des nächsten Satzes definieren wir jedes $x \in \mathbb{R}$ die Aufrundungsfunktion

$$\lceil x \rceil := min\{z \in \mathbb{Z} \,|\, x \leq z\}.$$

Außerdem betrachten wir für beliebige $n \in \mathbb{N}$ die *Ordnungsstatistiken* $U_{(1:n)}, ..., U_{(n:n)}$ der Beobachtungen $U_1, ..., U_n$. Das vorhergehende Lemma kann nun verwendet werden um das folgende wichtige Resultat zu beweisen:

Satz 2.1. *Für $n \in \mathbb{N}$ sei die endliche Folge von $\mathbb{R}$-wertigen Zufallsvariablen $U_1, ..., U_{n+1}$ bezüglich P vertauschbar, wobei unter diesen fast sicher keine Bindungen vorkommen. Für jedes $\alpha \in (0,1)$, welches die Ungleichung*

$$(1-\alpha)(n+1) \leq n \tag{2.1}$$

erfüllt, gilt dann

$$P(U_{n+1} \in (-\infty, U_{(h:n)})) \geq 1-\alpha \tag{2.2}$$

und

$$P(U_{n+1} \in (-\infty, U_{(h:n)})) < 1-\alpha+\frac{1}{n+1}, \tag{2.3}$$

wobei $h := \lceil (1-\alpha)(n+1) \rceil$. Darüber hinaus ist h die kleinste Zahl aus der Menge $\{1, ..., n+1\}$, für die Ungleichung (2.2) erfüllt ist, d.h. für jedes $m \in \{1, ..., n+1\}$ mit $m < h$ gilt

$$P(U_{n+1} \in (-\infty, U_{(m:n)})) < 1-\alpha.$$

Beweis. Aus den Ungleichungen $U_{(1:n)} \leq U_{(2:n)} \leq ... \leq U_{(n:n)}$, die mit jedem $n \in \mathbb{N}$ auf ganz Ω gültig sind, folgen

$$\{U_{n+1} \in (-\infty, U_{(1:n)})\} = \{rang(U_{n+1}) = 1\},$$

$$\{U_{n+1} \in [U_{((m-1):n)}, U_{(m:n)})\} = \{rang(U_{n+1}) = m\}$$

für alle $m \in \{2, ..., n\}$. Es gilt also für jedes $m \in \{1, ..., n\}$ die Identität

$$\{U_{n+1} \in (-\infty, U_{(m:n)})\} = \{rang(U_{n+1}) \leq m\}.$$

Beachte, dass $h \in \{1, ..., n\}$ wegen Ungleichung (2.1) gilt. Unter Anwendung der σ-Additivität und Lemma 2.3 erhalten wir damit

$$P(U_{n+1} \in (-\infty, U_{(h:n)})) = P(rang(U_{n+1}) \leq h) = P\left(\bigcup_{i=1}^{h} \{rang(U_{n+1}) = i\}\right)$$

$$= \sum_{i=1}^{h} P(rang(U_{n+1}) = i) = \sum_{i=1}^{h} \frac{1}{n+1} = \frac{h}{n+1}.$$

Aus $h \geq (1-\alpha)(n+1)$ ergibt sich dann einerseits

$$P(U_{n+1} \in (-\infty, U_{(h:n)})) \geq \frac{(1-\alpha)(n+1)}{n+1} = 1-\alpha$$

und aus $h < (1-\alpha)(n+1)+1$ andererseits

$$P(U_{n+1} \in (-\infty, U_{(h:n)})) < 1-\alpha+\frac{1}{n+1}.$$

Betrachte nun ein $m \in \{1, ..., n\}$ mit $m < h$. Analog zu oben zeigt man, dass

$$P(U_{n+1} \in (-\infty, U_{(m:n)})) = \frac{m}{n+1}.$$

Da jedoch h die Kleinste unter den ganzen Zahlen ist, welche größer als oder gleich $(1-\alpha)(n+1)$ sind, gilt gleichzeitig $m < (1-\alpha)(n+1)$ und somit

$$P(U_{n+1} \in (-\infty, U_{(m:n)})) < \frac{(1-\alpha)(n+1)}{n+1} = 1-\alpha.$$

□

3 Zwei Methoden konformale Konfidenzbereiche zu bestimmen: Betrachtung bei finiten Stichproben

In diesem Kapitel wollen wir die in der Einleitung erwähnten Methoden zur Konstruktion von konformalen Konfidenzbereichen diskutieren, nämlich die *Full-Methode* und die *Split-Methode*. Dabei soll gezeigt werden, dass bei fester Stichprobengröße die Überdeckungswahrscheinlichkeit der Konfidenzbereiche das gewünschte nominale Konfidenzniveau nicht unterschreitet. Um dies zu bewerkstelligen, werden die Resultate aus Kapitel 2 Verwendung finden. In Ergänzung dazu werden von uns *parametrische Konfidenzbereiche* sowie eine Variante der *Bootstrap-Methode* als Alternativen zu den konformalen Techniken kurz diskutiert. Wie schon eingangs erwähnt, finden die folgenden Überlegungen im Rahmen einer Regressionsanalyse statt.

3.1 Die Full-Methode

Wir betrachten in diesem Unterkapitel eine Folge $(X_i, Y_i)_{i \in \mathbb{N}}$ von Zufallsvariablen, wobei für jedes $i \in \mathbb{N}$ die Zufallsvariable X_i Werte aus $\mathbb{R}^d$ mit $d \in \mathbb{N}$ annehme und $\mathscr{A}$-$\mathscr{B}_d$-messbar sei. Die Zufallsvariable Y_i nehme Werte aus $\mathbb{R}$ an und sei $\mathscr{A}$-$\mathscr{B}_1$-messbar. Für jedes $i \in \mathbb{N}$ ist (X_i, Y_i) also eine $\mathbb{R}^{d+1}$-wertige Zufallsvariable und $\mathscr{A}$-$\mathscr{B}_{d+1}$-messbar. Im Bezug auf Regressionsanalysen werden wir X_i als (mehrdimensionalen) *Regressor* und Y_i als *Regressand* auffassen. Zum Zwecke der Übersichtlichkeit werden wir an manchen Stellen statt (X_i, Y_i) den kürzeren Ausdruck Z_i verwenden, d.h. für jedes $i \in \mathbb{N}$ sei $Z_i \equiv (X_i, Y_i)$. Beachte außerdem, dass wir in diesem Kapitel die Stichprobengröße n überwiegend als Konstante betrachten und damit das On-Line Setting sozusagen

„einfrieren“.

Für beliebiges $n \in \mathbb{N}$ steht $(\mathbb{R}^{d+1})^{n+1} \times \mathbb{R}^d$ für das kartesische Produkt zwischen dem $(n+1)$-fachen kartesischen Produkt von $\mathbb{R}^{d+1}$ und $\mathbb{R}^d$, welches somit als $[(d+1)(n+1)+d]$-dimensionaler reeller euklidischer Raum ausgestattet mit der dazugehörigen Borelschen σ-Algebra $\mathscr{B}_{(d+1)(n+1)+d}$ aufgefasst werden kann. Wir führen zunächst folgende Klasse von messbaren Funktionen ein:

Definition 3.1. *Für gegebenes $n \in \mathbb{N}$ heißt eine $\mathscr{B}_{(d+1)(n+1)+d}$-$\mathscr{B}_1$-messbare Funktion*

$$a_n : (\mathbb{R}^{d+1})^{n+1} \times \mathbb{R}^d \to \mathbb{R}$$

Algorithmus zur Full-Methode.

Eine Funktion a_n im Sinne von Definition 3.1 bildet also das kartesische Produkt zwischen dem $(n+1)$-fachen kartesischen Produkt von $\mathbb{R}^{d+1}$ und der Menge $\mathbb{R}^d$ in die Menge der reellen Zahlen $\mathbb{R}$ ab. Für beliebige $z_1, ..., z_{n+1} \in \mathbb{R}^{d+1}$ und $x \in \mathbb{R}^d$ schreiben wir außerdem den Funktionswert eines Algorithmus a_n in der Form

$$a_n(z_1, ..., z_{n+1}, x) \equiv a_n(z_1, ..., z_{n+1})(x).$$

Bei festen $z_1, \dots, z_{n+1}$ können wir also den Ausdruck $a_n(z_1, ..., z_{n+1})$ als Funktion, welche $\mathbb{R}^d$ in $\mathbb{R}$ abbildet, betrachten. Weil a_n als $\mathscr{B}_{(d+1)(n+1)+d}$-$\mathscr{B}_1$-messbar postuliert wird, ist $a_n(Z_1, ..., Z_{n+1})(X_i)$ für jedes $i \in \{1, ..., n+1\}$ eine $\mathscr{A}$-$\mathscr{B}_1$-messbare Funktion. Im Bezug auf $Z_1, ..., Z_{n+1}$ können wir damit für jedes $i \in \{1, ..., n+1\}$ die Zufallsvariable

$$U_{i,Z_{n+1}} := |Y_i - a_n(Z_1, ..., Z_{n+1})(X_i)|$$

definieren, welche $\mathscr{A}$-$\mathscr{B}_1$-messbar ist. Die Zufallsvariablen $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ sind also *Residuen* bezüglich der Variablen $Z_1, ..., Z_{n+1}$ und der Funktion a_n. Die Notation $U_{i,Z_{n+1}}$ soll dabei zum Ausdruck bringen, dass die Residuen insbesondere von der Zufallsvariable $Z_{n+1} = (X_{n+1}, Y_{n+1})$ abhängen, welche im On-Line Setting nicht vollständig beobachtbar ist, da der Wert von Y_{n+1} im Allgemeinen *unbekannt* ist und daher *prognostiziert* werden soll.

Im Kontext einer Regressionsanalyse steht ein Algorithmus zur Full-Methode a_n also für ein Verfahren, durch das eine *Regressionsfunktion* $a_n(z_1, ..., z_{n+1})$ auf gegebenen $\mathbb{R}^{d+1}$-wertigen Daten $z_1, ..., z_{n+1}$, welche in diesem Kontext eine Realisation der *vollen* Stichprobe darstellen, trainiert wird.[3] Über diese Regressionsfunktion können dann die entsprechenden Residuen berechnet werden.

Der nächste Begriff nimmt Bezug auf eine Eigenschaft, die viele wichtige Algorithmen aufweisen (etwa die *Methode der kleinsten Quadrate* oder das

Maximum-Likelihood-Verfahren): die *Permutationsinvarianz*. Diese definieren wir wie folgt:

Definition 3.2. *Ein Algorithmus a_n zur Full-Methode heißt für festes $n \in \mathbb{N}$ permutationsinvariant, wenn für alle $(z_1, ..., z_{n+1}) \in (\mathbb{R}^{d+1})^{n+1}$ und $x \in \mathbb{R}^d$ sowie für beliebige Permutationen $\sigma \in S_{n+1}$ gilt:*

$$a_n(z_1, ..., z_{n+1})(x) = a_n(z_{\sigma(1)}, ..., z_{\sigma(n+1)})(x).$$

In diesem Kontext bedeutet Permuationsinvarianz also, dass wir unabhängig von der Reihenfolge, in welche die Daten $z_1, ..., z_{n+1}$ als Argumente in einen Algorithmus a_n eingehen, immer die gleiche Regressionsfunktion erhalten. Eine wichtige Konsequenz, die sich daraus ergibt, kann im folgenden Lemma zusammengefasst werden:

Lemma 3.1. *Betrachte für $n \in \mathbb{N}$ eine endliche Folge von bezüglich P vertauschbaren $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen $Z_1, ..., Z_{n+1}$ und einen Algorithmus a_n zur Full-Methode. Weiters seien $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ die Residuen bezüglich $Z_1, ..., Z_{n+1}$ und a_n. Wenn der Algorithmus a_n permutationsinvariant ist, dann sind die Zufallsvariablen $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ bezüglich P vertauschbar.*

Beweis. Sei $n \in \mathbb{N}$ beliebig. Wir definieren zunächst eine Hilfsfunktion

$$f : (\mathbb{R}^{d+1})^{n+1} \to \mathbb{R}^{n+1}$$

der folgenden Bauart: für jedes $(z_1, ..., z_{n+1}) \in (\mathbb{R}^{d+1})^{n+1}$, wobei $z_i \equiv (x_i, y_i)$ mit $x_i \in \mathbb{R}^d$ und $y_i \in \mathbb{R}$ für jedes $i \in \{1, ..., n+1\}$, sei

$$f(z_1, ..., z_{n+1})$$

$$:= \left(|y_1 - a_n(z_1, ..., z_{n+1})(x_1)|, ..., |y_{n+1} - a_n(z_1, ..., z_{n+1})(x_{n+1})|\right).$$

Damit können wir den Vektor der Residuen wie folgt anschreiben:

$$(U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}})$$

$$= (|Y_1 - a_n(Z_1, ..., Z_{n+1})(X_1)|, ..., |Y_{n+1} - a_n(Z_1, ..., Z_{n+1})(X_{n+1})|)$$

$$= f(Z_1, ..., Z_{n+1}).$$

Betrachte nun eine Permutation $\sigma \in S_{n+1}$. Für den permutierten Vektor der Residuen gilt dann

$$(U_{\sigma(1),Z_{n+1}}, ..., U_{\sigma(n+1),Z_{n+1}})$$

$$= \left(|Y_{\sigma(1)} - a_n(Z_1, ..., Z_{n+1})(X_{\sigma(1)})|, ..., |Y_{\sigma(n+1)} - a_n(Z_1, ..., Z_{n+1})(X_{\sigma(n+1)})|\right).$$

Da a_n nach Voraussetzung permutationsinvariant ist, gilt gleichzeitig

$$a_n(Z_1, ..., Z_{n+1}) = a_n(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)}).$$

Somit ist

$$\begin{aligned}&(U_{\sigma(1),Z_{n+1}}, ..., U_{\sigma(n+1),Z_{n+1}})\\ &= \Big(|Y_{\sigma(1)} - a_n(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})(X_{\sigma(1)})|, ..., |Y_{\sigma(n+1)} - a_n(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})(X_{\sigma(n+1)})|\Big)\\ &= f(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)}).\end{aligned}$$

Da die Zufallsvariablen $Z_1, ..., Z_{n+1}$ vertauschbar sind, besitzen die Zufallsvektoren $(Z_1, ..., Z_{n+1})$ und $(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})$ bezüglich P die gleiche Verteilung. Gleichzeitig ist die Funktion f deterministisch, also unabhängig von $Z_1, ..., Z_{n+1}$ definiert, womit schließlich die Zufallsvektoren $(U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}})$ und $(U_{\sigma(1),Z_{n+1}}, ..., U_{\sigma(n+1),Z_{n+1}})$ bezüglich P in Verteilung gleich sind. □

Um einen konformalen Konfidenzbereich nach der Full-Methode zu konstruieren, betrachten wir für festes $n \in \mathbb{N}$ wieder einen Algorithmus a_n im Sinne von Definition 3.1 sowie $\mathbb{R}^{d+1}$-wertige Zufallsvariablen $Z_1, ..., Z_n, Z_{n+1}$ mit den dazu korrespondierenden Residuen $U_{1,Z_{n+1}}, ..., U_{n,Z_{n+1}}, U_{n+1,Z_{n+1}}$; vgl. G'Sell et al. (2016, S.6-9). Für beliebige $y \in \mathbb{R}$ sei außerdem

$$U_{i,(X_{n+1},y)} := |Y_i - a_n(Z_1, ..., Z_n, (X_{n+1}, y))(X_i)|,$$

wobei $i \in \{1, ..., n\}$, sowie

$$U_{n+1,(X_{n+1},y)} := |y - a_n(Z_1, ..., Z_n, (X_{n+1}, y))(X_{n+1})|.$$

Zudem stehen $U_{(1:n),(X_{n+1},y)}, ..., U_{(n:n),(X_{n+1},y)}$ für die Ordnungsstatistiken der Zufallsvariablen $U_{1,(X_{n+1},y)}, ..., U_{n,(X_{n+1},y)}$. Für beliebige $\alpha \in (0, 1)$, welche die Bedingung

$$(1 - \alpha)(n + 1) \leq n$$

erfüllen, kann also der nach der *Full-Methode* konstruierte konformale Konfidenzbereich für die unbekannte Variable Y_{n+1} in Abhängigkeit von der beobachtbaren Variable X_{n+1} wie folgt definiert werden:

$$C^{full}_{n,\alpha}(X_{n+1}) := \{y \in \mathbb{R} \,|\, U_{n+1,(X_{n+1},y)} \in [0, U_{(h:n),(X_{n+1},y)})\}, \tag{3.1}$$

wobei

$$h := \lceil (1 - \alpha)(n + 1) \rceil.$$

Die Menge $C_{n,\alpha}^{full}(X_{n+1})$ dient also als *Prognosebereich* für Y_{n+1} und setzt sich aus allen reellen Zahlen y zusammen, welche jeweils im offen Intervall

$$(a_n - U_{(h:n),(X_{n+1},y)}, a_n + U_{(h:n),(X_{n+1},y)})$$

mit der Abkürzung $a_n \equiv a_n(Z_1, ..., Z_n, (X_{n+1}, y))(X_{n+1})$ enthalten sind. Die Länge des Intervalls entspricht $2 \cdot U_{(h:n),(X_{n+1},y)}$ und wird durch die Zufallsvariable $2 \cdot U_{(h:n),(X_{n+1},Y_{n+1})}$, welche bei den asymptotische Betrachtungen in Unterkapitel 4.1 von Bedeutung sein wird, realisiert. In diesem Zusammenhang könnten man bei $2 \cdot U_{(h:n),(X_{n+1},Y_{n+1})}$ auch von der „Breite“ des Konfidenzbereichs $C_{n,\alpha}^{full}(X_{n+1})$ sprechen, wobei an dieser Stelle anzumerken ist, dass es sich bei $C_{n,\alpha}^{full}(X_{n+1})$ nicht zwingend um ein zusammenhängendes Intervall reeller Zahlen handelt. Wir sehen zudem, dass bei der Full-Methode das Residuum $U_{n+1,(X_{n+1},y)}$ die Rolle des *Nichtkonformitätsmaßes* spielt: ein gegebenes y fällt also nicht in die Menge $C_{n,\alpha}^{full}(X_{n+1})$, wenn die „Diskrepanz“ zu den in der Regressionsfunktion $a_n(Z_1, ..., Z_n, (X_{n+1}, y))(X_{n+1})$ zusammengefassten Daten zu „ausgeprägt“ ist, was in diesem Kontext nichts anderes heißt, als dass $U_{n+1,(X_{n+1},y)}$ einen Wert, der nicht unter der Ordnungsstatistik $U_{(h:n),(X_{n+1},y)}$ liegt, realisiert. Für den Fall

$$(1 - \alpha)(n + 1) > n$$

definieren wir noch

$$C_{n,\alpha}^{full}(X_{n+1}) := \mathbb{R}.$$

Dies ist jedoch nur für endlich viele Stichprobengrößen $n \in \mathbb{N}$ von Bedeutung, da die Bedingung $(1 - \alpha)(n + 1) > n$ äquivalent ist zu

$$\frac{1 - \alpha}{\alpha} > n,$$

und die Menge der natürlichen Zahlen $\mathbb{N}$ nach oben unbeschränkt ist.

Der nächste Satz fasst die Bedienungen zusammen, unter denen $C_{n,\alpha}^{full}(X_{n+1})$ ein Konfidenzbereich für Y_{n+1} zum nominellen Konfidenzniveau $1 - \alpha$ ist:

Satz 3.1. *Betrachte für $n \in \mathbb{N}$ eine endliche Folge von bezüglich P vertauschbaren $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen $Z_1, ..., Z_{n+1}$ und einen permutationsinvarianten Algorithmus zur Full-Methode a_n, wobei unter den dazugehörigen Residuen $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ bezüglich P fast sicher keine Bindungen vorkommen. Für jedes $\alpha \in (0, 1)$ ist dann $C_{n,\alpha}^{full}(X_{n+1})$ ein Konfidenzbereich für Y_{n+1} zum nominellen Konfidenzniveau $1 - \alpha$, d.h.*

$$P(Y_{n+1} \in C_{n,\alpha}^{full}(X_{n+1})) \geq 1 - \alpha. \tag{3.2}$$

Es gilt außerdem

$$P(Y_{n+1} \in C_{n,\alpha}^{full}(X_{n+1})) < 1 - \alpha + \frac{1}{n+1}. \tag{3.3}$$

Beweis. Betrachte ein beliebiges $n \in \mathbb{N}$. Im Fall

$$(1-\alpha)(n+1) > n$$

ist $C_{n,\alpha}^{full}(X_{n+1}) = \mathbb{R}$, womit die Ungleichungen (3.2) und (3.3) trivialerweise erfüllt sind. Wir untersuchen den Fall

$$(1-\alpha)(n+1) \leq n.$$

Es seien $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ die Residuen bezüglich der Variablen $Z_1, ..., Z_n, Z_{n+1}$ und des Algorithmus a_n, sowie $U_{(1:n),Z_{n+1}}, ..., U_{(n:n),Z_{n+1}}$ die Ordnungsstatistiken zu $U_{1,Z_{n+1}}, ..., U_{n,Z_{n+1}}$. Es gilt

$$\{Y_{n+1} \in C_{n,\alpha}^{full}(X_{n+1})\}$$

$$= \{U_{n+1,(X_{n+1},Y_{n+1})} \in [0, U_{(h:n),(X_{n+1},Y_{n+1})})\}$$

$$= \{U_{n+1,Z_{n+1}} \in [0, U_{(h:n),Z_{n+1}})\}.$$

Nach Voraussetzung sind die Zufallsvariablen $Z_1, ..., Z_{n+1}$ vertauschbar und a_n ein permutationsinvarianter Algorithmus. Wegen Lemma 3.1 sind also $U_{1,Z_{n+1}}, ..., U_{n,Z_{n+1}}, U_{n+1,Z_{n+1}}$ vertauschbar. Nach Anwendung von Satz 2.1 erhalten wir

$$P(U_{n+1,Z_{n+1}} \in (-\infty, U_{(h:n),Z_{n+1}})) \geq 1 - \alpha$$

und

$$P(U_{n+1,Z_{n+1}} \in (-\infty, U_{(h:n),Z_{n+1}})) < 1 - \alpha + \frac{1}{n+1}.$$

Da $U_{n+1,Z_{n+1}}$ auf ganz Ω nicht-negativ ist, gilt natürlich

$$\{U_{n+1,Z_{n+1}} \in [0, U_{(h:n),Z_{n+1}})\}$$

$$= \{U_{n+1,Z_{n+1}} \in (-\infty, U_{(h:n),Z_{n+1}})\},$$

und somit ergeben sich die Ungleichungen (3.2) und (3.3). □

Für unendliche Folgen $(Z_i)_{i\in\mathbb{N}}$ von vertauschbaren Zufallsvariablen und einen permutationsinvarianten Algorithmus zur Full-Methode a_n, wobei für jedes $n \in \mathbb{N}$ die Residuen bezüglich $Z_1, ..., Z_{n+1}$ und a_n keine Bindungen aufweisen, ergibt sich aus den Ungleichungen (3.2) und (3.3) unmittelbar

$$\lim_{n\to\infty} P(Y_{n+1} \in C_{n,\alpha}^{full}(X_{n+1})) = 1 - \alpha,$$

d.h. die Überdeckungswahrscheinlichkeit des konformalen Konfidenzbereiches $C_{n,\alpha}^{full}(X_{n+1})$ konvergiert für wachsende Stichprobengröße n gegen das nominelle Konfidenzniveau.

Um einen konformalen Konfidenzbereich nach der Full-Methode zu bestimmen, muss für jedes $y \in \mathbb{R}$ bzw. jedes $y \in \mathscr{Y} \subseteq \mathbb{R}$, falls für jedes $i \in \mathbb{N}$ die Zufallsvariable Y_i auf ganz Ω Werte aus der Teilmenge $\mathscr{Y}$ annimmt, zu gegebenen Umfang $n \in \mathbb{N}$ der Algorithmus a_n auf den gesamten vorgegebenen Datensatz $(x_1, y_1), ..., (x_n, y_n), (x_{n+1}, y)$ von Elementen aus der Menge $\mathbb{R}^{d+1}$ angewendet werden, um eine Regressionsfunktion zu bilden. Im Anschluss müssen dann die entsprechenden $n + 1$ Residuen bestimmt werden. Diese Vorgehensweise berücksichtigt bei der Berechnung der Regressionsfunktion auch die möglichen Werte y, die von der zu prognostizierenden Variable Y_{n+1} angenommen werden können, was vermutlich einen geringeren Fehler in der Schätzung einer möglichen Relation zwischen Regressoren und Regressanden durch die Full-Methode nach sich ziehen könnte. Allerdings kann diese Prozedur bei großem n oder bei einer hohen Anzahl an Elementen in $\mathscr{Y}$ mit *hohem* Rechenaufwand verbunden sein (siehe auch Unterkapitel 5.5). Eine Alternative bietet die Split-Methode, welche im nächsten Unterkapitel diskutiert wird.

3.2 Die Split-Methode

Wir betrachten für festes $d \in \mathbb{N}$ wieder eine Folge $(X_i, Y_i)_{i \in \mathbb{N}}$, wobei wieder für alle $i \in \mathbb{N}$ die Zufallsvariable X_i als $\mathbb{R}^d$-wertig angenommen sei und die Zufallsvariable Y_i als $\mathbb{R}$-wertig (wobei diese Zufallsvariablen die gleichen Messbarkeitseigenschaften erfüllen wie in Unterkapietl 3.1). Die Variable X_i werden wir wieder als *Regressor* und die Variable Y_i als *Regressand* auffassen. Zur besseren Übersicht wird auch in diesem Unterkapitel an einigen Stellen der Ausdruck Z_i statt (X_i, Y_i) Verwendung finden.

Das Besondere an der Split-Methode ist, dass sie für jedes $n \in \mathbb{N}$ mit $n \geq 2$ auf der Indexmenge $I_n := \{1, ..., n\}$ eine Partition, bestehend aus zwei nicht leeren disjunkten Teilmengen $I_{n,1}$ und $I_{n,2}$, bestimmt. In dieser Arbeit wird folgende Prozedur zur Darstellung der Mengen $I_{n,1}$ und $I_{n,2}$ umgesetzt: es sei $\delta \in [\frac{1}{2}, 1)$ eine feste Zahl und für jedes $n \in \mathbb{N}$ mit $n \geq 2$ definieren wir

$$m := \lfloor \delta n \rfloor,$$

wobei

$$\lfloor x \rfloor := max\{z \in \mathbb{Z} \,|\, z \leq x\}$$

die *Abrundungsfunktion* für beliebige $x \in \mathbb{R}$ sei. Die Mengen $I_{n,1}$ und $I_{n,2}$ stellen wir dann wie folgt dar:

$$I_{n,1} := \{i_1, ..., i_m\}$$

und

$$I_{n,2} := \{j_1, ..., j_k\},$$

wobei

$$k := n - m.$$

Die (paarweise verschiedenen) Elemente von I_n, welche in $I_{n,1}$ liegen, seien also durch $i_1, ..., i_m$ bezeichnet, während wir jene, die in $I_{n,2}$ liegen (und verschieden sind von jenen Elementen, die in $I_{n,1}$ enthalten sind), durch $j_1, ..., j_k$ darstellen. Beachte, dass aus $\delta < 1$ die Ungleichung $m < n$ und somit $k \geq 1$ folgt. Wegen $\delta \geq \frac{1}{2}$ gilt gleichzeitig $m \geq 1$. Die Bedingung $\delta \in [\frac{1}{2}, 1)$ stellt also sicher, dass für jedes $n \in \mathbb{N}$ mit $n \geq 2$ sowohl $I_{n,1}$ als auch $I_{n,2}$ nicht leer sind. Bei einer gegebenen endlichen Folge von $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen $Z_1, ..., Z_n$ erfolgt also ein „Split“ derselben in $Z_{i_1}, ..., Z_{i_m}$ und $Z_{j_1}, ..., Z_{j_k}$. Zur Formierung von Regressionsfunktionen sei ein *Algorithmus* verwendet, den wir im Rahmen der Split-Methode wie folgt einführen:

Definition 3.3. *Für $n \in \mathbb{N}$ mit $n \geq 2$ und $\delta \in [\frac{1}{2}, 1)$, wobei $m := \lfloor \delta n \rfloor$, heißt eine $\mathscr{B}_{(d+1)m+d}$-$\mathscr{B}_1$-messbare Funktion*

$$a_{n,\delta} : (\mathbb{R}^{d+1})^m \times \mathbb{R}^d \to \mathbb{R}$$

Algorithmus zur Split-Methode.

Für alle $z_1, ..., z_m \in \mathbb{R}^{d+1}$ und $x \in \mathbb{R}^d$ schreiben wir den Funktionswert eines Algorithmus zur Split-Methode $a_{n,\delta}$ als

$$a_{n,\delta}(z_1, ..., z_m, x) \equiv a_{n,\delta}(z_1, ..., z_m)(x),$$

womit wir $a_{n,\delta}(z_1, ..., z_m)$ für feste $z_1, ..., z_m$ als Abbildung von $\mathbb{R}^d$ in $\mathbb{R}$ auffassen können. Da wir außerdem $a_{n,\delta}$ als $\mathscr{B}_{(d+1)m+d}$-$\mathscr{B}_1$-messbar angenommen haben, ist $a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_i)$ für jedes $i \in I_{n,2} \cup \{n+1\}$ eine $\mathscr{A}$-$\mathscr{B}_1$-messbare Funktion. Damit können wir für jedes $i \in I_{n,2} \cup \{n+1\}$ das *Residuum*

$$U_{i,\delta} := |Y_i - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_i)|$$

als $\mathscr{A}$-$\mathscr{B}_1$-messbare Zufallsvariable definieren.

Um eine Regressionsfunktion zu formieren, verwendet ein Algorithmus $a_{n,\delta}$ im Sinne von Definition 3.3 von vorgegebenen $\mathbb{R}^{d+1}$-wertigen Daten $z_1, ..., z_n$

also nur einen *Teil* als Trainingsdaten, nämlich $z_{i_1}, ..., z_{i_m}$. Residuen werden nach der Bestimmung von $a_{n,\delta}(z_{i_1}, ..., z_{i_m})$ hingegen nur bezüglich $z_{j_1}, ..., z_{j_k}$ sowie bezüglich z_{n+1} berechnet (im Gegensatz zur Full-Methode, die den *vollen* Datensatz als Trainingsmenge verwendet und zur Berechnung der Residuen heranzieht).

Für die Residuen erhalten wir folgendes Resultat:

Lemma 3.2. *Für $n \in \mathbb{N}$ mit $n \geq 2$ und $\delta \in [\frac{1}{2}, 1)$, wobei $m := \lfloor \delta n \rfloor$, betrachte eine Folge von $\mathbb{R}^{d+1}$-wertigen bezüglich P vertauschbaren Zufallsvariablen $Z_1, ..., Z_{n+1}$ und einen Algorithmus zur Split-Methode $a_{n,\delta}$. Die mittels $Z_1, ..., Z_n, Z_{n+1}$ und $a_{n,\delta}$ bestimmten Residuen $U_{j_1,\delta}, ..., U_{j_k,\delta}, U_{n+1,\delta}$ sind vertauschbar.*

Beweis. Sei $n \in \mathbb{N}$ mit $n \geq 2$ beliebig. Zu Hilfszwecken definieren wir eine Funktion

$$f : (\mathbb{R}^{d+1})^{n+1} \to \mathbb{R}^{k+1}.$$

An jeder Stelle $(z_1, ..., z_{n+1}) \in (\mathbb{R}^{d+1})^{n+1}$, wobei $z_i \equiv (x_i, y_i)$ mit $x_i \in \mathbb{R}^d$ und $y_i \in \mathbb{R}$ für alle $i \in \{1, ..., n+1\}$, sei f definiert durch

$$f(z_1, ..., z_m, z_{m+1}, ..., z_{n+1})$$

$$:= (|y_{m+1} - a_{n,\delta}(z_1, ..., z_m)(x_{m+1})|, ..., |y_{n+1} - a_{n,\delta}(z_1, ..., z_m)(x_{n+1})|)\,.$$

Sei wieder $k := n - m$ definiert. Wenn wir $j_{k+1} := n + 1$ festsetzen, dann können wir den Vektor der Risiduen wie folgt schreiben:

$$(U_{j_1,\delta}, ..., U_{j_{k+1},\delta})$$

$$= (|Y_{j_1} - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{j_1})|, ..., |Y_{j_{k+1}} - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{j_{k+1}})|)$$

$$= f(Z_{i_1}, ..., Z_{i_m}, Z_{j_1}, ..., Z_{j_{k+1}}).$$

Für jede Permutation $\sigma \in S_{k+1}$ gilt dann

$$(U_{j_{\sigma(1)},\delta}, ..., U_{j_{\sigma(k+1)},\delta})$$

$$= \left(|Y_{j_{\sigma(1)}} - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{j_{\sigma(1)}})|, ..., |Y_{j_{\sigma(k+1)}} - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{j_{\sigma(k+1)}})|\right)$$

$$= f(Z_{i_1}, ..., Z_{i_m}, Z_{j_{\sigma(1)}}, ..., Z_{j_{\sigma(k+1)}}).$$

Da die Zufallsvariablen $Z_1, ..., Z_{n+1}$ bezüglich P vertauschbar sind, besitzen die beiden Zufallsvektoren

$$(Z_{i_1}, ..., Z_{i_m}, Z_{j_1}, ..., Z_{j_{k+1}})$$

und

$$(Z_{i_1}, ..., Z_{i_m}, Z_{j_{\sigma(1)}}, ..., Z_{j_{\sigma(k+1)}})$$

bezüglich P die gleiche Verteilung. Gleichzeitig ist die Funktion f deterministisch, also unabhängig von $Z_1, ..., Z_{n+1}$ definiert, womit schließlich die Zufallsvektoren $(U_{j_1,\delta}, ..., U_{j_{k+1},\delta})$ und $(U_{j_{\sigma(1)},\delta}, ..., U_{j_{\sigma(k+1)},\delta})$ bezüglich P in Verteilung gleich sind.

□

Bemerkenswert ist dabei der Umstand, dass im Gegensatz zu Lemma 3.1 aus dem vorhergehenden Unterkapitel in Lemma 3.2 Permuationsinvarianz des gegebenen Algorithmus nicht gefordert wird.

Wir sind nun endlich in der Lage konformale Konfidenzbereiche nach der Split-Methode zu bilden; vgl. G'Sell et al. (2016, S.9-10). Dazu betrachten wir für festes $n \in \mathbb{N}$ mit $n \geq 2$ einen Algorithmus $a_{n,\delta}$ sowie $\mathbb{R}^{d+1}$-wertige Zufallsvariablen $Z_1, ..., Z_n, Z_{n+1}$ mit den dazugehörigen Residuen $U_{j_1,\delta}, ..., U_{j_k,\delta}, U_{n+1,\delta}$. Dabei stehen die Zufallsvariablen $U_{(1:k),\delta}, ..., U_{(k:k),\delta}$ für die Ordnungsstatistiken von $U_{j_1,\delta}, ..., U_{j_k,\delta}$. Für beliebiges $y \in \mathbb{R}$ definieren wir außerdem

$$U_{n+1,\delta,y} := |y - a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{n+1})|\,.$$

Bei gegebenem $\alpha \in (0, 1)$, welches die Bedingung

$$(1 - \alpha)(k + 1) \leq k$$

erfüllt, kann also der nach der *Split-Methode* konstruierte konformale Konfidenzbereich, welcher als Prognosebereich für die unbekannte Variable Y_{n+1} in Abhängigkeit von der beobachtbaren Variable X_{n+1} dient, wie folgt definiert werden:

$$C^{split}_{n,\delta,\alpha}(X_{n+1}) := \{y \in \mathbb{R} \,|\, U_{n+1,\delta,y} \in [0, U_{(\tilde{h}:k),\delta})\}, \tag{3.4}$$

wobei

$$\tilde{h} := \lceil (1 - \alpha)(k + 1) \rceil.$$

Hier kann nun $U_{n+1,\delta,y}$ als *Nichtkonformitätsmaß* aufgefasst werden: für gegebenes y bedeutet eine zu ausgeprägte „Diskrepanz“ zu $a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{n+1})$ bzw. ein Wert von $U_{n+1,\delta,y}$ größer als $U_{(\tilde{h}:k),\delta}$, dass y nicht in der Menge $C^{split}_{n,\delta,\alpha}(X_{n+1})$ liegt. Für den Fall

$$(1 - \alpha)(k + 1) > k$$

definieren wir noch

$$C^{split}_{n,\delta,\alpha}(X_{n+1}) := \mathbb{R}.$$

Dieser Spezialfall betrifft wieder nur endlich viele $n \in \mathbb{N}$, weil $(1-\alpha)(k+1) > k$ äquivalent zu

$$\frac{1-\alpha}{\alpha} > k$$

ist und k als Folge natürlicher Zahlen bezüglich n nach oben unbeschränkt ist. Um das zu sehen sei daran erinnert, dass $m = \lfloor \delta n \rfloor$. Es gilt also $m \leq \delta n$ und damit

$$n - m \geq n(1-\delta).$$

Aus $k = n - m$ folgt nun

$$k \geq n(1-\delta).$$

Weil $\delta < 1$ konvergiert der rechte Term in dieser Ungleichung für $n \to \infty$ gegen ∞, womit auch k als Folge in n für $n \to \infty$ gegen ∞ strebt.

Der nächste Satz besagt, dass unter der Bedingung der Vertauschbarkeit $C^{split}_{n,\delta,\alpha}(X_{n+1})$ ein Konfidenzbereich für Y_{n+1} zum nominellen Konfidenzniveau $1-\alpha$ ist:

Satz 3.2. *Für $n \in \mathbb{N}$ mit $n \geq 2$ und $\delta \in [\frac{1}{2}, 1)$, wobei $k := n - \lfloor \delta n \rfloor$, betrachte eine Folge von vertauschbaren $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen $Z_1, ..., Z_{n+1}$ und einen Algorithmus zur Split-Methode $a_{n,\delta}$, wobei unter den dazugehörigen Residuen $U_{j_1,\delta}, ..., U_{j_k,\delta}, U_{n+1,\delta}$ bezüglich P fast sicher keine Bindungen vorkommen. Für jedes $\alpha \in (0,1)$ ist dann $C^{split}_{n,\delta,\alpha}(X_{n+1})$ ein Konfidenzbereich für Y_{n+1} zum nominellen Konfidenzniveau $1-\alpha$, d.h.*

$$P(Y_{n+1} \in C^{split}_{n,\delta,\alpha}(X_{n+1})) \geq 1 - \alpha. \tag{3.5}$$

Gleichzeitig gilt die Ungleichung

$$P(Y_{n+1} \in C^{split}_{n,\delta,\alpha}(X_{n+1})) < 1 - \alpha + \frac{1}{k+1}. \tag{3.6}$$

Beweis. Für den Fall

$$(1-\alpha)(k+1) > k$$

ist $C^{split}_{n,\delta,\alpha}(X_{n+1}) = \mathbb{R}$, womit die Ungleichungen (3.5) und (3.6) trivialerweise erfüllt sind. Wir betrachten also den Fall

$$(1-\alpha)(k+1) \leq k.$$

Es seien $U_{(1:k),\delta}, ..., U_{(k:k),\delta}$ die Ordnungsstatistiken von $U_{j_1,\delta}, ..., U_{j_k,\delta}$. Es gilt dann

$$\begin{aligned} &\{Y_{n+1} \in C^{split}_{n,\delta,\alpha}(X_{n+1})\} \\ &= \{U_{n+1,\delta,Y_{n+1}} \in [0, U_{(\tilde{h}:k),\delta})\}. \end{aligned}$$

Nach Voraussetzung ist $Z_1, ..., Z_{n+1}$ eine Folge bezüglich P vertauschbarer Zufallsvariablen. Wegen Lemma 3.2 sind somit $U_{j_1,\delta}, ..., U_{j_k,\delta}, U_{n+1,\delta}$ bezüglich P vertauschbar. Nach Voraussetzung weisen außerdem die Residuen $U_{j_1,\delta}, ..., U_{j_k,\delta}, U_{n+1,\delta}$ bezüglich P keine Bindungen auf. Durch Anwendung von Satz 2.1 erhalten wir

$$P(U_{n+1,\delta,Y_{n+1}} \in (-\infty, U_{(\tilde{h}:k),\delta})) \geq 1 - \alpha$$

und

$$P(U_{n+1,\delta,Y_{n+1}} \in (-\infty, U_{(\tilde{h}:k),\delta})) < 1 - \alpha + \frac{1}{k+1}.$$

Da $U_{n+1,\delta,Y_{n+1}}$ auf ganz Ω nicht-negativ ist, gilt natürlich

$$\{U_{n+1,\delta,Y_{n+1}} \in [0, U_{(\tilde{h}:k),\delta})\}$$

$$= \{U_{n+1,\delta,Y_{n+1}} \in (-\infty, U_{(\tilde{h}:k),\delta})\},$$

womit wir die Ungleichungen (3.5) und (3.6) erhalten. □

Für eine unendliche Folge $(Z_i)_{i\in\mathbb{N}}$ vertauschbarer Zufallsvariablen und einen Algorithmus zur Split-Methode $a_{n,\delta}$, wobei für jedes $n \in \mathbb{N}$ mit $n \geq 2$ die entsprechenden Residuen bezüglich $Z_1, ..., Z_{n+1}$ und $a_{n,\delta}$ fast sicher keine Bindungen aufweisen, ergibt sich aus den Ungleichungen (3.5) und (3.6) sowie aus dem Umstand, dass $k \to \infty$ für $n \to \infty$ gilt, unmittelbar

$$\lim_{n\to\infty} P(Y_{n+1} \in C^{split}_{n,\delta,\alpha}(X_{n+1})) = 1 - \alpha,$$

d.h. wie bei der Full-Methode konvergiert auch die Überdeckungswahrscheinlichkeit des konformalen Konfidenzbereichs $C^{split}_{n,\delta,\alpha}(X_{n+1})$ für wachsende Stichprobengrößen n gegen das nominelle Konfidenzniveau. Außerdem lässt sich $C^{split}_{n,\delta,\alpha}(X_{n+1})$ auch als (im topologischen Sinn) offenes Intervall schreiben. Wegen der Definition von $U_{n+1,\delta,y}$, die wir für beliebige $y \in \mathbb{R}$ eingeführt haben, gilt nämlich

$$C^{split}_{n,\delta,\alpha}(X_{n+1}) = (a_{n,\delta} - U_{(\tilde{h}:k),\delta}, a_{n,\delta} + U_{(\tilde{h}:k),\delta}),$$

mit der Abkürzung $a_{n,\delta} \equiv a_{n,\delta}(Z_{i_1}, ..., Z_{i_m})(X_{n+1})$. Wir können also im Bezug auf $C^{split}_{n,\delta,\alpha}(X_{n+1})$ berechtigterweise von einem konformalen Konfidenzintervall für Y_{n+1} mit der Breite $2 \cdot U_{(\tilde{h}:k),\delta}$ sprechen. Dessen Asymptotik wird genauer in Unterkapitel 4.3 untersucht.

An dieser Stelle sei noch einmal hervorgehoben, dass in Satz 3.2 Permutationsinvarianz des gegebenen Algorithmus nicht gefordert wird. Dies hat zur Folge, dass die Split-Methode im Allgemeinen für eine *breitere* Klasse

von Algorithmen, welche auch nicht permutationsinvariante umfasst, gültige Konfidenzbereiche hervorbringt. Dies steht im Gegensatz zur Full-Methode, die im Allgemeinen für permuationsinvariante Algorithmen Konfidenzbereiche mit der gewünschten Überdeckungswahrscheinlichkeit hervorbringt (vergleiche auch Satz 3.1). Dies könnte man als „Vorteil“ der Split-Methode gegenüber der Full-Methode auffassen. Die praktische Überlegenheit der Split-Methode liegt aber vor allem darin, dass das Trainieren der Regressionsfunktion nur mehr noch auf einer *Teilmenge* der Daten beruht, während die restlichen Daten zur Bestimmung der Residuen herangezogen werden. Bei hohem Stichprobenumfang n kann dies in Abhängigkeit von der Anzahl der Elemente im Wertebereich $\mathscr{Y}$ des Regressanden und der Art des Trainingsalgorithmus (sowie bei Anwendung einer nicht zu komplexen Methode zur Separierung der Trainingsdaten von den restlichen Daten) eine erhebliche Ersparnis an Rechenaufwand gegenüber der Full-Methode bedeuten (siehe auch Unterkapitel 5.5). Ein Nachteil der Split-Methode ist hingegen ein unter Umständen höherer Schätzfehler durch Anwendung der Regressionsfunktion, da für deren Formierung nicht der volle Datensatz als Trainingsmenge zur Verfügung steht. Dieser Umstand kann bei gegebenen n zu „breiteren“ bzw. „unschärferen“ Konfidenzbereichen für die zu prognostizierende Variable Y_{n+1} führen.

3.3 Parametrische Konfidenzintervalle und eine nichtparametrische Boostrap-Methode als Alternativen

In diesem Unterkapietl wollen kurz zwei Alternativen zu den in den beiden vorhergehenden Unterkapiteln beschriebenen konformalen Methoden zur Generierung von Konfidenzbereichen diskutieren, nämlich die auf *parametrischen* Verteilungen basierenden Konfidenzintervalle und das von Efron (1979) vorgeschlagene *Boostrap-Verfahren*, welches auf *Resampling* der originären Stichprobe beruht und sich ebenfalls zur Erzeugung von Konfidenzintervallen eignet. Dazu betrachten wir für gegebene Stichprobengröße $n \in \mathbb{N}$ wieder die Zufallsvariablen $Z_1, ..., Z_n$ mit $Z_i \equiv (X_i, Y_i)$ sowie eine Regressionsfunktion

$$\hat{r}_n \equiv \hat{r}_n(Z_1, ..., Z_n),$$

welche aus einem wohldefinierten Algorithmus, der die Realisationen der obigen Zufallsvariablen als Trainingsdaten nutzt, hervorgeht. Des Weiteren betrachten wir für Y_{n+1} den *Prognosefehler*

$$PF_{n+1} := Y_{n+1} - \hat{r}_n(X_{n+1}).$$

Wir haben in den letzten Unterkapiteln gesehen, dass die Full- und die Split-Methode gültige Konfidenzbereiche hervorbringt unter der Bedingung, dass bezüglich des gegebenen Wahrscheinlichkeitsmaßes P Vertauschbarkeit gegeben ist und die auf Grundlage einer Regressionsfunktion bestimmten Residuen keine Bindungen aufweisen (wobei im Bezug auf die Full-Methode noch zusätzlich Permutationsinvarianz des verwendeten Trainingsalgorithmus gefordert wird). Die beiden konformalen Methoden können als *nichtparametrische* Verfahren angesehen werden und zwar in dem Sinne, dass dem Prognosefehler keine spezifische parametrische Verteilung zu Grunde liegen muss. Anders verhält es sich bei *parametrischen* Konfidenzintervallen, von denen wir einige kurz betrachten; vgl. auch Efron und Tibshirani (1993, S.153-159). Es sei ρ_n^2 die von n abhängige Varianz der Zufallsvariable PF_{n+1}, welche wir unter P als endlich annehmen. Für gegebenes $\alpha \in (0,1)$ betrachten wir außerdem das $(1-\frac{\alpha}{2})$-Quantil der *Standardnormalverteilung*, welches wir mit $z_{1-\frac{\alpha}{2}}$ bezeichnen. Wir definieren das Intervall

$$I_{z,n,\alpha} := \left[\hat{r}_n(X_{n+1}) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\rho_n^2}, \hat{r}_n(X_{n+1}) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\rho_n^2}\right].$$

Unter der Annahme, dass PF_n einer *Normalverteilung* mit Erwartungswert 0 folgt, gilt für den standardisierten Prognosefehler

$$\frac{PF_{n+1}}{\sqrt{\rho_n^2}} \sim N(0,1),$$

wobei $N(0,1)$ für die Standardnormalverteilung stehe. Daraus folgt

$$P(Y_{n+1} \in I_{z,n,\alpha}) = 1 - \alpha,$$

d.h. $I_{z,n,\alpha}$ ist ein (exakt gültiges) Konfidenzintervall zum Konfidenzniveau $1-\alpha$. In nahezu allen praktischen Fällen werden wir ρ_n^2 als unbekannt vorfinden. Ein Ausweg bietet sich an, wenn ein Schätzer $\hat{\rho}_n^2$ für die Varianz ρ_n^2 verfügbar ist, sodass

$$\frac{PF_{n+1}}{\sqrt{\hat{\rho}_n^2}} \sim t(k_n)$$

gilt. Der Ausdruck $t(k_n)$ steht für die *t-Verteilung* mit k_n Freiheitsgraden, welche von n abhängen. Unter diesen Umständen ist

$$I_{t,n,\alpha} := \left[\hat{r}_n(X_{n+1}) - t_{k_n,1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\rho}_n^2}, \hat{r}_n(X_{n+1}) + t_{k_n,1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\rho}_n^2}\right]$$

mit $t_{k_n,1-\frac{\alpha}{2}}$ als $(1-\frac{\alpha}{2})$-Quantil der obigen t-Verteilung ein (exakt gültiges) Konfidenzintervall für Y_{n+1} zum nominellen Niveau $1-\alpha$. Dieses entspricht

der Form nach jenen im Rahmen von linearen Modellen formulierten Prognoseintervallen, welche in ökonometrischen Lehrbücher häufig anzutreffen sind (siehe zum Beispiel von Auer (2003) oder Wooldridge (2003)). Sofern außerdem die Zufallsvariable

$$\frac{PF_{n+1}}{\sqrt{\rho_n^2}}$$

asymptotisch *standardnormalverteilt* ist (also bezüglich P für $n \to \infty$ *in Verteilung* gegen eine standardnormalverteilte Zufallsvariable konvergiert) und für einen Varianzschätzer $\hat{\rho}_n^2$ der Quotient $\rho_n^2/\hat{\rho}_n^2$ *in Wahrscheinlichkeit* gegen den Wert 1 konvergiert, dann erhalten wir mit

$$I_{\tilde{z},n,\alpha} := \left[\hat{r}_n(X_{n+1}) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\rho}_n^2}, \hat{r}_n(X_{n+1}) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\rho}_n^2}\right]$$

ein asymptotisch gültiges Konfidenzintervall für Y_{n+1} zum Konfidenzniveau $1-\alpha$ im Sinne von

$$\liminf_{n\to\infty} P(Y_{n+1} \in I_{\tilde{z},n,\alpha}) \geq 1-\alpha.$$

Die Validität der soeben besprochenen Konfidenzintervalle speist sich aus dem Postulat, dass der Prognosefehler auf die eine oder andere Weise einer parametrischen Verteilung folgt.

Eine Alternative zu den konformalen Methoden, die nicht auf parametrische Verteilungen beruht, basiert auf ein *nichtparametrisches* Bootstrap-Verfahren, das durch Anwendung der *Monte-Carlo Methode* seine praktische Umsetzung erfährt. Im Zusammenhang mit *Regressionsanalysen* bieten sich zur Umsetzung dieser Methode zwei Alternativen an; vgl. Efron (1982, S.35-36). In einer davon wird ein gegebener Trainingsalgorithmus unmittelbar auf eine sogenannte *Bootstrap-Stichprobe* angewendet, die durch *Resampling* der gegebenen Stichprobe erzeugt wird; siehe auch Friedman et al. (2001, S.217-220). Die andere Alternative beruht auf Resampling der Residuen, die mit einer von der Ausgangsstichprobe abhängigen Regressionsfunktion bestimmt werden. In dieser Arbeit beschränken wir uns auf die Skizzierung dieses Ansatzes. Wir folgen dabei einer Vorgehensweise, die von Davison und Hinkley (1997, S.284-286) vorgeschlagen wird und welche im Rahmen einer linearen Regression erfolgt. Zu diesem Zweck stellen wir für alle $i \in \mathbb{N}$ den Regressor X_i als d-dimensionalen Vektor

$$X_i \equiv (X_{i,1}, ..., X_{i,d})$$

dar und unterstellen die Existenz eines linearen Modells der Gestalt

$$Y_i \equiv \beta_0 + \sum_{k=1}^{d} \beta_k \cdot X_{i,k} + \epsilon_i,$$

wobei $\beta_0, \beta_1, ..., \beta_d \in \mathbb{R}$ feste Koeffizienten seien und $(\epsilon_i)_{i\in\mathbb{N}}$ sei eine Folge von Störgrößen, die bezüglich P unabhängig und identisch verteilt sind mit 0 als gemeinsamen Erwartungswert und σ^2 als gemeinsame Varianz. Zur Vereinfachung nehmen Davison und Hinkley (1997) noch an, dass der Regressor X_i nicht-zufällig sei. Wir fassen

$$\mathbf{X}_i := \begin{pmatrix} 1 & X_{1,1} & ... & X_{1,d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{i,1} & ... & X_{i,d} \end{pmatrix}$$

als Designmatrix des Modells auf und definieren den i-dimensionalen Spaltenvektor

$$\mathbf{Y}_i := (Y_1, ..., Y_i)^T.$$

Dabei sei angenommen, dass die Matrix $\mathbf{X}_i$ den Rang $d+1$ besitzt. Für festes $n \in \mathbb{N}$ schätzen wir die Koeffizienten nach der *Kleinsten-Quadrate-Methode* (welche dem hier angewendeten Trainingsalgorithmus entspricht) und erhalten

$$\hat{\beta}_n := ((\mathbf{X}_n)^T \cdot \mathbf{X}_n)^{-1} \cdot (\mathbf{X}_n)^T \cdot \mathbf{Y}_n.$$

Unsere Regressionsfunktion $\hat{r}_n$ hat somit die Form

$$\hat{r}_n(x) = x \cdot \hat{\beta}_n,$$

wobei $x \in \mathbb{R}^{d+1}$ als $(d+1)$-dimensionaler Zeilenvektor aufzufassen sei. Da in unserem Fall die *Kleinste-Quadrate-Methode* noch die Schätzung des Ordinatenabschnitts β_0 beinhaltet, führen wir zum Zwecke der Darstellung noch für jedes $i \in \mathbb{N}$ den $(d+1)$-dimensionalen Zeilenvektor

$$\mathbf{X}_{0,i} := (1, X_{i,1}, ..., X_{i,d})$$

ein, welcher den Eintrag 1 sowie die Einträge des Regressors X_i enthält. Damit können wir für alle $i \in \{1, ..., n\}$ das Residuum

$$E_i := Y_i - \hat{r}_n(\mathbf{X}_{0,i})$$

ermitteln. Wir wenden eine *Modifikation* auf die Residuen durch

$$R_i := \frac{E_i}{\sqrt{1-h_i}}$$

an, wobei h_i der i-te Eintrag in der Hauptdiagonale der Matrix

$$\mathbf{X}_n \cdot ((\mathbf{X}_n)^T \cdot \mathbf{X}_n)^{-1} \cdot (\mathbf{X}_n)^T$$

sei. Die modifizierten Residuen $R_1, ..., R_n$ weisen die gleiche Varianz auf. Diese stimmt mit der gemeinsamen Varianz σ^2 der Störgrößen überein; vgl. Davison und Hinkley (1997, S.259 und S.261). Es sei außerdem $\bar{R}$ das arithmetische Mittel über $R_1, ..., R_n$. Das Ziel ist nun die Verteilung bezüglich des „wahren" Wahrscheinlichkeitsmaßes P des *Prognosefehlers*

$$pf_{n+1} := \hat{r}_n(\mathbf{X}_{0,n+1}) - Y_{n+1}$$

mit der Bootstrap-Methode zu schätzen. Um dies zu erreichen, simulieren wir die Verteilung der Störgrößen mittels Resampling der modifizierten zentrierten Residuen $R_1 - \bar{R}, ..., R_n - \bar{R}$. Es erfolgt eine Ziehung von n Elementen mit Zurücklegen aus der Menge $\{R_1 - \bar{R}, ..., R_n - \bar{R}\}$. Zur konzeptionellen Veranschaulichung dieses Vorganges sei für jede Menge $A \in \mathscr{B}_1$ der (zufällige) Wert

$$P_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(R_i - \bar{R})$$

der *empirischen Verteilung* bezüglich der $\mathbb{R}$-wertigen Zufallsvariablen $R_1 - \bar{R}, ..., R_n - \bar{R}$ betrachtet, wobei die Funktion $\mathbb{1}_A$ für jedes $x \in \mathbb{R}$ definiert werde durch

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

und damit für die charakteristische Funktion des Ereignisses A stehe. Beachte, dass P_n für gegebene Realisationen $r_1, ..., r_n$ von $R_1 - \bar{R}, ..., R_n - \bar{R}$ ein (festes) Wahrscheinlichkeitsmaß auf den Meßraum $(\mathbb{R}, \mathscr{B}_1)$ ist und nur positive Werte für Mengen A, welche mindestens eine der $r_1, ..., r_n$ enthält, annimmt. Weiters seien $\epsilon_1^*, ..., \epsilon_n^*$ unabhängige und identisch verteilte Zufallsvariablen mit Verteilung P_n (gegeben feste Realisationen $r_1, ..., r_n$ von $R_1 - \bar{R}, ..., R_n - \bar{R}$). Wir erhalten also eine sogenannte *Bootstrap-Stichprobe*

$$\mathbf{E}^* \equiv (\epsilon_1^*, ..., \epsilon_n^*),$$

indem wir n Elemente aus $\{R_1 - \bar{R}, ..., R_n - \bar{R}\}$ mit Zurücklegen ziehen. Für jedes $i \in \{1, ..., n\}$ bestimmen wir damit

$$Y_i^* := \hat{r}_n(\mathbf{X}_{0,i}) + \epsilon_i^*$$

und definieren den n-dimensionalen Spaltenvektor

$$\mathbf{Y}_n^* := (Y_1^*, ..., Y_n^*)^T.$$

Wir führen anschließend eine Kleinste-Quadrate-Schätzung durch, indem wir

$$\hat{\beta}_n^* := ((\mathbf{X}_n)^T \cdot \mathbf{X}_n)^{-1} \cdot (\mathbf{X}_n)^T \cdot \mathbf{Y}_n^*$$

bestimmen und erhalten die Bootstrap-Regressionsfunktion $\hat{r}_n^*$, die durch

$$\hat{r}_n^*(x) := x \cdot \hat{\beta}_n^*$$

für alle $(d+1)$-dimensionalen Zeilenvektoren $x \in \mathbb{R}^{d+1}$ definiert ist. Wir ziehen anschließend M weitere Elemente aus $\{R_1 - \bar{R}, ..., R_n - \bar{R}\}$, die wir als $\epsilon_{n+1,1}^*, ..., \epsilon_{n+1,M}^*$ bezeichnen und bestimmen damit den Bootstrap-Prognosefehler

$$pf_{n+1,m}^* := \hat{r}_n^*(\mathbf{X}_{0,n+1}) - (\hat{r}_n(\mathbf{X}_{0,n+1}) + \epsilon_{n+1,m}^*)$$

für alle $m \in \{1, ..., M\}$. Wenn wir für jedes $pf_{n+1,m}^*$ die geeigneten Quantile bezüglich der empirischen Verteilung P_n (gegeben feste Werte $r_1, ..., r_n$ der Zufallsvariablen $R_1 - \bar{R}, ..., R_n - \bar{R}$), mit der wir die „wahre“ Verteilung des Prognosefehlers pf_{n+1} schätzen, ermitteln könnten, dann wären wir in der Lage die Grenzen eines Bootstrap-Konfidenzintervalls für Y_{n+1} zu bestimmen. Es würde dabei genügen $M = 1$ zu setzen. Um die Quantile zu bestimmen, müssten wir alle möglichen Werte für $pf_{n+1,1}^*$ berechnen und dazu wäre es unter anderem nötig, alle möglichen Ausprägungen des Zufallsvektors $\mathbf{E}^*$ sowie $\epsilon_{n+1,1}^*$ zu bestimmen. Die Zufallsvariablen $\mathbf{E}^*$, $\epsilon_{n+1,1}^*$ und somit auch $pf_{n+1,1}^*$ können zwar bei gegebenen Realisationen $r_1, ..., r_n$ von $R_1 - \bar{R}, ..., R_n - \bar{R}$ nur endlich viele Werte annehmen, bei hohem n wäre jedoch die vollständige Bestimmung der (unter P_n diskreten) Verteilung von $pf_{n+1,1}^*$ zu aufwändige, da schon allein E^* bis zu n^n verschiedene Ausprägungen annehmen kann.[4] Wir greifen somit auf das Monte-Carlo Verfahren zurück und legen eine Anzahl $B \in \mathbb{N}$ an Monte-Carlo Iterationen, welche *unabhängig* voneinander ausgeführt werden, fest. In jeder Iteration folgen wir der oben beschriebenen Vorgehensweise. Für alle Iterationsläufe $b \in \{1, ..., B\}$ ziehen wir mit Zurücklegen n Elemente aus der Menge $\{R_1 - \bar{R}, ..., R_n - \bar{R}\}$ und erhalten die Bootstrap-Stichprobe $\mathbf{E}_b^*$. Wir berechnen damit den Spaltenvektor

$$\mathbf{Y}_{n,b}^* := (Y_{1,b}^*, ..., Y_{n,b}^*)^T,$$

und bestimmen den Kleinste-Quadrate-Schätzer $\hat{\beta}_{n,b}^*$, womit wir die Regressionsfunktion $\hat{r}_{n,b}^*$ erhalten. Für alle Iterationen b ziehen wir als Nächstes M Elemente $\epsilon_{n+1,1,b}^*, ..., \epsilon_{n+1,M,b}^*$ aus $\{R_1 - \bar{R}, ..., R_n - \bar{R}\}$ und bestimmen schließlich

$$pf_{n+1,m,b}^* := \hat{r}_{n,b}^*(\mathbf{X}_{0,n+1}) - (\hat{r}_n(\mathbf{X}_{0,n+1}) + \epsilon_{n+1,m,b}^*)$$

für jedes $m \in \{1, ..., M\}$. Insgesamt erhalten wir also $B \cdot M$ Bootstrap-Prognosen. Die Zahl $B \cdot M$ sollte möglichst groß sein. Davison und Hinkley

(1997, S.285) betonen daher, dass $M = 1$ ausreicht, sofern B bereits groß gewählt ist (weil das Monte-Carlo Verfahren auf den *Gesetz der großen Zahlen* beruht, schätzt es unter geeigneten Bedingungen um so genauer, je mehr Iterationen durchgeführt werden). Die Ordnungsstatistiken $pf^*_{n+1,(1)}, ..., pf^*_{n+1,(B\cdot M)}$ der Zufallsvariablen

$$pf^*_{n+1,1,1}, ..., pf^*_{n+1,M,1}, ..., pf^*_{n+1,1,B}, ..., pf^*_{n+1,M,B}$$

beschreiben nun eine empirische Verteilung, mit der wir die Verteilung von pf_{n+1} schätzen. Wir betrachten ein $\alpha \in (0, 1)$. Zur besseren Übersicht seien

$$m_1 := (B \cdot M + 1) \cdot \left(1 - \frac{\alpha}{2}\right),$$

$$m_2 := (B \cdot M + 1) \cdot \frac{\alpha}{2}$$

definiert und wir nehmen vereinfachend an, dass $m_1, m_2 \in \{1, ..., B \cdot M\}$. Die Ordnungsstatistiken $pf^*_{n+1,(m_1)}$ und $pf^*_{n+1,(m_2)}$ versetzen uns schließlich in die Lage ein auf der Monte-Carlo Methode basierendes Bootstrap-Konfidenzintervall für Y_{n+1} durch

$$I_{B,n,\alpha} := \left[\hat{r}_n(\mathbf{X}_{0,n+1}) - pf^*_{n+1,(m_1)}, \hat{r}_n(\mathbf{X}_{0,n+1}) - pf^*_{n+1,(m_2)}\right]$$

zu formulieren. Ein etwas anderer Zugang zur Gewinnung eines Bootstrap-Konfidenzintervalls besteht darin, als Ausgangspunkt

$$\frac{pf_{n+1}}{\sqrt{s_n^2}}$$

mit

$$s_n^2 := \frac{1}{n - d - 1} \sum_{i=1}^{n} (E_i)^2$$

als Varianzschätzer für pf_{n+1} zu wählen, also den Prognosefehler pf_{n+1} zu „standardisieren“. Dazu bestimmen wir in der Monte-Carlo Simulation in jedem Iterationsschritt $b \in \{1, ..., B\}$ den Varianzschätzer

$$s^{*2}_{n,b} := \frac{1}{n - d - 1} \sum_{i=1}^{n} (E^*_{i,b})^2,$$

wobei für jedes $i \in \{1, ..., n\}$ die Zufallsvariable

$$E^*_{i,b} := Y^*_{i,b} - \hat{r}^*_{n,b}(\mathbf{X}_{0,i})$$

für das Bootstrap-Residuum im Iterationsschritt b stehe. Damit können wir

$$spf^*_{n+1,m,b} := \frac{pf^*_{n+1,m,b}}{\sqrt{s^{*2}_{n,b}}}$$

für jedes $m \in \{1, ..., M\}$ berechnen. Wir betrachten nun die Ordnungsstatistiken $spf^*_{n+1,(m_1)}$ und $spf^*_{n+1,(m_2)}$ der Zufallsvariablen

$$spf^*_{n+1,1,1}, ..., spf^*_{n+1,M,1}, ..., spf^*_{n+1,1,B}, ..., spf^*_{n+1,M,B}$$

und erhalten das auf den „standardisierten“ Prognosefehler basierende Konfidenzintervall

$$SI_{B,n,\alpha} := \left[\hat{r}_n(\mathbf{X}_{0,n+1}) - spf^*_{n+1,(m_1)} \cdot \sqrt{s_n^2}, \hat{r}_n(\mathbf{X}_{0,n+1}) - spf^*_{n+1,(m_2)} \cdot \sqrt{s_n^2}\right].$$

Wie wir gesehen haben, hängt die Qualität der von uns betrachteten Bootstrap-Technik im Wesentlichen davon ab, wie gut wir mit der empirische Verteilung bezüglich der modifizierten Residuen die „wahre“ Verteilung der Störgrößen zu schätzen vermögen. Im Allgemeinen ist zu erwarten, dass dieses Verfahren gültige Konfidenzintervalle für Y_{n+1} (mit der Eigenschaft, dass die Überdeckungswahrscheinlichkeit der Intervalle nicht kleiner als das gewünschte Konfidenzniveau $1 - \alpha$ ist bzw. mit wachsendem Stichprobenumfang zu $1 - \alpha$ konvergiert) unter *restriktiveren* Bedingungen erzeugt als die konformalen Ansätze. Wir sagen das schon deswegen, weil wir bei der Präsentation dieser Methode explizit von der Existenz eines Modells, das spezifische Annahmen über die Störgrößen beinhaltet, ausgegangen sind (was bisher für die konformalen Methoden nicht von Nöten war) und es darüber hinaus durch Anwendung der Monte-Carlo Methode zu zusätzlichen Schätzfehlern kommt. Der Frage, unter welchen Umständen die Bootstrap-Methode geeignete Konfidenzintervalle für Y_{n+1} hervorbringt, wollen wir an dieser Stelle nicht weiter nachgehen, da dies den Rahmen dieser Arbeit in erheblicher Weise sprengen würde. Wir verweisen stattdessen auf Beran (1992), der die mathematisch-technischen Details bezüglich der Konstruktion von auf der Bootstrap-Technik basierenden Prognosebereichen in den Fokus stellt. Dem interessierten Leser sei außerdem Stine (1985) empfohlen, der sich ebenfalls mit der Formierung von Bootstrap-Prognosebereichen im Rahmen von linearen Regressionsmodellen mit nicht stochastischen Regressoren beschäftigt, sowie Horowitz (2001), der sich mit der alllgemeinen Anwendung der Bootstrap-Methode im Rahmen ökonometrischer Analysen auseinandersetzt.

4 Asymptotisches Verhalten der nach der Full- und der Split-Methode bestimmten konformalen Konfidenzbereiche

Im Kapitel 3 haben wir uns mit zwei Methoden zur Bestimmung konformaler Konfidenzbereiche beschäftigt, nämlich der Full- und der Split-Methode. Wir haben gesehen, dass bei *vertauschbaren* Zufallsvariablen beide Methoden bei *festem*, aber *beliebig hohem* Stichprobenumfang Konfidenzbereiche hervorbringen, deren Überdeckungswahrscheinlichkeit von unten durch das vorgegebene nominelle Konfidenzniveau beschränkt ist und bei wachsender Stichprobengröße sich diesen immer weiter annähert. In diesem Kapitel werden wir uns mit dem *asymptotischen* Verhalten zweier für die praktische Anwendung von Konfidenzbereichen auf dem Gebiet der Regression wichtigen Größen beschäftigen, nämlich einerseits mit der *empirischen Überdeckungsrate* und andererseits mit den *Grenzen* bzw. der *Breite* der Konfidenzbereiche.

4.1 Ein Referenzmodell

Viele ökonometrische Lehrbücher (zum Beispiel von Auer (2003) oder Wooldridge (2003)) gehen bei der Einführung in das Gebiet der Regression von der Existenz eines *Referenzmodells* (manchmal auch als „wahres“ Modell bezeichnet) aus, welches häufig als lineares Modell spezifiziert ist und mit einer Regressionsfunktionen (bestimmt durch Algorithmen wie etwa der Methode der kleinsten Quadrate) geschätzt wird. Auch G'Sell et al. (2016, S.28-34) gehen bei der Untersuchung der asymptotischen Eigenschaften der Konfidenzbereiche von einem Referenzmodell aus. Wir folgen dieser „Tradition“ und gehen zunächst von einem allgemein gehaltenen Modell aus.[5] Um ein solches zu formulieren verwenden wir wieder die gleiche Notation wie im Kapitel 3. Sei $d \in \mathbb{N}$. Wir betrachten eine Folge von Zufallsvariablen $(Z_i)_{i\in\mathbb{N}}$, wobei wieder $Z_i \equiv (X_i, Y_i)$ mit $X_i \in \mathbb{R}^d$ und $Y_i \in \mathbb{R}$ für jedes $i \in \mathbb{N}$. Weiters seien gegeben eine Folge von reellwertigen $\mathscr{A}$-$\mathscr{B}_1$-messbaren Zufallsvariablen $(\epsilon_i)_{i\in\mathbb{N}}$ sowie eine Funktion

$$a : \mathbb{R}^d \to \mathbb{R},$$

welche wir als $\mathscr{B}_d$-$\mathscr{B}_1$-messbar voraussetzen. Unser Referenzmodell setzt sich aus folgenden *Postulaten* zusammen:

(P1) Der Prozess $(Z_i)_{i\in\mathbb{N}}$ sei eine Folge von Zufallsvariablen, die bezüglich P *unabhängigen* und *identisch verteilt* sind.

(P2) Für jedes $i \in \mathbb{N}$ sei die Bedingung

$$Y_i \equiv a(X_i) + \epsilon_i \tag{4.1}$$

erfüllt.

(P3) Für jedes $i \in \mathbb{N}$ *existiere* bezüglich P der Erwartungswert und die Varianz der Zufallsvariable ϵ_i. Der Erwartungswert von ϵ_i sei außerdem mit dem Wert 0 ident.

(P4) Für beliebige $i \in \mathbb{N}$ seien die Zufallsvariablen X_i und ϵ_i bezüglich P *unabhängig*.

Jede das Postulat (P1) erfüllende Folge von Zufallsvariablen $(Z_i)_{i\in\mathbb{N}}$ ist nach Lemma 2.2 vertauschbar, womit alle Resultate aus Kapitel 3 gültig sind. Aus den Postulaten (P1) bis (P4) folgt außerdem, dass $(\epsilon_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen ist. Im Referenzmodell fungiert diese also als Folge unabhängiger und identisch verteilter *Störgrößen* mit verschwindenden Erwartungswert, welche darüber hinaus von den Regressoren unabhängig sind. Für jedes $i \in \mathbb{N}$ definieren wir

$$U_i := |\epsilon_i| \,, \tag{4.2}$$

also den *Absolutbetrag* der Störgröße ϵ_i. Damit ist $(U_i)_{i\in\mathbb{N}}$ ebenfalls eine Folge unabhängiger und identisch verteilter Zufallsvariablen. Deren gemeinsame Verteilungsfunktion sei für den Rest dieses Kapitels mit F bezeichnet. Der Ausdruck F^{-1} stehe für die *verallgemeinerte Inverse* von F, die wir für jedes $\alpha \in (0,1)$ als

$$F^{-1}(\alpha) := \inf\{x \in \mathbb{R} \,|\, F(x) \geq \alpha\} \tag{4.3}$$

festlegen. Im Rahmen des Referenzmodells können wir bei gegebenen $\alpha \in (0,1)$ und $n \in \mathbb{N}$ das Intervall

$$C_{n,\alpha}(X_{n+1}) := [a(X_{n+1}) - F^{-1}(1-\alpha), a(X_{n+1}) + F^{-1}(1-\alpha)]$$

als Konfidenzbereich bzw. Konfidenzintervall für Y_{n+1} zum Konfidenzniveau $1-\alpha$ auffassen, da unter (4.1) und (4.2) gilt:

$$P(Y_{n+1} \in C_{n,\alpha}(X_{n+1})) = P(U_{n+1} \leq F^{-1}(1-\alpha)) = F(F^{-1}(1-\alpha)) \geq 1-\alpha.$$

Dabei entspricht $2 \cdot F^{-1}(1-\alpha)$, also 2 mal dem $(1-\alpha)$-Quantil der Verteilung F, der Breite von $C_{n,\alpha}(X_{n+1})$. Die Konfidenzbereiche, die wir in Kapitel 3 diskutiert haben, sind auf Regressionsfunktionen aufgebaut, mittels derer die

Funktion *a geschätzt* werden soll. Die „Breite“ der Konfidenzbereiche beinhalten damit neben der *Variation*, die von der Störgröße ausgeht, auch den von den Regressionsfunktionen begangenen *Schätzfehler*. Dieser Schätzfehler verschwindet bei $C_{n,\alpha}(X_{n+1})$ jedoch vollständig.[6] Falls F darüber hinaus stetig und auf der Menge der nicht negativen reellen Zahlen streng monoton steigend ist, dann ist der obige Konfidenzbereich sogar *exakt*, d.h.

$$P(Y_{n+1} \in C_{n,\alpha}(X_{n+1})) = F(F^{-1}(1-\alpha)) = 1-\alpha.$$

Unter dieser Bedingung könnte man $C_{n,\alpha}(X_{n+1})$ als „optimalen“ Konfidenzbereich für Y_{n+1} auffassen. In diesem Kontext sprechen G'Sell et al. (2016, S.30) auch von einem „super-oracle“, welches in einer Situation *vollständiger Information* über die Funktion a und die Verteilung F bestimmt werden kann. Wir werden sehen, dass das Quantil $F^{-1}(1-\alpha)$ eine zentrale Rolle bei der Asymptotik der Full- wie auch der Split-Methode spielt.

In den folgenden Unterkapiteln werden wir für jede Folge von reellwertigen Zufallsvariablen $(A_i)_{i\in\mathbb{N}}$, welche bezüglich P *in Wahrscheinlichkeit* gegen eine feste reelle Zahl $a \in \mathbb{R}$ konvergiert, abkürzend

$$\operatorname*{plim}_{n\to\infty} A_n = a$$

schreiben. Außerdem schreiben wir für zwei reellwertige Zufallsvariablen A und B, die bezüglich P *fast sicher* ident sind, den Ausdruck

$$A \overset{f.s.}{=} B.$$

4.2 Asymptotik der Full-Methode

In diesem Unterkapitel verwenden wir die gleiche Notation wie in Unterkapitel 3.1. Betrachte eine Folge von Zufallsvariable $(Z_i)_{i\in\mathbb{N}}$ mit $Z_i \equiv (X_i, Y_i)$ für jedes $i \in \mathbb{N}$. Für beliebige Stichprobengrößen $n \in \mathbb{N}$ seien $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ die mittels $Z_1, ..., Z_{n+1}$ und eines Algorithmus zur Full-Methode a_n bestimmten Residuen und $C^{full}_{n,\alpha}(X_{n+1})$ sei der dazu korrespondierende Konfidenzbereich. Für gegebenes $\alpha \in (0,1)$ definieren wir außerdem

$$h_{n,\alpha} := \lceil (1-\alpha)(n+1) \rceil,$$

und unter der Bedingung $h_{n,\alpha} \leq n$ betrachten wir $U_{(h_{n,\alpha}:n),Z_{n+1}}$ als die $h_{n,\alpha}$-te Ordnungsstatistik unter den Residuen $U_{1,Z_{n+1}}, ..., U_{n,Z_{n+1}}$. Um in den folgenden Überlegungen die Notation möglichst einfach zu halten definieren wir

$$E^{full}_{n,\alpha} := \{Y_{n+1} \in C^{full}_{n,\alpha}(X_{n+1})\},$$

also das Ereignis, dass sich die Zielvariable Y_{n+1} im Konfidenzbereich $C^{full}_{n,\alpha}(X_{n+1})$ realisiert. Damit können wir für jedes $n \in \mathbb{N}$ die *empirische Überdeckungsrate* definieren als

$$A^{full}_{n,\alpha} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{E^{full}_{i,\alpha}},$$

wobei für jedes $i \in \mathbb{N}$ der Ausdruck $\mathbb{1}_{E^{full}_{i,\alpha}}$ für die Indikatorfunktion des Ereignisses $E^{full}_{i,\alpha}$ stehe. Die Größe $A^{full}_{n,\alpha}$ ist also der relative Anteil jener Zufallsvariablen unter den $Y_2, ..., Y_{n+1}$, die sich in ihren jeweiligen Konfidenzbereichen $C^{full}_{1,\alpha}(X_2), ..., C^{full}_{n,\alpha}(X_{n+1})$ realisieren. Dieser kann als Schätzer für die nicht beobachtbare Überdeckungswahrscheinlichkeit $P(E^{full}_{n,\alpha})$ zur Stichprobengröße n herangezogen werden. Beachte, dass wir hier die Stichprobengröße n wachsen lassen und damit das Konzept des On-Line Settings im vollen Umfang zum Tragen kommt. Die Menge an Daten, welche zur Bestimmung von $C^{full}_{n,\alpha}(X_{n+1})$ dienen, werden also bei jeder Erhöhung von n um die neue Instanz erweitert.

Der nächsten Satz macht eine Aussage über das *asymptotische* Verhalten von $A^{full}_{n,\alpha}$ (eine Beweisidee zu einer Version dieses Satzes, die die gleiche Schlussfolgerung beinhaltet aber unter allgemeineren Voraussetzungen formuliert ist, findet sich in Shafer und Vovk (2008, S.386 und S.413-414)):

Satz 4.1. *Sei $(Z_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen, welche das Postulat (P1) erfülle, und betrachte für jedes $n \in \mathbb{N}$ einen permutationsinvarianten Algorithmus zur Full-Methode a_n, wobei unter den mit $Z_1, ..., Z_{n+1}$ und a_n korrespondierenden Residuen $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ bezüglich P fast sicher keine Bindungen vorkommen. Dann gilt für beliebiges $\alpha \in (0, 1)$, dass*

$$\operatorname*{plim}_{n\to\infty} A^{full}_{n,\alpha} = 1 - \alpha.$$

Beweis. Wir setzen zunächst voraus, dass $(\mathbb{1}_{E^{full}_{i,\alpha}})_{i\in\mathbb{N}}$ eine Folge von bezüglich P unabhängigen Zufallsvariablen ist. Für alle $n \in \mathbb{N}$ gilt außerdem

$$Var(\mathbb{1}_{E^{full}_{n,\alpha}}) = P(E^{full}_{n,\alpha}) \cdot (1 - P(E^{full}_{n,\alpha})),$$

wobei $Var(\mathbb{1}_{E^{full}_{n,\alpha}})$ für die Varianz von $\mathbb{1}_{E^{full}_{n,\alpha}}$ bezüglich P stehe. Wir haben in Unterkapitel 3.1 gesehen, dass

$$\lim_{n\to\infty} P(E^{full}_{n,\alpha}) = 1 - \alpha, \tag{4.4}$$

woraus unmittelbar

$$\lim_{n\to\infty} Var(\mathbb{1}_{E^{full}_{n,\alpha}}) = \alpha(1 - \alpha)$$

folgt. Weil $Var(\mathbb{1}_{E_{n,\alpha}^{full}})$ endlich ist für jedes $n \in \mathbb{N}$ und $\alpha(1-\alpha) \in \mathbb{R}$, gilt

$$\sup_{n\in\mathbb{N}} Var(\mathbb{1}_{E_{n,\alpha}^{full}}) < \infty.$$

Wir können damit auf $(\mathbb{1}_{E_{i,\alpha}^{full}})_{i\in\mathbb{N}}$ ein schwaches Gesetz der großen Zahlen anwenden, dass für Folgen unkorrelierter Zufallsvariablen bewiesen werden kann; vgl. Kusolitsch (2011, S.254). Es folgt

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{E_{i,\alpha}^{full}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbb{1}_{E_{i,\alpha}^{full}})\right) = 0.$$

Wegen (4.4) gilt wegen des Grenzwertsatzes von Cauchy

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbb{1}_{E_{i,\alpha}^{full}}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}P(E_{i,\alpha}^{full}) = 1-\alpha,$$

vgl. Heuser (2006, S.177). Wir erhalten schließlich

$$\operatorname*{plim}_{n\to\infty}(A_{n,\alpha}^{full} - (1-\alpha)) = 0.$$

Es bleibt noch zu zeigen, dass es sich bei $(\mathbb{1}_{E_{i,\alpha}^{full}})_{i\in\mathbb{N}}$ um eine Folge unabhängiger Zufallsvariablen handelt. Zu diesem Zweck werden wir uns die Rechenregeln für bedingte Erwartungswerte zu Nutze machen. Bezüglich der Definition und der Existenz bedingter Erwartungswerte sowie die sich daraus ergebenden Rechenregeln sei auf die Literatur, zum Beispiel Chow und Teicher (1978, S.198-209) oder Kusolitsch (2011, S.234-242), verwiesen.Wir beginnen für beliebiges $n \in \mathbb{N}$ mit der Einführung der Menge

$$M_n := \{A \subseteq \mathbb{R}^{d+1} \,|\, |A| \leq n+1\},$$

wobei der Ausdruck $|A|$ für die Mächtigkeit von Mengen $A \subseteq \mathbb{R}^{d+1}$ stehe. Wir formulieren damit die Funktion

$$f_n : \Omega \to M_n$$

mit

$$f_n(\omega) := \{Z_1(\omega), ..., Z_{n+1}(\omega)\}$$

für jedes $\omega \in \Omega$. Wir definieren die Mengensysteme

$$\mathscr{M}_n := \{B \subseteq M_n \,|\, f_n^{-1}(B) \in \mathscr{A}\}$$

und

$$\mathscr{F}_n := \{f_n^{-1}(B) \,|\, B \in \mathscr{M}_n\},$$

wobei

$$f_n^{-1}(B) := \{\omega \in \Omega \,|\, f_n(\omega) \in B\}$$

für das Urbild einer Menge $B \subseteq M_n$ unter f_n stehe. Weil $\mathscr{A}$ eine σ-Algebra auf Ω ist, folgt aus der Operationstreue des Urbilds, dass $\mathscr{M}_n$ eine σ-Algebra auf M_n ist. Damit ist $\mathscr{F}_n$ eine σ-Algebra auf Ω, also die von der Funktion f_n *induzierte* σ-Algebra; vgl. auch Hinderer (1972, S.78-79). Wir schreiben

$$\sigma(\{Z_1, ..., Z_{n+1}\}) := \mathscr{F}_n.$$

Jedes Ereignis aus dieser σ-Algebra enthält zwar Informationen über die Zufallsmenge

$$\{Z_1, ..., Z_{n+1}\},$$

jedoch nicht über den Zufallsvektor

$$(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})$$

mit beliebiger Permutation $\sigma \in S_{n+1}$ (mit anderen Worten verfügen wir über keine genaueren Informationen über die einzelnen Zufallsvariablen $Z_1, ..., Z_{n+1}$). Aus der Definition von $\mathscr{M}_n$ und $\mathscr{F}_n$ geht unmittelbar hervor, dass

$$\sigma(\{Z_1, ..., Z_{n+1}\}) \subseteq \mathscr{A},$$

womit der bedingte Erwartungswert $\mathbb{E}(X \,|\, \sigma(\{Z_1, ..., Z_{n+1}\}))$ für beliebige $\mathscr{A}$-$\mathscr{B}_1$-messbare und bezüglich P integrierbare Funktionen X existiert.

Nun gilt es für beliebige $n \in \mathbb{N}$ zu zeigen, dass

$$\mathbb{E}\left(\mathbb{1}_{E_{n,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{n+1}\})\right) \overset{f.s.}{=} P(E_{n,\alpha}^{full}). \tag{4.5}$$

Da nach Voraussetzung unter den Residuen $U_{1,Z_{n+1}}, ..., U_{n+1,Z_{n+1}}$ fast sicher keine Bindungen vorkommen, gilt für beliebiges $k \in \{1, ..., n+1\}$, dass

$$\sum_{i=1}^{n+1} \mathbb{1}_{\left\{rang(U_{i,Z_{n+1}})=k\right\}} \overset{f.s.}{=} 1.$$

Für jedes Ereignis $A \in \sigma(\{Z_1, ..., Z_{n+1}\})$ folgt daraus

$$\sum_{i=1}^{n+1} \mathbb{E}(\mathbb{1}_{\{rang(U_{i,Z_{n+1}})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_A), \tag{4.6}$$

wobei der Operator $\mathbb{E}(\cdot)$ für die Bildung des Erwartungswertes einer integrierbaren Zufallsvariable unter P stehe. Im Beweis von Lemma 3.1 haben wir gesehen, dass aufgrund der geforderten Permutationsinvarianz des Algorithmus a_n für beliebige $\sigma \in S_{n+1}$ der Zufallsvektor der Residuen

$$(U_{\sigma(1),Z_{n+1}}, ..., U_{\sigma(n+1),Z_{n+1}})$$

als eine deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})$$

geschrieben werden kann. Damit ist auch der Zufallsvektor

$$(\mathbb{1}_{\{rang(U_{\sigma(1),Z_{n+1}})=k\}}, ..., \mathbb{1}_{\{rang(U_{\sigma(n+1),Z_{n+1}})=k\}})$$

ein deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)}).$$

Da $A \in \sigma(\{Z_1, ..., Z_{n+1}\})$, gibt es gleichzeitig eine (feste) Menge $B \in \mathscr{M}_n$, sodass

$$A = f_n^{-1}(B).$$

Somit ist für alle $\omega \in \Omega$ die Aussage

$$\mathbb{1}_A(\omega) = 1$$

genau dann erfüllt, wenn

$$f_n(\omega) \in B.$$

Damit gilt

$$\mathbb{1}_A(\omega) = \mathbb{1}_B(f_n(\omega)).$$

Beachte, dass $B \subseteq M_n$. Definiere nun die (feste) Funktion

$$g : M_n \to \mathbb{R}$$

mit

$$g(x) := \mathbb{1}_B(x)$$

für alle $x \in M_n$. Für alle $\omega \in \Omega$ gilt dann

$$\mathbb{1}_A(\omega) = g(f_n(\omega)) = g(\{Z_1(\omega), ..., Z_{n+1}(\omega)\}),$$

womit die Zufallsvariable $\mathbb{1}_A$ eine Funktion der Zufallsmenge

$$\{Z_1, ..., Z_{n+1}\}$$

ist. Da

$$\{Z_1(\omega), ..., Z_{n+1}(\omega)\} = \{Z_{\sigma(1)}(\omega), ..., Z_{\sigma(n+1)}(\omega)\}$$

für beliebige $\omega \in \Omega$ gilt, kann also der Zufallsvektor

$$(\mathbb{1}_{\{rang(U_{\sigma(1),Z_{n+1}})=k\}} \cdot \mathbb{1}_A, ..., \mathbb{1}_{\{rang(U_{\sigma(n+1),Z_{n+1}})=k\}} \cdot \mathbb{1}_A)$$

als deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{\sigma(1)}, ..., Z_{\sigma(n+1)})$$

und der Zufallsmenge

$$\{Z_{\sigma(1)}, ..., Z_{\sigma(n+1)}\}$$

betrachtet werden. Die Folge $(Z_i)_{i\in\mathbb{N}}$ ist nach Voraussetzung unabhängig und identisch verteilt, womit sie nach Lemma 2.2 vertauschbar ist. Die Zufallsvariablen

$$\mathbb{1}_{\{rang(U_{1,Z_{n+1}})=k\}} \cdot \mathbb{1}_A, ..., \mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})=k\}} \cdot \mathbb{1}_A$$

sind also ebenfalls vertauschbar. Nach Lemma 2.1 sind diese Zufallsvariablen somit in Verteilung gleich. Für beliebige $i \in \{1, ..., n\}$ gilt folglich

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{i,Z_{n+1}})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})=k\}} \cdot \mathbb{1}_A).$$

Aus Gleichung (4.6) erhalten wir

$$(n+1) \cdot \mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_A)$$

und damit

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}\left(\frac{1}{n+1}\mathbb{1}_A\right).$$

Weil auf ganz Ω die Gleichheit

$$\mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})\leq h_{n,\alpha}\}} = \sum_{i=1}^{h_{n,\alpha}} \mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})=i\}}$$

gilt, haben wir

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,Z_{n+1}})\leq h_{n,\alpha}\}} \cdot \mathbb{1}_A) = \mathbb{E}\left(\frac{h_{n,\alpha}}{n+1} \cdot \mathbb{1}_A\right).$$

Im Fall $h_{n,\alpha} \leq n$ gilt außerdem

$$E_{n,\alpha}^{full} = \{U_{n+1,Z_{n+1}} \in [0, U_{(h_{n,\alpha}:n),Z_{n+1}})\},$$

also

$$E_{n,\alpha}^{full} = \{rang(U_{n+1,Z_{n+1}}) \leq h_{n,\alpha}\},$$

sowie

$$P(E_{n,\alpha}^{full}) = \frac{h_{n,\alpha}}{n+1}$$

(siehe auch den Beweis von Satz 2.1 und Unterkapitel 3.1). Wir haben also

$$\mathbb{E}(\mathbb{1}_{E_{n,\alpha}^{full}} \cdot \mathbb{1}_A) = \mathbb{E}\left(P(E_{n,\alpha}^{full}) \cdot \mathbb{1}_A\right)$$

für beliebige $A \in \sigma(\{Z_1, ..., Z_{n+1}\})$. Aus der Definition bedingter Erwartungswerte erhalten wir schließlich

$$\mathbb{E}\left(\mathbb{1}_{E_{n,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{n+1}\})\right) \overset{f.s.}{=} P(E_{n,\alpha}^{full}).$$

Für den Fall $h_{n,\alpha} > n$ haben wir $C_{n,\alpha}^{full}(X_{n+1}) = \mathbb{R}$ festgelegt (siehe Unterkapitel 3.1), womit $E_{n,\alpha}^{full} = \Omega$ und schließlich $P(E_{n,\alpha}^{full}) = 1$ erfüllt sind. Damit gilt

$$\mathbb{E}(\mathbb{1}_{E_{n,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{n+1}\})) \overset{f.s.}{=} P(E_{n,\alpha}^{full}).$$

Für beliebiges $i \in \mathbb{N}$ sei nun

$$\sigma(Z_i) := \{Z_i^{-1}(B) \,|\, B \in \mathscr{B}_1\}$$

die von der Zufallsvariable Z_i *erzeugte* σ-Algebra. Außerdem schreiben wir für beliebige integrierbare Zufallsvariablen X abkürzend

$$\mathbb{E}(X \,|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1})$$

$$:= \mathbb{E}(X \,|\, \sigma[\sigma(\{Z_1, ..., Z_{m+1}\}) \cup \sigma[\sigma(Z_{m+2}) \cup ... \cup \sigma(Z_{n+1})]]),$$

wobei $\sigma[\mathscr{C}]$ für die von einer beliebigen Teilmenge $\mathscr{C}$ der Potenzmenge von Ω erzeugten σ-Algebra stehe. Für beliebige $n \in \mathbb{N}$ mit $n \geq 2$ und $m \in \{1, ..., n-1\}$ gilt

$$\mathbb{E}\left(\prod_{i=m}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right)$$

$$= \mathbb{E}\left(\mathbb{E}\left(\mathbb{1}_{E_{m,\alpha}^{full}} \cdot \prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}} \,\middle|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1}\right)\right).$$

Die Zufallsvariable $\prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}$ hängt von der Zufallsmenge $\{Z_1, ..., Z_{m+1}\}$ sowie von den einzelnen Zufallsvariablen $Z_{m+2}, ..., Z_{n+1}$ ab, womit

$$\mathbb{E}\left(\mathbb{1}_{E_{m,\alpha}^{full}} \cdot \prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}} \,\middle|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1}\right)$$

$$\overset{f.s.}{=} \prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}} \cdot \mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1}).$$

Gleichzeitig ist $(Z_i)_{i\in\mathbb{N}}$ nach Voraussetzung eine Folge unabhängiger Zufallsvariablen. Da $\mathbb{1}_{E_{m,\alpha}^{full}}$ unabhängig von den Zufallsvariablen $Z_{m+2}, ..., Z_{n+1}$ ist, gilt

$$\mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1})$$

$$\overset{f.s.}{=} \mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}} \,|\, \sigma\left[\sigma(\{Z_1, ..., Z_{m+1}\})\right]) = \mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{m+1}\}))$$

und wegen (4.5) damit

$$\mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}} \,|\, \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1}) \overset{f.s.}{=} P(E_{m,\alpha}^{full}).$$

Wir erhalten

$$\mathbb{E}\left(\mathbb{1}_{E_{m,\alpha}^{full}} \cdot \prod_{i=m+1}^{n} \mathbb{1}_{E_i} \middle| \sigma(\{Z_1, ..., Z_{m+1}\}), Z_{m+2}, ..., Z_{n+1}\right)$$

$$\overset{f.s.}{=} P(E_{m,\alpha}^{full}) \cdot \prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}} = \mathbb{E}\left(\mathbb{1}_{E_{m,\alpha}^{full}}\right) \prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}$$

und schließlich

$$\mathbb{E}\left(\prod_{i=m}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right) = \mathbb{E}(\mathbb{1}_{E_{m,\alpha}^{full}}) \cdot \mathbb{E}\left(\prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right). \tag{4.7}$$

Da m aus $\{1, ..., n-1\}$ beliebig gewählt war, erhalten wir aus Gleichung (4.7)

$$\mathbb{E}\left(\prod_{i=1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right) = \mathbb{E}(\mathbb{1}_{E_{1,\alpha}^{full}}) \cdot \mathbb{E}\left(\prod_{i=2}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right).$$

Nach iterativer Anwendung von (4.7) erhalten wir schließlich

$$\mathbb{E}\left(\prod_{i=1}^{n} \mathbb{1}_{E_{i,\alpha}^{full}}\right) = \prod_{i=1}^{n} \mathbb{E}(\mathbb{1}_{E_{i,\alpha}^{full}}).$$

Diese Aussage gilt für alle $n \in \mathbb{N}$, da sie für $n = 1$ trivialerweise erfüllt ist. Da es sich bei

$$\mathbb{1}_{E_{1,\alpha}^{full}}, ..., \mathbb{1}_{E_{n,\alpha}^{full}}$$

um bernoulliverteilte Zufallsvariablen handelt, sind diese folglich bezüglich P unabhängig. Da gleichzeitig diese Aussage für beliebige $n \in \mathbb{N}$ erfüllt ist, gilt schließlich, dass die gesamte Folge von Zufallsvariablen $(\mathbb{1}_{E_{i,\alpha}^{full}})_{i \in \mathbb{N}}$ unabhängig ist. □

Es konvergiert also nicht nur die *nicht beobachtbare* Überdeckungswahrscheinlichkeit $P(E_{n,\alpha}^{full})$ als deterministische Folge reeller Zahlen gegen das nominelle Konfidenzniveau $1 - \alpha$, sondern auch die aus der Stichprobe *berechenbare* empirische Überdeckungsrate $A_{n,\alpha}^{full}$ konvergiert als Folge von Zufallsvariablen in Wahrscheinlichkeit gegen den gleichen Wert. Damit kann $P(E_{n,\alpha}^{full})$ durch $A_{n,\alpha}^{full}$ für entsprechend große n gut geschätzt werden.

Wir betrachten nun die Funktion a, die wir in Unterkapitel 4.1 eingeführt haben, sowie für jedes $n \in \mathbb{N}$ einen Algorithmus zur Full-Methode a_n. Bezüglich der Funktion a und der Folge von Zufallsvariablen $(Z_i)_{i \in \mathbb{N}}$ formulieren wir

für die Funktionenfolge $(a_n)_{n\in\mathbb{N}}$ unter dem Wahrscheinlichkeitsmaß P folgende *Konsistenzbedingung*:

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{i\in\{1,\dots,n\}}|a_n(Z_1,\dots,Z_{n+1})(X_i)-a(X_i)|\right)=0. \tag{KF}$$

Im Lichte dieser Konsistenzbedingung formulieren wir den nächsten Satz:

Satz 4.2. *Sei $(Z_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen und $(\epsilon_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}$-wertigen Zufallsvariablen, welche die Postulate (P1) bis (P4) erfüllen. Für jedes $n\in\mathbb{N}$ betrachte einen permutationsinvarianten Algorithmus zur Full-Methode a_n, sowie die zu $Z_1,\dots,Z_n$ und a_n korrespondierenden Residuen $U_{1,Z_{n+1}},\dots,U_{n,Z_{n+1}}$, wobei unter diesen bezüglich P fast sicher keine Bindungen vorkommen. Darüber hinaus sei die Zufallsvariable U_n durch (4.2) definiert. Die gemeinsame Verteilungsfunktion F der Folge $(U_i)_{i\in\mathbb{N}}$ sei stetig und auf dem Intervall $[0,\infty)$ streng monoton steigend. Falls die Funktionenfolge $(a_n)_{n\in\mathbb{N}}$ die Konsistenzbedingung (KF) bezüglich a und $(Z_i)_{i\in\mathbb{N}}$ erfüllt, dann gilt für beliebiges $\alpha\in(0,1)$ die Aussage*

$$\operatorname*{plim}_{n\to\infty}U_{(h_{n,\alpha}:n),Z_{n+1}}=F^{-1}(1-\alpha),$$

wobei $U_{(h_{n,\alpha}:n),Z_{n+1}}$ die $h_{n,\alpha}$-Ordungsstatistik unter den Residuen sei und F^{-1} stehe für die durch (4.3) definierte verallgemeinerte Inverse von F.

Beweis. Sei $n\in\mathbb{N}$ beliebig. Für jedes $i\in\{1,\dots,n\}$ gilt

$$\begin{aligned}|U_{i,Z_{n+1}}-U_i| &= ||Y_i-a_n(Z_1,\dots,Z_{n+1})(X_i)|-|\epsilon_i||\\ &= ||Y_i-a_n(Z_1,\dots,Z_{n+1})(X_i)|-|Y_i-a(X_i)||\\ &\le |(Y_i-a_n(Z_1,\dots,Z_{n+1})(X_i))-(Y_i-a(X_i))|\\ &= |a_n(Z_1,\dots,Z_{n+1})(X_i)-a(X_i)|.\end{aligned}$$

Es gilt also

$$\sup_{i\in\{1,\dots,n\}}|U_{i,Z_{n+1}}-U_i|\le\sup_{i\in\{1,\dots,n\}}|a_n(Z_1,\dots,Z_{n+1})(X_i)-a(X_i)|.$$

Aus der Konsistenzbedingung (KF) folgt

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{i\in\{1,\dots,n\}}|U_{i,Z_{n+1}}-U_i|\right)=0. \tag{4.8}$$

Sei nun $\epsilon>0$ beliebig gewählt. Wir definieren die Menge

$$A_{n,\epsilon}:=\bigcap_{i=1}^{n}\{|U_{i,Z_{n+1}}-U_i|\le\epsilon\}.$$

Da

$$\left\{ \sup_{i\in\{1,\dots,n\}} |U_{i,Z_{n+1}} - U_i| \leq \epsilon \right\} = \bigcap_{i=1}^{n} \{|U_{i,Z_{n+1}} - U_i| \leq \epsilon\},$$

gilt wegen (4.8), dass

$$\lim_{n\to\infty} P(A_{n,\epsilon}) = 1. \tag{4.9}$$

Wir betrachten jetzt ein beliebiges $x \in \mathbb{R}$ und führen die empirischen Verteilungsfunktionen

$$\tilde{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_{i,Z_{n+1}} \leq x\}}$$

und

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq x\}}$$

zur Stichprobengröße n ein. Für alle $i \in \{1, ..., n\}$ gilt auf der Menge $A_{n,\epsilon}$ die Ungleichung

$$\mathbb{1}_{\{U_{i,Z_{n+1}} \leq x\}} \leq \mathbb{1}_{\{U_i \leq x+\epsilon\}},$$

wegen

$$\{U_{i,Z_{n+1}} \leq x\} \cap \{|U_{i,Z_{n+1}} - U_i| \leq \epsilon\} \subseteq \{U_i \leq x + \epsilon\}.$$

Gleichzeitig gilt auf $A_{n,\epsilon}$ die Ungleichung

$$\mathbb{1}_{\{U_i \leq x-\epsilon\}} \leq \mathbb{1}_{\{U_{i,Z_{n+1}} \leq x\}},$$

da

$$\{U_i \leq x - \epsilon\} \cap \{|U_{i,Z_{n+1}} - U_i| \leq \epsilon\} \subseteq \{U_{i,Z_{n+1}} \leq x\}.$$

Auf der Menge $A_{n,\epsilon}$ sind also die Ungleichungen

$$\hat{F}_n(x - \epsilon) \leq \tilde{F}_n(x) \leq \hat{F}_n(x + \epsilon).$$

erfüllt, womit

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| \leq \sup_{x\in\mathbb{R}} |\hat{F}_n(x - \epsilon) - \hat{F}_n(x + \epsilon)|$$

$$= \sup_{x\in\mathbb{R}} |\hat{F}_n(x - \epsilon) - F(x - \epsilon) + F(x - \epsilon) - F(x + \epsilon) + F(x + \epsilon) - \hat{F}_n(x + \epsilon)|$$

$$\leq \sup_{x\in\mathbb{R}} |\hat{F}_n(x-\epsilon) - F(x-\epsilon)| + \sup_{x\in\mathbb{R}} |F(x-\epsilon) - F(x+\epsilon)| + \sup_{x\in\mathbb{R}} |F(x+\epsilon) - \hat{F}_n(x+\epsilon)|.$$

Betrachte nun ein beliebiges $\gamma > 0$. Die Verteilungsfunktion F ist auf ganz $\mathbb{R}$ stetig, womit sie auch gleichmäßig stetig ist. Es existiert also ein (hinreichend kleines) $\epsilon_\gamma > 0$, sodass

$$\sup_{x\in\mathbb{R}} |F(x - \epsilon_\gamma) - F(x + \epsilon_\gamma)| \leq \frac{\gamma}{6}.$$

Beachte, dass diese Aussage auf ganz Ω gilt. Also folgt auf der Menge A_{n,ϵ_γ} aus den Ungleichungen

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x-\epsilon_\gamma) - F(x-\epsilon_\gamma)| \leq \frac{\gamma}{6}$$

und

$$\sup_{x\in\mathbb{R}} |F(x+\epsilon_\gamma) - \hat{F}_n(x+\epsilon_\gamma)| \leq \frac{\gamma}{6}$$

die Aussage

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| \leq 3\frac{\gamma}{6} = \frac{\gamma}{2}.$$

Wir erhalten also

$$\left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| > \frac{\gamma}{2}\right\} \cap A_{n,\epsilon_\gamma}$$

$$\subseteq \left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x-\epsilon_\gamma) - F(x-\epsilon_\gamma)| > \frac{\gamma}{6}\right\} \cup \left\{\sup_{x\in\mathbb{R}} |F(x+\epsilon_\gamma) - \hat{F}_n(x+\epsilon_\gamma)| > \frac{\gamma}{6}\right\}$$

$$= \left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \frac{\gamma}{6}\right\}$$

und somit

$$P\left(\left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| > \frac{\gamma}{2}\right\} \cap A_{n,\epsilon_\gamma}\right)$$

$$\leq P\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \frac{\gamma}{6}\right).$$

Da $(U_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger und identisch verteilter Zufallsvariablen ist, gilt der Satz von Glivenko-Cantelli. Wir haben also

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|\right) = 0. \tag{4.10}$$

Daraus folgt

$$\lim_{n\to\infty} P\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \frac{\gamma}{6}\right) = 0$$

und schließlich

$$\lim_{n\to\infty} P\left(\left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| > \frac{\gamma}{2}\right\} \cap A_{n,\epsilon_\gamma}\right) = 0. \tag{4.11}$$

Sei nun A^c_{n,ϵ_γ} das Komplement von A_{n,ϵ_γ} bezüglich Ω. Es gilt

$$P\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \tilde{F}_n(x)| > \frac{\gamma}{2}\right)$$

$$= P\left(\left\{\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-\tilde{F}_n(x)|>\frac{\gamma}{2}\right\}\cap(A_{n,\epsilon_\gamma}\cup A^c_{n,\epsilon_\gamma})\right)$$

$$= P\left(\left\{\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-\tilde{F}_n(x)|>\frac{\gamma}{2}\right\}\cap A_{n,\epsilon_\gamma}\right)+P\left(\left\{\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-\tilde{F}_n(x)|>\frac{\gamma}{2}\right\}\cap A^c_{n,\epsilon_\gamma}\right)$$

$$\leq P\left(\left\{\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-\tilde{F}_n(x)|>\frac{\gamma}{2}\right\}\cap A_{n,\epsilon_\gamma}\right)+P(A^c_{n,\epsilon_\gamma}).$$

Aus (4.9) folgt gleichzeitig

$$\lim_{n\to\infty}P(A^c_{n,\epsilon_\gamma})=0,$$

und wegen (4.11) gilt damit

$$\lim_{n\to\infty}P\left(\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-\tilde{F}_n(x)|>\frac{\gamma}{2}\right)=0. \tag{4.12}$$

Wir betrachten jetzt die auf ganz Ω gültige Ungleichung

$$\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|$$

$$=\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-\hat{F}_n(x)+\hat{F}_n(x)-F(x)|$$

$$\leq\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-\hat{F}_n(x)|+\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-F(x)|.$$

Daraus folgt

$$\left\{\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|>\gamma\right\}$$

$$\subseteq\left\{\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-\hat{F}_n(x)|>\frac{\gamma}{2}\right\}\cup\left\{\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-F(x)|>\frac{\gamma}{2}\right\},$$

und unter Anwendung der endlichen Subadditivität

$$P\left(\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|>\gamma\right)$$

$$\leq P\left(\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-\hat{F}_n(x)|>\frac{\gamma}{2}\right)+P\left(\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-F(x)|>\frac{\gamma}{2}\right).$$

Wegen (4.10) gilt

$$\lim_{n\to\infty}P\left(\sup_{x\in\mathbb{R}}|\hat{F}_n(x)-F(x)|>\frac{\gamma}{2}\right)=0,$$

und wegen (4.12) erhalten wir

$$\lim_{n\to\infty}P\left(\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|>\gamma\right)=0.$$

Da wir $\gamma > 0$ beliebig gewählt haben, gilt also

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|\right)=0. \tag{4.13}$$

Beachte nun, dass für (alle hinreichend große) $n \in \mathbb{N}$ mit $h_{n,\alpha} \leq n$ auf ganz Ω die Ungleichung

$$|\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})-F(U_{(h_{n,\alpha}:n),Z_{n+1}})|\leq\sup_{x\in\mathbb{R}}|\tilde{F}_n(x)-F(x)|$$

erfüllt ist. Aus (4.13) folgt somit

$$\operatorname*{plim}_{n\to\infty}|\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})-F(U_{(h_{n,\alpha}:n),Z_{n+1}})|=0. \tag{4.14}$$

Wir betrachten die auf ganz Ω gültige Ungleichung

$$\begin{aligned}
&|F(U_{(h_{n,\alpha}:n),Z_{n+1}})-(1-\alpha)|\\
&=|F(U_{(h_{n,\alpha}:n),Z_{n+1}})-\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})+\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})-(1-\alpha)|\\
&\leq|F(U_{(h_{n,\alpha}:n),Z_{n+1}})-\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})|+|\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})-(1-\alpha)|.
\end{aligned}$$

Beachte, dass

$$\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})\overset{f.s.}{=}\frac{h_{n,\alpha}}{n}$$

gilt, da $U_{(h_{n,\alpha}:n),Z_{n+1}}$ die $h_{n,\alpha}$-Ordnungsstatistik unter den Residuen $U_{1,Z_{n+1}},...,U_{n,Z_{n+1}}$ ist (beachte, dass das Vorkommen von Bindungen unter den Residuen per Annahme ausgeschlossen ist). Wir erinnern uns, dass gleichzeitig

$$h_{n,\alpha}=\lceil(1-\alpha)(n+1)\rceil,$$

womit

$$\frac{(1-\alpha)(n+1)}{n}\leq\frac{h_{n,\alpha}}{n}<\frac{(1-\alpha)(n+1)+1}{n}.$$

Da sowohl der linke wie auch der rechte Term in der obigen Ungleichungskette für $n \to \infty$ gegen $1-\alpha$ konvergiert, gilt

$$\lim_{n\to\infty}\frac{h_{n,\alpha}}{n}=1-\alpha,$$

und somit

$$\lim_{n\to\infty}|\tilde{F}_n(U_{(h_{n,\alpha}:n),Z_{n+1}})-(1-\alpha)|\overset{f.s.}{=}\lim_{n\to\infty}\left|\frac{h_{n,\alpha}}{n}-(1-\alpha)\right|=0.$$

In Kombination mit (4.14) erhalten wir also

$$\operatorname*{plim}_{n\to\infty}F(U_{(h_{n,\alpha}:n),Z_{n+1}})=1-\alpha.$$

Da nach Voraussetzung die Verteilungsfunktion F stetig sowie auf den Intervall $[0, \infty)$ streng monoton steigend ist und auf ganz Ω außerdem die Ungleichung $U_i \geq 0$ für beliebiges $i \in \mathbb{N}$ erfüllt ist, gilt $F(0) = 0$. Wenn wir nun $F^{-1}(0) := 0$ definieren, dann ist F^{-1} die inverse Funktion der Einschränkung von F auf $[0, \infty)$, und somit eine stetige Funktion mit der auf ganz Ω gültigen Eigenschaft

$$F^{-1}(F(U_{(h_{n,\alpha}:n),Z_{n+1}})) = U_{(h_{n,\alpha}:n),Z_{n+1}}.$$

Wir erhalten schließlich mit dem Satz von den stetigen Abbildungen

$$\operatorname*{plim}_{n\to\infty} U_{(h_{n,\alpha}:n),Z_{n+1}} = \operatorname*{plim}_{n\to\infty} F^{-1}(F(U_{(h_{n,\alpha}:n),Z_{n+1}})) = F^{-1}(1-\alpha).$$

□

Wenn wir für beliebige $\alpha \in (0,1)$ und $n \in \mathbb{N}$ mit $h_{n,\alpha} \leq n$ die Zufallsvariable $2 \cdot U_{(h_{n,\alpha}:n),Z_{n+1}}$ als „Breite“ des Konfidenzbereiches $C^{full}_{n,\alpha}(X_{n+1})$ auffassen, dann konvergiert diese unter den in Satz 4.2 festgelegten Bedingungen in Wahrscheinlichkeit gegen die Länge von $C_{n,\alpha}(X_{n+1})$, welche dem Doppelten des $(1-\alpha)$-Quantils der Verteilung F entspricht. Beachte, dass $U_{(h_{n,\alpha}:n),Z_{n+1}}$ das empirische $\frac{h_{n,\alpha}}{n}$-Quantil unter den Zufallsvariablen $U_{1,Z_{n+1}}, ..., U_{n,Z_{n+1}}$ ist und

$$\lim_{n\to\infty} \frac{h_{n,\alpha}}{n} = 1 - \alpha$$

(wie im Beweis von Satz 4.2 gezeigt).

4.3 Asymptotik der Split-Methode

Wie auch in Unterkapitel 4.2 führen wir wieder eine Folge von Zufallsvariablen $(Z_i)_{i\in\mathbb{N}}$ ein. Wir betrachten außerdem zu gegebenem und festem $\delta \in [\frac{1}{2}, 1)$ eine Folge $(I_{n,1}, I_{n,2})_{n\geq 2}$. Für jedes $n \in \mathbb{N}$ mit $n \geq 2$ seien dabei $I_{n,1}$ und $I_{n,2}$ disjunkte Mengen, welche eine Partition auf der Menge $\{1, ..., n\}$ darstellen, wobei wir

$$I_{n,1} := \{i(n,1), ..., i(n,m_n)\}$$

und

$$I_{n,2} := \{j(n,1), ..., j(n,k_n)\}$$

mit

$$m_n := \lfloor \delta n \rfloor$$

und

$$k_n := n - m_n$$

schreiben (siehe auch Unterkapitel 3.2). Bezüglich verschiedener $u, v \in \mathbb{N}$ wird dabei nicht gefordert, dass $i(u,s)$ und $i(v,s)$ für natürliche Indizes s mit $s \leq \min\{m_u, m_v\}$ bzw. $j(u,s)$ und $j(v,s)$ mit $s \leq \min\{k_u, k_v\}$ gleiche Werte annehmen. Ferner sei $a_{n,\delta}$ ein Algorithmus zur Split-Methode nach Definition 3.2, wobei für $n \in \mathbb{N}$ mit $n \geq 2$ die Realisationen von $Z_{i(n,1)}, ..., Z_{i(n,m_n)}$ dem Algorithmus $a_{n,\delta}$ als Trainingsmenge dienen, während $Z_{j(n,1)}, ..., Z_{j(n,k_n)}$ wieder zur Berechnung der Residuen $U_{j(n,1),\delta}, ..., U_{j(n,k_n),\delta}$ herangezogen werden. Zur Beobachtung Z_{n+1} sei noch das Residuum $U_{n+1,\delta}$ zu betrachten. Für gegebenes $\alpha \in (0,1)$ definieren wir außerdem

$$\tilde{h}_{n,\alpha} := \lceil (1-\alpha)(k_n+1) \rceil,$$

und unter der Bedingung $\tilde{h}_{n,\alpha} \leq k_n$ betrachten wir $U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$ als die $\tilde{h}_{n,\alpha}$-te Ordnungsstatistik unter den Residuen $U_{j(n,1),\delta}, ..., U_{j(n,k_n),\delta}$. Der Ausdruck $C^{split}_{n,\delta,\alpha}(X_{n+1})$ stehe wieder für den Konfidenzbereich, welcher durch die Split-Methode zur Stichprobengröße n bestimmt wird. Auch hier soll die Notation einfach gehalten werden. Wir definieren daher

$$E^{split}_{n,\delta,\alpha} := \{Y_{n+1} \in C^{split}_{n,\delta,\alpha}(X_{n+1})\},$$

also das Ereignis, dass die Zielvariable Y_{n+1} im Konfidenzbereich $C^{split}_{n,\delta,\alpha}(X_{n+1})$ liegt. Für jedes $n \in \mathbb{N}$ mit $n \geq 2$ sei dann die *empirische Überdeckungsrate* definiert durch

$$A^{split}_{n,\delta,\alpha} := \frac{1}{n-1} \sum_{i=2}^{n} \mathbb{1}_{E^{split}_{i,\delta,\alpha}},$$

wobei für jedes natürliche $i \geq 2$ der Ausdruck $\mathbb{1}_{E^{split}_{i,\delta,\alpha}}$ die Indikatorfunktion des Ereignisses $E^{split}_{i,\delta,\alpha}$ darstelle. Die Größe $A^{split}_{n,\delta,\alpha}$ hat die selbe Interpretation wie $A^{full}_{n,\alpha}$ aus dem vorhergehenden Unterkapitel. Sie steht für den relativen Anteil der Zufallsvariablen $Y_3, ..., Y_{n+1}$, die sich in ihren jeweiligen Konfidenzbereichen $C^{split}_{2,\delta,\alpha}(X_3), ..., C^{split}_{n,\delta,\alpha}(X_{n+1})$ realisieren, und kann demnach als Schätzer für die Überdeckungswahrscheinlichkeit $P(E^{split}_{n,\delta,\alpha})$ betrachtet werden. Wir betonen an dieser Stelle, dass auch hier die auf wachsendes n Bezug nehmende Asymptotik sich an einem On-Line Setting orientiert. Die Menge an Daten zur Bestimmung von $C^{split}_{n,\delta,\alpha}(X_{n+1})$ wird also wieder mit anwachsendem n entsprechend um die neuen Instanzen erweitert.

Wir betrachten nun den Satz:

Satz 4.3. *Sei $(Z_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen, welche das Postulat (P1) erfülle, und betrachte für festes $\delta \in [\frac{1}{2}, 1)$ eine beliebige Folge von Partitionen $(I_{n,1}, I_{n,2})_{n\geq 2}$ wie oben beschrieben. Für jedes $n \in \mathbb{N}$*

mit $n \geq 2$ sei darüber hinaus $a_{n,\delta}$ ein Algorithmus zur Split-Methode und $U_{j(n,1),\delta}, ..., U_{j(n,k_n),\delta}, U_{n+1,\delta}$ seien die mit $Z_1, ..., Z_{n+1}$ und $a_{n,\delta}$ korrespondierenden Residuen, wobei unter diesen bezüglich P fast sicher keine Bindungen vorkommen. Dann gilt für beliebiges $\alpha \in (0,1)$, dass

$$\operatorname*{plim}_{n\to\infty} A^{split}_{n,\delta,\alpha} = 1 - \alpha.$$

Beweis. Der Beweis dieses Satzes wird analog zu jenem des Satzes 4.1 geführt. Wir werden daher auf die Präsentation der analogen Schritte verzichten und uns auf die wichtigsten Zwischenresultate, deren Formulierungen für die Split-Methode spezifisch sind, beschränken.

Wir nehmen zunächst an, dass $(\mathbb{1}_{E^{split}_{i,\delta,\alpha}})_{i\geq 2}$ eine Folge bezüglich P unabhängiger Zufallsvariablen ist. Wir haben in Unterkapitel 3.2 gesehen, dass

$$\lim_{n\to\infty} P(E^{split}_{n,\delta,\alpha}) = 1 - \alpha, \tag{4.15}$$

woraus unmittelbar

$$\lim_{n\to\infty} Var(\mathbb{1}_{E^{split}_{n,\delta,\alpha}}) = \lim_{n\to\infty} \Big(P(E^{split}_{n,\delta,\alpha}) \cdot (1 - P(E^{split}_{n,\delta,\alpha}))\Big) = \alpha(1-\alpha)$$

folgt. Weil $Var(\mathbb{1}_{E^{split}_{n,\delta,\alpha}})$ endlich ist für jedes $n \in \mathbb{N}$ mit $n \geq 2$ und $\alpha(1-\alpha) \in \mathbb{R}$, gilt

$$\sup_{n\geq 2} Var(\mathbb{1}_{E^{split}_{n,\delta,\alpha}}) < \infty.$$

Für $(\mathbb{1}_{E^{split}_{i,\delta,\alpha}})_{i\geq 2}$ gilt also ein schwaches Gesetz der großen Zahlen; vgl. Kusolitsch (2011, S.254). Wir erhalten

$$\operatorname*{plim}_{n\to\infty} \left(\frac{1}{n-1} \sum_{i=2}^{n} \mathbb{1}_{E^{split}_{i,\delta,\alpha}} - \frac{1}{n-1} \sum_{i=2}^{n} \mathbb{E}(\mathbb{1}_{E^{split}_{i,\delta,\alpha}}) \right) = 0.$$

Wegen (4.15) gilt wegen des Grenzwertsatzes von Cauchy auch

$$\lim_{n\to\infty} \frac{1}{n-1} \sum_{i=2}^{n} \mathbb{E}(\mathbb{1}_{E^{split}_{i,\delta,\alpha}}) = \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=2}^{n} P(E^{split}_{i,\delta,\alpha}) = 1 - \alpha;$$

vgl. Heuser (2006, S.177). Es folgt schließlich

$$\operatorname*{plim}_{n\to\infty} (A^{split}_{n,\delta,\alpha} - (1-\alpha)) = 0.$$

Es bleibt zu zeigen, dass $(\mathbb{1}_{E^{split}_{i,\delta,\alpha}})_{i\geq 2}$ eine Folge unabhängiger Zufallsvariablen ist. Sei $n \in \mathbb{N}$ mit $n \geq 2$. Die σ-Algebra $\sigma(\{Z_1, ..., Z_{n+1}\})$ sei gleich definiert wie im Beweis von Satz 4.1. Jedes Ereignis in dieser σ-Algebra enthält Informationen über die Zufallsmenge

$$\{Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,1)}, ..., Z_{j(n,k_n+1)}\},$$

wobei $j(n, k_n + 1) := n + 1$, jedoch nicht über den konkreten Zufallsvektor

$$(Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,\sigma(1))}, ..., Z_{j(n,\sigma(k_n+1))})$$

mit beliebiger Permutation $\sigma \in S_{k_n+1}$. Nach Voraussetzung gibt es unter den Residuen $U_{j(n,1),\delta}, ..., U_{j(n,k_n+1),\delta}$ fast sicher keine Bindungen. Für beliebiges $k \in \{1, ..., k_n + 1\}$ gilt daher

$$\sum_{i \in I_{n,2} \cup \{n+1\}} \mathbb{1}_{\{rang(U_{i,\delta})=k\}} \overset{f.s.}{=} 1$$

und für jedes Ereignis $A \in \sigma(\{Z_1, ..., Z_{n+1}\})$ somit

$$\sum_{i \in I_{n,2} \cup \{n+1\}} \mathbb{E}(\mathbb{1}_{\{rang(U_{i,\delta})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_A). \tag{4.16}$$

Im Beweis von Lemma 3.2 haben wir gesehen, dass für beliebige $\sigma \in S_{k_n+1}$ der Zufallsvektor der Residuen

$$(U_{j(n,\sigma(1)),\delta}, ..., U_{j(n,\sigma(k_n+1)),\delta})$$

als eine deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,\sigma(1))}, ..., Z_{j(n,\sigma(k_n+1))})$$

geschrieben werden kann, womit auch der Zufallsvektor

$$(\mathbb{1}_{\{rang(U_{j(n,\sigma(1)),\delta})=k\}}, ..., \mathbb{1}_{\{rang(U_{j(n,\sigma(k_n+1)),\delta})=k\}})$$

eine deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,\sigma(1))}, ..., Z_{j(n,\sigma(k_n+1))})$$

ist. Gleichzeitig gilt wegen $A \in \sigma(\{Z_1, ..., Z_{n+1}\})$, dass die Zufallsvariable $\mathbb{1}_A$ eine Funktion der Zufallsmenge

$$\{Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,1)}, ..., Z_{j(n,k_n+1)}\}$$

ist (siehe auch den Beweis von Satz 4.1). Da außerdem

$$\{Z_{i(n,1)}(\omega), ..., Z_{i(n,m_n)}(\omega), Z_{j(n,1)}(\omega), ..., Z_{j(n,k_n+1)}(\omega)\}$$

$$= \{Z_{i(n,1)}(\omega), ..., Z_{i(n,m_n)}(\omega), Z_{j(n,\sigma(1))}(\omega), ..., Z_{j(n,\sigma(k_n+1))}(\omega)\}$$

für beliebige $\omega \in \Omega$ gilt, kann auch der Zufallsvektor

$$(\mathbb{1}_{\{rang(U_{j(n,\sigma(1)),\delta})=k\}} \cdot \mathbb{1}_A, ..., \mathbb{1}_{\{rang(U_{j(n,\sigma(k_n+1)),\delta})=k\}} \cdot \mathbb{1}_A)$$

als deterministische und von σ unabhängige Funktion des Zufallsvektors

$$(Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,\sigma(1))}, ..., Z_{j(n,\sigma(k_n+1))})$$

und der Zufallsmenge

$$\{Z_{i(n,1)}, ..., Z_{i(n,m_n)}, Z_{j(n,\sigma(1))}, ..., Z_{j(n,\sigma(k_n+1))}\}$$

betrachtet werden. Die Folge $(Z_i)_{i\in\mathbb{N}}$ ist nach Voraussetzung unabhängig und identisch verteilt, womit sie nach Lemma 2.2 vertauschbar ist. Die Zufallsvariablen

$$\mathbb{1}_{\{rang(U_{j(n,1),\delta})=k\}} \cdot \mathbb{1}_A, ..., \mathbb{1}_{\{rang(U_{j(n,k_n+1),\delta})=k\}} \cdot \mathbb{1}_A$$

sind also ebenfalls vertauschbar. Nach Lemma 2.1 sind diese somit in Verteilung gleich. Für beliebige $i \in I_{n,2}$ gilt also

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{i,\delta})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,\delta})=k\}} \cdot \mathbb{1}_A).$$

Unter Ausnutzung der Gleichung (4.16) erhalten wir damit

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{n+1,\delta})=k\}} \cdot \mathbb{1}_A) = \mathbb{E}\left(\frac{1}{k_n+1} \cdot \mathbb{1}_A\right).$$

Weil

$$\mathbb{1}_{\{rang(U_{n+1,\delta})\leq\tilde{h}_{n,\alpha}\}} = \sum_{i=1}^{\tilde{h}_{n,\alpha}} \mathbb{1}_{\{rang(U_{n+1,\delta})=i\}},$$

haben wir

$$\mathbb{E}(\mathbb{1}_{\{rang(U_{i,\delta})\leq\tilde{h}_{n,\alpha}\}} \cdot \mathbb{1}_A) = \mathbb{E}\left(\frac{\tilde{h}_{n,\alpha}}{k_n+1} \cdot \mathbb{1}_A\right). \tag{4.17}$$

Falls $\tilde{h}_{n,\alpha} \leq k_n$, dann gilt

$$E^{split}_{n,\delta,\alpha} = \{U_{n+1,\delta} \in [0, U_{(\tilde{h}_{n,\alpha}:k_n),\delta})\},$$

also

$$E^{split}_{n,\delta,\alpha} = \{rang(U_{n+1,\delta}) \leq \tilde{h}_{n,\alpha}\}$$

sowie

$$P(E^{split}_{n,\delta,\alpha}) = \frac{\tilde{h}_{n,\alpha}}{k_n+1}$$

(siehe auch den Beweis von Satz 2.1 und Unterkapitel 3.2). Für den Fall $\tilde{h}_{n,\alpha} > k_n$, haben wir außerdem $C^{split}_{n,\delta,\alpha}(X_{n+1}) = \mathbb{R}$ festgelegt (siehe Unterkapitel 3.2), womit $E^{split}_{n,\delta,\alpha} = \Omega$ und damit $P(E^{split}_{n,\delta,\alpha}) = 1$ gilt. Unter Berücksichtigung von (4.17) erhalten wir für beide Fälle schließlich

$$\mathbb{E}(\mathbb{1}_{E^{split}_{n,\delta,\alpha}} \mid \sigma(\{Z_1, ..., Z_{n+1}\})) \overset{f.s.}{=} P(E^{split}_{n,\delta,\alpha}). \tag{4.18}$$

Sei nun $n \in \mathbb{N}$ mit $n \geq 3$ und betrachte $m \in \{2, ..., n-1\}$. Die Zufallsvariable $\prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}$ hängt von der Zufallsmenge $\{Z_1, ..., Z_{m+1}\}$ sowie von den einzelnen Zufallsvariablen $Z_{m+2}, ..., Z_{n+1}$ ab. Gleichzeitig ist $(Z_i)_{i\in\mathbb{N}}$ nach Voraussetzung eine Folge unabhängiger Zufallsvariablen. Die Zufallsvariable $\mathbb{1}_{E_{m,\delta,\alpha}^{split}}$ ist damit unabhängig von den einzelnen Zufallsvariablen $Z_{m+2}, ..., Z_{n+1}$. Mittels dieser Tatsachen kann unter Anwendung von (4.18) sowie der Rechenregeln für bedingte Erwartungswerte gezeigt werden, dass

$$\mathbb{E}\left(\prod_{i=m}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}\right) = \mathbb{E}(\mathbb{1}_{E_{m,\delta,\alpha}^{split}}) \cdot \mathbb{E}\left(\prod_{i=m+1}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}\right) \tag{4.19}$$

(siehe auch den Beweis von Satz 4.1). Da m aus $\{2, ..., n-1\}$ beliebig gewählt ist, folgern wir aus (4.19) die Gleichung

$$\mathbb{E}\left(\prod_{i=2}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}\right) = \mathbb{E}(\mathbb{1}_{E_{2,\delta,\alpha}^{split}}) \cdot \mathbb{E}\left(\prod_{i=3}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}\right).$$

Nach iterativer Anwendung von (4.19) erhalten wir schließlich

$$\mathbb{E}\left(\prod_{i=2}^{n} \mathbb{1}_{E_{i,\delta,\alpha}^{split}}\right) = \prod_{i=2}^{n} \mathbb{E}(\mathbb{1}_{E_{i,\delta,\alpha}^{split}}).$$

Da diese Gleichung trivialerweise auch für $n = 2$ erfüllt ist, gilt sie über alle $n \in \mathbb{N}$ mit $n \geq 2$. Somit sind die bernoulliverteilten Zufallsvariablen

$$\mathbb{1}_{E_{2,\delta,\alpha}^{split}}, ..., \mathbb{1}_{E_{n,\delta,\alpha}^{split}}$$

bezüglich P unabhängig. Weil diese Aussage für beliebige $n \in \mathbb{N}$ mit $n \geq 2$ erfüllt ist, gilt, dass die gesamte Folge von Zufallsvariablen $(\mathbb{1}_{E_{i,\delta,\alpha}^{split}})_{i\geq 2}$ unabhängig ist. □

Ebenso wie bei der Full-Methode konvergiert auch hier die aus den Daten berechenbare empirische Überdeckungsrate $A_{n,\delta,\alpha}^{split}$ in Wahrscheinlichkeit gegen das nominelle Konfidenzniveau $1-\alpha$ und eignet sich für große n zum Schätzen der Überdeckungswahrscheinlichkeit $P(E_{n,\delta,\alpha}^{split})$, welche als deterministische Folge gegen $1-\alpha$ konvergiert.

Sei a die in Unterkapitel 4.1 eingeführte messbare Funktion. Für $\delta \in [\frac{1}{2}, 1)$ betrachten wir eine beliebige Folge von Partitionen $(I_{n,1}, I_{n,2})_{n\geq 2}$ wie wir sie am Anfang dieses Unterkapitels beschrieben haben, sowie einen Algorithmus $a_{n,\delta}$ zur Split-Methode für beliebige $n \in \mathbb{N}$ mit $n \geq 2$. Bezüglich der Funktion a, der Folge von Zufallsvariablen $(Z_i)_{i\in\mathbb{N}}$ und der Folge $(I_{n,1}, I_{n,2})_{n\geq 2}$ formulieren wir für die Funktionenfolge $(a_{n,\delta})_{n\geq 2}$ unter dem Wahrscheinlichkeitsmaß P

die *Konsistenzbedingung*:

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{i\in I_{n,2}}\left|a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i)-a(X_i)\right|\right)=0. \tag{KS}$$

Mit dieser formulieren wir den Satz:

Satz 4.4. *Sei $(Z_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}^{d+1}$-wertigen Zufallsvariablen und $(\epsilon_i)_{i\in\mathbb{N}}$ eine Folge von $\mathbb{R}$-wertigen Zufallsvariablen, welche die Postulate (P1) bis (P4) erfüllen. Für beliebige $n\in\mathbb{N}$ mit $n\geq 2$ und festen $\delta\in[\frac{1}{2},1)$ seien die disjunkten Mengen $I_{n,1}$ und $I_{n,2}$ eine Partition vom oben beschriebenen Typus der Menge $\{1,...,n\}$ und $a_{n,\delta}$ sei ein Algorithmus zur Split-Methode. Seien außerdem $U_{j(n,1),\delta},...,U_{j(n,k_n),\delta}$ die zu $a_{n,\delta}$ und $Z_{j(n,1)},...,Z_{j(n,k_n)}$ korrespondierenden Residuen, wobei unter diesen bezüglich P fast sicher keine Bindungen vorkommen. Darüber hinaus sei für jedes $n\in\mathbb{N}$ die Zufallsvariable U_n durch (4.2) definiert. Die gemeinsame Verteilungsfunktion F der Folge $(U_i)_{i\in\mathbb{N}}$ sei außerdem stetig und auf dem Intervall $[0,\infty)$ streng monoton steigend. Falls die Funktionenfolge $(a_{n,\delta})_{n\geq 2}$ die Konsistenzbedingung (KS) bezüglich a, $(I_{n,1},I_{n,2})_{n\geq 2}$ und $(Z_i)_{i\in\mathbb{N}}$ erfüllt, dann gilt für beliebiges $\alpha\in(0,1)$ die Aussage*

$$\operatorname*{plim}_{n\to\infty}U_{(\tilde{h}_{n,\alpha}:k_n),\delta}=F^{-1}(1-\alpha),$$

wobei $U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$ die $\tilde{h}_{n,\alpha}$-te Ordnungsstatistik unter den Residuen sei und F^{-1} stehe für die durch (4.3) definierte verallgemeinerte Inverse von F.

Beweis. Der Beweis dieses Satzes ist größtenteils analog zu jenem des Satzes 4.2. Auch hier werden wir nur die wichtigsten Zwischenresultate formulieren.

Sei $n\in\mathbb{N}$ mit $n\geq 2$ beliebig. Für jedes $i\in I_{n,2}$ gilt

$$\begin{aligned}|U_{i,\delta}-U_i| &= ||Y_i-a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i)|-|\epsilon_i||\\ &= ||Y_i-a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i)|-|Y_i-a(X_i)||\\ &\leq |(Y_i-a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i))-(Y_i-a(X_i))|\\ &= |a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i)-a(X_i)|.\end{aligned}$$

Es gilt also

$$\sup_{i\in I_{n,2}}|U_{i,\delta}-U_i|\leq\sup_{i\in I_{n,2}}|a_{n,\delta}(Z_{i(n,1)},...,Z_{i(n,m_n)})(X_i)-a(X_i)|.$$

Aus der Konsistenzbedingung (KS) folgt

$$\operatorname*{plim}_{n\to\infty}\left(\sup_{i\in I_{n,2}}|U_{i,\delta}-U_i|\right)=0. \tag{4.20}$$

Sei nun $\epsilon > 0$ beliebig gewählt. Wir definieren die Menge

$$\tilde{A}_{n,\epsilon} := \bigcap_{i \in I_{n,2}} \{|U_{i,\delta} - U_i| \leq \epsilon\}.$$

Da

$$\left\{ \sup_{i \in I_{n,2}} |U_{i,\delta} - U_i| \leq \epsilon \right\} = \bigcap_{i \in I_{n,2}} \{|U_{i,\delta} - U_i| \leq \epsilon\},$$

gilt wegen (4.20), dass

$$\lim_{n \to \infty} P(\tilde{A}_{n,\epsilon}) = 1. \tag{4.21}$$

Wir betrachten jetzt ein beliebiges $x \in \mathbb{R}$ und führen die empirischen Verteilungsfunktionen

$$\tilde{F}_{n,\delta}(x) := \frac{1}{k_n} \sum_{i \in I_{n,2}} \mathbb{1}_{\{U_{i,\delta} \leq x\}}$$

und

$$\hat{F}_{n,\delta}(x) := \frac{1}{k_n} \sum_{i \in I_{n,2}} \mathbb{1}_{\{U_i \leq x\}}$$

zur Stichprobengröße n ein. Wir erinnern uns außerdem, dass

$$\lim_{n \to \infty} k_n = \infty \tag{4.22}$$

(siehe Unterkapitel 3.2). Beachte nun, dass $|I_{n,2}| = k_n$, d.h. die Anzahl der Elemente in $I_{n,2}$ strebt gegen ∞ für größer werdendes n. Da $(U_i)_{i \in \mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen ist, gilt[7]

$$\operatorname*{plim}_{n \to \infty} \left(\sup_{x \in \mathbb{R}} |\hat{F}_{n,\delta}(x) - F(x)| \right) = 0. \tag{4.23}$$

Weil darüber hinaus die Verteilungsfunktion F an jeder beliebigen Stelle $x \in \mathbb{R}$ stetig ist, kann unter Anwendung von (4.21) und (4.23) gezeigt werden, dass

$$\operatorname*{plim}_{n \to \infty} \left(\sup_{x \in \mathbb{R}} |\tilde{F}_{n,\delta}(x) - F(x)| \right) = 0 \tag{4.24}$$

(siehe auch den Beweis von Satz 4.2). Für (alle hinreichend große) $n \in \mathbb{N}$ mit $\tilde{h}_{n,\alpha} \leq k_n$ gilt auf ganz Ω die Ungleichung

$$|\tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta})| \leq \sup_{x \in \mathbb{R}} |\tilde{F}_{n,\delta}(x) - F(x)|$$

Aus (4.24) folgt somit

$$\operatorname*{plim}_{n \to \infty} |\tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta})| = 0. \tag{4.25}$$

Wir betrachten die auf ganz Ω gültige Ungleichung

$$\begin{aligned}
&|F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - (1-\alpha)| \\
&= |F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - \tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) + \tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - (1-\alpha)| \\
&\leq |F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - \tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta})| + |\tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - (1-\alpha)|.
\end{aligned}$$

Beachte, dass

$$\tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) \overset{f.s.}{=} \frac{\tilde{h}_{n,\alpha}}{k_n}$$

gilt, da $U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$ die $\tilde{h}_{n,\alpha}$-Ordnungsstatistik unter den Residuen $U_{j(n,1),\delta}, ..., U_{j(n,k_n),\delta}$ ist (beachte, dass hier das Vorkommen von Bindungen unter den Residuen per Annahme ausgeschlossen ist) . Wir erinnern uns, dass gleichzeitig

$$\tilde{h}_{n,\alpha} = \lceil (1-\alpha)(k_n+1) \rceil,$$

womit

$$\frac{(k_n+1)(1-\alpha)}{k_n} \leq \frac{\tilde{h}_{n,\alpha}}{k_n} < \frac{(k_n+1)(1-\alpha)+1}{k_n}.$$

Wegen (4.22) konvergiert sowohl der linke als auch der rechte Term in der obigen Ungleichungskette für $n \to \infty$ gegen $1-\alpha$. Es gilt also

$$\lim_{n\to\infty} \frac{\tilde{h}_{n,\alpha}}{k_n} = 1-\alpha,$$

und somit

$$\lim_{n\to\infty} |\tilde{F}_{n,\delta}(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) - (1-\alpha)| \overset{f.s.}{=} \lim_{n\to\infty} \left| \frac{\tilde{h}_{n,\alpha}}{n} - (1-\alpha) \right| = 0.$$

In Kombination mit (4.25) erhalten wir

$$\operatorname*{plim}_{n\to\infty} F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta}) = 1-\alpha.$$

Da nach Voraussetzung F stetig sowie auf den Intervall $[0,\infty)$ streng monoton steigend ist und auf ganz Ω die Ungleichung $U_i \geq 0$ für beliebiges $i \in \mathbb{N}$ gilt, ist $F(0) = 0$ erfüllt. Wenn wir nun $F^{-1}(0) := 0$ definieren, dann ist F^{-1} die inverse Funktion der Einschränkung von F auf $[0,\infty)$, und somit eine stetige Funktion mit

$$F^{-1}(F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta})) = U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$$

auf ganz Ω. Wir erhalten schließlich gemäß des Satzes von den stetigen Abbildungen

$$\operatorname*{plim}_{n\to\infty} U_{(\tilde{h}_{n,\alpha}:k_n),\delta} = \operatorname*{plim}_{n\to\infty} F^{-1}(F(U_{(\tilde{h}_{n,\alpha}:k_n),\delta})) = F^{-1}(1-\alpha).$$

□

Die Konsequenzen von Satz 4.4 für die Split-Methode sind die gleichen wie jene, die sich aus Satz 4.2 für die Full-Methode ergeben. Für $\alpha \in (0,1)$ und $n \in \mathbb{N}$ mit $\tilde{h}_{n,\alpha} \leq k_n$ entspricht die Zufallsvariable $2 \cdot U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$ der Breite des Konfidenzintervalls $C^{split}_{n,\delta,\alpha}(X_{n+1})$. Diese konvergiert dann unter den in Satz 4.4 festgelegten Bedingungen in Wahrscheinlichkeit gegen die Länge des Konfidenzintervalls $C_{n,\alpha}(X_{n+1})$, also gegen $2 \cdot F^{-1}(1-\alpha)$. Wir erinnern uns, dass $U_{(\tilde{h}_{n,\alpha}:k_n),\delta}$ das empirische $\frac{\tilde{h}_{n,\alpha}}{k_n}$-Quantil unter den Zufallsvariablen $U_{j(n,1),\delta}, ..., U_{j(n,k_n),\delta}$ mit

$$\lim_{n\to\infty} \frac{\tilde{h}_{n,\alpha}}{k_n} = 1 - \alpha$$

gilt (wie wir im vorhergehenden Beweis gesehen haben).

5 Exemplarische Demonstration der Asymptotik der Full- und der Split-Methode durch Simulation eines On-Line Settings

In diesem Kapitel sollen durch *Simulation* von Stichproben innerhalb eines On-Line Settings Daten gewonnen werden, die zur Bestimmung von konformalen Konfidenzbereichen sowohl nach der Full- wie auch nach der Split-Methode geeignet sind. Die Simulation soll dabei im Rahmen eines *linearen Modells* erfolgen, wobei das asymptotische Verhalten der Konfidenzbereiche, die auf Grundlage des *Kleinsten-Quadrate-Schätzers* und der *Lasso-Methode* berechnet werden, im Mittelpunkt steht. Konkret geht es dabei um das asymptotische Verhalten der empirischen Überdeckungsrate sowie die Breite der Konfidenzbereiche in Abhängigkeit von der Größe der simulierten Stichproben.

5.1 Ein lineares Modell als Grundlage der Simulation

Um ein geeignetes ein On-Line Setting zu erhalten, simulieren wir unabhängig voneinander eine festgelegte Anzahl $s \in \mathbb{N}$ an Stichproben vom Umfang $n+1$, wobei $n \in \mathbb{N}$ fest gewählt sei. Um dies zu bewerkstelligen betrachten wir für jede Iteration $t \in \{1, ..., s\}$ eine Folge von Zufallsvariablen $(X_i^t, Y_i^t)_{i\in\mathbb{N}}$ mit den folgenden Eigenschaften: für alle $i \in \mathbb{N}$ seien die Regressoren X_i^t für $d \in \mathbb{N}$ wieder $\mathbb{R}^d$-wertige Zufallsvariablen, wobei

$$X_i^t \equiv (X_{i,1}^t, ..., X_{i,d}^t) \in \mathbb{R}^{1\times d}$$

als Zeilenvektor aufzufassen sei. Dabei sei $(X_i^t)_{i\in\mathbb{N}}$ eine Folge unabhängiger und identisch verteilter Zufallsvariablen, wobei für alle $i \in \mathbb{N}$, $j \in \{1, ..., d\}$ und

s	n	d	α	β_0	β_1	β_2	β_3	β_4	β_5
400	1000	5	0.1	0.12	-0.12	-0.03	0.132	0.72	-0.10

Tabelle 1: *Die in der Simulation verwendeten Werte für die Parameter s, n, d, α, β_0, β_1, β_2, β_3, β_4 und β_5.*

$t \in \{1, ..., s\}$ die $\mathbb{R}$-wertigen Regressoren $X_{i,j}^t$ einer *Dreiecksverteilung* folgen mit der Wahrscheinlichkeitsdichte

$$f_\Delta(x) := \frac{1}{6} \cdot \max\{\sqrt{6} - |x|, 0\}$$

für alle $x \in \mathbb{R}$. Diese Verteilung weist einen Erwartungswert von 0 und eine Varianz von 1 auf. Beachte, dass die Regressoren damit unter dem gegebenen Wahrscheinlichkeitsmaß fast sicher Werte im beschränkten Intervall $(-\sqrt{6}, \sqrt{6})$ annehmen. Die Regressanden Y_i^t seien wieder $\mathbb{R}$-wertige Zufallsvariablen. Wir nehmen ferner an, dass zwischen Regressoren und Regressand ein *lineares Modell* bestehe. Es gelte also für alle $i \in \mathbb{N}$ und $t \in \{1, ..., s\}$ die Relation

$$Y_i^t \equiv \beta_0 + \sum_{k=1}^{d} \beta_k \cdot X_{i,k}^t + \epsilon_i^t \tag{5.1}$$

mit $\beta_0, \beta_1, ..., \beta_d \in \mathbb{R}$ als festen Koeffizienten und $(\epsilon_i^t)_{i \in \mathbb{N}}$ als Folge von unabhängigen und identisch verteilten Störgrößen. Diese werden als *standardnormalverteilt* vorausgesetzt, also

$$\epsilon_i^t \sim N(0, 1)$$

für jedes $i \in \mathbb{N}$ und $t \in \{1, ..., s\}$. Darüber hinaus seien die Zufallsvariablen $X_{i,1}^t, ..., X_{i,d}^t$ und ϵ_i^t über die Indizes i und t als unabhängig vorausgesetzt. Beachte, dass die Folgen $(X_i^t, Y_i^t)_{i \in \mathbb{N}}$ und $(\epsilon_i^t)_{i \in \mathbb{N}}$ die Postulate (P1) bis (P4) des Referenzmodells aus Unterkapitel 4.1 erfüllen, wobei die Funktion a die Form

$$a(x) = \beta_0 + \sum_{k=1}^{d} \beta_k \cdot x_k$$

für alle $x \equiv (x_1, ..., x_d) \in \mathbb{R}^d$ aufweist. An dieser Stelle legen wir wieder F als die gemeinsame Verteilungsfunktion der Absolutbeträge der Störgrößen fest mit F^{-1} als durch (4.3) definierte verallgemeinerte Inverse. Bei standardnormalverteilten Störgrößen entspricht $F^{-1}(1-\alpha)$ damit dem $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung für beliebige $\alpha \in (0, 1)$.

Unter den obigen Bedingungen simulieren wir für festes $n \in \mathbb{N}$ zu jeder Iteration $t \in \{1, ..., s\}$ eine Stichprobe $(X_1^t, Y_1^t), ..., (X_{n+1}^t, Y_{n+1}^t)$. Zur Berechnung

von Konfidenzbereichen dienen die dabei realisierten Zahlen sowohl der Full- als auch der Split-Methode als Daten. Die konkreten Zahlenwerte für s, n, d und α sowie für die Koeffizienten $\beta_0, ..., \beta_d$, die in der Simulation verwendet werden, sind der Tabelle 1 zu entnehmen.[8] Mit $\alpha = 0.1$ entspricht $F^{-1}(1-\alpha)$ dem 0.95-Quantil der Standardnormalverteilung. Es gilt also

$$F^{-1}(1-\alpha) \approx 1.645.$$

Somit fällt die Länge des „optimalen“ Konfidenzintervalls aus dem Referenzmodell mit

$$2 \cdot F^{-1}(1-\alpha) \approx 3.29$$

zusammen. Bei Durchführung der Full- wie auch der Split-Methode werden die Koeffizienten *geschätzt*. Das tun wir zum einen mit der *Kleinste-Quadrate-Methode*, wobei wir damit sowohl eine *korrekt spezifizierte* Schätzung als auch eine *fehlspezifizierte* Schätzung mit unter anderem fehlenden Regressoren durchführen. Zum anderen schätzen wir die Koeffizienten mit dem in Tibshirani (1996) beschriebenen *Lasso-Verfahren*. Dabei handelt es sich um eine *Shrinkage-Methode*, welche in der Schätzung *Variablenselektion* inkorporiert. Zur Ausführung der Lasso-Schätzung führen wir zwei weitere reellwertige Zufallsvariablen $X^t_{i,d+1}$ und $X^t_{i,d+2}$ ein, welche für alle $i \in \mathbb{N}$ der obigen Dreiecksverteilung folgen und nicht im Modell (5.1) als Regressoren enthalten sind (also Variablen, die für die Schätzung von (5.1) als „redundante“ Information zu betrachten sind und durch Variablenselektion ausgesondert werden sollten). In jedem Iterationslauf wird zudem noch darauf geachtet, dass bei der Anwendung der oben genannten Verfahren die Anzahl der als Regressoren verwendeten Variablen kleiner ist als die Gesamtzahl an verfügbaren Beobachtungen, um so mögliche Schwierigkeiten, die im umgekehrten Fall auftauchen können (wie etwa die Problematik, dass der Kleinsten-Quadrate-Schätzer nicht eindeutig bestimmt werden kann), zu vermeiden. Wir betonen an dieser Stelle noch, dass wir dabei nicht untersuchen, ob der von uns gewählte Modellrahmen und die verwendeten Schätzverfahren die Voraussetzungen der Sätze aus Kapitel 4 erfüllen, sondern ob die Ergebnisse der Simulation mit den Schlussfolgerungen, die aus diesen Sätze gewonnen werden, in Einklang stehen. In den folgenden Unterkapiteln werden wir an einigen Stelle auf die kürzere Notation $Z^t_i \equiv (X^t_i, Y^t_i)$ für $i \in \{1, ..., n+1\}$ und $t \in \{1, ..., s\}$ zurückgreifen.

5.2 Endliche Netze als Mittel zur Umsetzung der Full-Methode

Bevor wir die Full-Methode anwenden, gilt es noch ein besonderes Problem, welches bei der Umsetzung dieser Methode auftaucht, zu betrachten; vgl. auch G'Sell et al. (2016, S.7 und S.9). Um nämlich zu entscheiden, ob ein möglicher Wert $y \in \mathbb{R}$, der vom Regressanden angenommen werden kann, den Konfidenzbereich angehört, muss bei gegebenen Daten $(x_1, y_1), ..., (x_n, y_n), x_{n+1}$ ein Algorithmus zur Bestimmung der Regressionsfunktion jedes mal auf $(x_1, y_1), ..., (x_n, y_n), (x_{n+1}, y)$ angewendet werden. Im Fall, dass der Regressand unter dem gegebenen Wahrscheinlichkeitsmaß fast sicher Werte aus einer endlichen Teilmenge von $\mathbb{R}$ annimmt, ist eine solche Prozedur von Prinzip her auch umsetzbar. Anders verhält es sich, wenn der Regressand fast sicher Werte aus ganz $\mathbb{R}$ oder aus einer (abzählbar oder überabzählbar) unendlichen Teilmenge annimmt. Und selbst im Fall, dass der Wertebereich des Regressanden endlich ist, aber eine sehr große Anzahl an Elementen enthält, könnte die Bestimmung eines konformalen Konfidenzbereiches nach der Full-Methode zu rechenintensiv und daher praktisch nicht umsetzbar sein. Einen möglichen Ausweg bietet ein Verfahren, dass im R-Paket `conformalInference` implementiert ist (siehe auch Anhang B). Es basiert auf die Erzeugung eines *Netzes endlich vieler äquidistanter Punkte*, aus dem Konfidenzintervalle bestimmt werden können. Zur Beschreibung dieser Prozedur verwenden wir die Notation aus Unterkapitel 3.1.

Wir betrachten die Zufallsvariablen $(X_1, Y_1), ..., (X_n, Y_n)$ und X_{n+1}, sowie für beliebige $y \in \mathbb{R}$ die Residuen $U_{1,(X_{n+1},y)}, ..., U_{n+1,(X_{n+1},y)}$, die bei gegebenem Algorithmus zur Full-Methode a_n über die Regressionsfunktion

$$a_n(Z_1, ..., Z_n, (X_{n+1}, y))$$

bestimmt werden. Dabei seien $U_{(1:n),(X_{n+1},y)}, ..., U_{(n:n),(X_{n+1},y)}$ die zu den Zufallsvariablen $U_{1,(X_{n+1},y)}, ..., U_{n,(X_{n+1},y)}$ gehörigen Ordnungsstatistiken. Weiters definieren wir für vorgegebenes $\alpha \in (0, 1)$ wieder

$$h := \lceil (1 - \alpha)(n + 1) \rceil.$$

Für die Bestimmung eines endlichen Netzes verwendet die Prozedur die kleinste und die größte Ordnungsstatistik von $Y_1, ..., Y_n$, also

$$Y_{(1:n)} := min\{Y_1, ..., Y_n\},$$

$$Y_{(n:n)} := max\{Y_1, ..., Y_n\}.$$

Wir definieren

$$D := Y_{(n:n)} - Y_{(1:n)} + 2 \cdot \eta$$

mit

$$\eta := g \cdot (Y_{(n:n)} - Y_{(1:n)}),$$

wobei wir $g \geq 0$ als *Grid-Faktor* bezeichnen. Für vorgegebenes $m \in \mathbb{N}$ mit $m \geq 2$ setzen wir zur Vereinfachung

$$\tilde{Y}_i := Y_{(1:n)} - \eta + i \cdot \frac{D}{m-1}$$

für alle $i \in \{0, 1, ..., m-1\}$. Beachte, dass $\eta \geq 0$. Es folgt

$$\tilde{Y}_0 = Y_{(1:n)} - \eta \leq Y_{(1:n)},$$

$$\tilde{Y}_{m-1} = Y_{(n:n)} + \eta \geq Y_{(n:n)}.$$

Wir definieren das *endliche Netz*

$$\mathscr{Y}_m := \{\tilde{Y}_i \,|\, 0 \leq i \leq m-1\}.$$

Durch diese Menge wird also um die realisierten Werte von $Y_1, ..., Y_n$ ein Netz bestehend aus endlich vielen äquidistanten Punkten „gespannt“, wobei die Distanz von einem zum nächsten Punkt genau $D/(m-1)$ beträgt. Die „Breite“ des Netzes wird durch den Grid-Faktor g bestimmt. Je größer g gewählt, umso „breiter“ das Netz, d.h. desto größer ist die Distanz des kleinsten Elements $\tilde{Y}_0$ der Menge $\mathscr{Y}_m$ zur Ordnungsstatistik $Y_{(1:n)}$ und desto größer auch die Distanz des größten Elements $\tilde{Y}_{m-1}$ zur Ordnungsstatistik $Y_{(n:n)}$. Die Anzahl der Punkte im Netz entspricht der Zahl m, d.h. je größer m gewählt wird, um so „dichter“ das Netz. Von diesem Netz betrachten wir nun die Teilmenge

$$\mathscr{Y}_{m,\alpha}(X_{n+1}) := \{y \in \mathscr{Y}_m \,|\, rang(U_{n+1,(X_{n+1},y)}) \leq h\}.$$

Falls

$$(1-\alpha)(n+1) > n$$

und somit $h = n+1$, dann gilt wegen der Ungleichung

$$rang(U_{n+1,(X_{n+1},y)}) \leq n+1,$$

die für alle $y \in \mathscr{Y}_m$ erfüllt ist, die Gleichheit

$$\mathscr{Y}_{m,\alpha}(X_{n+1}) = \mathscr{Y}_m$$

(beachte, dass wir für diesen Spezialfall $C_{n,\alpha}^{full}(X_{n+1}) = \mathbb{R}$ festgelegt haben; siehe Unterkapitel 3.1). Für den Fall

$$(1-\alpha)(n+1) \leq n,$$

also wenn $h \leq n$, gilt

$$\mathscr{Y}_{m,\alpha}(X_{n+1}) = \{y \in \mathscr{Y}_m \,|\, U_{n+1,(X_{n+1},y)} \in [0, U_{(h:n),(X_{n+1},y)})\}$$

(siehe auch den Beweis von Satz 2.1). In beiden Fällen folgt unmittelbar

$$\mathscr{Y}_{m,\alpha}(X_{n+1}) \subseteq C_{n,\alpha}^{full}(X_{n+1}),$$

d.h. die Menge $\mathscr{Y}_{m,\alpha}(X_{n+1})$ enthält alle Elemente aus $\mathscr{Y}_m$, welche im konformalen Konfidenzbereich $C_{n,\alpha}^{full}(X_{n+1})$ enthalten sind. Falls nun kein Element aus $\mathscr{Y}_m$ in der Menge $\mathscr{Y}_{m,\alpha}(X_{n+1})$ liegt, setzt die Prozedur einfach

$$C_{n,\alpha}^{full,\mathscr{Y}_m}(X_{n+1}) := \{0\} \tag{5.2}$$

fest. Im entgegengesetzten Fall, dass Elemente aus $\mathscr{Y}_m$ in $\mathscr{Y}_{m,\alpha}(X_{n+1})$ enthalten sind, gibt es $i, j \in \{0, 1, ..., m-1\}$ mit $i \leq j$, sodass

$$\tilde{Y}_i = min(\mathscr{Y}_{m,\alpha}(X_{n+1})),$$

$$\tilde{Y}_j = max(\mathscr{Y}_{m,\alpha}(X_{n+1})),$$

d.h. $\tilde{Y}_i$ ist das kleinste Element in $\mathscr{Y}_{m,\alpha}(X_{n+1})$ und $\tilde{Y}_j$ das Größte. Der Kern der Prozedur besteht in diesem Fall darin, dass sie den gesuchten Konfidenzbereich $C_{n,\alpha}^{full}(X_{n+1})$ mit einem Intervall der Form

$$C_{n,\alpha}^{full,\mathscr{Y}_m}(X_{n+1}) := [l, r]$$

„approximiert“, wobei $l, r \in \mathbb{R}$. Zur Bestimmung der Intervallsgrenzen stehen der Prozedur drei Methoden zur Verfügung. Die erste wollen wir als *hypokonservative* Methode bezeichnen. Bei dieser wird einfach

$$l := \tilde{Y}_i,$$

$$r := \tilde{Y}_j$$

gesetzt. Die Zweite bezeichnen wir als die *hyperkonservative* Methode. Bei $i = 0$ wird von dieser Methode $l := \tilde{Y}_i$ festgesetzt. Falls $i > 0$, dann wird

$$l := \tilde{Y}_{i-1}$$

festgelegt. Bei $j = m-1$ definiert diese Methode $r := \tilde{Y}_j$. Für den Fall $j < m-1$ wird

$$r := \tilde{Y}_{j+1}$$

festgelegt. Die dritte Methode, welche wir als *moderat* bezeichnen, geht wie folgt vor: sofern $i = 0$ gilt, wird von dieser Methode $l := \tilde{Y}_i$ festgelegt. Für den Fall $i > 0$ wird

$$l := \tilde{Y}_{i-1} + s_l \cdot (\tilde{Y}_i - \tilde{Y}_{i-1})$$

festgelegt, wobei

$$s_l := \frac{rang(U_{n+1,(X_{n+1},\tilde{Y}_{i-1})}) - h}{rang(U_{n+1,(X_{n+1},\tilde{Y}_{i-1})}) - rang(U_{n+1,(X_{n+1},\tilde{Y}_i)})}.$$

Da $\tilde{Y}_i \in \mathscr{Y}_{m,\alpha}(X_{n+1})$ und gleichzeitig $\tilde{Y}_{i-1} \notin \mathscr{Y}_{m,\alpha}(X_{n+1})$, gilt

$$rang(U_{n+1,(X_{n+1},\tilde{Y}_{i-1})}) > h \geq rang(U_{n+1,(X_{n+1},\tilde{Y}_i)}), \tag{5.3}$$

und somit

$$s_l \in (0, 1].$$

Die untere Intervallsgrenze l ist hiermit eine konvexe Linearkombination von $\tilde{Y}_{i-1}$ und $\tilde{Y}_i$. Falls $j = m-1$, dann ist $r := \tilde{Y}_j$. Sofern $j < m-1$ erfüllt ist, setzt die Methode

$$r := \tilde{Y}_j + s_r \cdot (\tilde{Y}_{j+1} - \tilde{Y}_j)$$

fest, wobei

$$s_r := \frac{rang(U_{n+1,(X_{n+1},\tilde{Y}_j)}) - h}{rang(U_{n+1,(X_{n+1},\tilde{Y}_j)}) - rang(U_{n+1,(X_{n+1},\tilde{Y}_{j+1})})}.$$

Da $\tilde{Y}_j \in \mathscr{Y}_{m,\alpha}(X_{n+1})$ und gleichzeitig $\tilde{Y}_{j+1} \notin \mathscr{Y}_{m,\alpha}(X_{n+1})$, gilt

$$rang(U_{n+1,(X_{n+1},\tilde{Y}_j)}) \leq h < rang(U_{n+1,(X_{n+1},\tilde{Y}_{j+1})}), \tag{5.4}$$

und somit

$$s_r \in [0, 1).$$

Die obere Intervallsgrenze r ist also eine konvexe Linearkombination von $\tilde{Y}_j$ und $\tilde{Y}_{j+1}$.

Abgesehen vom Spezialfall $i = 0$ und $j = m-1$, in welchem alle Elemente von $\mathscr{Y}_m$ in der Menge $\mathscr{Y}_{m,\alpha}(X_{n+1})$ bzw. in $C^{full}_{n,\alpha}(X_{n+1})$ liegen und folglich die drei Methoden jeweils das gleiche Konfidenzintervall bestimmen, gehen aus der hypokonservativen Methode die kürzesten Intervalle und aus der hyperkonservativen Methode die längsten Intervalle hervor, während die Länge

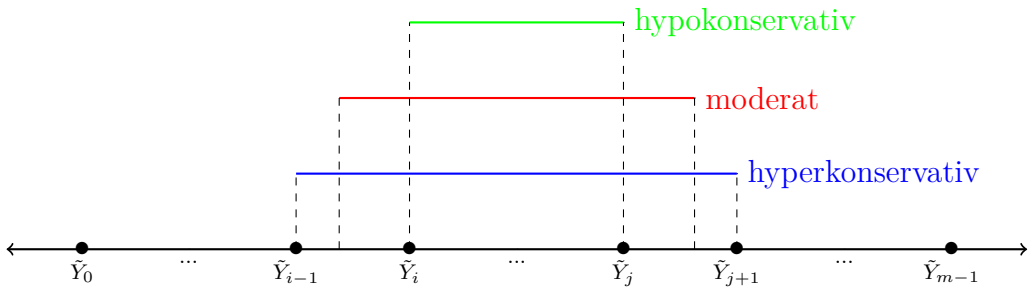

Abbildung 1: *Graphische Illustration der Längen der durch Erzeugung eines endlichen Netzes konstruierten Konfidenzintervalle für den Fall* $i > 0$ *und* $j < m - 1$*. Dabei repräsentiert die schwarze horizontale Linie die reelle Zahlenachse, auf der die relevanten Netzpunkte aus der Menge* $\mathscr{Y}_m$ *aufgetragen sind.*

der durch die moderate Methode erzeugten Intervalle in der Mitte liegt (siehe auch Abbildung 1 zur Veranschaulichung). Es ist daher zu erwarten, dass im Zuge der Simulation die empirische Überdeckungsrate bei der hypokonservativen Methode am kleinsten ausfällt und jene der hyperkonservativen Methode am größten (siehe Unterkapitel 5.3). Wenn bei fester Stichprobengröße n die Größe h unter Beibehaltung der Gültigkeit der Gleichungen (5.3) und (5.4) variiert wird, dann ergibt sich ein besonderes Phänomen. Bei größerem h, also größerem nominellen Konfidnezniveau $1 - \alpha$, kommt es zu einer Abnahme der Größe s_l und gleichzeitig zu einer Zunahme von s_r, womit die Intervallsgrenzen der moderaten Methode näher an jene der hyperkonservativen Methode heranrücken (die moderate Methode wird also „konservativer“). Bei sinkendem h, also bei geringerem nominellen Konfidenzniveau, wird s_l hingegen größer und s_r kleiner. Die Intervallsgrenzen der moderaten Methode nähren sich den Grenzen der hypokonservativen Methode an (die moderate Methode wird also „weniger“ konservativ). Die Art und Weise, wie die Intervallsgrenzen bei der moderaten Methode definiert werden, ist also mit dem Effekt verbunden, dass diese Methode sich relativ flexibel an Änderungen des nominellen Konfidenzniveaus anpasst. Mit höherem (niedrigerem) $1 - \alpha$ neigt nämlich bei fester Stichprobengröße der Konfidenzbereich $C_{n,\alpha}^{full}(X_{n+1})$ „breiter“ („schmaler“) zu werden, womit „konservativere“ („weniger“ konservative) Intervalle zu einer besseren Abschätzung von $C_{n,\alpha}^{full}(X_{n+1})$ führen. Damit im Allgemeinen der Wertebereich des Regressanden in befriedigender Weise „umspannt“ werden kann, muss das Netz $\mathscr{Y}_m$ entsprechend „breit“ und „dicht“ gewählt sein. Um also den Konfidenzbereich $C_{n,\alpha}^{full}(X_{n+1})$ möglichst gut durch ein Intervall $C_{n,\alpha}^{full,\mathscr{Y}_m}(X_{n+1})$

„abzuschätzen“, müssen die Parameter g und m entsprechend groß gewählt werden. Bei großem m, also bei vielen Punkten in der Menge $\mathscr{Y}_m$, bringen die drei Methode außerdem Intervalle hervor, deren Länge näher beieinander liegen als wenn m relativ klein ist (d.h. der Unterschied zwischen den drei Methoden wird bei wachsendem m geringer). Allerdings bedeutet ein hohes m für die Umsetzung der Full-Methode einen höheren Rechenaufwand (siehe auch die Unterkapitel 5.3 und 5.5).

5.3 Umsetzung und Resultate der Full-Methode

Wir betrachten für jede Wiederholung $t \in \{1, ..., s\}$ und für alle Stichprobengrößen $i \in \{1, ..., n+1\}$ die in Unterkapitel 5.1 eingeführten Zufallsvariablen (X_i^t, Y_i^t) sowie die „überflüssigen“ Variablen $X_{i,d+1}^t$und $X_{i,d+2}^t$. Die „Approximation“ des Konfidenzbereiches $C_{i,\alpha}^{full}(X_{i+1}^t)$ durch die auf die Erzeugung eines endlichen Netzes basierenden Full-Methoden wird für einzelne Suchprobengrößen i ausgeführt. Die Menge $\mathscr{Y}_{m,i}^t$ sei dabei ein endliches Netz, dass mittels der in Unterkapitel 5.2 vorgestellten Weise aus den Zufallsvariablen $Y_1^t, ..., Y_i^t$ gewonnen wird, wobei wir die Konfidenzintervalle fallweise unter $m = 25$ oder $m = 50$ bestimmen. Für alle Fälle sei $g = 0.25$ gewählt. Für die Kleinste-Quadrate-Methode legen wir nun eine Designmatrix $\mathbf{X}_i^t$ fest, wobei diese für die korrekt spezifizierte Schätzung die Form

$$\mathbf{X}_i^t = \begin{pmatrix} 1 & X_{1,1}^t & ... & X_{1,d}^t \\ \vdots & \vdots & & \vdots \\ 1 & X_{i,1}^t & ... & X_{i,d}^t \\ 1 & X_{i+1,1}^t & ... & X_{i+1,d}^t \end{pmatrix}$$

und für die fehlspezifizierte Schätzung die Form

$$\mathbf{X}_i^t = \begin{pmatrix} 1 & (X_{1,1}^t)^2 & X_{1,3}^t & X_{1,d}^t \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_{i,1}^t)^2 & X_{i,3}^t & X_{i,d}^t \\ 1 & (X_{i+1,1}^t)^2 & X_{i+1,3}^t & X_{i+1,d}^t \end{pmatrix}$$

annimmt. Die Fehlspezifikation zeichnet sich aus durch fehlende relevante Regressoren sowie durch den Umstand, dass zwischen dem Regressanden und einem der Regressoren ein quadratischer Zusammenhang unterstellt wird (was im „wahren“ Modell (5.1) nicht der Fall ist). Für alle $y \in \mathscr{Y}_{m,i}^t$ definieren wir den $(i+1)$-dimensionalen Spaltenvektor

$$\mathbf{Y}_{i,y}^t := (Y_1^t, ..., Y_i^t, y)^T.$$

Wir berechnen den Schätzer

$$\hat{\beta}^t_{i,y} \equiv (\hat{\beta}^t_{i,y,0}, \hat{\beta}^t_{i,y,1}, ..., \hat{\beta}^t_{i,y,q})^T$$

als $(q+1)$-dimensionalen Spaltenvektor, wobei $q = d$ für die korrekten Spezifikation, $q = 3$ für die Fehlspezifikation und $q = d+2$ für die Lasso-Schätzung festgesetzt sei. Unter Anwendung der Kleinste-Quadrate-Methode tritt uns dieser dann für alle $i \in \{q+1, ..., n\}$ in der wohlbekannten Gestalt

$$\hat{\beta}^t_{i,y} := ((\mathbf{X}^t_i)^T \cdot \mathbf{X}^t_i)^{-1} \cdot (\mathbf{X}^t_i)^T \cdot \mathbf{Y}^t_{i,y}$$

entgegen. Im Rahmen der mit Hilfe der Lasso-Methode durchgeführten Schätzung ergibt sich $\hat{\beta}^t_{i,y}$ aus der Minimierung der Funktion

$$f^{lasso}_{full}(\beta_0, ..., \beta_q) := \sum_{j=1}^{i+1} \left(Y^t_j - \beta_0 - \sum_{k=1}^{q} \beta_k \cdot X^t_{j,k} \right)^2 + \lambda \cdot 2 \cdot (i+1) \cdot \sum_{k=1}^{q} |\beta_k|$$

über alle $\beta_0, ..., \beta_q \in \mathbb{R}$ für $Y^t_{i+1} = y$ und gegebenen Realisationen der Zufallsvariablen $Y^t_1, ..., Y^t_i$ sowie $X^t_{1,k}, ..., X^t_{i+1,k}$ mit $k \in \{1, ..., q\}$, wobei $\lambda \geq 0$ gilt; vgl. Friedman et al. (2010, S.3). Bei jenem im Funktionswert von f^{lasso}_{full} vorkommenden Ausdruck, welcher die Summe der Absolutbeträge der Argumente $\beta_1, ..., \beta_q$ enthält, handelt es sich um einen *Strafterm*, der den Funktionswert bei Argumenten ungleich 0 *zusätzlich* erhöht und die Lasso-Methode damit „zwingt“ gegenüber den entsprechenden Kleinste-Quadrate-Schätzer die Beträge der geschätzten Koeffizienten gegen 0 schrumpfen zu lassen. Im Extremfall werden Koeffizienten sogar auf 0 gesetzt (d.h. die dazugehörigen Regressoren werden im Zuge der Variablenselektion „entfernt“), was bei anderen Shrinkage-Methoden wie der Ridge-Methode in einem vergleichsweise eher geringen Ausmaß der Fall ist.[9] Der Idealfall wäre natürlich, dass im Zuge der mit der Ausführung der Lasso-Methode verbundenen Variablenselektion die geschätzten Koeffizienten der „überflüssigen“ Variablen auf 0 „geschrumpft“ und die zu den relevanten Regressoren gehörigen Koeffizienten (zumindest asymptotisch) korrekt geschätzt werden. Der Parameter λ steuert zudem, in welchem Ausmaß die „Schrumpfung“ der einzelnen Koeffizienten zu erfolgen hat. Je höher λ gewählt wird, um so stärker neigt unter Umständen das Verfahren die Koeffizienten zu „schrumpfen“ bzw. um so größer neigt die Zahl der auf 0 gesetzten Koeffizienten zu werden. Im Fall $\lambda = 0$ kann das Lasso-Verfahren auf die Kleinste-Quadrate-Methode reduziert werden. Es ist zu beachten, dass es sich bei λ um einen *Tuningparameter* handelt, der üblicherweise durch datengestützte Methoden wie Kreuzvalidierung bestimmt wird. Wir verzichten in dieser Arbeit jedoch darauf, da λ ansonsten von den Daten auf eine Weise

abhängt, die eine Änderung der Qualität des Trainingsalgorithmus nach sich zieht und die Gültigkeit der Konfidenzintervalle in Frage stellt. Im Fall der Full-Methode könnte beispielsweise die Verwendung eines datengetriebenen Tuningparameters für den Algorithmus den Verlust der Permutationsinvarianz bedeuten.[10] Wir wählen daher *ad hoc* einen festen Wert $\lambda = 0.15$. Wir bemerken, dass wir die Lasso-Schätzung mit allen im Modell (5.1) enthaltenen Regressoren sowie zwei überflüssigen Variablen durchführen und uns der Frage zuwenden, ob unter den von uns festgelegten Umständen das Verfahren durch mögliche Unzulänglichkeiten in der Art der Variablenselektion (z.B. durch „Weglassen“ *relevanter* Regressoren) die Qualität der konformalen Konfidenzbereiche beeinträchtigt. Im Gegensatz zur Kleinste-Quadrate-Methode existiert bei der Lasso-Schätzung im Allgemeinen keine eindeutige analytisch geschlossene Form für $\hat{\beta}^t_{i,y}$, sondern wir müssen auf geeignete *heuristische Näherungsverfahren* zurückgreifen (interessierte Leser seien auf die in dieser Arbeit angegebenen Literatur verwiesen). Die numerische Umsetzung erfolgt in dieser Arbeit mit dem Algorithmus des *covariance updating*, welcher in der Funkion `glmnet` aus dem R-Paket `glmnet` implementiert ist. Die von uns verwendeten Schätzverfahren zur Bestimmung von $\hat{\beta}^t_{i,y}$ korrespondieren jeweils mit einem Algorithmus zur Full-Methode a^t_i, der uns eine Regressionsfunktion liefert, mittels derer wir die entsprechenden Residuen bestimmen. Diese nimmt an der Stelle

$$x = (x_1, ..., x_q) \in \mathbb{R}^q$$

für die korrekte Spezifikation und die Lasso-Schätzung die Form

$$a^t_i(Z^t_1, ..., Z^t_i, (X^t_{i+1}, y))(x) = \hat{\beta}^t_{i,y,0} + \sum_{k=1}^{q} \hat{\beta}^t_{i,y,k} \cdot x_k,$$

und für die Fehlspezifikation die Form

$$a^t_i(Z^t_1, ..., Z^t_i, (X^t_{i+1}, y))(x) = \hat{\beta}^t_{i,y,0} + \hat{\beta}^t_{i,y,1} \cdot (x_1)^2 + \hat{\beta}^t_{i,y,2} \cdot x_2 + \hat{\beta}^t_{i,y,3} \cdot x_3$$

an.[11] Unter Anwendung der in Unterkapitel 5.2 diskutierten Methode ermitteln wir die Menge $\mathscr{Y}_{m,\alpha}(X^t_{i+1})$ und berechnen daraus das Konfidenzintervall

$$C^{full,\mathscr{Y}_m}_{i,\alpha}(X^t_{i+1}) = [l^t_i, r^t_i]$$

mit den entsprechenden Intervallsgrenzen l^t_i und r^t_i, welche wir einmal gemäß der moderaten Methode, einmal gemäß der hyperkonservativen Methode und einmal gemäß der hypokonservativen Methode bestimmen. Damit sind wir in der Lage jene zwei für diese Analyse wichtigen Größen zu bestimmen, nämlich

die *gemittelte empirische Überdeckungsrate*, mit der wir die Überdeckungswahrscheinlichkeit des Konfidenzbereiches schätzen, und die *gemittelte Länge* der Intervalle. Die gemittelte empirische Überdeckungsrate $A_{i,\alpha}^{full,gem}$ zur Stichprobengröße $i \in \{q+1, ..., n\}$ sei definiert durch

$$A_{i,\alpha}^{full,gem} := \frac{1}{s} \sum_{t=1}^{s} \frac{1}{i-q} \sum_{j=q+1}^{i} B_{j,\alpha,t}^{full},$$

wobei für jedes $j \in \{q+1, ..., i\}$ und jedes $t \in \{1, ..., s\}$ der Indikator $B_{j,\alpha,t}^{full}$ definiert sei durch

$$B_{j,\alpha,t}^{full} := \begin{cases} 1 & \text{falls } Y_{j+1}^t \in C_{j,\alpha}^{full,\mathscr{Y}_m}(X_{j+1}^t), \\ 0 & \text{sonst.} \end{cases}$$

Für jedes $i \in \{q+1, ..., n\}$ und jedes $t \in \{1, ..., s\}$ entspricht die Größe

$$\frac{1}{i-q} \sum_{j=q+1}^{i} B_{j,\alpha,t}^{full}$$

der empirischen Überdeckungsrate zur Stichprobengröße i im Iterationsdurchlauf t. Die gemittelte Intervallslänge $I_{i,\alpha}^{full}$ ist bestimmt durch

$$I_{i,\alpha}^{full} := \frac{1}{s} \sum_{t=1}^{s} (r_i^t - l_i^t),$$

wobei $r_i^t - l_i^t$ der Breite des Konfidenzintervalls zur Stichprobengröße i im Iterationsdurchlauf t entspricht.[12] Um eine Vorstellung zu bekommen in welchem Ausmaß für einzelne Stichprobengrößen i die empirische Überdeckungsrate und die Länge der Full-Konfidenzintervalle um ihre gemittelten Werte streuen, wollen wir die obigen Berechnungen mit einfachen *Intervallschätzern* ergänzen. Dazu bestimmen wir zunächst die korrigierte Stichprobenstandardabweichung für die empirische Überdeckungsrate

$$sd(A_{i,\alpha}^{full,gem}) := \sqrt{\frac{1}{s-1} \sum_{t=1}^{s} \left(\frac{1}{i-q} \sum_{j=q+1}^{i} B_{j,\alpha,t}^{full} - A_{i,\alpha}^{full,gem} \right)^2}$$

und

$$sd(I_{i,\alpha}^{full}) := \sqrt{\frac{1}{s-1} \sum_{t=1}^{s} \left(r_i^t - l_i^t - I_{i,\alpha}^{full} \right)^2}$$

für die Intervallslänge. Wir verwenden für die empirische Überdeckungsrate das zweiseitige Konfidenzintervall

$$\left[A_{i,\alpha}^{full,gem} - \frac{F^{-1}(1-\alpha) \cdot sd(A_{i,\alpha}^{full,gem})}{\sqrt{s}}, A_{i,\alpha}^{full,gem} + \frac{F^{-1}(1-\alpha) \cdot sd(A_{i,\alpha}^{full,gem})}{\sqrt{s}} \right]$$

als Intervallschätzer.[13] Für die Intervallsbreite wählen wir das zweiseitige Konfidenzintervall

$$\left[I_{i,\alpha}^{full} - \frac{F^{-1}(1-\alpha) \cdot sd(I_{i,\alpha}^{full})}{\sqrt{s}} \, , \, I_{i,\alpha}^{full} + \frac{F^{-1}(1-\alpha) \cdot sd(I_{i}^{full})}{\sqrt{s}} \right].$$

Wir erinnern uns, dass $F^{-1}(1-\alpha)$ dem 0.95-Quantil der Standardnormalverteilung entspricht (siehe Unterkapitel 5.1). Beachte, dass beide Konfidenzintervalle nicht exakt sondern asymptotisch gültig zum Konfidenzniveau $1-\alpha$ sind.

Von Interesse ist also, wie sich die gemittelte empirische Überdeckungsrate und die Länge der Konfidenzintervalle in Abhängigkeit von der Stichprobengröße i entwickeln. Die Ergebnisse der Simulation unter der korrekt spezifizierten Kleinste-Quadrate-Schätzung sind für $m = 25$ in den Abbildungen 2 und 3 sowie für $m = 50$ in den Abbildungen 4 und 5 dargestellt. Tabelle 2 enthält außerdem Werte, die für die betrachteten Größen im Zuge der Simulation unter $m = 25$ bzw. unter $m = 50$ für eine Auswahl von Stichprobengrößen beobachtet wurden, wobei die Werte der Stichprobenstandardabweichungen der empirischen Überdeckungsrate und der Intervallslänge jeweils in Klammern gesetzt sind. Wir sehen sowohl für $m = 25$ als auch für $m = 50$ ähnliche Muster. Wie nicht anders zu erwarten, liefert die hyperkonservative Methode sowohl bei der gemittelten empirischen Überdeckungsrate als auch bei der gemittelten Länge der Konfidenzintervalle die größten Werte, während wir bei der hypokonservativen Methode bei den gleichen Größen die kleinsten Werte erhalten. Die gemittelten empirischen Überdeckungsraten, die bei der Anwendung der moderaten Methode berechnet werden, weisen dabei die geringste Distanz zum nominellen Konfidenzniveau $1-\alpha$ auf, was darauf hinweist, dass wir mit dieser Methode am ehesten Konfidenzintervalle mit der gewünschten Überdeckungswahrscheinlichkeit zu ermitteln in der Lage sind. Bei kleinen Stichprobengrößen liefern die drei Methoden gemittelte empirische Überdeckungsraten, die relativ nahe beieinander liegen und außerdem noch relativ starken Schwankungen unterworfen sind. Mit wachsender Stichprobengröße gehen die gemittelten empirischen Überdeckungsraten dann stark auseinander. Dabei stabilisieren sich die gemittelten Überdeckunsgraten der hyperkonsarvativen und der moderaten Methode auf einem Niveau, welches über dem nominellen Konfidenzniveau liegt. Bei der hypokonservativen Methode vollzieht sich die Stabilisierung der gemittelten empirischen Überdeckungsrate deutlich unter dem Konfidenzniveau, wobei sich diese bei $m = 50$ sehr klar zeigt, während sich bei $m = 25$ im Bereich größerer Stichprobenumfänge noch ein leicht negativer Trend zu

vollziehen scheint. Hervorzuheben ist noch einmal, dass bei der moderaten Methode die gemittelte Überdeckungsrate sich in relativer Nähe zum nominellen Konfidenzniveau stabilisiert. Die moderate Methode scheint also von den drei Methoden jene zu sein, die der Schlussfolgerung von Satz 4.1 am ehesten gerecht zu werden scheint. Ähnlich wie bei der Überdeckungsrate liefern die drei Methoden bei kleinen Stichprobenumfängen noch ungefähr gleichlange Konfidenzintervalle, welche sich mit wachsender Stichprobe stark auseinander entwickeln. Bei allen drei Methoden sind bei niedrigem Stichprobenumfang die Intervallslängen noch sehr hoch, sinken mit wachsendem Stichprobenumfang und scheinen sich dann ab einem ausreichend hohen Stichprobenniveau zu stabilisieren, wobei sich die Stabilisierung der Intervallslänge bei der moderate Methode in relativ enger Umgebung von $2 \cdot F^{-1}(1-\alpha)$ vollzieht. Die Resultate bei der moderaten Methode scheinen sich also relativ gut mit der Konklusion von Satz 4.2 zu vertragen. Bemerkenswert ist dabei, dass wir mit wachsender Stichprobengröße in der Entwicklung der gemittelten Intervallslängen einen regulären Zyklus aufeinanderfolgender Wachstums- und Kontraktionsphasen beobachten (man könnte versucht sein, diese oszillierenden Schwankungen auf die Netzkonstruktion der Full-Methoden zurückzuführen, aber in Unterkapitel 5.4 werden wir sehen, dass diese auch bei der Entwicklung der Länge der durch die Split-Methode konstruierten Konfidenzintervalle auftreten). Im Fall $m = 50$ sehen wir außerdem, dass bei größeren Stichprobenumfängen bezüglich der gemittelten empirischen Überdeckungsrate und der gemittelten Intervallslänge die Differenz zwischen den drei Methoden deutlich geringer ausfällt als bei $m = 25$. Bei 50 Netzpunkten ist darüber hinaus bei großen Stichproben die Distanz der durch die moderate Methode hervorgebrachten Werte zu den durch die hyperkonservativen Methode generierten Werten ungefähr so groß wie die Distanz zu den durch die hypokonservativen Methode erzeugten Werten. Die Werte der gemittelten empirischen Überdeckungsrate, die wir durch die Anwendung der moderaten Methode erhalten, liegen im Fall $m = 50$ zudem deutlich näher beim nominellen Konfidenzniveau als im Fall $m = 25$. Bei dichter werdenden Netzen scheint also die moderate Methode Konfidenzintervalle zu liefern, deren Überdeckungswahrscheinlichkeit näher zum vorgegebenen Konfidenzniveau heranrückt. Auffallend ist auch, dass mit 25 Netzpunkten bei entsprechend hohen Stichprobenumfängen die Werte der Überdeckungsrate und der Intervallslänge, die durch die moderate Methode hervorgebracht werden, näher bei jenen der hyperkonservativen Methode liegt als bei jenen der hypokonservativen Methode. Wie schon vorher erwähnt, stabilisiert sich bei $m = 25$ die gemittelte empirische Überdeckunsgrate, die sich bei der mo-

deraten Methode ergibt, im Vergleich zum Fall $m = 50$ deutlich über dem Konfidenzniveau. Dies könnte ein Hinweis sein, dass man bei $m = 25$ mit den drei Methoden noch eine relativ schlechte Abschätzungen von $C_{i,\alpha}^{full}(X_{i+1}^t)$ für die betrachteten Stichprobengrößen i und Iterationsläufe t erhält bzw. die moderate Methode in diesem Fall noch zu konservative Intervalle hervorbringt.

Die Anwendung der Full-Methoden, welche auf der fehlspezifierten Kleinste-Quadrate-Schätzung beruhen, ist auf den Fall $m = 50$ beschränkt. Die Simulationsergebnisse sind in den Abbildungen 6 und 7 illustriert und für ausgewählte Stichprobengrößen in Tabelle 3 zusammengefasst. Bezüglich der empirischen Überdeckungsraten ergibt sich das gleiche Bild wie bei der korrekt spezifizierten Schätzung. Auch das Verhalten der Intervallsbreiten ergibt einen ähnlichen Sachverhalt. Der einzige aber dafür gravierende Unterschied ist, dass für alle drei Full-Methoden die Breite des Konfidenzintervalls zu kleinen wie auch zu großen Stichproben deutlich größer ausfällt als die Breite seines jeweiligen mit der korrekten Spezifikation bestimmten Pendants. Ebenso wie bei der korrekten Spezifikation kommt es auch hier mit wachsendem Stichprobenumfang zu einer Stabilisierung der Intervallsbreiten, welche sich jedoch deutlich über $2 \cdot F^{-1}(1 - \alpha)$ vollzieht. Dies deutet auf eine Verzerrung in der Schätzung des linearen Modells (5.1) durch die fehlspezifizierte Kleinste-Quadrate-Methode hin, die einen Schätzfehler ergibt, welcher mit wachsender Stichprobengröße zwar kleiner wird aber eben nicht verschwindet. Im Gegensatz zur korrekten Spezifikation scheint also hier die Konsistenzbedingung (KF) aus Unterkapitel 4.2 verletzt zu sein, welche eine hinreichende Bedingung für die Konvergenz der Intervallsbreite zur Länge des „optimalen“ Konfidenzintervalls aus Unterkapitel 4.1 darstellt (siehe Satz 4.2). Was also die empirische Überdeckungsrate anbelangt, sind die auf der fehslpezifizierten Schätzung basierenden Full-Methoden gegenüber jenen, die sich einer korrekt spezifizierten Schätzung bedienen, qualitativ gleichwertig, jedoch bringen diese systematisch längere und für Prognosezwecke damit „unschärfere“ Konfidenzintervalle hervor.

Die Umsetzung der auf der Lasso-Schätzung beruhenden Full-Methoden beschränkt sich ebenfalls auf den Fall $m = 50$. Die Resultate werden in den Abbildungen 8 und 9 grafisch dargestellt und Tabelle 3 weist Werte für ausgewählte Stichprobengrößen aus. Die Entwicklung der gemittelten empirischen Überdeckungsrate zeichnet für die Full-Methoden unter der Lasso-Schätzung beinahe das gleiche Bild wie in den beiden davor betrachten Fällen. Insbesondere beobachten wir auch hier eine deutliche Annäherung der über die moderate Methode bestimmten Überdeckungsrate zum nominellen Konfidenzniveau mit wachsender Stichprobe. Ein Unterschied zu den davor diskutierten Fällen zeigt

sich jedoch hinsichtlich der gemittelten Intervallslänge. Unter Anwendung der Lasso-Methode beobachten wir wieder oszillierende Muster, welche sich ein wenig über den von den korrekt spezifizierten Full-Methoden generierten Gegenstücken befinden aber noch immer deutlich unter den Intervallslängen, die auf den fehlspezifizierten Schätzungen beruhen. Dabei beobachten wir eine Stabilsierung der von der moderaten Full-Methode, die sich der Lasso-Schätzung bedient, produzierten Intervallslänge auf einem Niveau, welches doch eine klar größere Distanz zu $2 \cdot F^{-1}(1-\alpha)$ aufweist als die unter der korrekt spezifierten Full-Methode (mit $m = 50$) hervorgebrachten Intervallsänge. Wir haben vorhin gesehen, dass bei einer Fehlspezifikation in der Kleinste-Quadrate-Schätzung, welche unter anderem auch relevante Regressoren unberücksichtigt lässt, die moderate Full-Methode zu deutlich längeren Konfidenzintervallen führt, wobei die Differenz der Intervallsbreiten zu $2 \cdot F^{-1}(1-\alpha)$ in allen Fällen asymptotisch nicht zu verschwinden scheint. Es könnte also sein, dass die in der Lasso-Schätzung verwendeten Variablenselektion dazu neigt, auch relevante Regressoren zu entfernen, was zu einer Verzerrung in der Schätzung von Modell (5.1) führt. Diese mündet in gegenüber der korrekten Spezifikation längeren Konfidenzintervallen und führt damit zu einer leichten Beeinträchtigung der „Schärfe“ der Prognosen für den Regressanden. Die Verschlechterung in der Qualität der unter der Lasso-Schätzung bestimmten Konfidnezbereiche scheint aber nicht so schwerwiegend zu sein wie bei der Fehlspezifikation.

Kleinste-Quadrate-Methode mit korrekter Spezifikation				
i	20	100	500	1000
Moderate Methode 25 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.893 (0.080)	0.901 (0.029)	0.906 (0.012)	0.906 (0.009)
$I_{i,\alpha}^{full}$	4.025 (1.115)	3.459 (0.303)	3.367 (0.130)	3.369 (0.094)
Hyperkonservative Methode 25 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.910 (0.073)	0.931 (0.025)	0.937 (0.011)	0.938 (0.008)
$I_{i,\alpha}^{full}$	4.413 (1.127)	3.869 (0.240)	3.776 (0.240)	3.818 (0.232)
Hypokonservative Methode 25 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.889 (0.083)	0.874 (0.036)	0.853 (0.018)	0.844 (0.015)
$I_{i,\alpha}^{full}$	3.921 (1.139)	3.085 (0.349)	2.839 (0.235)	2.824 (0.229)
Moderate Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.896 (0.076)	0.900 (0.030)	0.903 (0.013)	0.902 (0.009)
$I_{i,\alpha}^{full}$	4.083 (1.123)	3.426 (0.312)	3.324 (0.136)	3.317 (0.092)
Hyperkonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.904 (0.075)	0.919 (0.027)	0.922 (0.012)	0.922 (0.009)
$I_{i,\alpha}^{full}$	4.297 (1.124)	3.662 (0.337)	3.558 (0.171)	3.546 (0.144)
Hypokonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.894 (0.076)	0.892 (0.031)	0.881 (0.015)	0.877 (0.012)
$I_{i,\alpha}^{full}$	4.056 (1.129)	3.278 (0.322)	3.099 (0.167)	3.059 (0.144)

Tabelle 2: *Werte, welche sich im Zuge der mittels Dreiecksverteilung durchgeführten Simulation bei den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate Methode verwenden, unter* $m = 25$ *bzw.* $m = 50$ *für die gemittelte empirische Überdeckungsrate* $A_{i,\alpha}^{full,gem}$ *und die gemittelte Intervallslänge* $I_{i,\alpha}^{full}$ *zu den Stichprobengrößen* 20, 100, 500 *und* 1000 *realisiert haben, wobei die Werte in den Klammern für die Stichprobenstandardabweichungen* $sd(A_{i,\alpha}^{full,gem})$ *und* $sd(I_{i,\alpha}^{full})$ *stehen.*

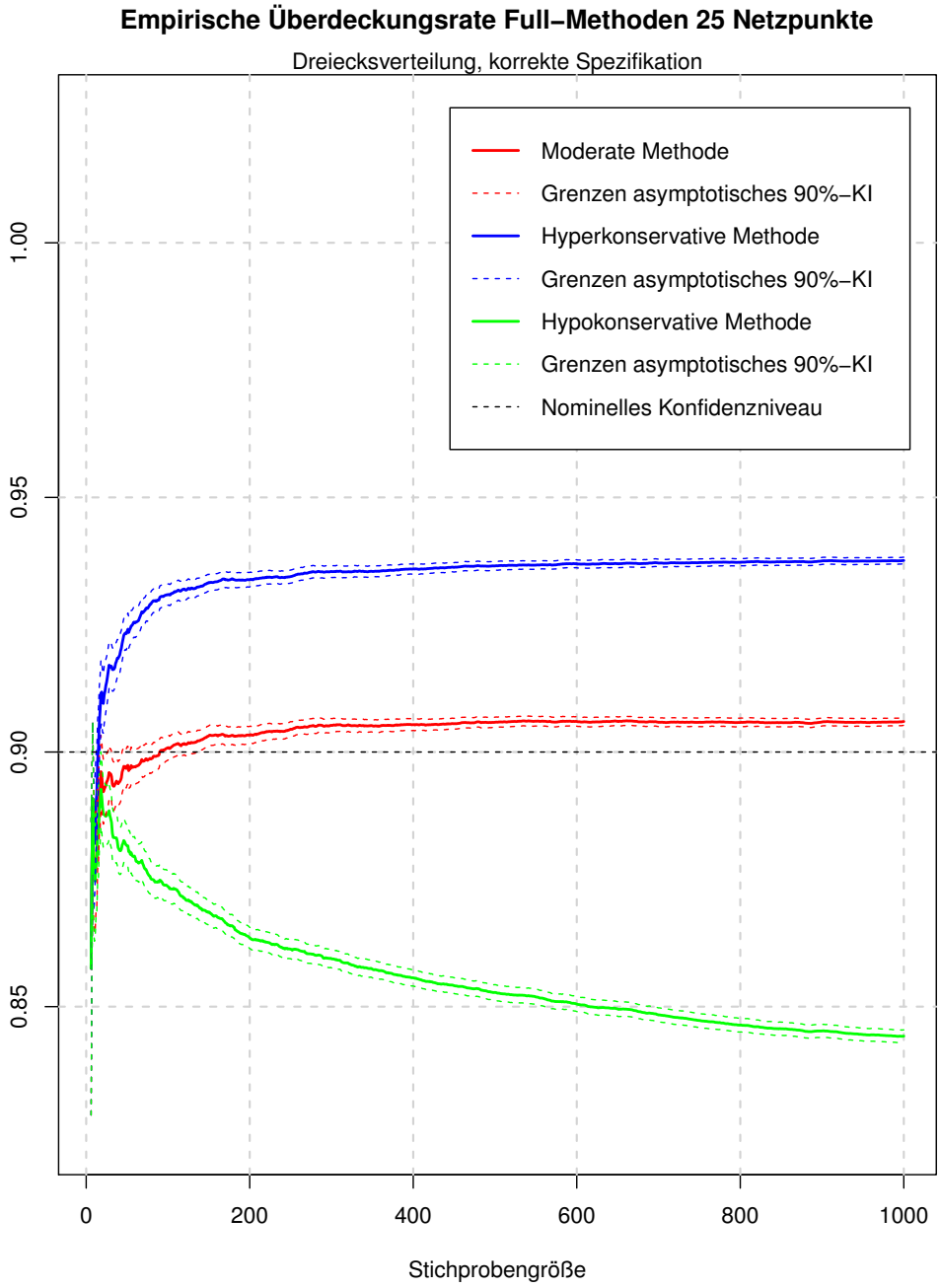

Abbildung 2: *Pfade für die nach den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate Methode verwenden, bestimmten gemittelten empirischen Überdeckungsraten, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 25$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

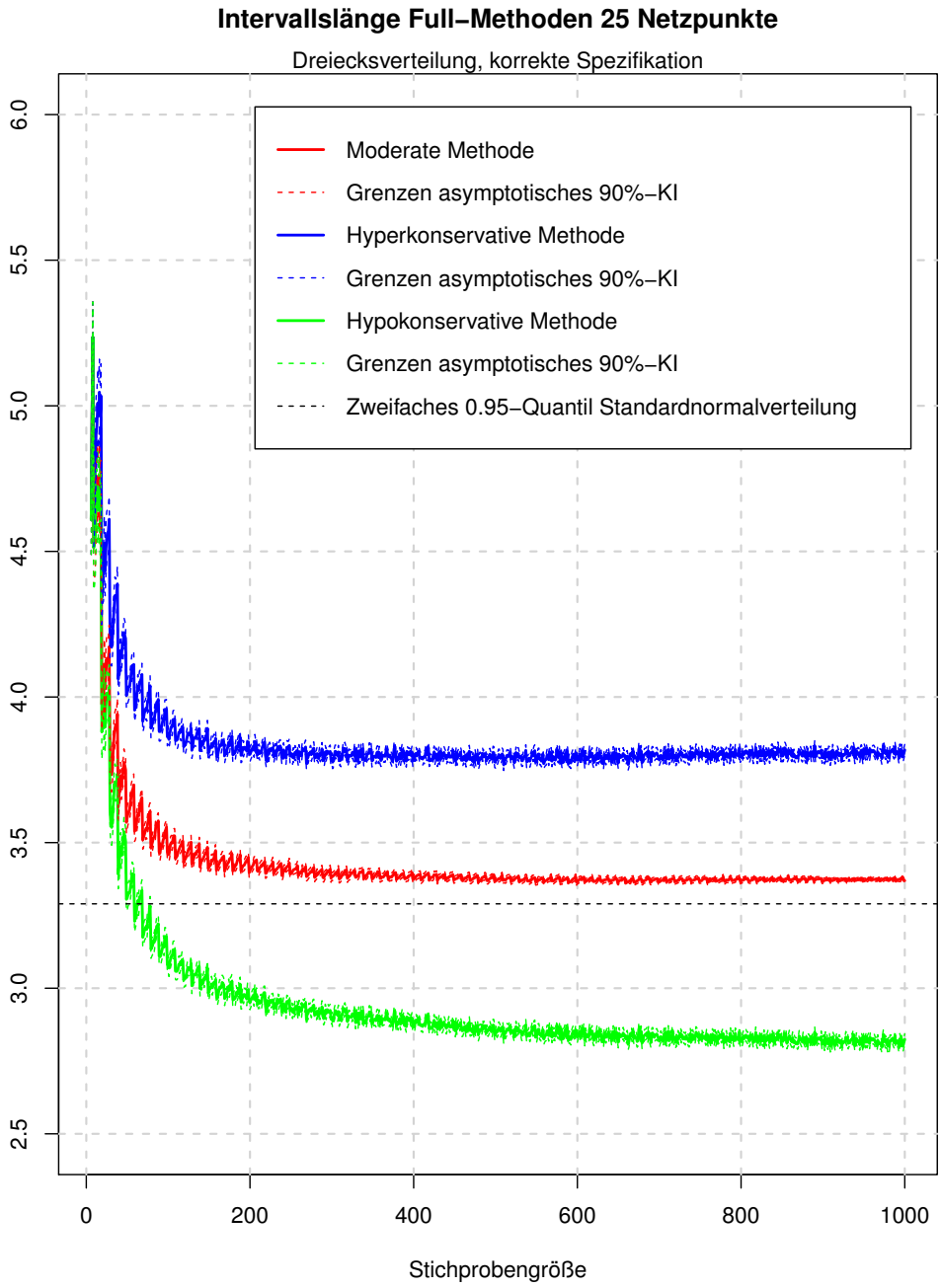

Abbildung 3: *Pfade für die nach den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate Methode verwenden, bestimmten gemittelten Intervallslängen, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 25$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

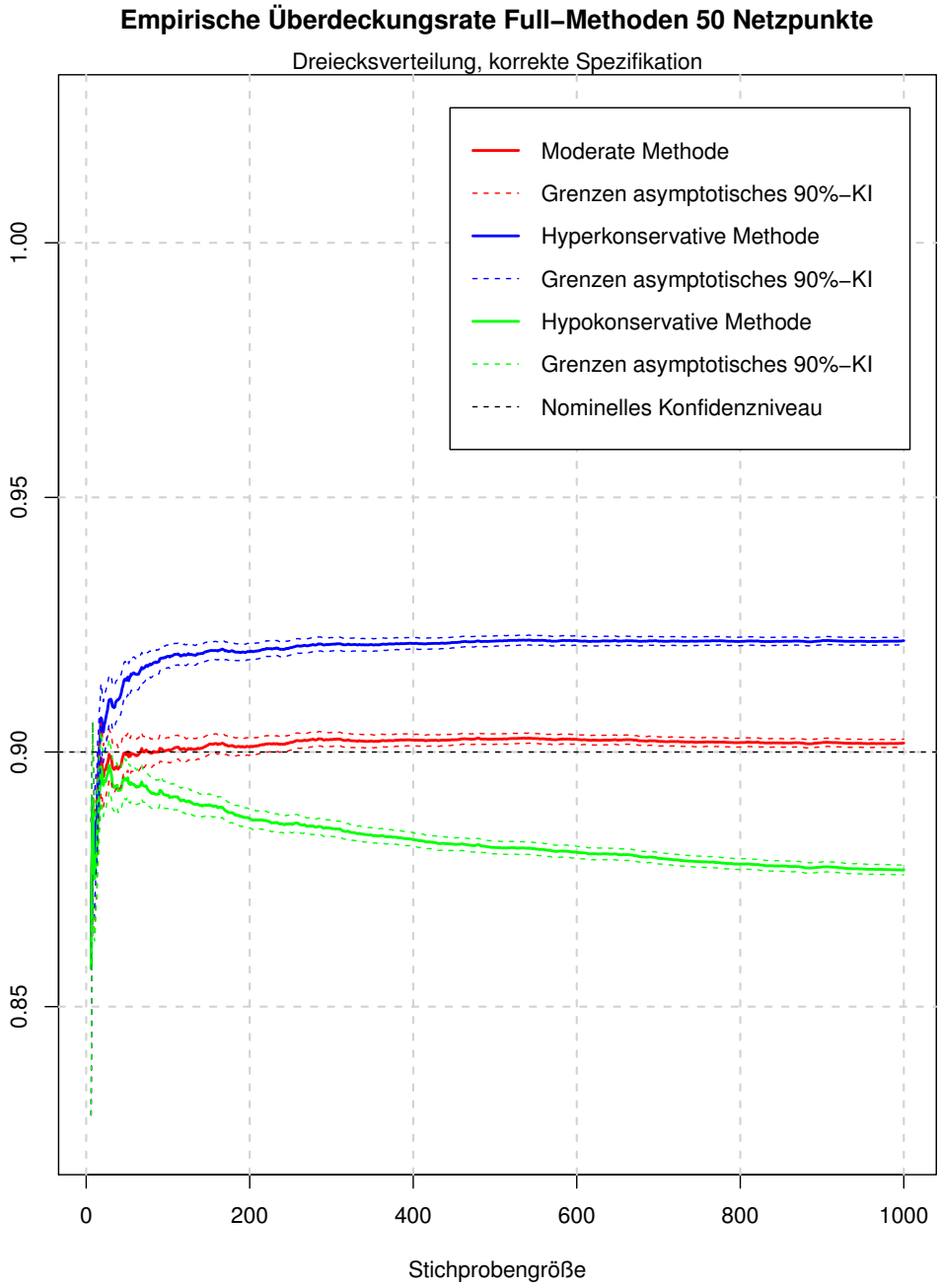

Abbildung 4: *Pfade für die nach den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate Methode verwenden, bestimmten gemittelten empirischen Überdeckungsraten, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 50$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

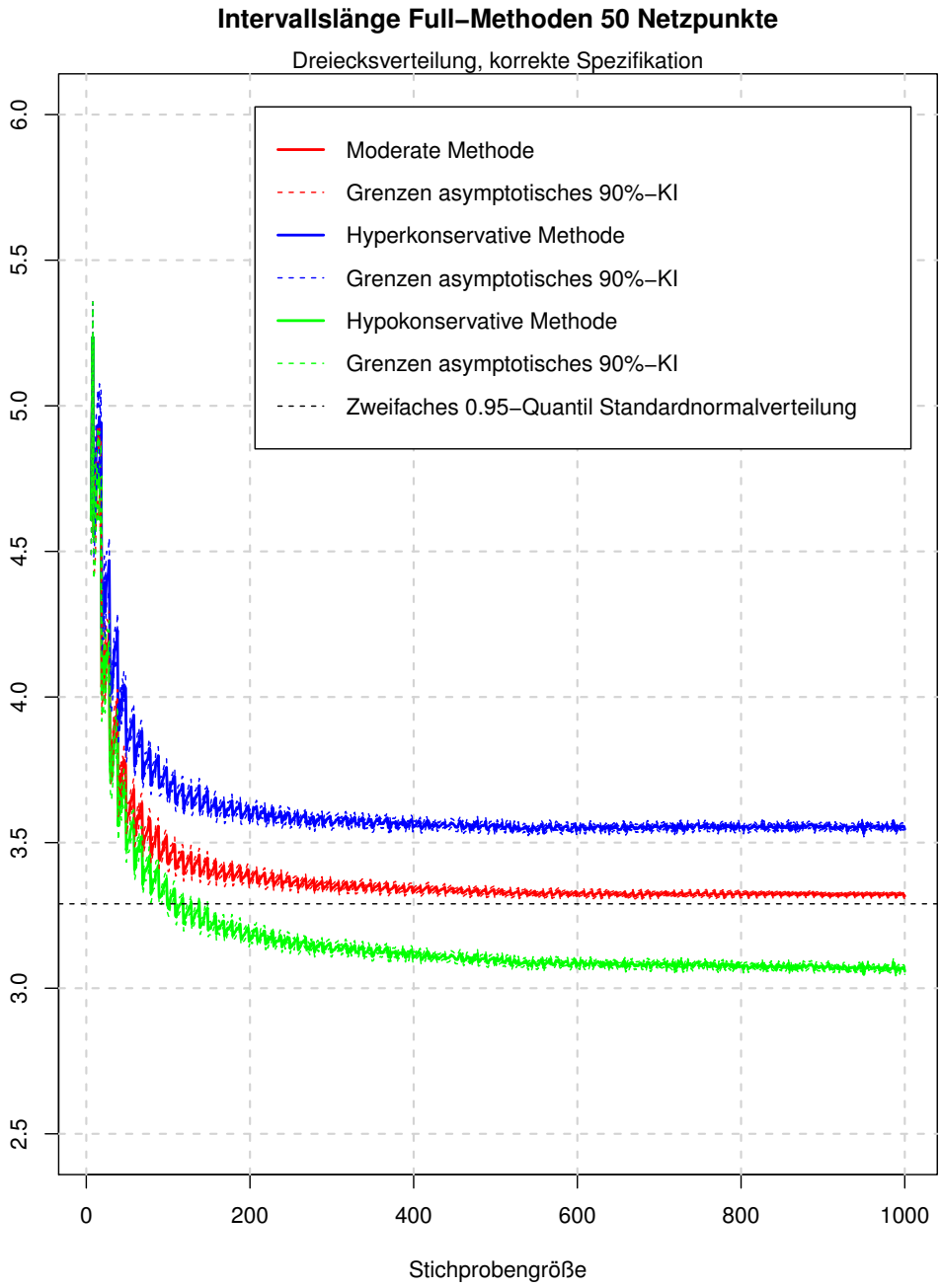

Abbildung 5: *Pfade für die nach den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate Methode verwenden, bestimmten gemittelten Intervallslängen, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 50$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

Kleinste-Quadrate-Methode mit Fehlspezifikation				
i	20	100	500	1000
Moderate Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.888 (0.078)	0.899 (0.030)	0.901 (0.013)	0.901 (0.009)
$I_{i,\alpha}^{full}$	4.621 (1.120)	4.188 (0.393)	4.104 (0.154)	4.100 (0.108)
Hyperkonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.895 (0.075)	0.914 (0.028)	0.917 (0.012)	0.918 (0.009)
$I_{i,\alpha}^{full}$	4.847 (1.152)	4.437 (0.424)	4.333 (0.177)	4.331 (0.154)
Hypokonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.888 (0.079)	0.893 (0.032)	0.885 (0.015)	0.882 (0.011)
$I_{i,\alpha}^{full}$	4.598 (1.126)	4.053 (0.407)	3.874 (0.176)	3.845 (0.151)
Lasso-Methode				
i	20	100	500	1000
Moderate Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.903 (0.080)	0.903 (0.030)	0.902 (0.013)	0.902 (0.009)
$I_{i,\alpha}^{full}$	3.897 (1.024)	3.446 (0.300)	3.410 (0.129)	3.414 (0.094)
Hyperkonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.916 (0.075)	0.922 (0.027)	0.922 (0.012)	0.921 (0.008)
$I_{i,\alpha}^{full}$	4.138 (1.053)	3.682 (0.323)	3.637 (0.165)	3.649 (0.140)
Hypokonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.902 (0.081)	0.893 (0.032)	0.882 (0.015)	0.878 (0.011)
$I_{i,\alpha}^{full}$	3.869 (1.032)	3.298 (0.310)	3.178 (0.160)	3.162 (0.138)

Tabelle 3: *Werte, welche sich im Zuge der mittels Dreiecksverteilung durchgeführten Simulation bei den Full-Methoden, welche die fehlspezifizierte Kleinste-Quadrate-Methode und die Lasso-Methode verwenden, unter* $m = 50$ *für die gemittelte empirische Überdeckungsrate* $A_{i,\alpha}^{full,gem}$ *und die gemittelte Intervallslänge* $I_{i,\alpha}^{full}$ *zu den Stichprobengrößen* 20, 100, 500 *und* 1000 *realisiert haben, wobei die Werte in den Klammern für die Stichprobenstandardabweichungen* $sd(A_{i,\alpha}^{full,gem})$ *und* $sd(I_{i,\alpha}^{full})$ *stehen.*

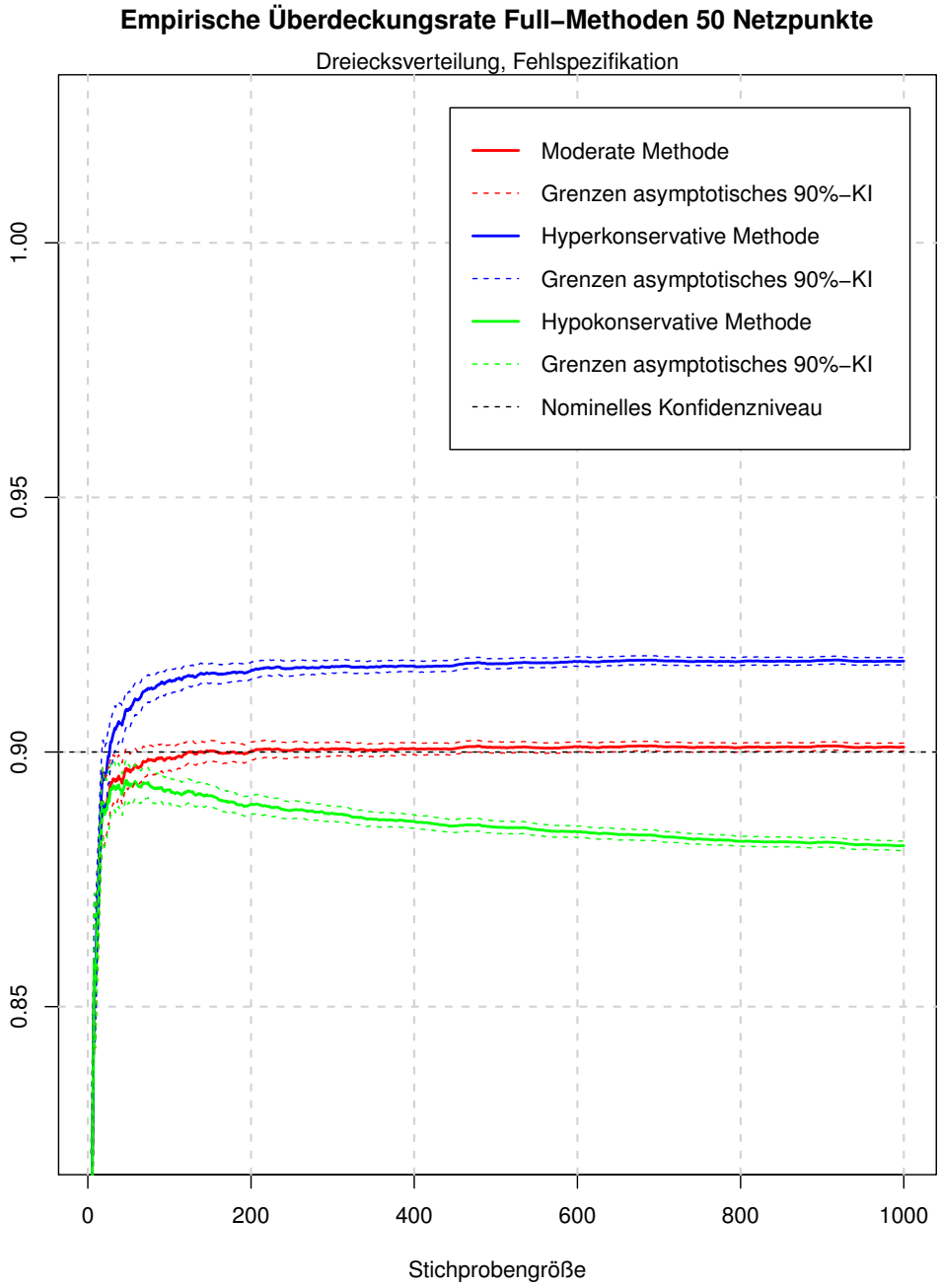

Abbildung 6: *Pfade für die nach den Full-Methoden, welche die fehlspezifizierte Kleinste-Quadrate-Methode verwenden, bestimmten gemittelten empirischen Überdeckungsraten, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q + 1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 50$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1 - \alpha)$*-Konfidenzintervalle.*

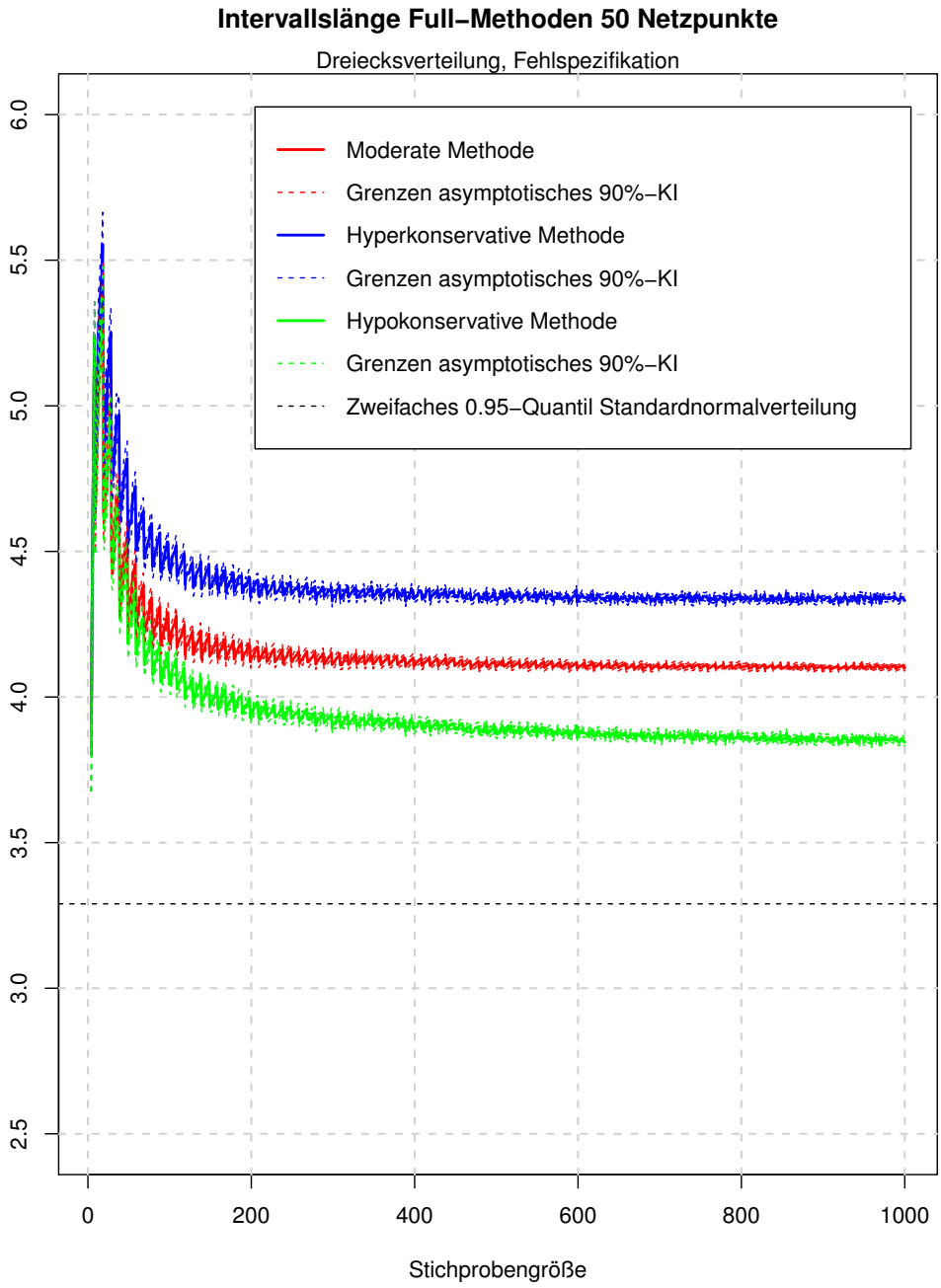

Abbildung 7: *Pfade für die nach den Full-Methoden, welche die fehlspezifizierte Kleinste-Quadrate-Methode verwenden, bestimmten gemittelten Intervallslängen, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 50$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

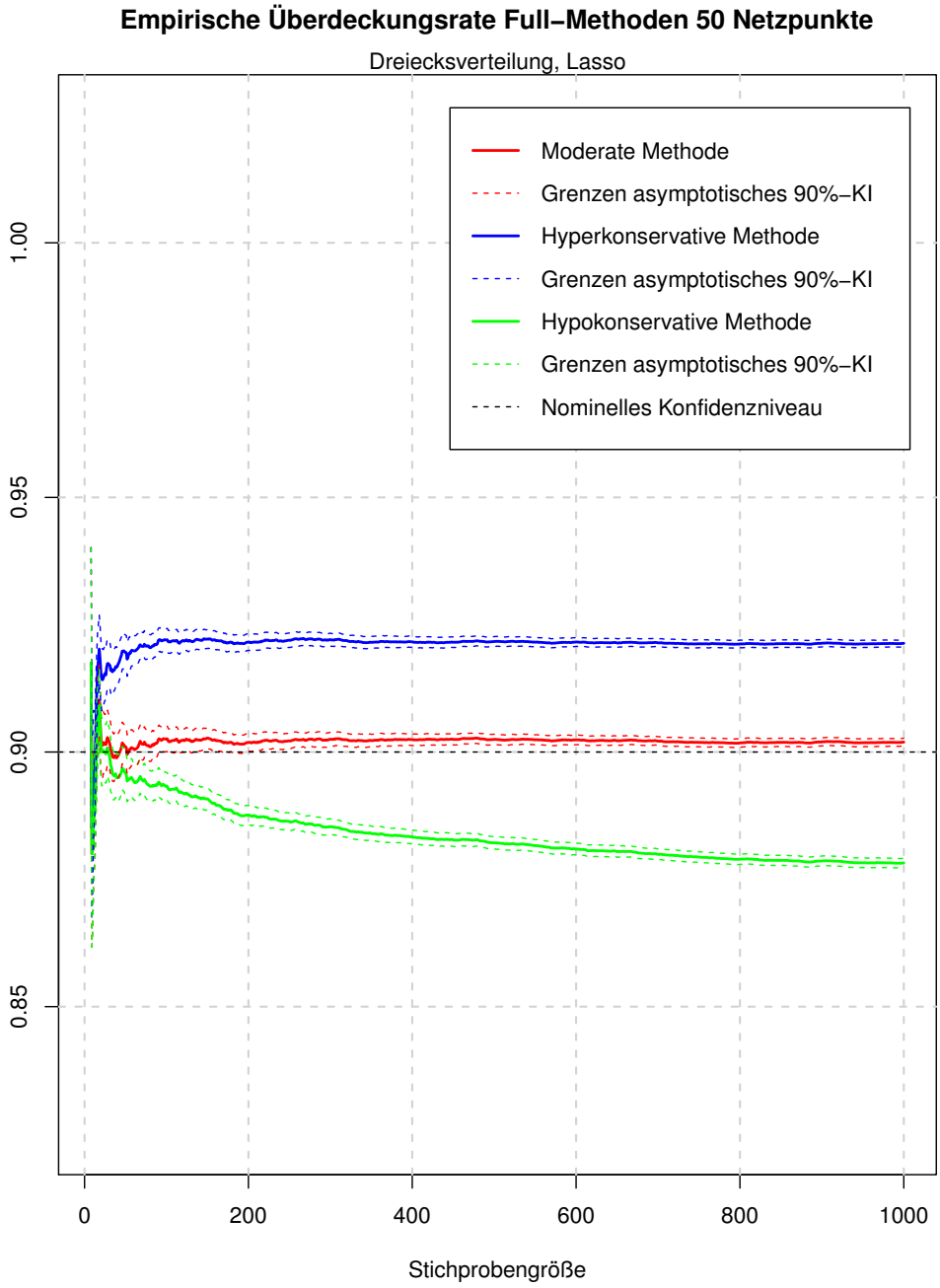

Abbildung 8: *Pfade für die nach den Full-Methoden, welche die Lasso-Methode verwenden, bestimmten gemittelten empirischen Überdeckungsraten, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $q+1$ *bis* n *reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter* $m = 50$ *realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

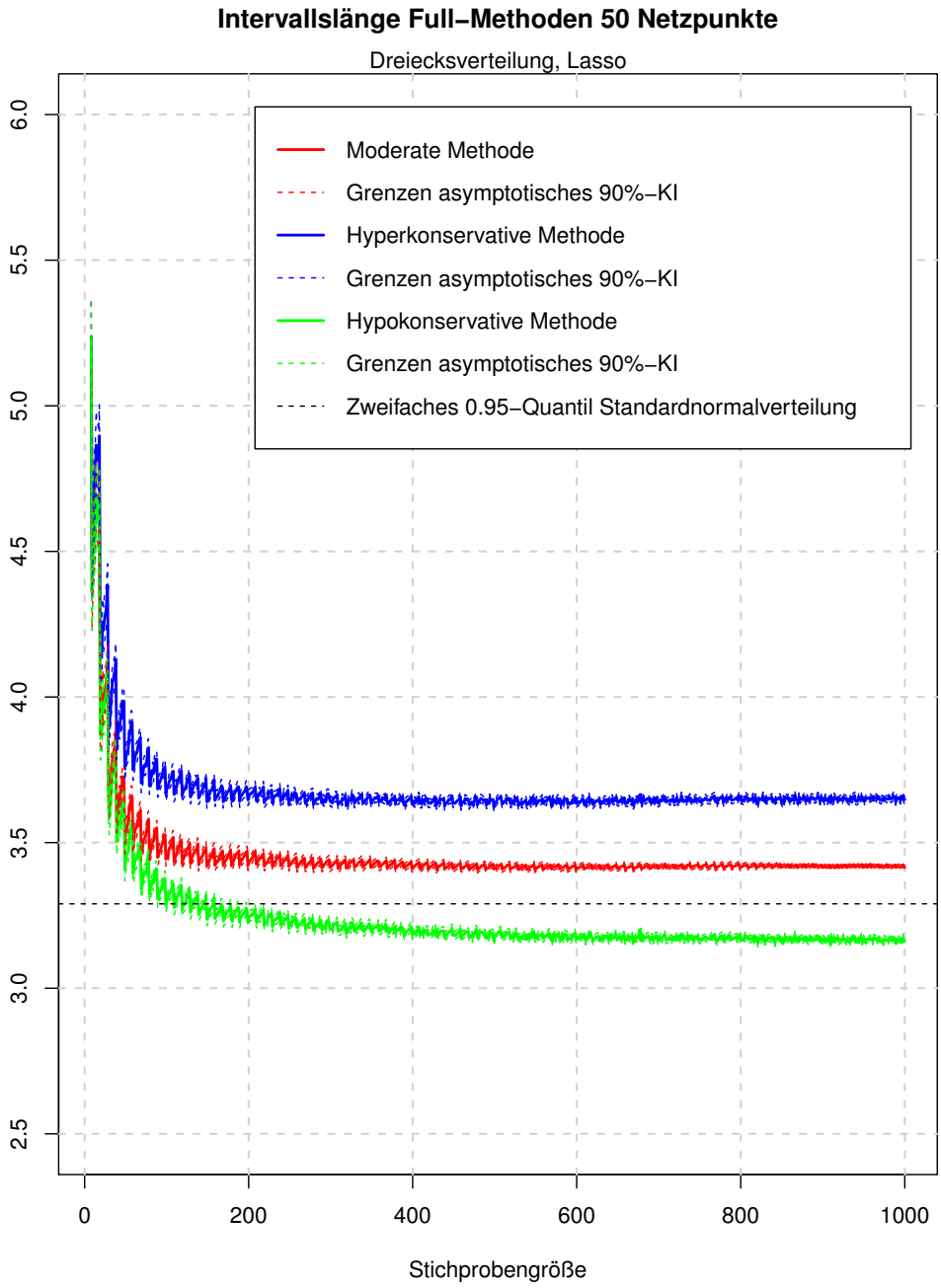

Abbildung 9: *Pfade für die nach den Full-Methoden, welche die Lasso-Methode verwenden, bestimmten gemittelten Intervallslängen, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von $q+1$ bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation unter $m = 50$ realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen $(1-\alpha)$-Konfidenzintervalle.*

5.4 Umsetzung und Resultate der Split-Methode

Wir betrachten für jede Wiederholung $t \in \{1, ..., s\}$ wieder die Zufallsvariablen (X_i^t, Y_i^t), $X_{i,d+1}^t$und $X_{i,d+2}^t$ für alle $i \in \{1, ..., n+1\}$. Ferner legen wir $\delta = 0.5$ fest, wonach die Aufteilung der in den Stichproben vorkommenden Instanzen in Trainingsmengen und in Teilmengen, auf welchen die Residuen berechnet werden, erfolgt. Zur folgenden Darstellung verwenden wir die Variable q, für welche wieder die Festsetzungen aus Unterkapitel 5.3 gelten. Für jedes $i \in \{p, ..., n\}$ mit

$$p := \left\lfloor \frac{q+1}{\delta} \right\rfloor$$

definieren wir

$$m_i := \lceil \delta i \rceil$$

und wählen zufällig eine Funktion k_i^t aus der Menge $S_{i,\delta}$ aller m_i-Permutationen von der Menge $\{1, ..., i\}$, d.h.

$$S_{i,\delta} := \{\sigma : \{1, ..., m_i\} \to \{1, ..., i\} \,|\, \sigma \text{ ist injektiv}\}.$$

Jede der $i!/(i - m_i)!$ Permutationen in $S_{i,\delta}$ kann dabei mit *gleicher Wahrscheinlichkeit* in die Auswahl fallen. Wir legen die Trainingsmenge durch

$$I_{i,1}^t := \{k_i^t(j) | \, 1 \leq j \leq m_i\}$$

fest (die Trainingsmenge bestimmen wir also durch zufälliges Ziehen aus der verfügbaren Stichprobe). Die Menge $I_{i,2}^t$ sei das Komplement von $I_{i,1}^t$ bezüglich $\{1, ..., i\}$. Wir definieren nun für die Kleinste-Quadrate-Schätzungen die Designmatrix $\mathbf{X}_{i,\delta}^t$ mit

$$\mathbf{X}_{i,\delta}^t := \begin{pmatrix} 1 & X_{k_i^t(1),1}^t & ... & X_{k_i^t(1),d}^t \\ \vdots & \vdots & & \vdots \\ 1 & X_{k_i^t(m_i),1}^t & ... & X_{k_i^t(m_i),d}^t \end{pmatrix}$$

für die korrekte Spezifikation und

$$\mathbf{X}_{i,\delta}^t := \begin{pmatrix} 1 & (X_{k_i^t(1),1}^t)^2 & X_{k_i^t(1),3}^t & X_{k_i^t(1),d}^t \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_{k_i^t(m_i),1}^t)^2 & X_{k_i^t(m_i),3}^t & X_{k_i^t(m_i),d}^t \end{pmatrix}$$

für die Fehlspezifikation, welche wie bei den Full-Methoden durch fehlende Regressoren und einen unterstellten quadratischen Zusammenhang gekennzeichnet ist. Wir definieren außerdem den m_i-dimensionalen Spaltenvektor

$$\mathbf{Y}_{i,\delta}^t := (Y_{k_i^t(1)}^t, ..., Y_{k_i^t(m_i)}^t)^T$$

und bestimmen den Schätzer

$$\hat{\beta}^t_{i,\delta} \equiv (\hat{\beta}^t_{i,\delta,0}, \hat{\beta}^t_{i,\delta,1}, ..., \hat{\beta}^t_{i,\delta,q})^T$$

als $(q+1)$-dimensionalen Spaltenvektor, welcher unter der Kleinste-Quadrate-Methode durch

$$\hat{\beta}^t_{i,\delta} := ((\mathbf{X}^t_{i,\delta})^T \cdot \mathbf{X}^t_{i,\delta})^{-1} \cdot (\mathbf{X}^t_{i,\delta})^T \cdot \mathbf{Y}^t_{i,\delta}$$

berechnet wird. Unter der Lasso-Methode ist $\hat{\beta}^t_{i,\delta}$ das Minimum von

$$f^{lasso}_{split}(\beta_0, ..., \beta_q) := \sum_{j=1}^{m_i} \left(Y^t_{k^t_i(j)} - \beta_0 - \sum_{k=1}^{q} \beta_k \cdot X^t_{k^t_i(j),k} \right)^2 + \lambda \cdot 2 \cdot m_i \cdot \sum_{k=1}^{q} |\beta_k|$$

über alle $\beta_0, ..., \beta_q \in \mathbb{R}$ für gegebenen Realisationen der Zufallsvariablen $Y^t_{k^t_i(1)}, ..., Y^t_{k^t_i(m_i)}$ und $X^t_{k^t_i(1),k}, ..., X^t_{k^t_i(m_i),k}$ mit $k \in \{1, ..., q\}$; vgl. Friedman et al. (2010, S.3). Im Rahmen des Lasso-Verfahrens verwenden wir auch hier für die näherungsweisen Bestimmung des Schätzers das in die R-Funktion `glmnet` integrierte *covariance updating*. Wieder setzen wir $\lambda = 0.15$ und verwenden alle im linearen Modell (5.1) vorkommenden Regressoren sowie zwei überflüssige Variablen. Die von uns verwendeten Schätzverfahren zur Bestimmung von $\hat{\beta}^t_{i,\delta}$ korrespondieren jeweils mit einem Algorithmus zur Split-Methode $a^t_{i,\delta}$, der uns an der Stelle

$$x = (x_1, ..., x_q) \in \mathbb{R}^q$$

die Regressionsfunktion

$$a^t_{i,\delta}(Z^t_{k^t_i(1)}, ..., Z^t_{k^t_i(m_i)})(x) = \hat{\beta}^t_{i,\delta,0} + \sum_{k=1}^{q} \hat{\beta}^t_{i,\delta,k} \cdot x_k$$

für die korrekte Spezifikation und den Lasso Fall sowie

$$a^t_{i,\delta}(Z^t_{k^t_i(1)}, ..., Z^t_{k^t_i(m_i)})(x) = \hat{\beta}^t_{i,\delta,0} + \hat{\beta}^t_{i,\delta,1} \cdot (x_1)^2 + \hat{\beta}^t_{i,\delta,2} \cdot x_2 + \hat{\beta}^t_{i,\delta,3} \cdot x_3$$

für die Fehlspezifikation liefert.[14] Wir bestimmen damit die Residuen $U^t_{j,\delta}$ für alle $j \in I^t_{i,2} \cup \{i+1\}$. Wir erhalten schließlich das auf der Split-Methode basierende Konfidenzintervall $C^{split}_{i,\delta,\alpha}(X^t_{i+1})$. Für alle zu betrachteten Stichprobengrößen $i \in \{p, ..., n\}$ schätzen wir die Überdeckungswahrscheinlichkeit des Konfidenzintervalls mit der gemittelten empirischen Überdeckungsrate $A^{split,gem}_{i,\delta,\alpha}$, welche durch

$$A^{split,gem}_{i,\delta,\alpha} := \frac{1}{s} \sum_{t=1}^{s} \frac{1}{i-p+1} \sum_{j=p}^{i} B^{split}_{j,\delta,\alpha,t}$$

bestimmt sei. Dabei wird für jedes $j \in \{p, ..., i\}$ und $t \in \{1, ..., s\}$ der Indikator $B^{split}_{j,\delta,\alpha,t}$ definiert als

$$B^{split}_{j,\delta,\alpha,t} := \begin{cases} 1 & \text{falls } Y^t_{j+1} \in C^{split}_{j,\delta,\alpha}(X^t_{j+1}), \\ 0 & \text{sonst.} \end{cases}$$

Für $j \in \{p, ..., i\}$ und $t \in \{1, ..., s\}$ entspricht also

$$\frac{1}{i-p+1} \sum_{j=p}^{i} B^{split}_{j,\delta,\alpha,t}$$

der empirischen Überdeckungsrate zur Stichprobengröße i im Iterationslauf t. Sei

$$\tilde{h}_i := \lceil (1-\alpha)(i-m_i+1) \rceil.$$

Falls $\tilde{h}_i > i - m_i$, dann sei die gemittelte Länge $I^{split}_{i,\delta,\alpha}$ der Konfidenzintervalle festgelegt durch

$$I^{split}_{i,\delta,\alpha} := \infty.$$

Für den Fall $\tilde{h}_i \leq i - m_i$ sei $U^t_{(\tilde{h}_i:(i-m_i)),\delta}$ die $\tilde{h}_i$-te Ordnungsstatistik der Residuen auf $I_{i,2}$ zur Iteration t. Die gemittelte Länge der Konfidenzintervalle ist somit bestimmt durch

$$I^{split}_{i,\delta,\alpha} := \frac{1}{s} \sum_{t=1}^{s} 2 \cdot U^t_{(\tilde{h}_i:(i-m_i)),\delta}.$$

Dabei entspricht $2 \cdot U^t_{(\tilde{h}_i:(i-m_i)),\delta}$ der Länge des Konfidenzintervalls $C^{split}_{i,\delta,\alpha}(X^t_{i+1})$ zur Stichprobengröße i im Iterationslauf t (vgl. Unterkapitel 3.2). Auch bei der Split-Methode wollen wir unsere Beobachtungen mit Intervallschätzern für die empirische Überdeckungsrate und für die Länge des Konfidenzintervalls zu jeder Stichprobengröße i ergänzen. Wir bestimmen jeweils die korrigierte Stichprobenstandardabweichung der beiden Größen, nämlich

$$sd(A^{split,gem}_{i,\delta,\alpha}) := \sqrt{\frac{1}{s-1} \sum_{t=1}^{s} \left(\frac{1}{i-p+1} \sum_{j=p}^{i} B^{split}_{j,\delta,\alpha,t} - A^{split,gem}_{i,\delta,\alpha} \right)^2}$$

und

$$sd(I^{split}_{i,\delta,\alpha}) := \sqrt{\frac{1}{s-1} \sum_{t=1}^{s} \left(2 \cdot U^t_{(\tilde{h}_i:(i-m_i)),\delta} - I^{split}_{i,\delta,\alpha} \right)^2},$$

um dann das asymptotisch gültige zweiseitige $1-\alpha$-Konfidenzintervall für die empirische Überdeckungsrate

$$\left[A^{split,gem}_{i,\delta,\alpha} - \frac{F^{-1}(1-\alpha) \cdot sd(A^{split,gem}_{i,\delta,\alpha})}{\sqrt{s}},\ A^{split,gem}_{i,\delta,\alpha} + \frac{F^{-1}(1-\alpha) \cdot sd(A^{split,gem}_{i,\delta,\alpha})}{\sqrt{s}} \right]$$

sowie das entsprechende zweiseitige Konfidenzintervall für die Intervallslänge

$$\left[I^{split}_{i,\delta,\alpha} - \frac{F^{-1}(1-\alpha) \cdot sd(I^{split}_{i,\delta,\alpha})}{\sqrt{s}} \, , \, I^{split}_{i,\delta,\alpha} + \frac{F^{-1}(1-\alpha) \cdot sd(I^{split}_{i,\delta,\alpha})}{\sqrt{s}} \right]$$

zu ermitteln.

Wieder interessieren wir uns, wie sich die gemittelte empirische Überdeckungsrate und die gemittelte Länge der Konfidenzintervalle in Abhängigkeit von der Stichprobengröße i entwickeln. Die Ergebnisse der Simulation für die Split-Methode, die sich der korrekt spezifizierten Kleinste-Quadrate-Schätzung bedient, sind in den Abbildungen 10 und 11 illustriert. Zusätzlich finden sich in Tabelle 4 Werte, die für die relevanten Größen für eine Auswahl von Stichprobengrößen aus $\{p, ..., n\}$ beobachtet wurden. Die Stichprobenstandartabweichungen der empirischen Überdeckungsrate und der Intervallslänge stehen in Klammern. Wir sehen, dass die gemittelte empirische Überdeckungsrate bei kleinen Stichprobengrößen relativ große Werte annimmt und sich mit wachsender Stichprobe dem nominellen Konfidenzniveau von $1 - \alpha$ annähert. Die gemittelte Intervallsbreite nimmt bei kleinen Stichproben ebenfalls größere Werte an. Mit wachsendem Stichprobenumfang nimmt sie jedoch deutlich ab und scheint sich dann ab einer Stichprobengröße von etwa 200 in einer Umgebung von $2 \cdot F^{-1}(1-\alpha)$ zu stabilisieren. Das hier beobachtete asymptotische Verhalten der gemittelten empirischen Überdeckungsrate und der gemittelten Intervallsbreite scheint gut mit dem Inhalt der Schlußfolgerungen der Sätze 4.3 und 4.4 zu harmonieren. Wie schon bei der Full-Methode beobachten wir auch hier in der Entwicklung der gemittelten Intervallslänge einen regelmäßigen Zyklus aufeinanderfolgender Wachstums- und Kontraktionsphasen, der sich mit wachsender Stichprobengröße realisiert.

Die Simulationsresultate der Split-Methode, die auf der fehlspezifizierten Schätzung basiert, sind durch die Abildungen 12 und 13 illustriert und für eine Auswahl an Stichprobenumfängen in Tabelle 4 zusammengefasst. Wie im Fall der Full-Methoden sehen wir auch bei der Split-Methode, dass eine Fehlspezifikation auf die Entwicklung der empirischen Überdeckungsrate keine Auswirkungen zu haben scheint. Auch die Intervallsbreite stabilisiert sich wieder mit zunehmender Stichprobengröße in einem oszillierenden Muster. Jedoch bringt die Fehlspezifikation einen Schätzfehler hervor, der im Vergleich zur korrekten Spezifikation zu deutlich breiteren Konfidenzintervallen führt. Die Differenz zwischen der Breite des Split-Intervalls und der Breite des „optimalen“ Konfidenzintervalls aus Unterkapitel 4.1 verschwindet außerdem nicht mit wachsender Stichprobengröße, sondern stabilisiert sich lediglich. Offenbar scheint die

Fehlspezifikation die Konsistenzbedingung (KS), welche hinreichend ist für das in Satz 4.4 behauptete Konvergenzverhalten der Intervallsbreite, zu verletzen.

Die Abbildungen 14 und 15 illustrieren die Resultate der auf das Lasso-Verfahren beruhenden Split-Methode und Tabelle 4 zeigt ausgewählte Werte für die emperische Überdeckungsrate und die Intervallslänge. Der realisierte Pfad der gemittelten empirischen Überdeckungsrate zeigt eine ähnliche Entwicklung wie schon in den Fällen davor. Insbesondere beobachten wir mit zunehmender Stichprobe wieder eine deutliche Annäherung zum nominellen Konfidenzniveau. Die gemittelte Intervallsänge zeigt auch unter Anwendung der Lasso-Methode einen oszilierenden Entwicklungspfad, welcher sicht leicht über jenem, der sich unter der korrekten Spezifikation realisert hat, befindet, aber noch immer klar unter dem mit der fehlspezifizierten Schätzung hervorgebrachten Pfad ist. Wir beobachten außerdem eine Entwicklung hin zu einem Niveau, welches ein wenig über $2 \cdot F^{-1}(1-\alpha)$ liegt. Im Verhältnis dazu liegt die unter der korrekt spezifizierten Split-Methode hervorgebrachten Intervallsänge im Bereich großer Stichproben deutlich näher bei $2 \cdot F^{-1}(1-\alpha)$. Insgesamt weisen diese Resultate (wie schon bei der Full-Methode) auf eine leichte Verzerrung hin, die möglicherweise von der in der Lasso-Schätzung inkorporierten Variablenselektion begangen wird und die zu einer gegenüber der korrekt spezifizierten Split-Methode leichten Verschlechterung in der Qualität der Prognoseintervalle führt. Diese bleibt jedoch im Verhältnis zur Fehlspezifikation noch immer überschaubar.

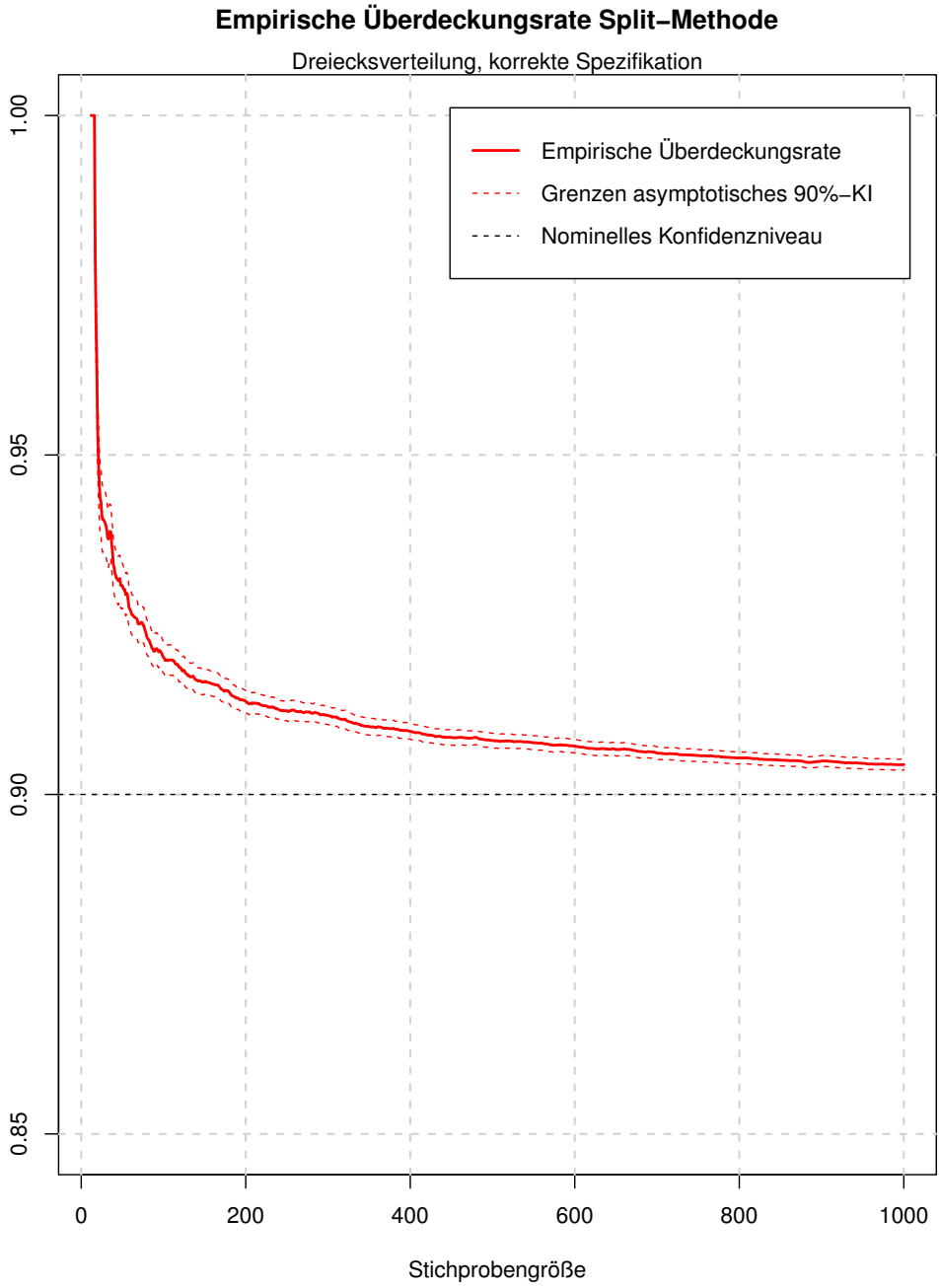

Abbildung 10: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten empirischen Überdeckungsrate* $A_{i,\delta,\alpha}^{split,gem}$ *realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

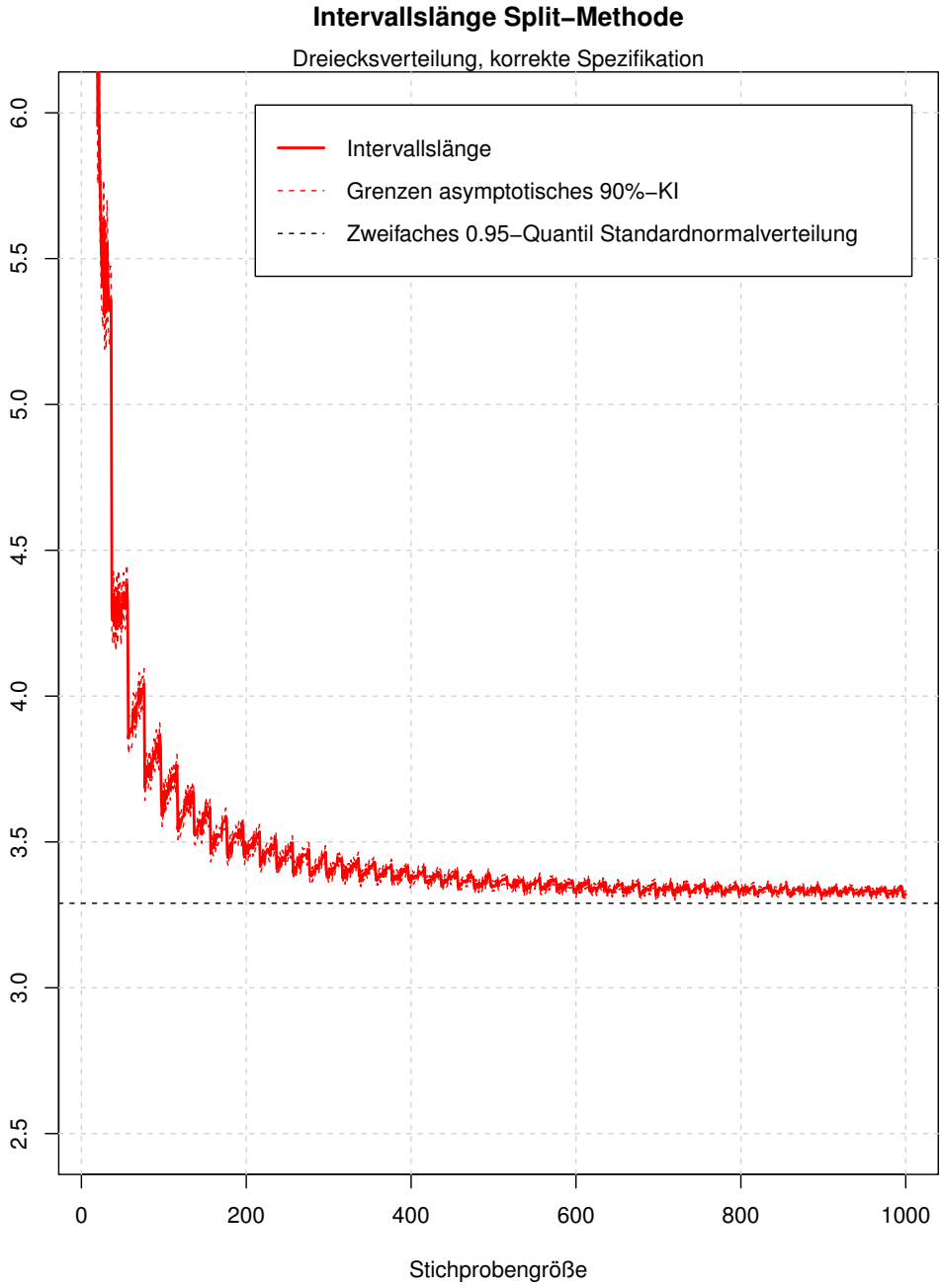

Abbildung 11: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten Intervallslänge* $I^{split}_{i,\delta,\alpha}$ *realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

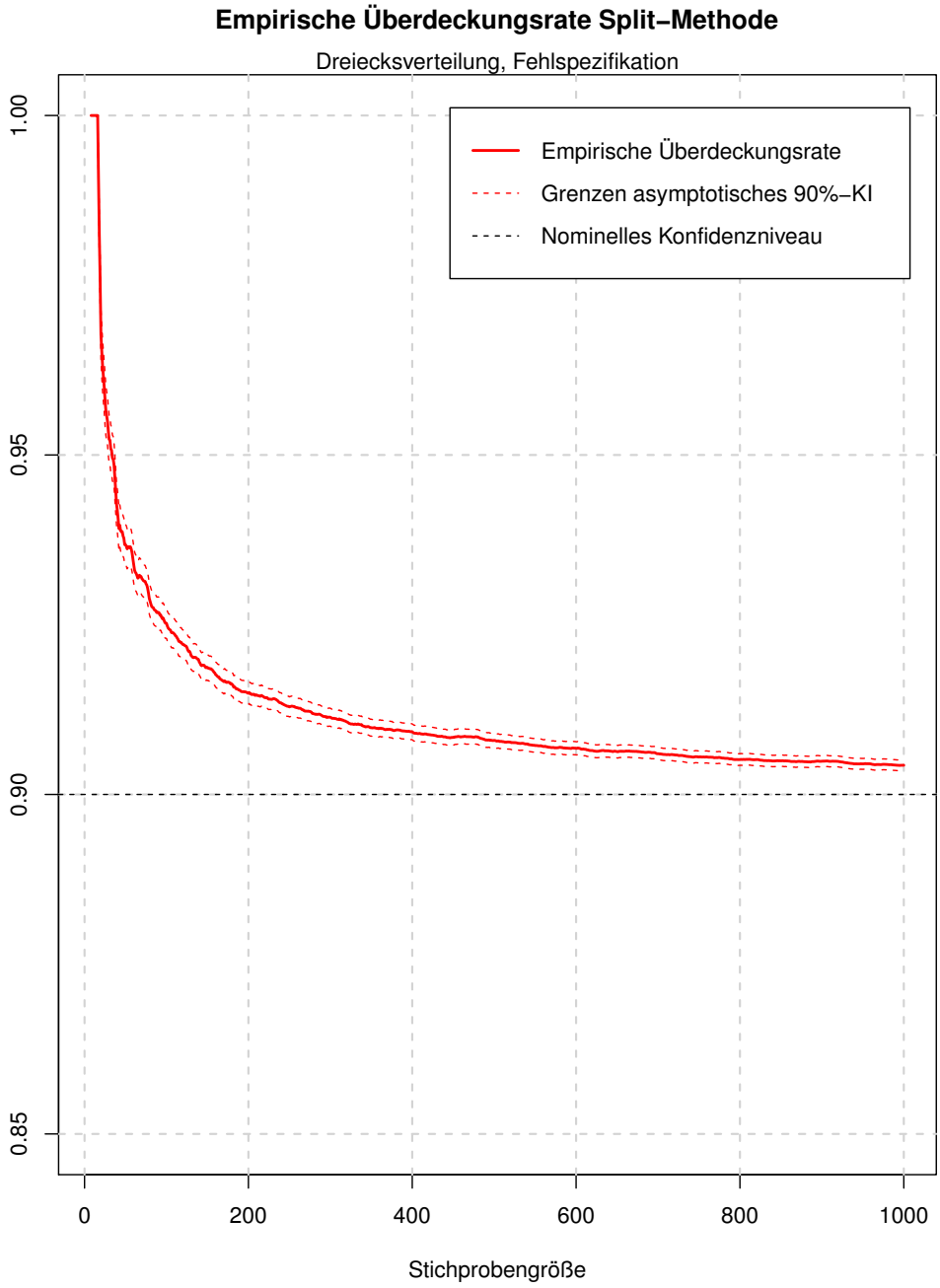

Abbildung 12: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die fehlspezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten empirischen Überdeckungsrate* $A_{i,\delta,\alpha}^{split,gem}$ *realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

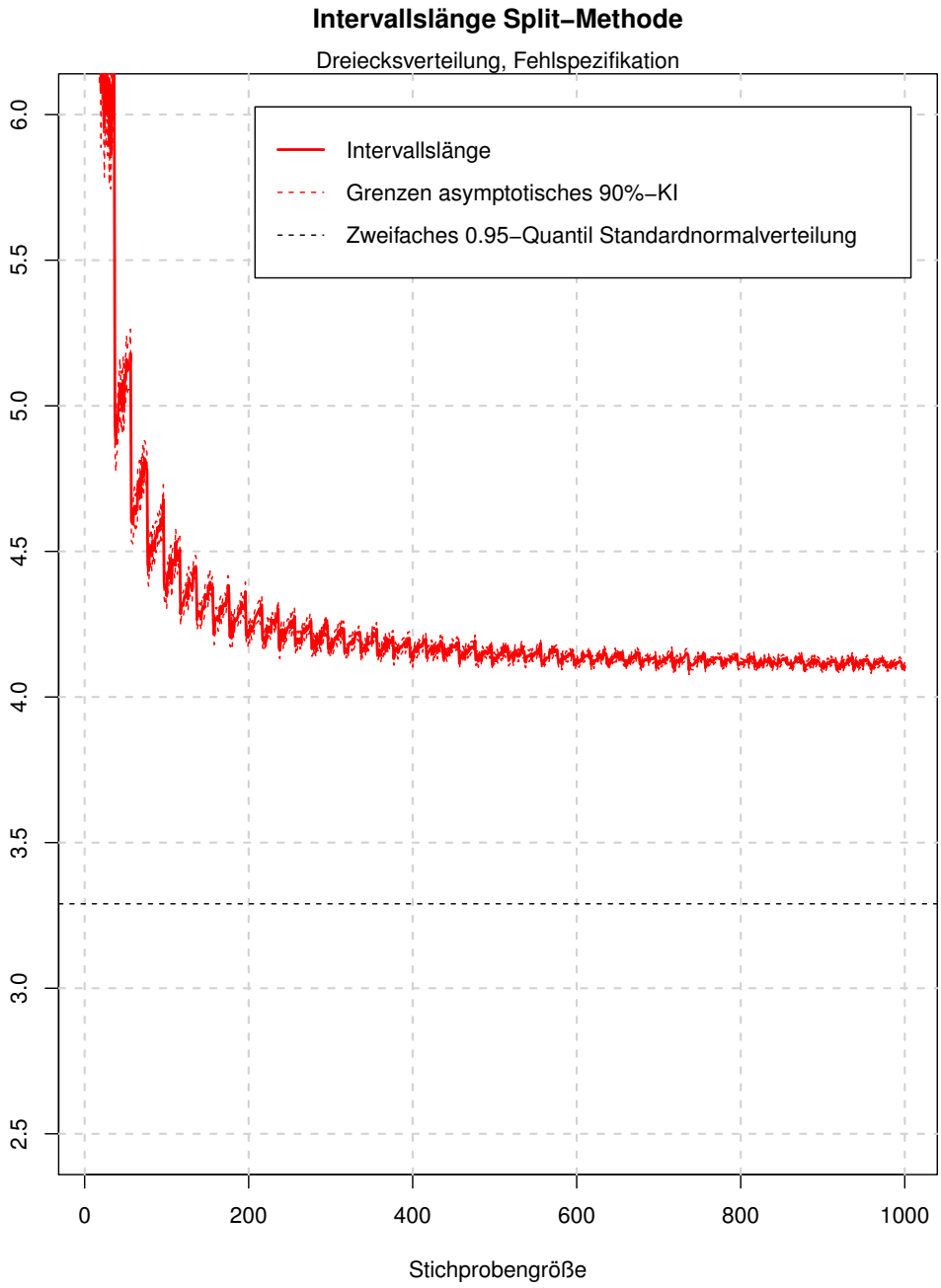

Abbildung 13: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die fehlspezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten Intervallslänge* $I^{split}_{i,\delta,\alpha}$ *realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

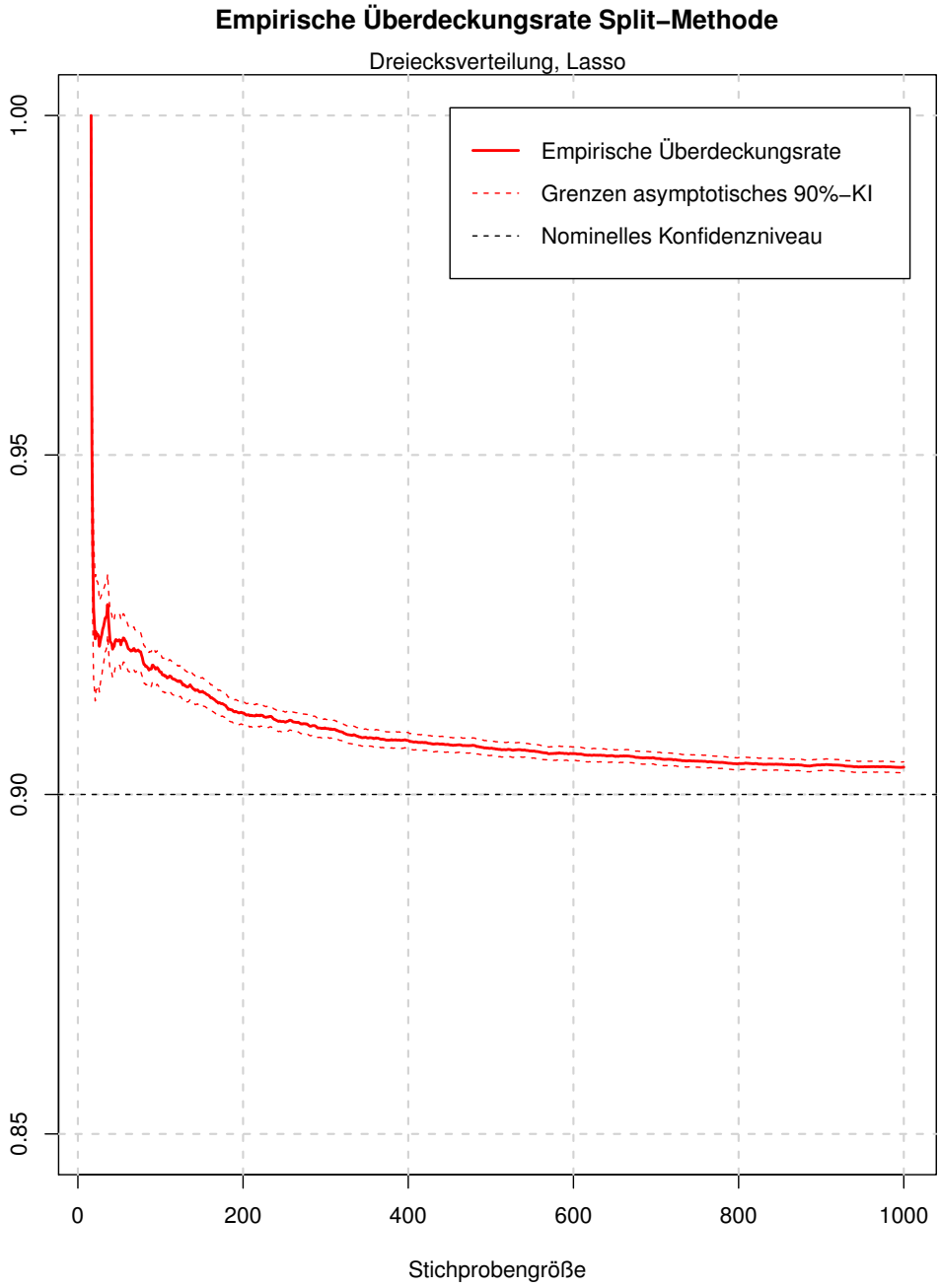

Abbildung 14: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die Lasso-Methode verwendet, bestimmten gemittelten empirischen Überdeckungsrate $A_{i,\delta,\alpha}^{split,gem}$ realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen $(1-\alpha)$-Konfidenzintervalle.*

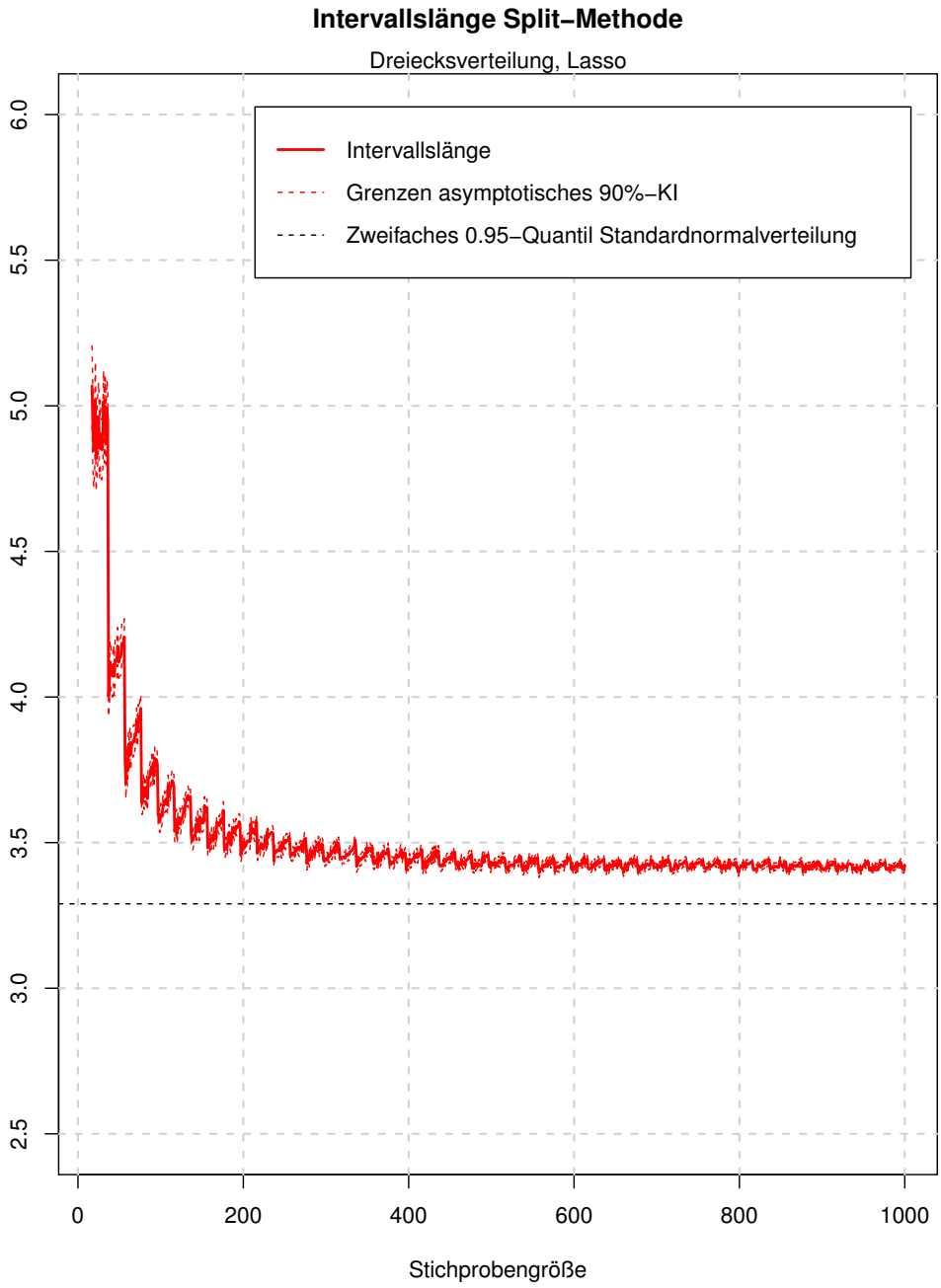

Abbildung 15: *Pfad, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von p bis n reicht, im Zuge der mittels Dreiecksverteilung durchgeführten Simulation für die nach der Split-Methode, welche die Lasso-Methode verwendet, bestimmten gemittelten Intervallslänge* $I^{split}_{i,\delta,\alpha}$ *realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle.*

Split-Methode				
i	20	100	500	1000
Kleinste-Quadrate-Methode mit korrekter Spezifikation				
$A_{i,\delta,\alpha}^{split,gem}$	0.953 (0.065)	0.920 (0.028)	0.908 (0.013)	0.904 (0.010)
$I_{i,\delta,\alpha}^{split}$	5.955 (2.426)	3.673 (0.468)	3.348 (0.204)	3.320 (0.135)
Kleinste-Quadrate-Methode mit Fehlspezifikation				
$A_{i,\delta,\alpha}^{split,gem}$	0.976 (0.042)	0.925 (0.027)	0.908 (0.013)	0.904 (0.009)
$I_{i,\delta,\alpha}^{split}$	6.429 (2.604)	4.384 (0.543)	4.135 (0.232)	4.095 (0.166)
Lasso-Methode				
$A_{i,\delta,\alpha}^{split,gem}$	0.925 (0.115)	0.918 (0.030)	0.907 (0.013)	0.904 (0.010)
$I_{i,\delta,\alpha}^{split}$	4.874 (1.406)	3.597 (0.466)	3.426 (0.201)	3.423 (0.142)

Tabelle 4: *Werte, welche sich im Zuge der mittels Dreiecksverteilung durchgeführten Simulation bei der Split-Methode, welche die korrekt spezifizierte bzw. fehlspezifizierte Kleinste-Quadrate-Methode und die Lasso-Methode verwendet, für die gemittelte empirische Überdeckungsrate* $A_{i,\delta,\alpha}^{split,gem}$ *und die gemittelte Intervallsbreite* $I_{i,\delta,\alpha}^{split}$ *zu den Stichprobengrößen* 20, 100, 500 *und* 1000 *realisiert haben, wobei die Werte in den Klammern für die Stichprobenstandardabweichungen* $sd(A_{i,\delta,\alpha}^{split,gem})$ *und* $sd(I_{i,\delta,\alpha}^{split})$ *stehen.*

5.5 Vergleich zwischen Full- und Split-Methode

Wir beschränken uns hier auf die korrekt spezifizierte Kleinste-Quadrate-Methode und das Lasso-Verfahren, und stellen hauptsächlich einen Vergleich zwischen der moderaten Full-Methode, welche unter $m = 50$ umgesetzt wird und von den alternativen Full-Methdoen bei großen Stichproben am ehesten Konfidenzintervalle mit der gewünschten Überdeckungswahrscheinlichkeit $1-\alpha$ und der geeigneten Länge $2 \cdot F^{-1}(1-\alpha)$ hervorbringt, und der Split-Methode an.

Grafische Gegenüberstellungen der beiden Methoden sind in den Abbildungen 16 und 17 sowie in den Abbildung 29 bis 32 aus Anhang A für die korrekt spezifizierte Kleinste-Quadrate-Schätzung zu finden. Bei kleinen Stichproben stellen wir demnach einen deutlichen Unterschied fest: sowohl die gemittelte empirische Überdeckungsrate als auch die gemittelte Intervallslänge sind bei der Split-Methode größer als bei der betrachteten Full-Methode, d.h. bei kleinen Stichproben bringt die Split-Methode vergleichsweise konservative Konfidenzintervalle hervor. Dies ist auch nicht verwunderlich, wenn man be-

denkt, dass zur Schätzung der Koeffizienten des linearen Modells (5.1) durch die Kleinste-Quadrate-Methode der Split-Methode weniger Daten zur Verfügung stehen als der Full-Methode und darüber hinaus bei insgesamt kleinen Stichprobengrößen der Schätzfehler noch relativ hoch ist. Erst mit wachsender Stichprobe nimmt der Schätzfehler sowohl für die Full- als auch für die Split-Methode entsprechend ab und bei hohen Stichprobengrößen fällt der Umstand, dass der Split-Methode nur einen Teil der Stichprobe für die Schätzung der Koeffizienten zur Verfügung steht, kaum noch ins Gewicht. Die gemittelte empirische Überdeckungsrate der durch die Split-Methode hervorgebrachten Konfidenzintervalle und die gemittelte empirische Überdeckungsrate der Konfidenzintervalle, die durch die betrachtete Full-Methode erzeugt werden, scheinen sich mit wachsendem Stichprobenumfang nach und nach anzugleichen. Ebenso beobachten wir mit größeren Stichproben eine Angleichung der durch die Split-Methode bestimmten gemittelten Intervallslänge zu der mit der Full-Methode ermittelten Länge. Die Breite der Split-Intervalle unterliegen dabei höheren Schwankungen, was womöglich auf die jedesmalig erfolgte Extraktion einer Trainingsmenge aus der Stichprobe, welche in unserem Modell zufällig erfolgt, zurückzuführen ist; siehe auch G'Sell et al. (2016, S.11). Im Großen und Ganzen suggerieren diese Ergebnisse, dass bei großen Stichproben die Split-Methode den Full-Methoden vorzuziehen ist, da der im Vergleich zur moderaten Full-Methode höhere Schätzfehler bei größeren Stichproben vernachlässigbar ist (und die Split-Methode damit nicht schlechtere Intervalle hervorbringt) und dabei je nach Anzahl an Netzpunkten, die die Full-Methoden zu verarbeiten haben, die Split-Methode weniger Rechenaufwand benötigt.[15] Dies schlägt sich auch in einer deutlich kürzeren mittleren Laufzeit pro Iteration, die wir unter den von uns festgesetzten Bedingungen und mit der uns zur Verfügung stehenden Rechenleistung für die Split-Methode erhalten, gegenüber jener mittleren Laufzeit pro Iteration, die wir mit der selben Rechenleistung für die (gemeinsame) Ausführung der drei Full-Methoden einmal unter $m = 50$ und einmal unter $m = 25$ erhalten, nieder (siehe Tabelle 5 und Abbildung 33 aus Anhang A). Bei den unter $m = 50$ ausgeführten Full-Methoden fällt die mittlere Laufzeit näherungsweise um den Faktor 2.017 höher aus als bei der Split-Methode, während der Faktor bei den unter $m = 25$ ausgeführten Full-Methoden näherungsweise 1.449 beträgt.

Die Split-Methode und die moderate Full-Methode werden für die Lasso-Schätzung in den Abbildungen 18 und 19 sowie in den Abbildungen 34 bis 37 aus Anhang A grafisch verglichen. Bezüglich der gemittelten empirischen Überdeckungsrate zeigt sich uns ein zur korrekten Spezifikation analoges Muster:

die Überdeckungsrate, welche uns als Geschöpf der Split-Methode gegenübertritt, ist bei kleinen Stichproben noch deutlich größer als das entsprechende von der moderaten Full-Methode gelieferte Gegenstück. Diese nähern sich dann mit zunehmender Stichprobengröße deutlich an. Der Unterschied zwischen der Split-Methode und der moderaten Full-Methode ist also wie bei der korrekten Spezifikation im Bereich kleiner Stichproben noch erheblich, verschwindet mit wachsender Stichprobe jedoch zusehends. Ähnlich verhält sich die Situation, wenn wir einen Blick auf die gemittelte Intervallslänge wagen. Wie bei der Überdeckungsrate ist auch hier im Bereich kleiner Stichprobengrößen die durch die Split-Methode geschaffenen Konfidenzintervalle länger als bei der moderaten Methode, die Differenz in den Intervallslängen verschwindet jedoch mit wachsender Stichprobe. Nun könnte man angesichts dieser Resultate meinen, dass bei einer Lasso-Schätzung die Unterschiede zwischen den beiden Methoden ähnlich gelagert seien wie bei der korrekt spezifizierten Kleinste-Quadrate-Methode, wäre da nicht die auf das Lasso-Verfahren aufbauenden Full-Methode mit einem Fluch belegt, die im Zusammenhang mit dem Rechenaufwand einen schärferen Kontrast zwischen der Full- und der Split-Methode bewirkt. Wir sehen in Tabelle 5, dass die (gemeinsame) Ausführung der Full-Methoden mit der durch die R-Funktion `glmnet` umgesetzten Lasso-Methode unter Ausnutzung der uns verfügbaren Rechenkapazität nicht nur eine höhere mittlere Laufzeit erfordert als es die Split-Methode zu den gleichen Bedingungen tut, sondern dass dieser Unterschied viel gravierender ausfällt als es bei Anwendung der korrekt spezifizierten Schätzung der Fall ist. Die mittlere Laufzeit ist bei den unter $m = 50$ ausgeführten Full-Methoden näherungsweise um den Faktor 27.668 höher als bei der Split-Methode. Es wird zwar im Bezug auf die Laufzeit die moderate Methode von uns nicht isoliert sondern in Verbund mit den anderen beiden Full-Methoden betrachtet, der Großteil des Rechenaufwandes ist jedoch dem Trainieren einer Regressionsfunktion, welches sich der in Unterkapitel 5.2 vorgestellten Netzkonstruktion bedient, geschuldet (was also auch bei alleiniger Ausführung der moderaten Methode eine entsprechend hohe Rechenzeit bedeuten würde). Und im Falle der Lasso-Schätzung ist die Formierung der Regressionsfunktion noch zusätzlich mit einem numerisch komplexeren Minimierungsproblem verbunden. Ohne an dieser Stelle eine vollständige Erklärung für den Unterschied in den Laufzeiten liefern zu wollen, weisen wir darauf hin, dass gerade das in der Lasso-Schätzung integrierte numerische Verfahren zur Lösung des Minimierungsproblems und die algorithmische Implementierung durch `glmnet` für den Großteil der gegenüber der Split-Methode hinzukommenden Zeitaufwand verantwortlich sein könnte.

Kleinste-Quadrate-Methode mit korrekter Spezifikation		
	Mittlere Laufzeit (in Sekunden)	Relative mittlere Laufzeit
Full-Methoden $m = 50$	6.125 (0.252)	2.017
Full-Methoden $m = 25$	4.400 (0.280)	1.449
Split-Methode	3.036 (0.342)	1
Lasso-Methode		
	Mittlere Laufzeit (in Sekunden)	Relative mittlere Laufzeit
Full-Methoden $m = 50$	304.628 (9.111)	27.668
Split-Methode	11.010 (0.312)	1

Tabelle 5: *Vergleich der über alle* 400 *Iterationsläufe gemittelten Laufzeiten für die Split-Methode sowie den Full-Methoden (bei gemeinsamer Ausführung der moderaten, der hyperkonservativen und der hypokonservativen Methode) unter jeweiliger Anwendung der korrekt spezifizierte Kleinste-Quadrate-Methode und der Lasso-Methode. Dabei beziehen sich die Werte in der zweiten Spalte auf den Mittelwert der in Sekunden gemessenen absoluten Laufzeit, wobei die dazugehörigen Stichprobenstandardabweichungen in Klammern gesetzt sind, und die Werte in der dritten Spalte repräsentieren jeweils für beide Schätzverfahren das Verhältnis der gemittelten Laufzeiten zu jener der Split-Methode, welche auf* 1 *normiert ist.*

Ein die Lasso-Methode ausführender konformaler Algorithmus muss nämlich im Fall der Split-Methode bedingt durch die Datenseparation das Minimierungsproblem nur einmal lösen und anschließend zur Berechnung der Residuen lediglich die geschätzten Koeffizienten, die nicht mit 0 gleichgesetzt werden, speichern. Im Gegensatz dazu muss bei der Full-Methode für jeden einzelnen Netzpunkt das Minimierungsproblem gelöst werden und anschließend die von 0 verschiedenen Koeffizienten gespeichert werden; vgl. auch G'Sell et al. (2016, S.9-10). Allein schon dieser Umstand führt zu deutlich höheren Rechenzeiten. Ein weiterer Faktor für den im Verhältnis zur korrekt spezifizierten Kleinste-Quadrate-Methode schärferen Kontrast in den Laufzeiten ist, dass wir unter der Lasso-Schätzung zur Umsetzung der Full-Methoden eine *Zählschleife* verwenden, während unter der Kleinste-Quadrate-Schätzung die Ausführung der Full-Methoden ausschließlich auf *Vektorarithmetik* basiert (siehe auch die im Anhang B präsentierten Quellcodes).

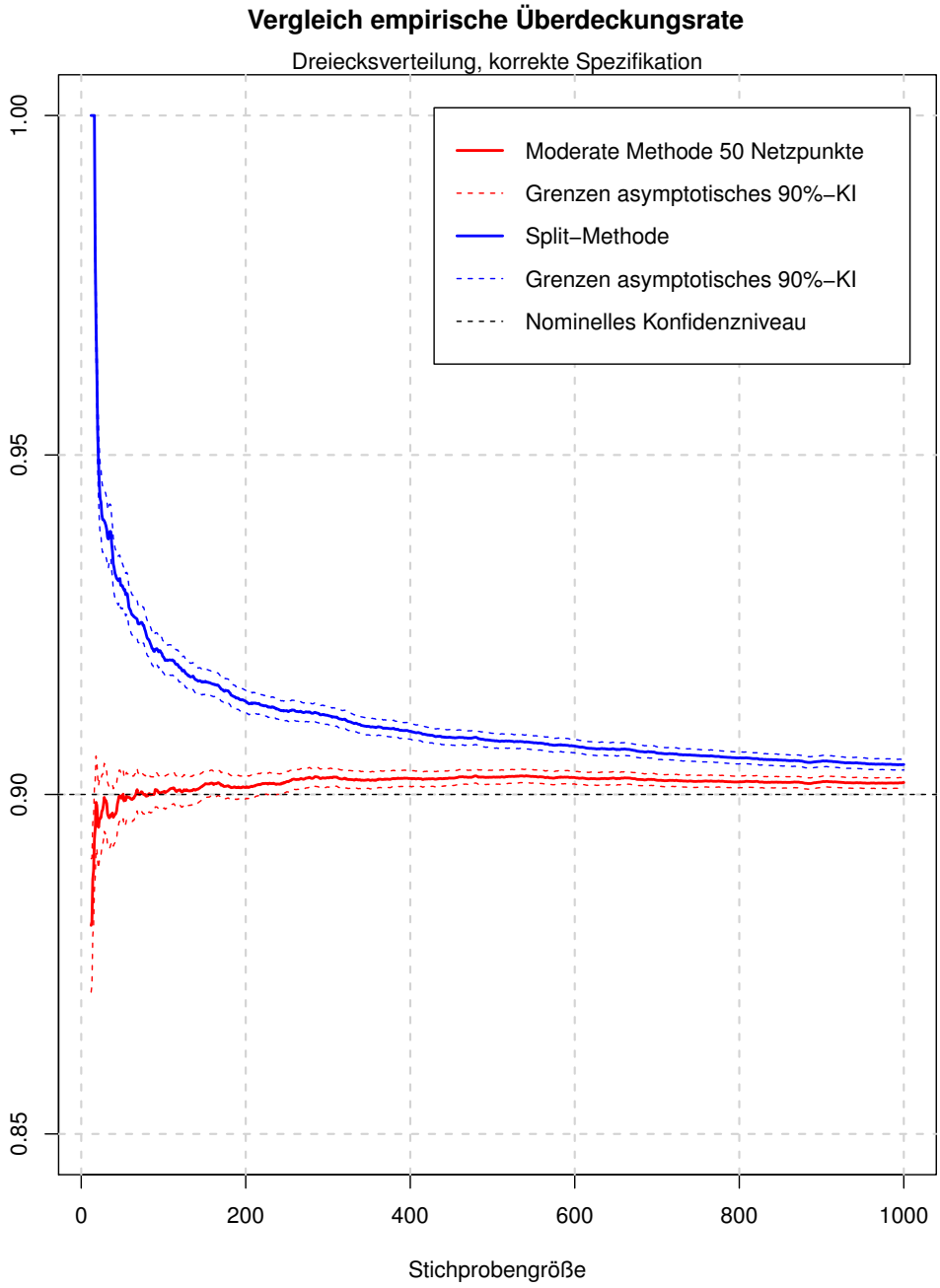

Abbildung 16: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate sowie der Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, mit wachsender Stichprobengröße, wobei diese von* p *bis* n *reicht.*

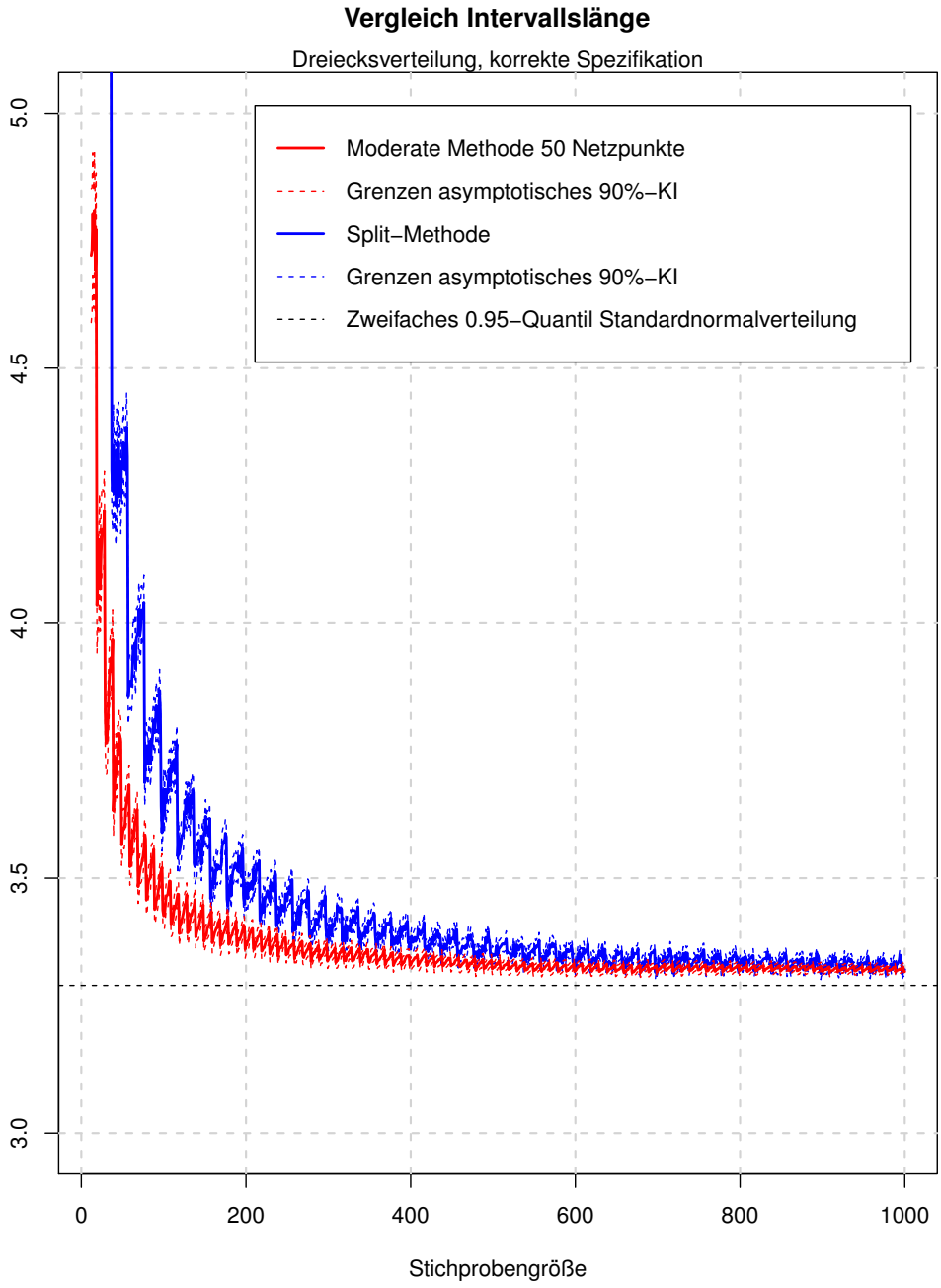

Abbildung 17: *Vergleich zur Entwicklung der gemittelten Intervallslänge sowie der Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, mit wachsender Stichprobengröße, wobei diese von* p *bis* n *reicht.*

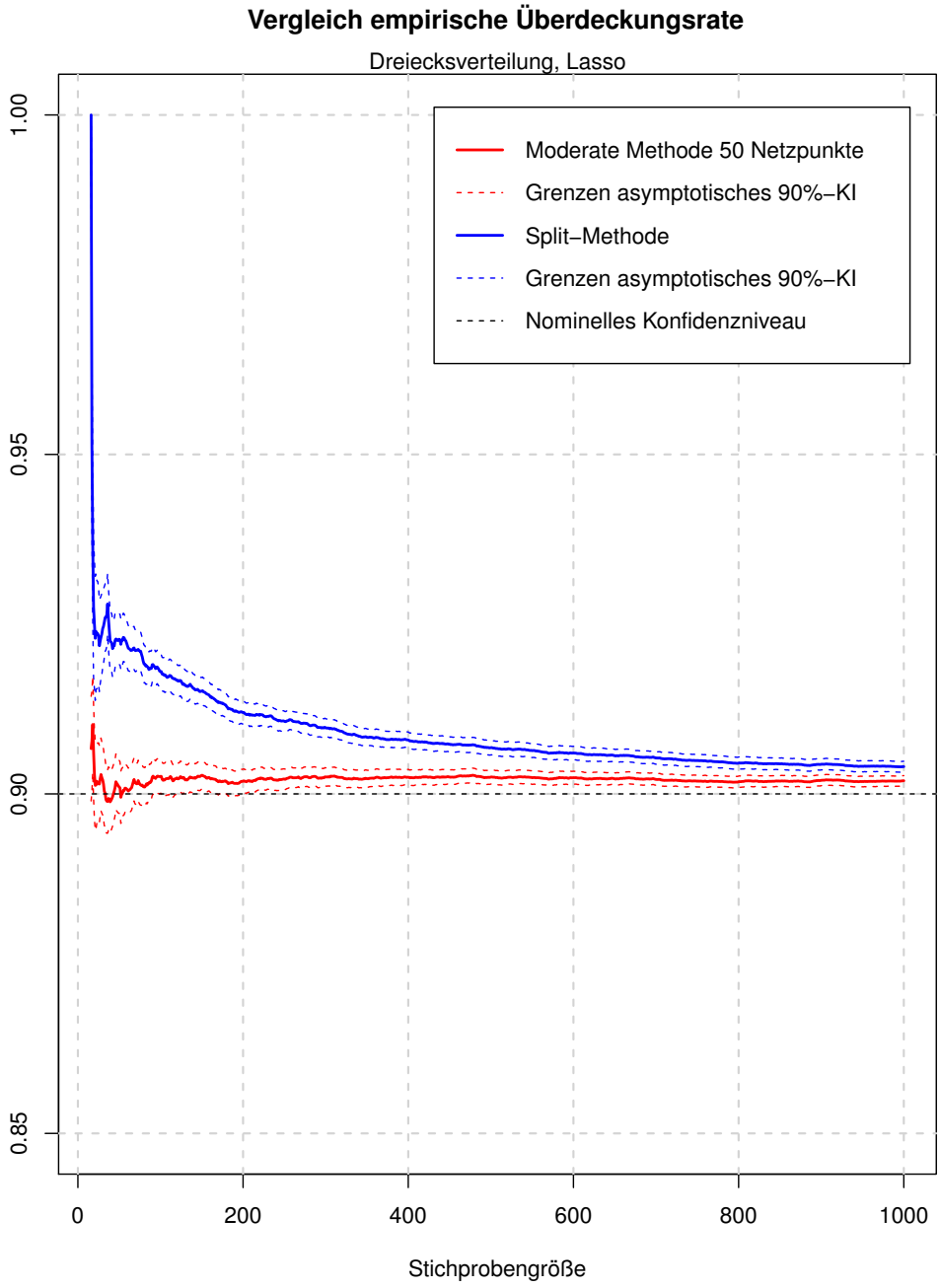

Abbildung 18: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate sowie der Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden, mit wachsender Stichprobengröße, wobei diese von* p *bis* n *reicht.*

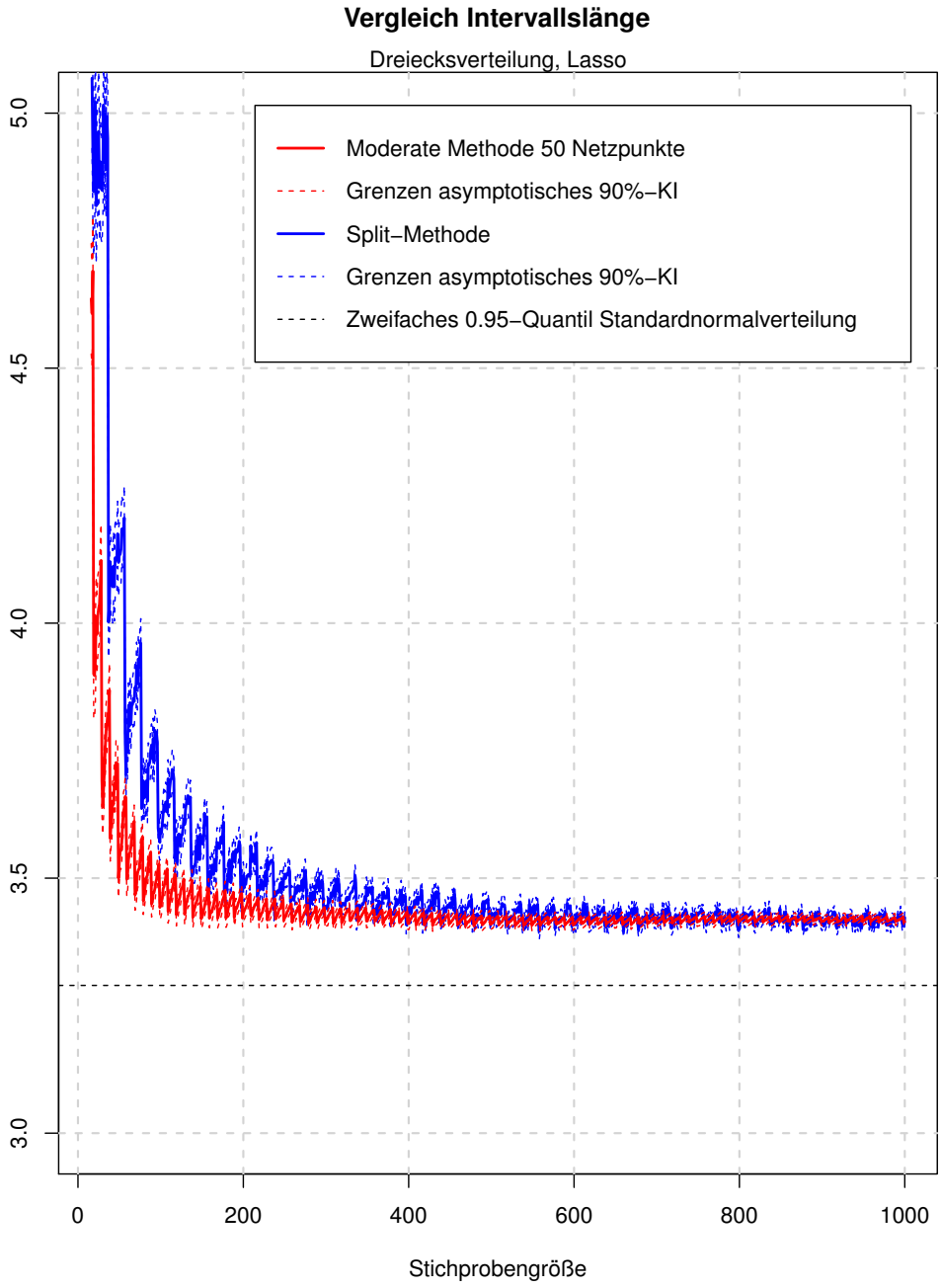

Abbildung 19: *Vergleich zur Entwicklung der gemittelten Intervallslänge sowie der Grenzen der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden, mit wachsender Stichprobengröße, wobei diese von* p *bis* n *reicht.*

5.6 Robustheit der Methoden bei Auftreten von Extremwerten

Wir ziehen noch einmal das lineare Modell (5.1) in Betracht mit dem Unterschied, dass die Regressoren diesmal *standardcauchyverteilt* sind mit der Wahrscheinlichkeitsdichte

$$f_C(x) := \frac{1}{\pi(1+x^2)}$$

für alle $x \in \mathbb{R}$. Diese Verteilung zeichnet sich gegenüber Verteilungen wie der Standardnormalverteilung oder der Dreiecksverteilung, die wir in den Unterkapiteln davor verwendet haben und durch einen beschränkten Träger charakterisiert ist, dadurch aus, dass sie aufgrund relativ „schwerer“ Tails dazu neigt *Extremwerte* (also gegenüber der Masse der von ihr erzeugten Daten sehr große und sehr kleine Zahlenwerte) hervorzubringen. Wir wiederholen die Simulation mit dieser Verteilung, wobei wir ansonsten wieder die Spezifikation aus Tabelle 1 verwenden. Für die Full-Methoden legen wir den Grid-Faktor wieder durch $g = 0.25$ fest und die Anzahl der Netzpunkte sei $m = 50$. Für die Split-Methode wählen wir wieder $\delta = 0.5$, wonach die Separation der Stichprobe zu erfolgen hat. Zur Schätzung der Koeffizienten aus Modell (5.1) verwenden wir diesmal ausschließlich die korrekt spezifizierte Kleinste-Quadrate-Methode. Wir fassen im Folgenden die von uns beobachteten Auswirkungen von Extremwerten, welche die Regressoren annehmen und sich über Modell (5.1) auf den Regressanden „fortpflanzen“, sowohl auf die Full-Methoden wie auch auf die Split-Methode kurz zusammen.

Die gemittelten Resultate, die wir im Zuge der Anwendung der drei Full-Methoden beobachten, sind für ausgewählte Stichprobengrößen in Tabelle 6 zusammengefasst. Die mit wachsender Stichprobengröße sich vollziehende Entwicklung der gemittelten Intervallslänge ist exemplarisch für die moderate Full-Methode in Abbildung 21 veranschaulicht. Hier sehen wir einen deutlichen Unterschied zum entsprechenden in Unterkapitel 5.3 präsentierten Bild. Wir beobachten unregelmäßige und zum Teil ausgeprägte Schwankungen, wobei sich keine klare Konvergenz zu $2 \cdot F^{-1}(1-\alpha)$, also das Zweifache des 0.95-Quantils der Standardnormalverteilung (welches der Länge des „optimale“ Konfidenzintervalls aus Unterkapitel 4.1 entspricht), ergibt. Das Verhalten der gemittelten empirischen Überdeckungsrate ist in Abbildung 20 illustriert und steht ebenfalls in einem scharfen Kontrast zu jenem, das wir in Unterkapitel 5.3 beobachtet haben. Anstatt dass es mit größer werdendem Stichprobenumfang zu einer Stabilisierung der gemittelten empirischen Überdeckungsrate in einer

engeren oder breiteren Umgebung des nominellen Konfidenzniveaus kommt, beobachten wir für alle drei Methoden bei größer werdenden Stichproben abnehmende Werte. Dies ist besonders ausgeprägt bei der hypokonservativen Methode, die im Vergleich zur hyperkonservativen und moderaten Methode besonders kleine Werte aufweist. Die gemittelten Überdeckungsraten, die von der hyperkonservativen und moderaten Methode hervorgebracht werden, laufen dagegen relativ synchron. Mit fortlaufendem Wachstum der Stichprobengröße scheint sich außerdem die Abnahme der gemittelten empirischen Überdeckungsrate bei allen drei Methoden zu verlangsamen. Um ein Verständnis für die möglichen Gründe des Verhaltens der empirischen Überdeckungsrate zu erlangen betrachten wir exemplarisch die Entwicklung der Intervallslänge, die wir für alle drei Methoden im Iterationschritt 100 beobachtet haben. Werte der zu diesem Iterationslauf realisierten Intervallslängen finden sich für ausgewählte Stichprobengrößen in Tabelle 7. Die Intervallslängen seien außerdem durch die Bilder 22 bis 24 veranschaulicht. Dort sind darüber hinaus die im Verlauf des Iterationsschritts realisierten Werte des Regressanden als Folge diskreter Punkte eingetragen sowie die Grenzen des „optimale“ Konfidenzintervalls, welche mit den entsprechenden Quantilen der Standardnormalverteilung zusammenfallen und innerhalb derer sich der Großteil der realisierten Werte des Regressanden befindet. Bei keiner der drei Methoden läßt sich ein klares Konvergenzverhalten der Intervallslängen zu $2 \cdot F^{-1}(1-\alpha)$, also der Länge des „optimale“ Konfidenzbereichs, beobachten. Vielmehr sehen wir Schwankungen der Konfidnezintervalle um den „optimale“ Konfidenzbereich, die sich scheinbar unsystematisch vollziehen. Wir beobachten auch, dass sich bei allen drei Methoden etliche Konfidenzintervalle mit der Länge 0 realisiert haben. Für eine große Zahl an Stichprobengrößen i zwischen den Werten $d+1$ und n nimmt das resultierende Intervall $C_{i,\alpha}^{full,\mathscr{Y}_m}(X_{i+1}^{100})$ für alle drei Full-Methoden die Form (5.2) an, d.h. es sind endliche Netze $\mathscr{Y}_m$ erzeugt worden, die keine Elemente aus dem konformalen Konfidenzbereich $C_{i,\alpha}^{full}(X_{i+1}^{100})$ enthalten (siehe auch Unterkapitel 5.2). Das trifft zum Beispiel auf die Stichprobengröße 1000 zu (siehe Tabelle 7). Dies hat zur Folge, dass zu vielen Stichprobenumfängen Konfidenzintervalle erzeugt werden, die die realisierten Werte des Regressanden nicht überdecken. Außerdem gibt es noch einige Fälle, in denen lediglich ein einziger Netzpunkt in $C_{i,\alpha}^{full}(X_{i+1}^{100})$ enthalten ist, was dazu führt, dass nur die hypokonservative Methode Intervalle der Länge 0 erzeugt (und deswegen gegenüber den anderen beiden Full-Methoden eine besonders kleine Überdeckungsrate aufweist). Auffallend ist noch, dass sich bei allen drei Methoden zur Stichprobengröße 845 jeweils ein sehr breites Intervall realisiert hat (siehe

Tabelle 7). Dessen Länge weist darüber hinaus in allen drei Fällen den gleichen Wert auf, womit wir es hier mit einer Situation zu tun haben, in der alle Netzpunkte aus $\mathscr{Y}_m$ in $C_{i,\alpha}^{full}(X_{i+1}^{100})$ liegen (siehe Unterkapitel 5.2). Offensichtlich scheinen also die auf endliche Netze basierenden Full-Methoden nicht besonders robust gegenüber in den Daten auftretenden Extremwerten zu sein. Wir wollen in dieser Arbeit nicht weiter den Versuch wagen, all die hier beschriebenen Phänomene einer vollständigen Erklärung zuzuführen, und verzichten auf eine vertiefende Untersuchung. Wir weisen lediglich darauf hin, dass für jede Stichprobengröße i die Bestimmung des endlichen Netzes $\mathscr{Y}_m$ von der Differenz $Y_{(i:i)} - Y_{(1:i)}$ zwischen der größten und der kleinsten Ordnungsstatistik unter den realisierten Werten des Regressanden abhängt. Diese auch als *Spannweite* bezeichnete Größe ist sehr sensitiv gegenüber Extremwerten.

Die aus der Split-Methode gewonnenen Resultate sind für ausgewählte Stichprobengrößen ebenfalls der Tabelle 6 zu entnehmen und sind in den Abbildungen 25 und 26 dargestellt. Außerdem bieten die Säulendiagramme 27 und 28 hinsichtlich der gemittelten empirischen Überdeckungsrate bzw. der gemittelten Intervallslänge einen Vergleich zwischen der Split-Methode und der moderaten Full-Methode mit $m = 50$ zu ausgewählten Stichprobengrößen. Bei der Split-Methode zeigt sowohl die gemittelte empirische Überdeckungsrate als auch die gemittelte Intervallslänge ein Verhalten, wie wir es schon in Unterkapitel 5.4 für die Dreiecksverteilung beschrieben haben. Lediglich bei kleinen Stichprobengrößen weist die gemittelte Intervalslänge mit den dazugehörigen Standardabweichungen noch vergleichsweise große Werte auf bevor es bei größeren Stichproben zu einer Stabilisierung in der Umgebung von $2 \cdot F^{-1}(1-\alpha)$ kommt. Die Resultate für die Split-Methode sind auch nicht von übermäßig großer Überraschung, wenn man berücksichtigt, dass wir in den Sätzen zur Asymptotik der Split-Methode aus dem Unterkapitel 4.3 keine konkrete Verteilung P gefordert haben. Es müssen nur die entsprechenden Stetigkeits- und Monotonieeigenschaften für die Verteilungsfunktion F sowie das Postulat, dass bezüglich P fast sicher keine Bindungen unter den Residuen vorkommen, erfüllt sein. Im Gegensatz zu den „approximierenden“ Full-Methoden ist also die Split-Methode gegenüber die von den Regressoren bzw. den Regressanden angenommenen Extremwerten relativ robust und ist somit auch in diesem Fall zur Erzeugung konformaler Prognosebereiche geeignet.

i	20	100	500	1000
Moderate Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.834 (0.127)	0.743 (0.212)	0.423 (0.212)	0.280 (0.161)
$I_{i,\alpha}^{full}$	13.781 (71.5)	11.595 (89.29)	3.901 (11.54)	8.312 (51.84)
Hyperkonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.864 (0.118)	0.786 (0.215)	0.445 (0.229)	0.229 (0.172)
$I_{i,\alpha}^{full}$	14.920 (71.92)	13.540 (89.40)	5.560 (15.05)	10.941 (61.85)
Hypokonservative Methode 50 Netzpunkte				
$A_{i,\alpha}^{full,gem}$	0.751 (0.176)	0.416 (0.227)	0.105 (0.078)	0.053 (0.039)
$I_{i,\alpha}^{full}$	12.467 (71.66)	7.562 (89.40)	0.092 (1.840)	0.000 (0.000)
Split-Methode				
$A_{i,\delta,\alpha}^{split,gem}$	0.966 (0.060)	0.919 (0.028)	0.907 (0.01)	0.904 (0.009)
$I_{i,\delta,\alpha}^{split}$	69.947 (308.7)	4.050 (1.059)	3.388 (0.204)	3.346 (0.133)

Tabelle 6: *Werte, welche sich im Zuge der mittels Cauchyverteilung durchgeführten Simulation für die gemittelte empirische Überdeckungsrate und die gemittelte Intervallslänge bestimmt nach den jeweiligen Full-Methoden unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, zu den Stichprobengrößen* 20, 100, 500 *und* 1000 *realisiert haben, wobei die Werte in den Klammern für die Stichprobenstandardabweichungen* $sd(A_{i,\alpha}^{full,gem})$ *und* $sd(I_{i,\alpha}^{full})$ *bzw.* $sd(A_{i,\delta,\alpha}^{split,gem})$ *und* $sd(I_{i,\delta,\alpha}^{split})$ *stehen.*

Intervallslängen für ausgewählte Stichprobengrößen zum Iterationslauf 100			
20	100	845	1000
Moderate Methode 50 Netzpunkte			
3.929	5.106	682.224	0.000
Hyperkonservative Methode 50 Netzpunkte			
4.715	7.385	682.224	0.000
Hypokonservative Methode 50 Netzpunkte			
2.829	2.462	682.224	0.000

Tabelle 7: *Intervallsängen, die im Zuge der mittels Cauchyverteilung durchgeführten Simulation nach den drei Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, unter* $m = 50$ *zu den Stichprobengrößen* 20, 100, 845 *und* 1000 *im Iterationslauf* 100 *bestimmt wurden.*

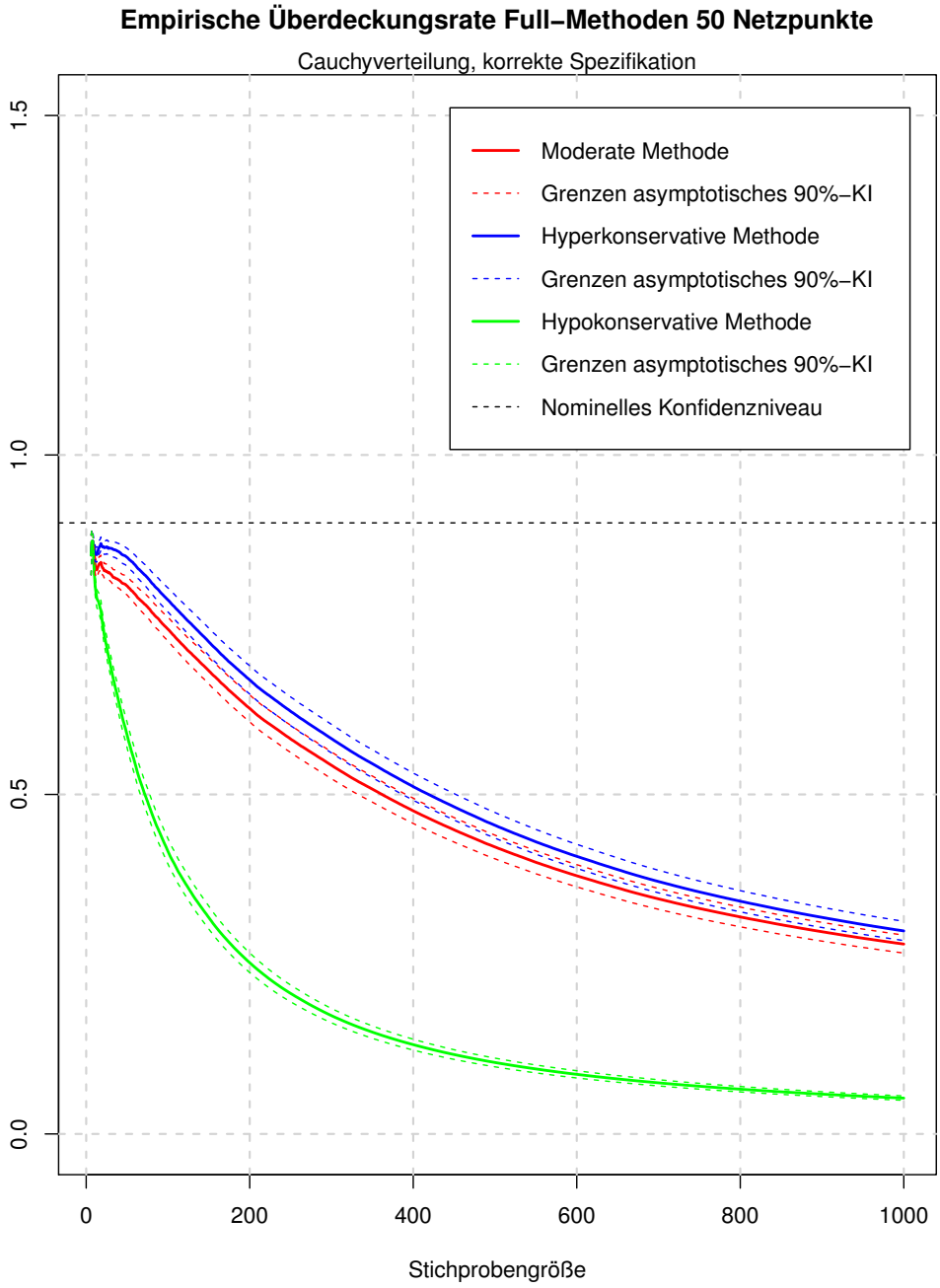

Abbildung 20: *Pfade für die nach den Full-Methoden, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, gemittelten empirischen Überdeckungsraten, welche sich mit wachsender Stichprobengröße, die in dieser Abbildung von $d+1$ bis n reicht, im Zuge der mittels Cauchyverteilung durchgeführten Simulation unter $m = 50$ realisiert haben, sowie die realisierten Grenzen der dazugehörigen asymptotischen $(1-\alpha)$-Konfidenzintervalle.*

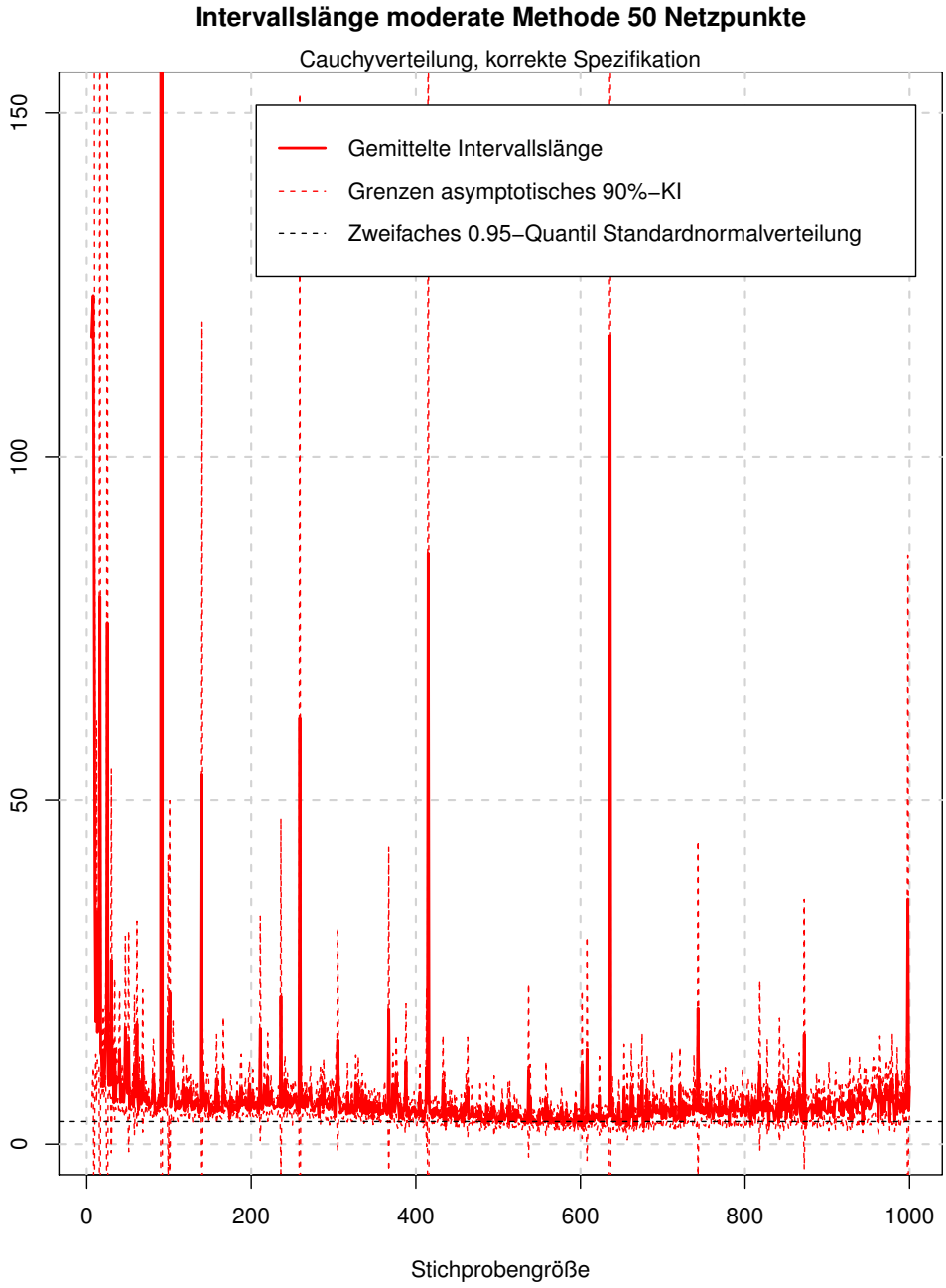

Abbildung 21: *Pfad für die über die moderate Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, unter* $m = 50$ *erzeugte gemittelte Intervallslänge, welcher sich mit wachsender Stichprobengröße, die in dieser Abbildung von* $d + 1$ *bis* n *reicht, im Zuge der mittels Cauchyverteilung durchgeführten Simulation realisiert hat, sowie die realisierten Grenzen der dazugehörigen asymptotischen* $(1 - \alpha)$*-Konfidenzintervalle.*

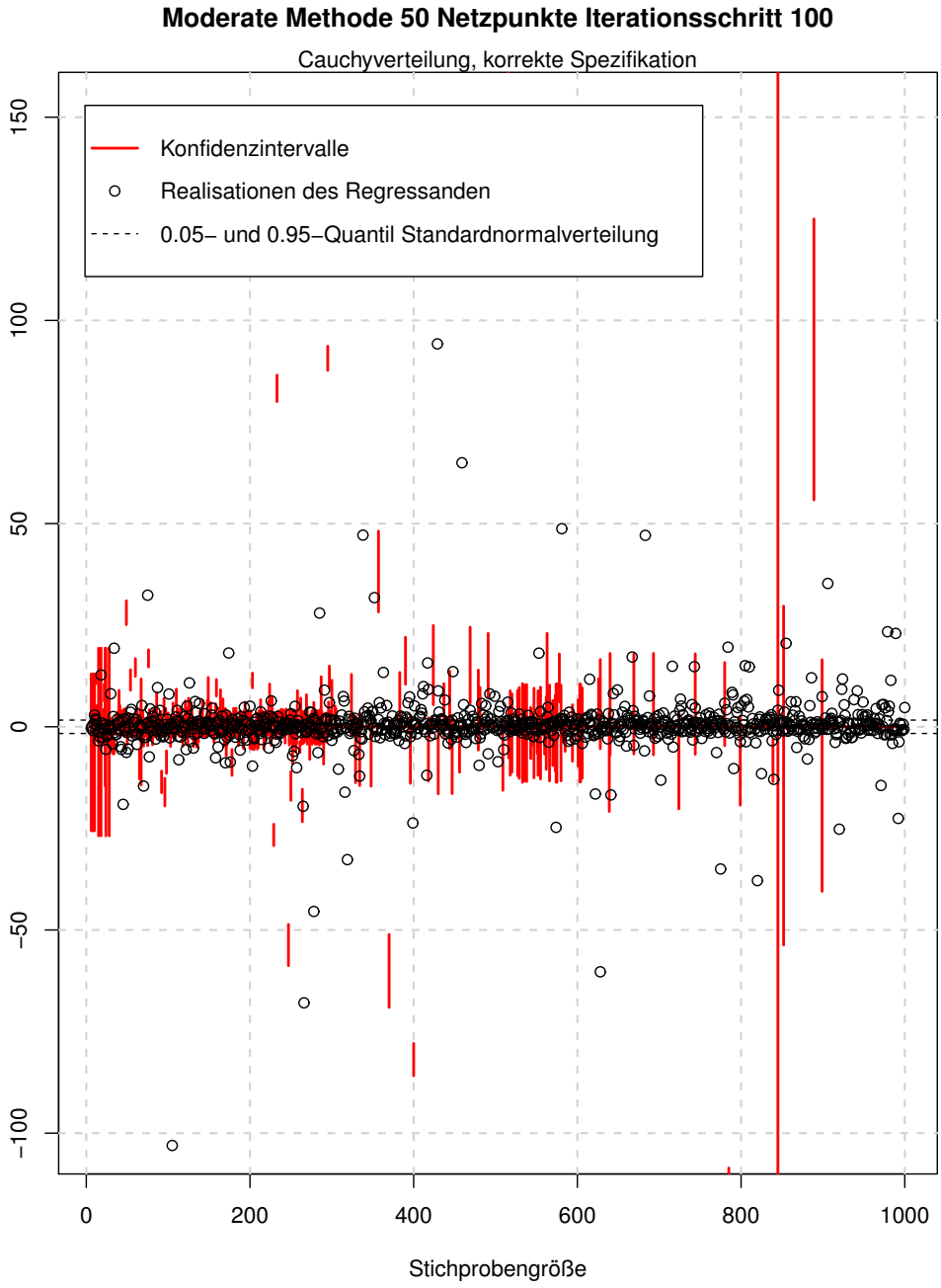

Abbildung 22: *Entwicklung der aus der moderaten Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, unter* $m = 50$ *hervorgegangen Intervallslänge, welche sich im Zuge der mittels Cauchyverteilung durchgeführten Simulation im Iterationslauf* 100 *mit wachsender Stichprobengröße, die in dieser Abbildung von* $d + 1$ *bis* n *reicht, realisiert hat.*

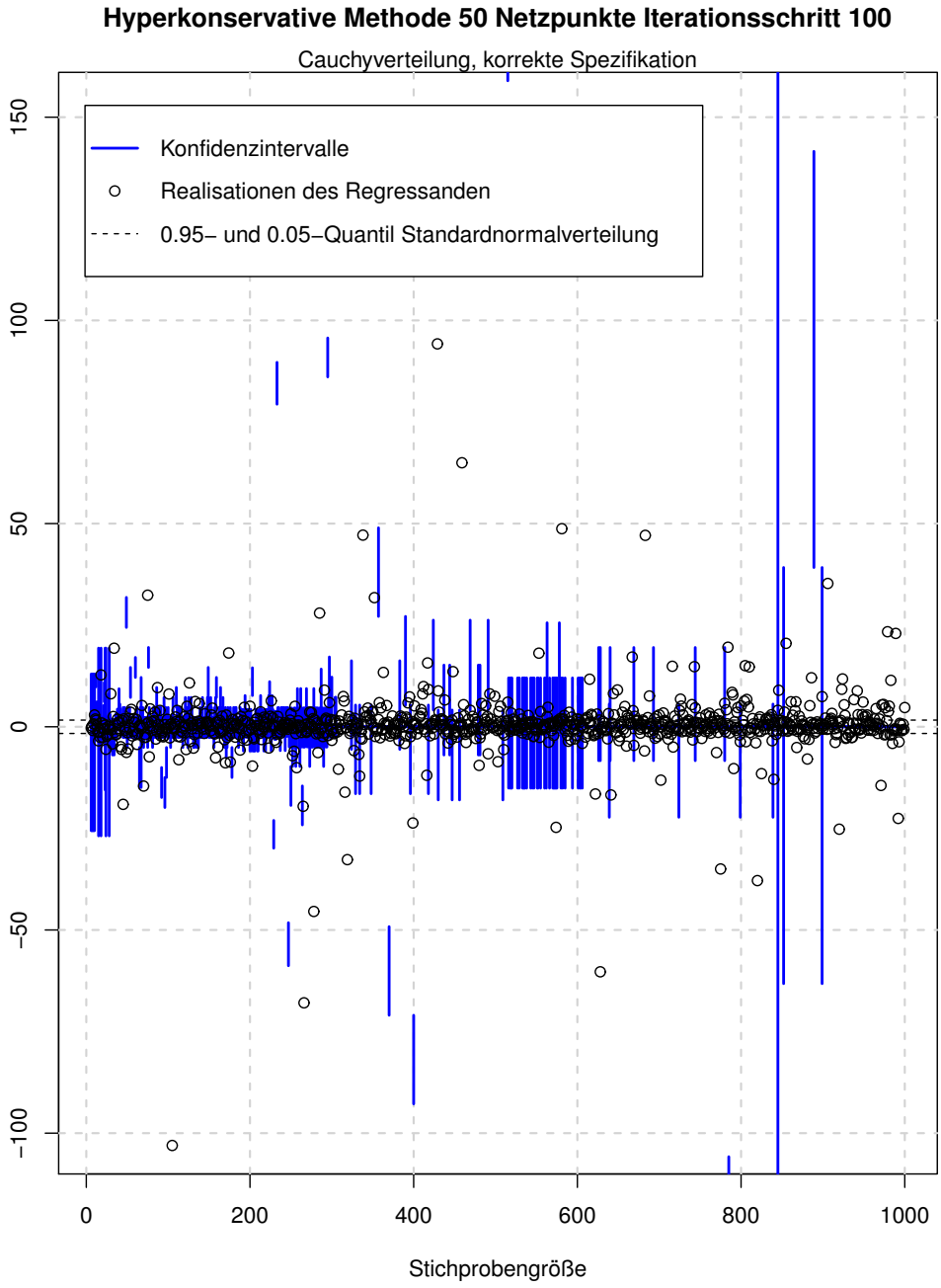

Abbildung 23: *Entwicklung der aus der hyperkonservativen Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, unter* $m = 50$ *hervorgegangen Intervallslänge, welche sich im Zuge der mittels Cauchyverteilung durchgeführten Simulation im Iterationslauf* 100 *mit wachsender Stichprobengröße, die in dieser Abbildung von* $d + 1$ *bis* n *reicht, realisiert hat.*

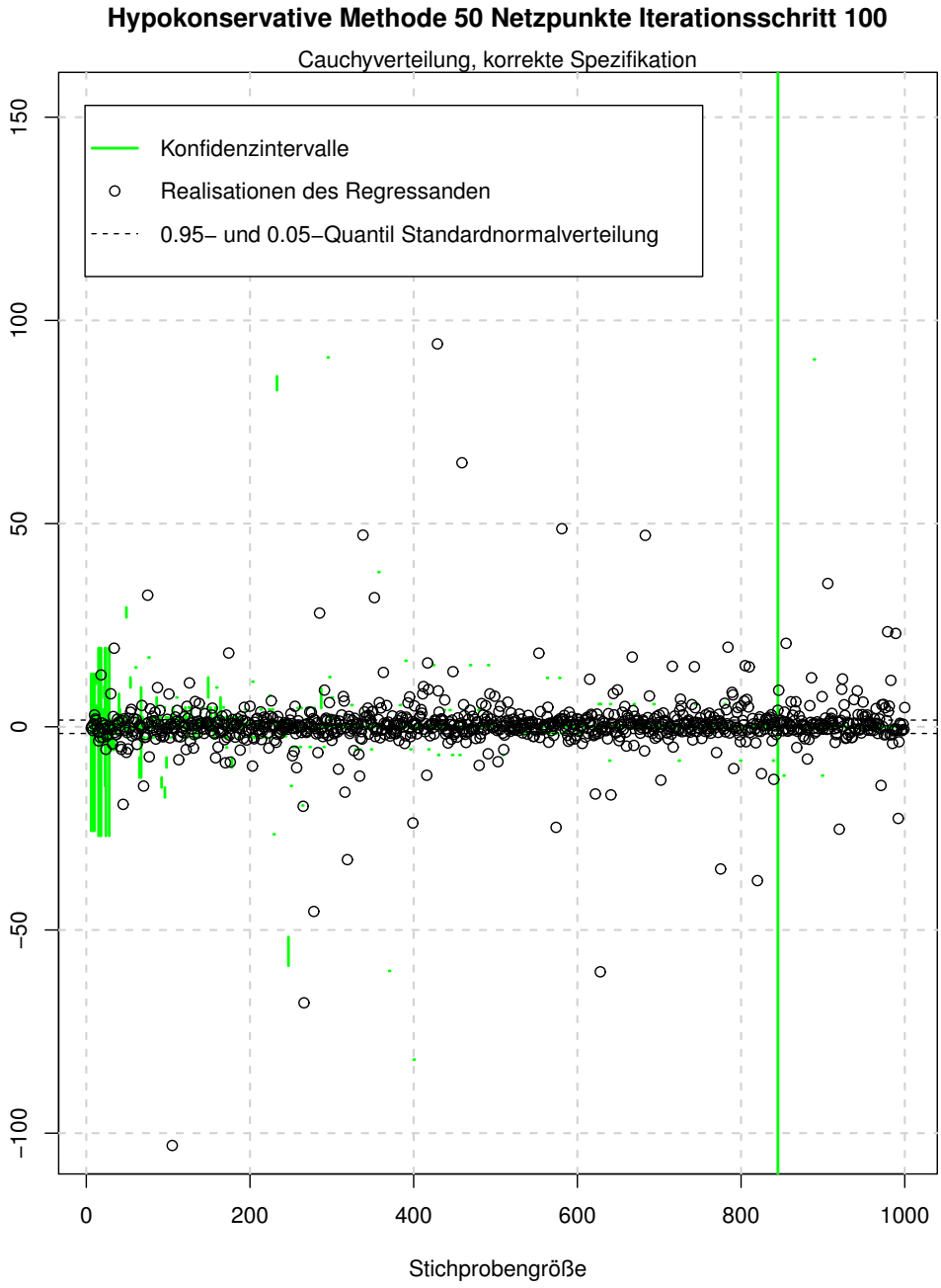

Abbildung 24: *Entwicklung der aus der hypokonservativen Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, unter* $m = 50$ *hervorgegangen Intervallslänge, welcher sich im Zuge der mittels Cauchyverteilung durchgeführten Simulation im Iterationslauf* 100 *mit wachsender Stichprobengröße, die in dieser Abbildung von* $d + 1$ *bis* n *reicht, realisiert hat.*

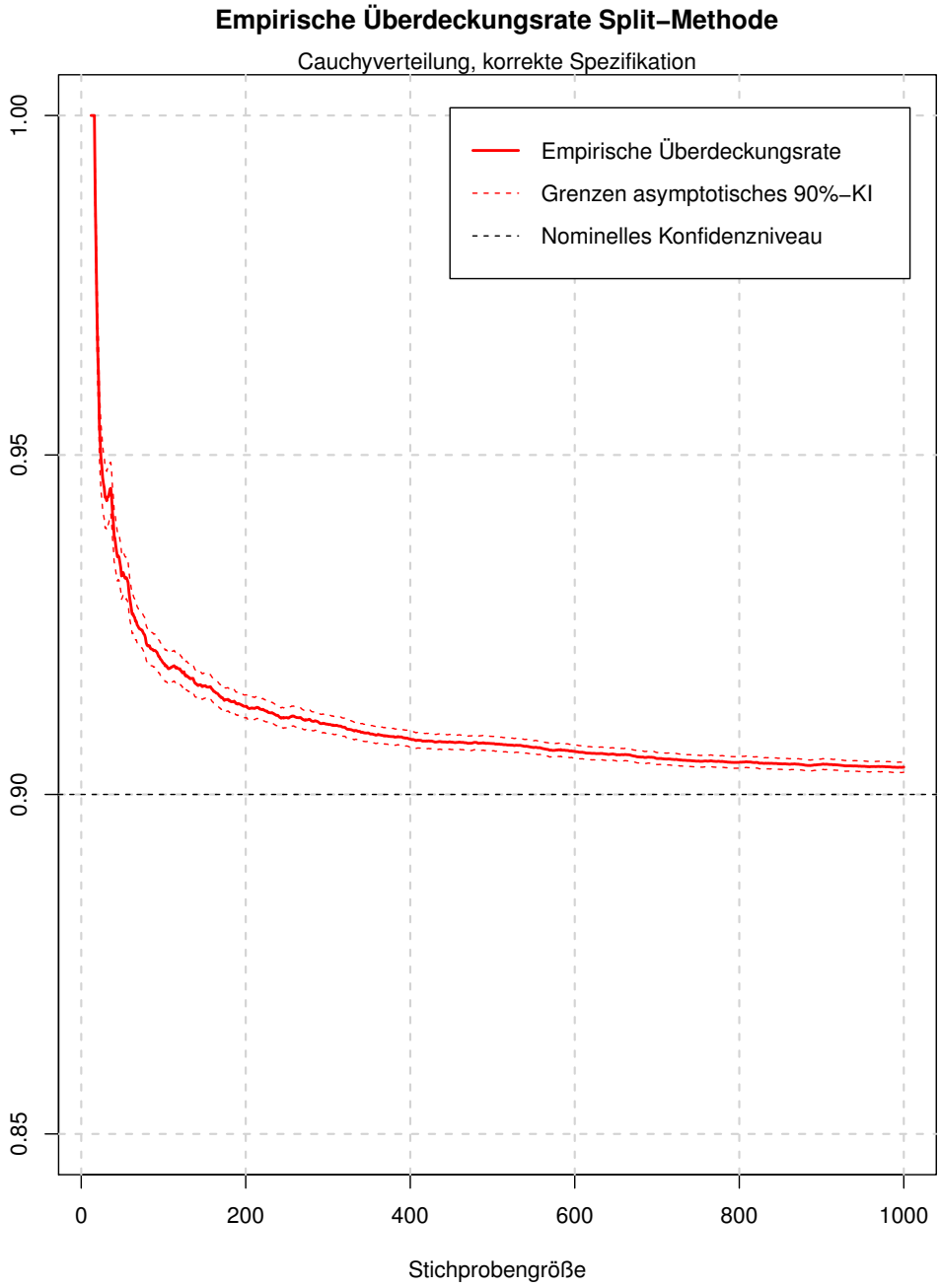

Abbildung 25: *Pfad der durch die Split-Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten empirischen Überdeckungsrate* $A_{i,\delta,\alpha}^{split,gem}$*, welcher sich für wachsende Stichprobengröße, die in dieser Abbildung von* p *bis* n *reicht, im Zuge der mittels Cauchyverteilung durchgeführten Simulation realisiert hat, sowie die realisierten Grenzen dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalls.*

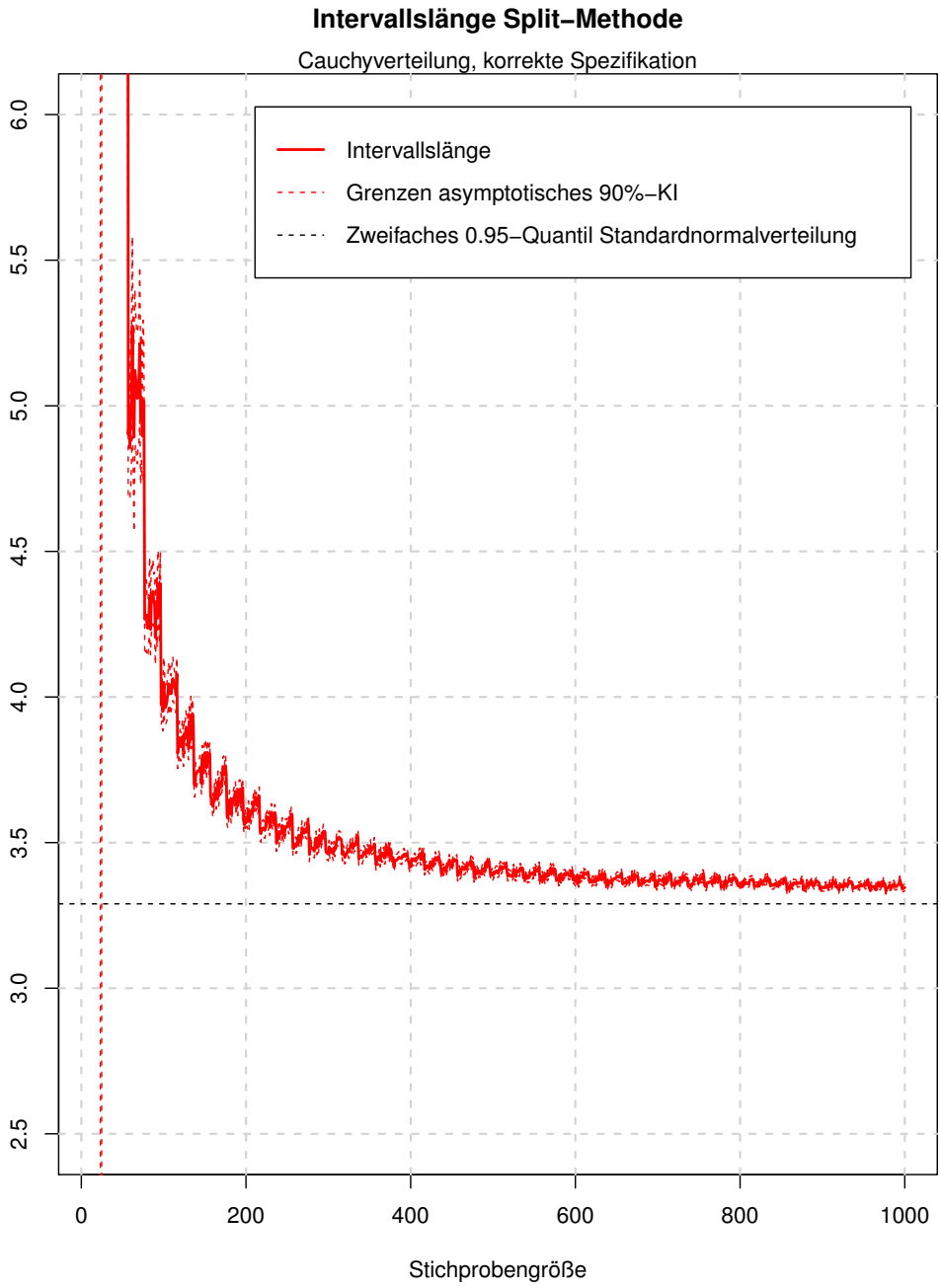

Abbildung 26: *Pfad der durch die Split-Methode, welche die korrekt spezifizierte Kleinste-Quadrate-Methode verwendet, bestimmten gemittelten empirischen Intervallslänge* $I^{split}_{i,\delta,\alpha}$*, welcher sich für wachsende Stichprobengröße, die in dieser Abbildung von* p *bis* n *reicht, im Zuge der mittels Cauchyverteilung durchgeführten Simulation realisiert hat, sowie die Grenzen des dazugehörigen realisierten asymptotischen* $(1-\alpha)$*-Konfidenzintervalls.*

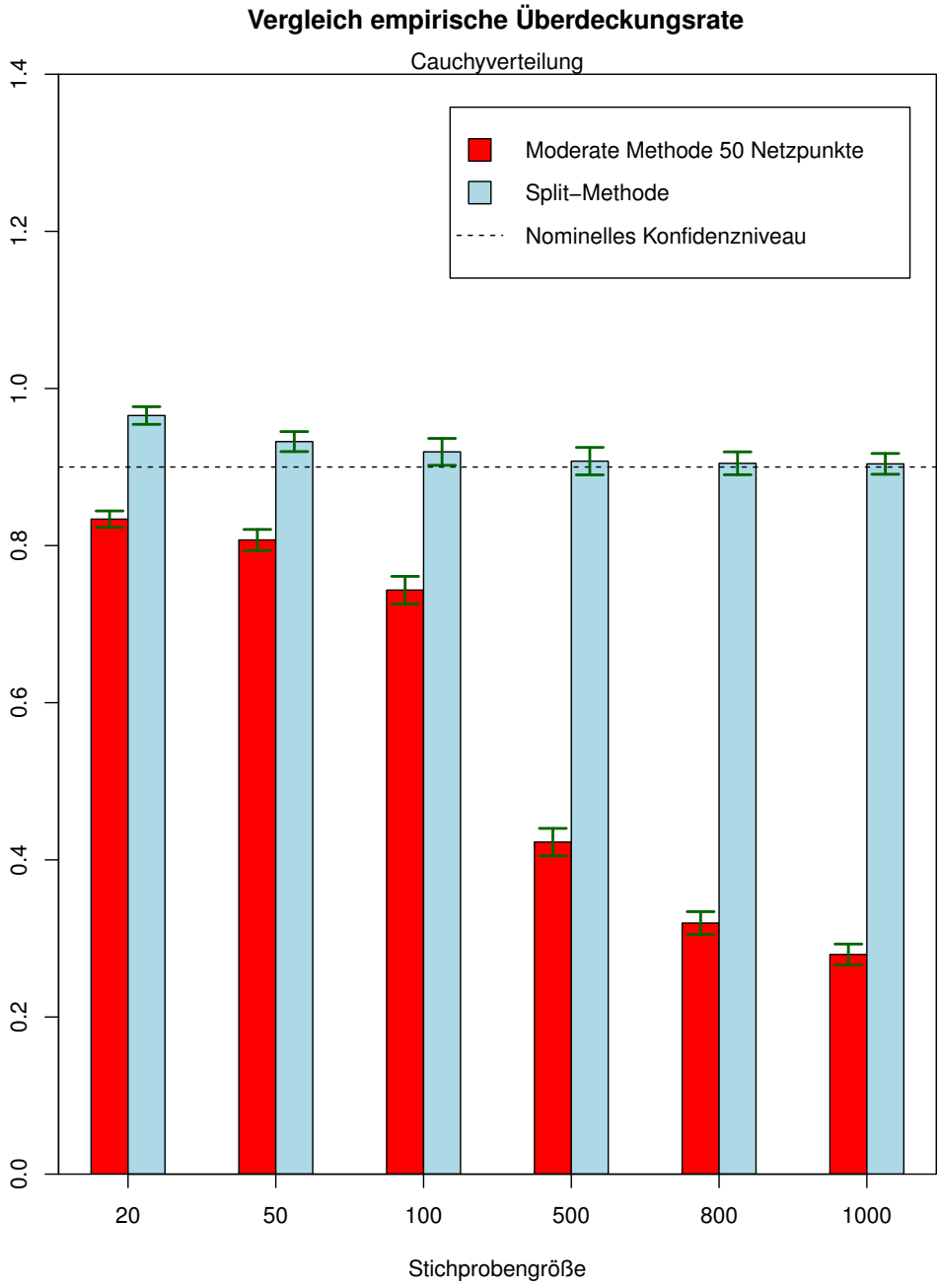

Abbildung 27: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

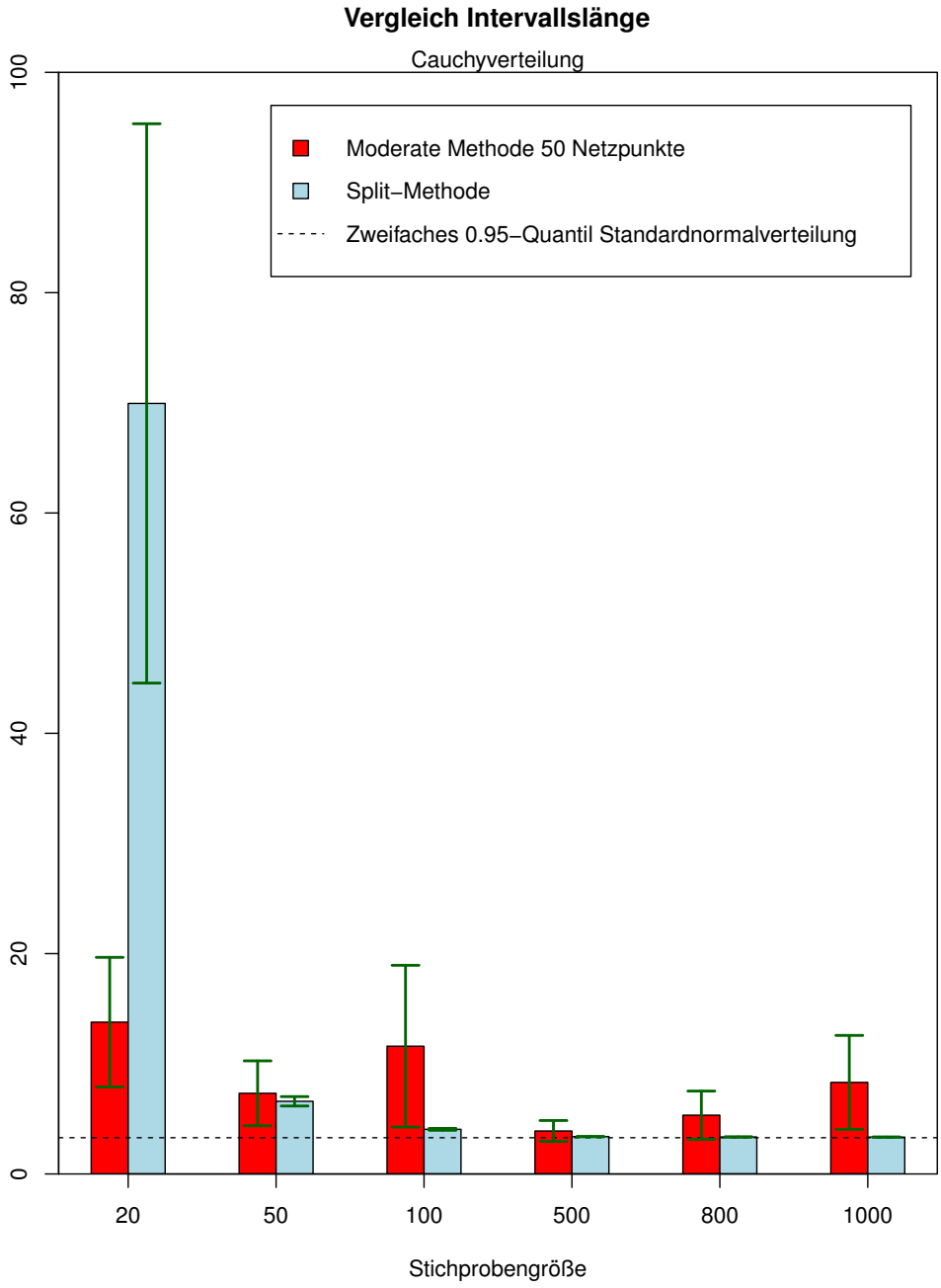

Abbildung 28: *Vergleich zur Entwicklung der gemittelten Intervallslänge sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

6 Konklusion

Wir haben in den Sätzen 3.1 und 3.2 gesehen, dass unter der Bedingung der *Vertauschbarkeit* sowohl die Full- wie auch die Split-Methode bei *endlichen* Stichproben *gültige* konformale Konfidenzbereiche hervorbringen, d.h. Prognosebereiche für die nächstfolgende Beobachtung des Regressanden, deren Überdeckungswahrscheinlichkeit das nominelle Konfidenzniveau nicht unterschreitet und die mit wachsender Stichprobengröße sogar zu diesem konvergiert. Mit der Vertauschbarkeit wird in der Regel weit weniger gefordert als andere gängige Methoden zur Erzeugung (asymptotisch) gültiger Konfidenzbereiche benötigen wie die in Unterkapitel 3.3 diskutierten Verfahren zur Erzeugung parametrischer Konfidenzintervalle oder die nichtparametrische Bootstrap-Methode. Bemerkenswert ist dabei auch, dass wir bei der Full-Methode zur Erlangung dieser Resultate Permutationsinvarianz des zur Konstruktion der Konfidenzbereiche zu verwendenden Algorithmus benötigten, während wir bei der Split-Methode auf diese Forderung verzichten konnten. Unter der stärkeren Annahme, dass die den Daten zugrundeliegenden Zufallsvariablen *unabhängig* und *identisch verteilt* sind, konnten wir außerdem für beide Methoden wünschenswerte asymptotische Eigenschaften nachweisen. Wir haben mit den Sätzen 4.1 und 4.3 gezeigt, dass die *empirische Überdeckungsrate* der Konfidenzbereiche *in Wahrscheinlichkeit* gegen das *nominelle Konfidenzniveau* konvergiert. Gleichzeitig konnten wir mit den Sätzen 4.2 und 4.4 auch sehen, dass die *Breite* der Konfidenzbereiche *in Wahrscheinlichkeit* gegen das Doppelte des entsprechenden *Quantils* jener Verteilung, die für den Absolutbetrag der Störgröße im *Referenzmodell* aus Unterkapitel 4.1 unterstellt ist, konvergiert, wobei von diesem die Grenzen des „optimalen“ Konfidenzintervalls konstituiert werden. Diese Resultate stehen auch in Einklang mit den Resultaten der *Simulation*, die wir unter den im Unterkapitel 5.1 festgelegten Bedingungen durchgeführt haben.

Weil die Split-Methode eine *Separation* des Datensatzes in eine Trainingsmenge und einer Datenmenge zur Berechnung der Residuen vornimmt, kann durch diese gegenüber der Full-Methode je nach Art des verwendeten konformalen Trainingsalgorithmus und Umfang des Wertebereichs des Regressanden eine erhebliche Reduzierung des Rechenaufwandes erzielt werden. Das scheint besonders dann der Fall zu sein, wenn der Trainingsalgorithmus ein numerisch komplexes Verfahren beinhaltet (wie beispielsweise die Lasso-Methode). Der Vorteil reduzierten Rechhenaufwands wird unter Umständen bei kleineren Stichproben mit einem höheren Fehler in der Schätzung des zugrundeliegenden

Referenzmodells erkauft. Die praktische Ausführung der Full-Methode hat uns außerdem vor einem zusätzlichen Problem gestellt. Im Gegensatz zur Split-Methode können wir bei nicht endlichem Wertebereich des Regressanden den Konfidenzbereich nicht direkt berechnen, sondern müssen einen Umweg gehen durch die Anwendung *endlicher Netze*, die wir aus den realisierten Daten des Regressanden erzeugen. Je nachdem, wie viele Netzpunkte erzeugt werden, erhalten wir eine mehr oder weniger genaue „Approximation“ des Konfidenzbereichs, was zu zusätzlichen Verzerrungen führen kann. Insbesondere scheint durch diesen Umstand die Full-Methode sensitiv auf *Extremwerte* in den Daten zu reagieren, während die Split-Methode in dieser Hinsicht *robust* zu sein scheint (siehe Unterkapitel 5.6). Wir kommen daher zum Schluss, dass für viele praktische Anwendungen, besonders dann, wenn sie durch große Stichproben, hohe numerische Komplexität oder mögliche Extremwerte in den Daten gekennzeichnet sind, die Split-Methode gegenüber der Full-Methode vorzuziehen ist.

Es bleiben noch einige offene Fragen, deren Klärung durchaus von Interesse sein könnte. Im Zuge der Simulation waren sowohl bei der Full- wie auch bei der Split-Methode in der Entwicklung der Intervallsgrenzen Schwankungen zu beobachten, die sich in einem *oszillierenden* Muster zu vollziehen scheinen (siehe Unterkapitel 5.3 und 5.4). Ähnliches konnten wir auch für die Split-Methode in der Simulation aus Unterkapitel 5.6 beobachten. Es war nicht ersichtlich, was diese Art von Schwankung verursachen könnte. Weiters ist offen, warum in Unterkapitel 5.3 bei der hypokonservativen Methode mit wachsender Stichprobengröße eine deutliche *Abnahme* der empirischen Überdeckungsrate zu beobachten war. Als interessant erachten wir auch eine genaue Analyse der *mangelnden* Robustheit der auf endlichen Netzen basierenden Full-Methoden, was einen Beitrag zu deren Verbesserung leisten könnte. Zum Abschluss möchten wir noch betonen, dass wir in dieser Arbeit mit der Full- und der Split-Methode nur einen *Teil* des Spektrums möglicher Verfahren zur Erzeugung konformaler Konfidenzbereiche abgedeckt haben. Vollkommen ausgeklammert wurden von uns Verfahren wie die *Jackknife-Methode* oder die *Rank-One-Out-Split-Methode*; vgl. G'Sell et al. (2016).

A Zusätzliche Abbildungen

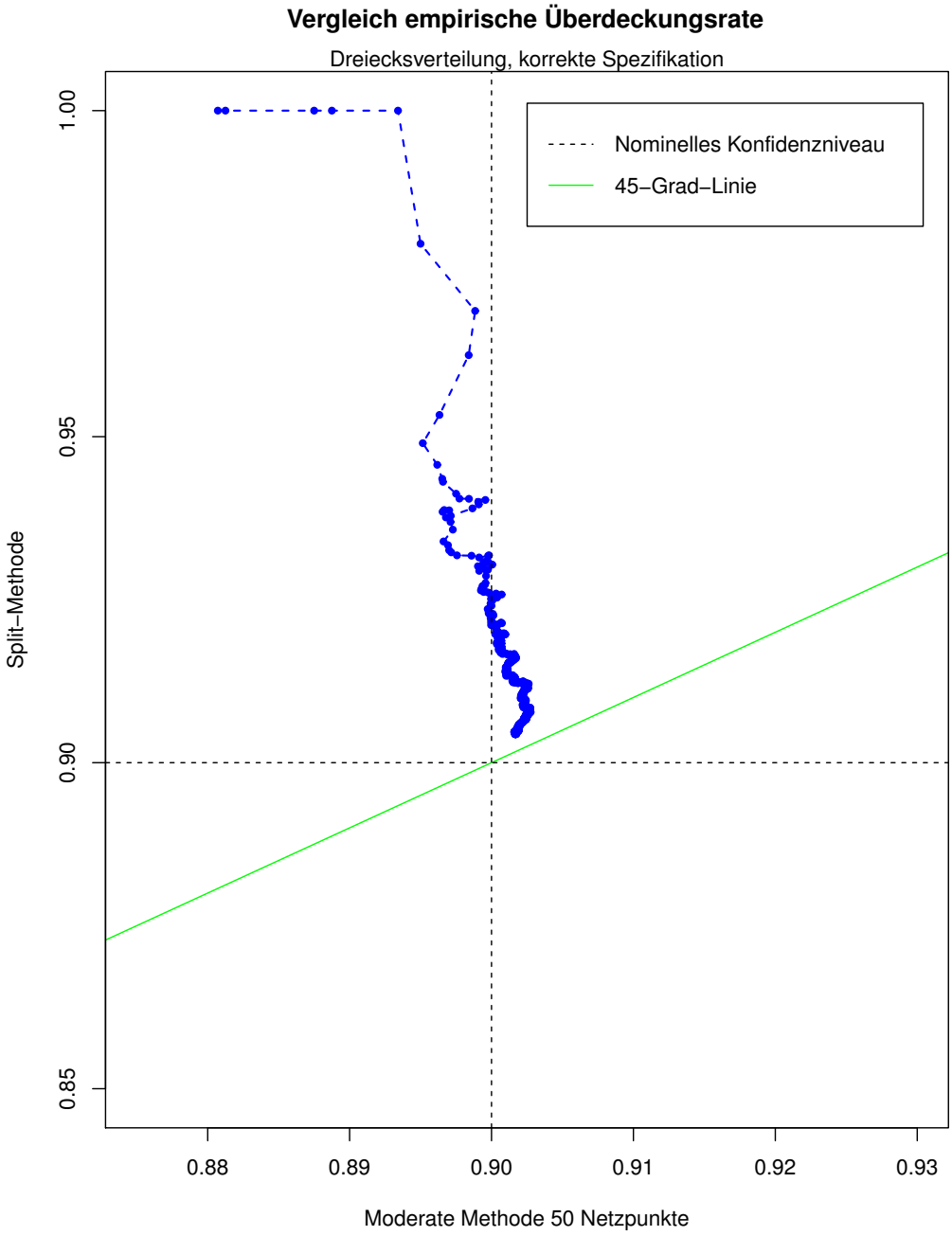

Abbildung 29: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden. Jeder Punkt repräsentiert dabei genau eine der Stichprobengrößen zwischen* p *(links oben im Bild) und* n *(Mitte des Bildes). Entlang der gestrichelten Linie, welche die Punkte verbindet, wächst die Stichprobengröße.*

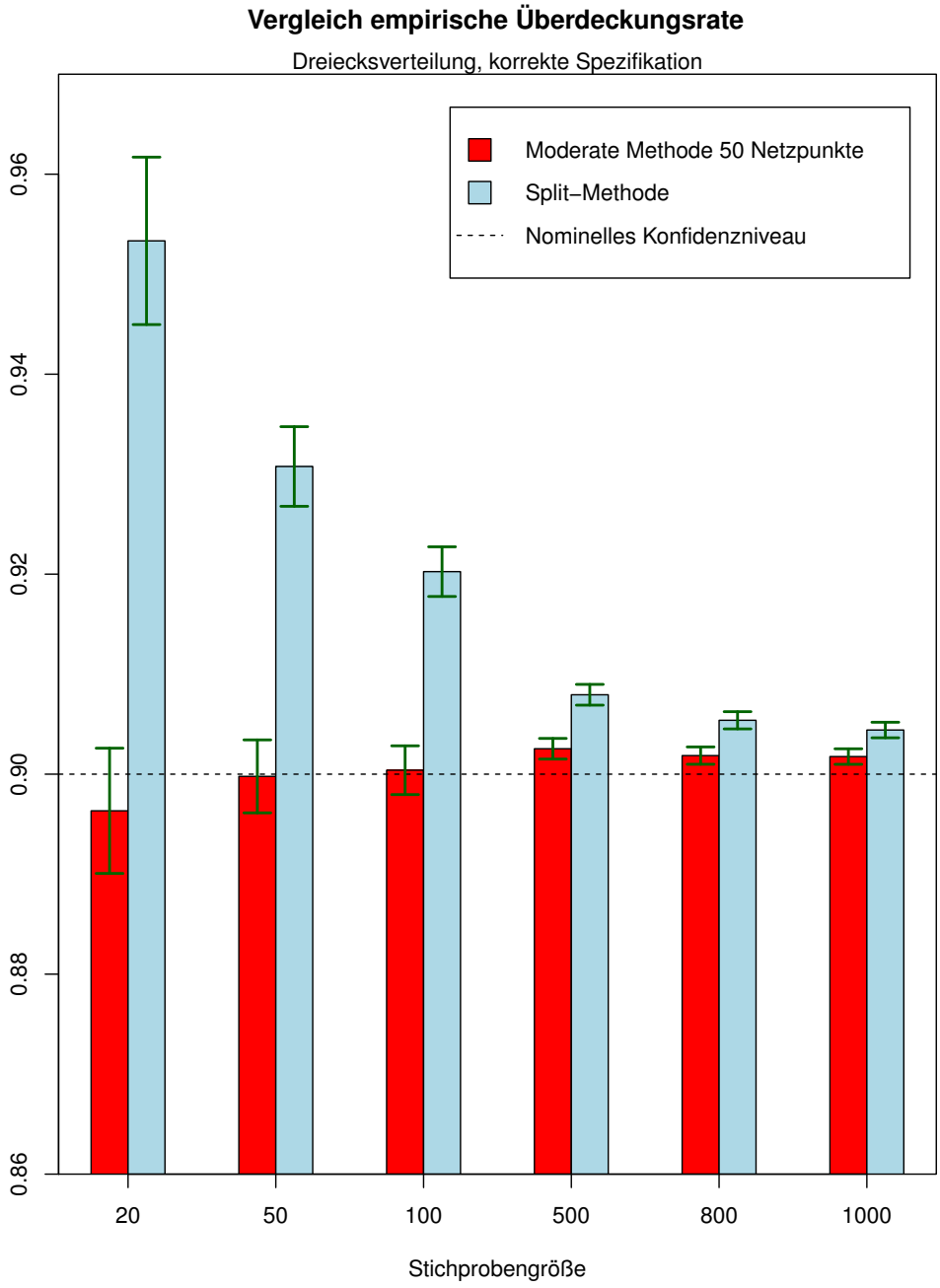

Abbildung 30: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

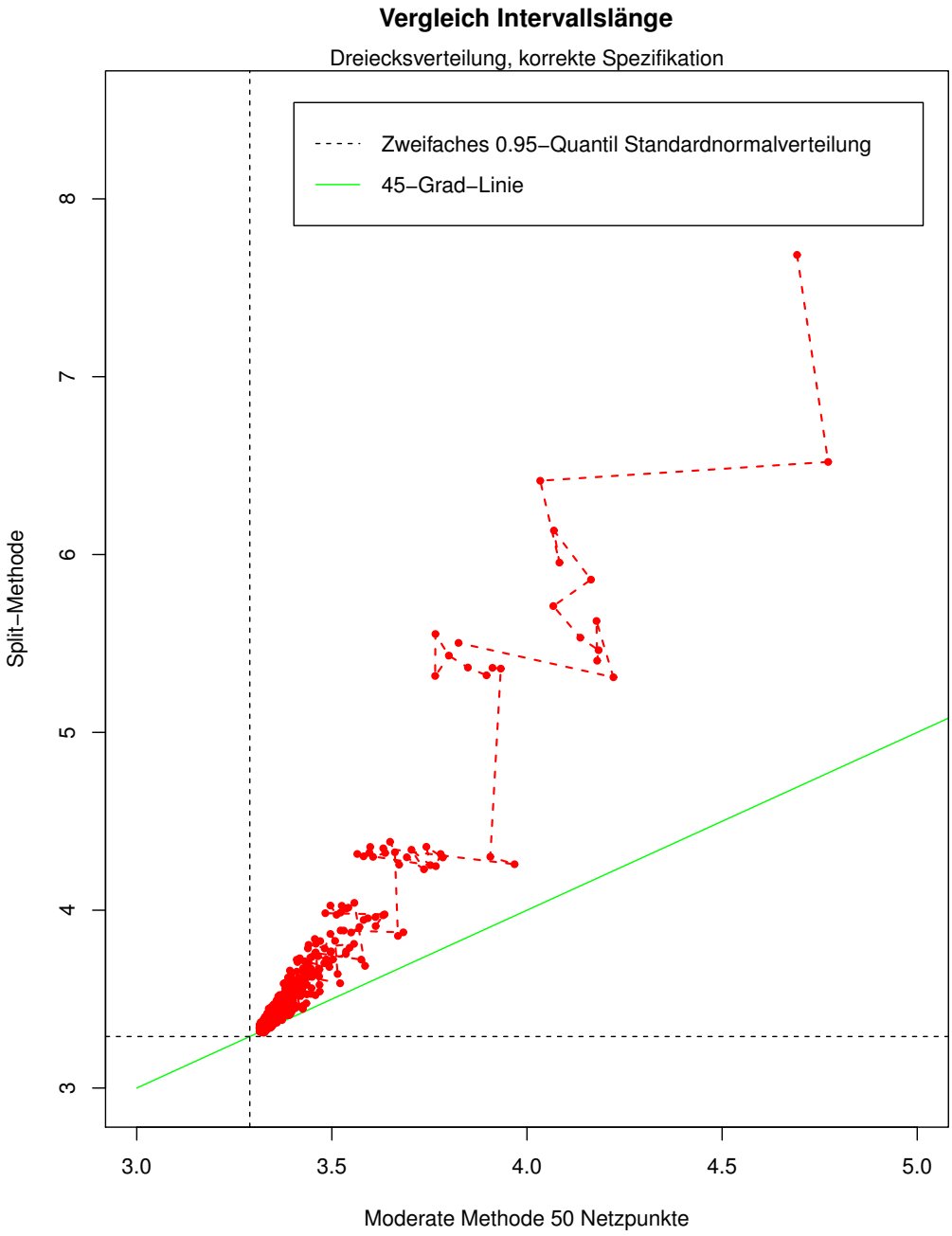

Abbildung 31: *Vergleich zur Entwicklung der gemittelten Intervallslänge zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden. Dabei repräsentiert jeder Punkt genau eine der Stichprobengrößen zwischen* p *(rechts oben im Bild) und* n *(Mitte des Bildes). Die gestrichelte Linie, welche die Punkte verbindet, markiert den Pfad, nach der die Stichprobengröße wächst.*

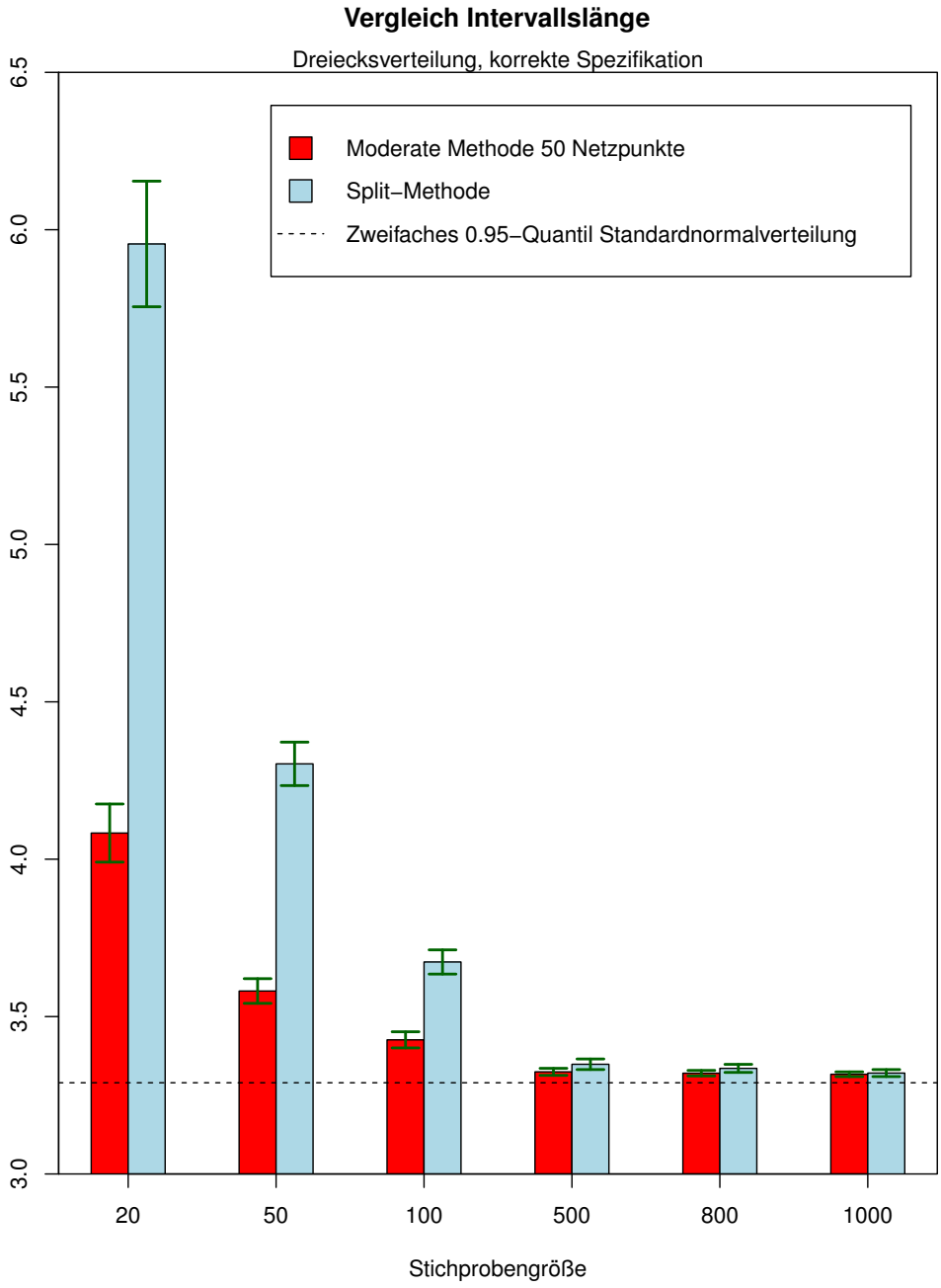

Abbildung 32: *Vergleich zur Entwicklung der gemittelten Intervallslänge sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die korrekt spezifizierte Kleinste-Quadrate-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

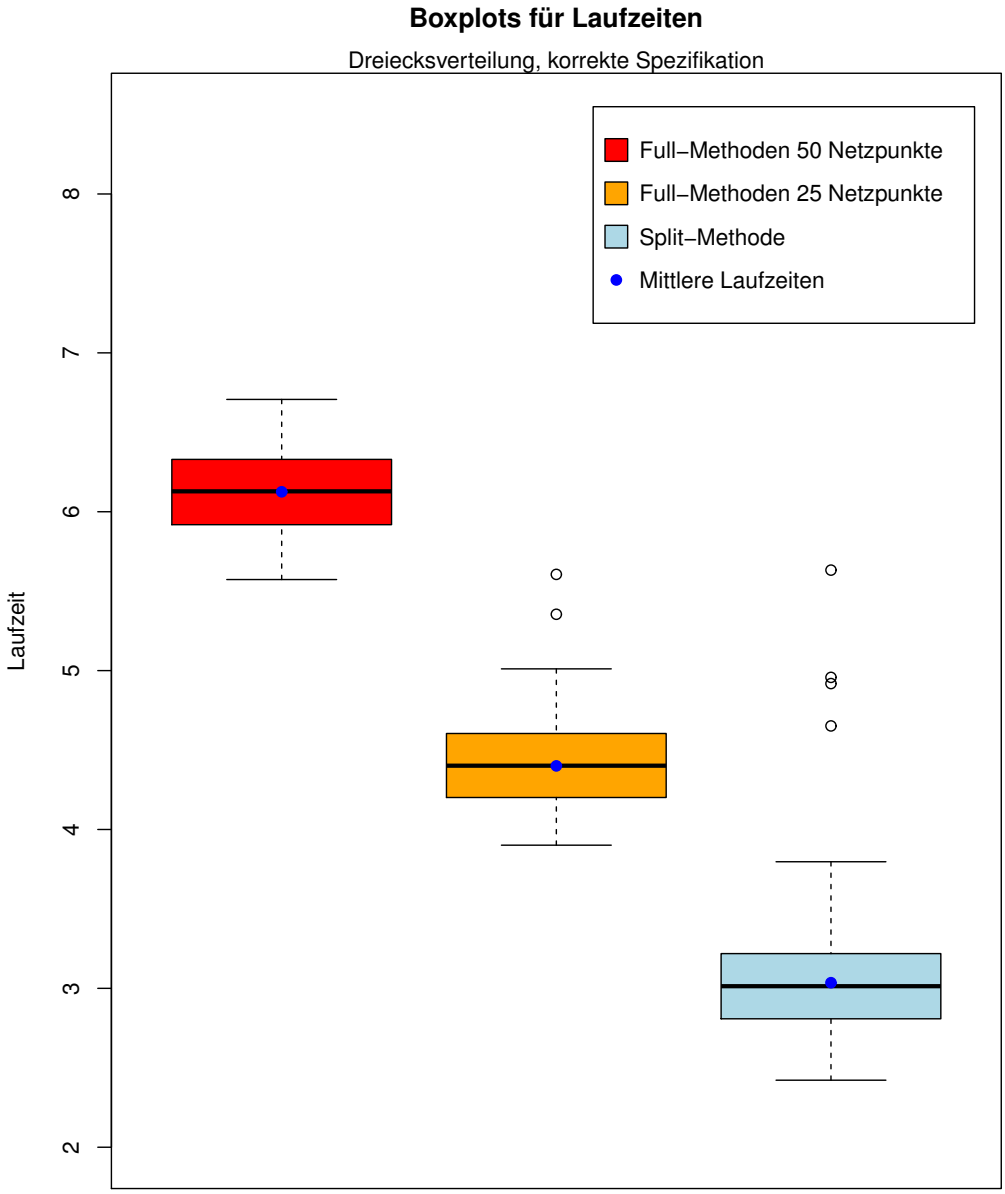

Abbildung 33: *Vergleich der empirischen Verteilung der in jedem der* 400 *Iterationen beobachteten Laufzeiten, welche für die (gemeinsame) Ausführung der Full-Methoden (mit* $m = 50$ *bzw.* $m = 25$*) sowie für die Ausführung der Split-Methode unter Anwendung der korrekt spezifizierte Kleinste-Quadrate-Schätzung erforderlich waren und als Boxplots dargestellt sind.*

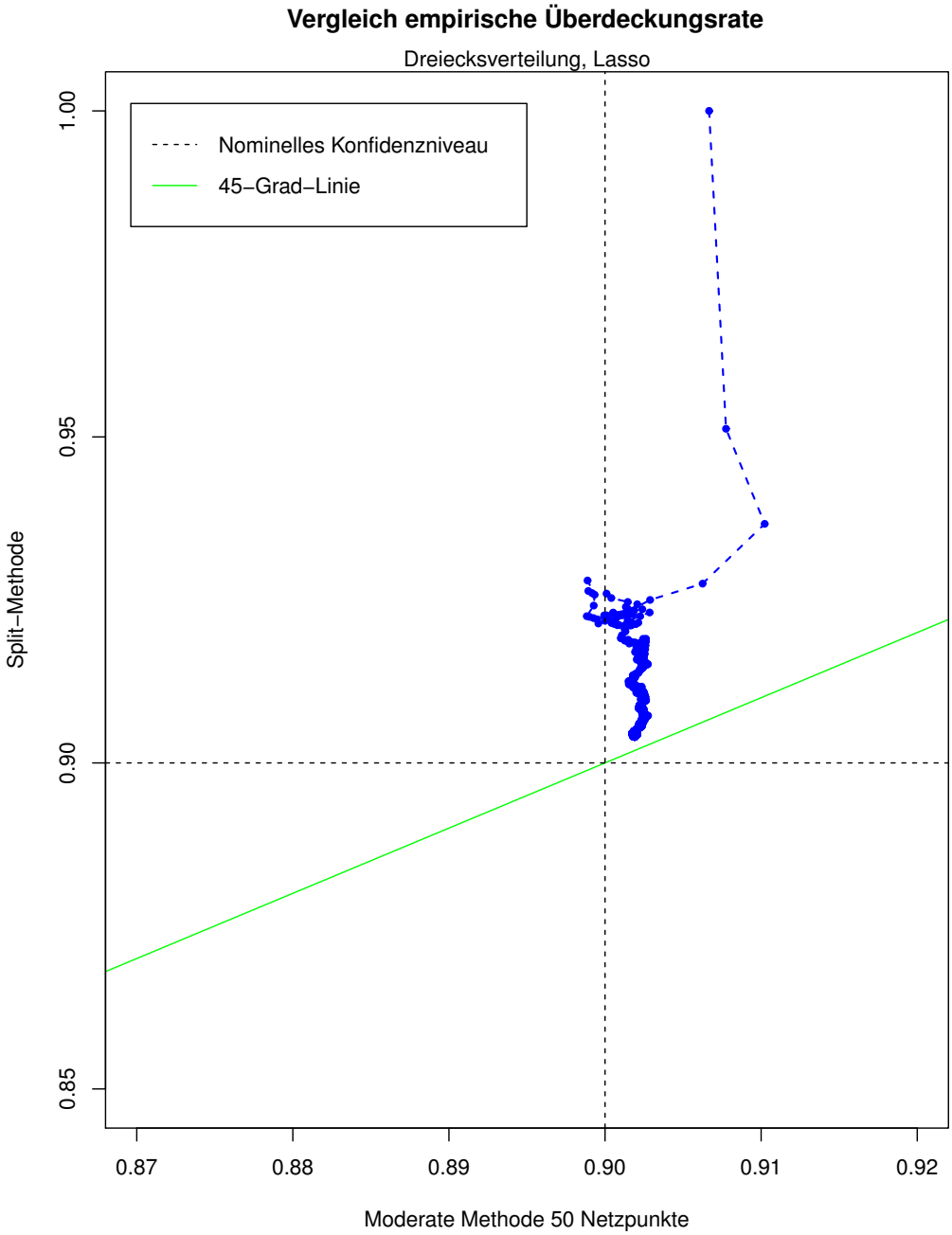

Abbildung 34: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden. Jeder Punkt repräsentiert dabei genau eine der Stichprobengrößen zwischen* p *(rechts oben im Bild) und* n *(Mitte des Bildes). Die gestrichelte Linie, welche die Punkte verbindet, markiert den Pfad, nach der die Stichprobengröße wächst.*

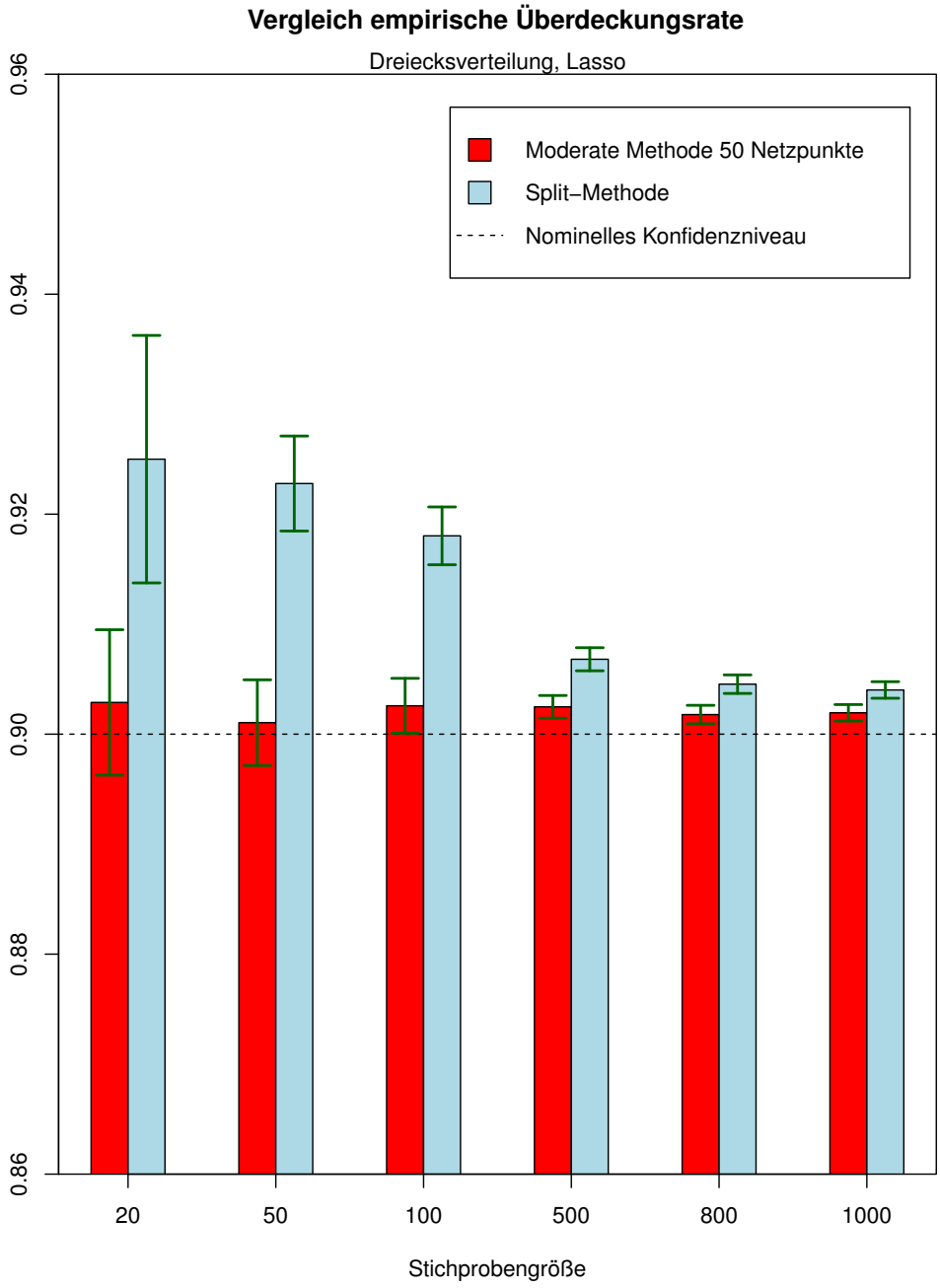

Abbildung 35: *Vergleich zur Entwicklung der gemittelten empirischen Überdeckungsrate sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

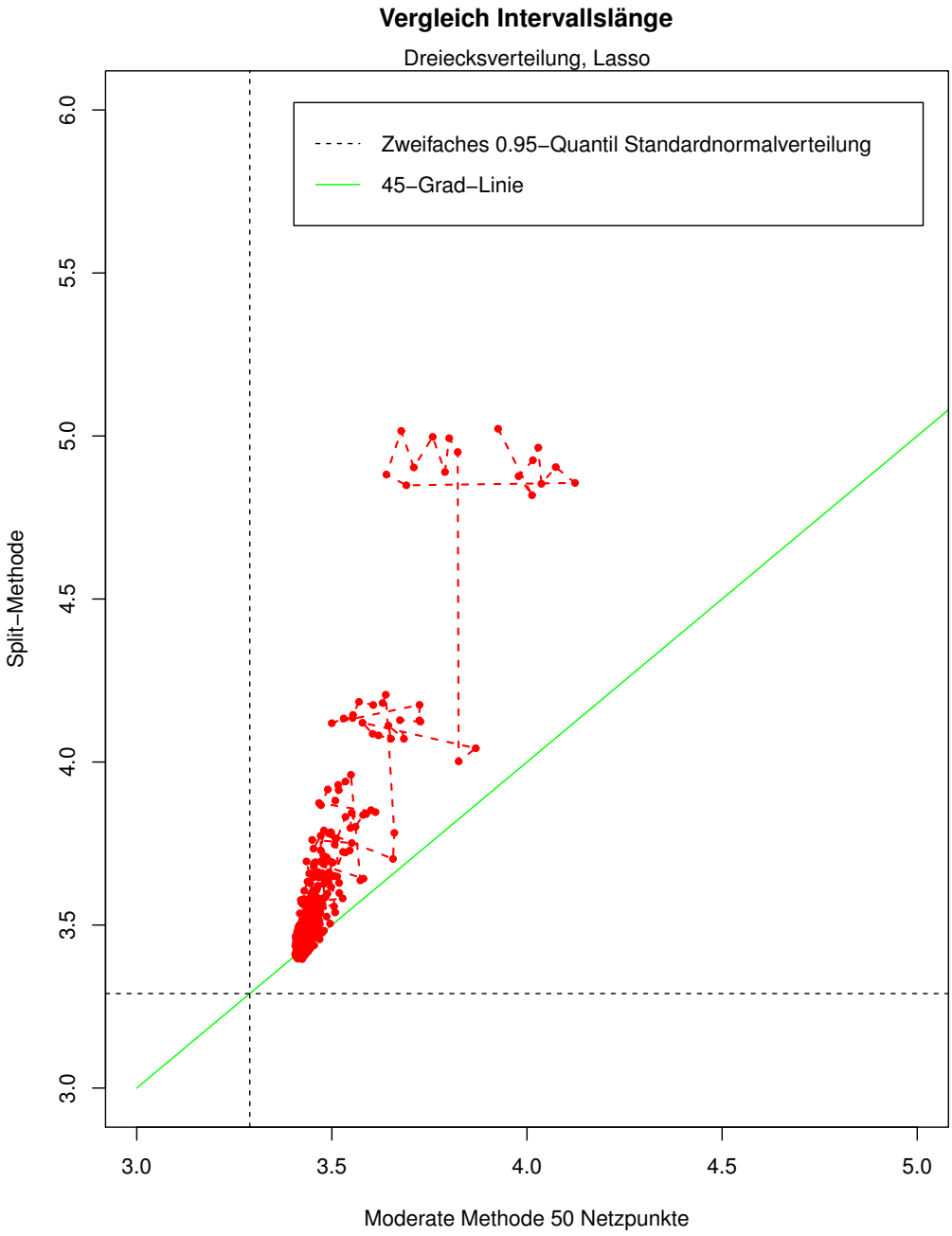

Abbildung 36: *Vergleich zur Entwicklung der gemittelten Intervallslänge zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden. Dabei repräsentiert jeder Punkt genau eine der Stichprobengrößen zwischen* p *(rechts oben im Bild) und* n *(links unten im Bild). Die gestrichelte Linie, welche die Punkte verbindet, markiert den Pfad, nach der die Stichprobengröße wächst.*

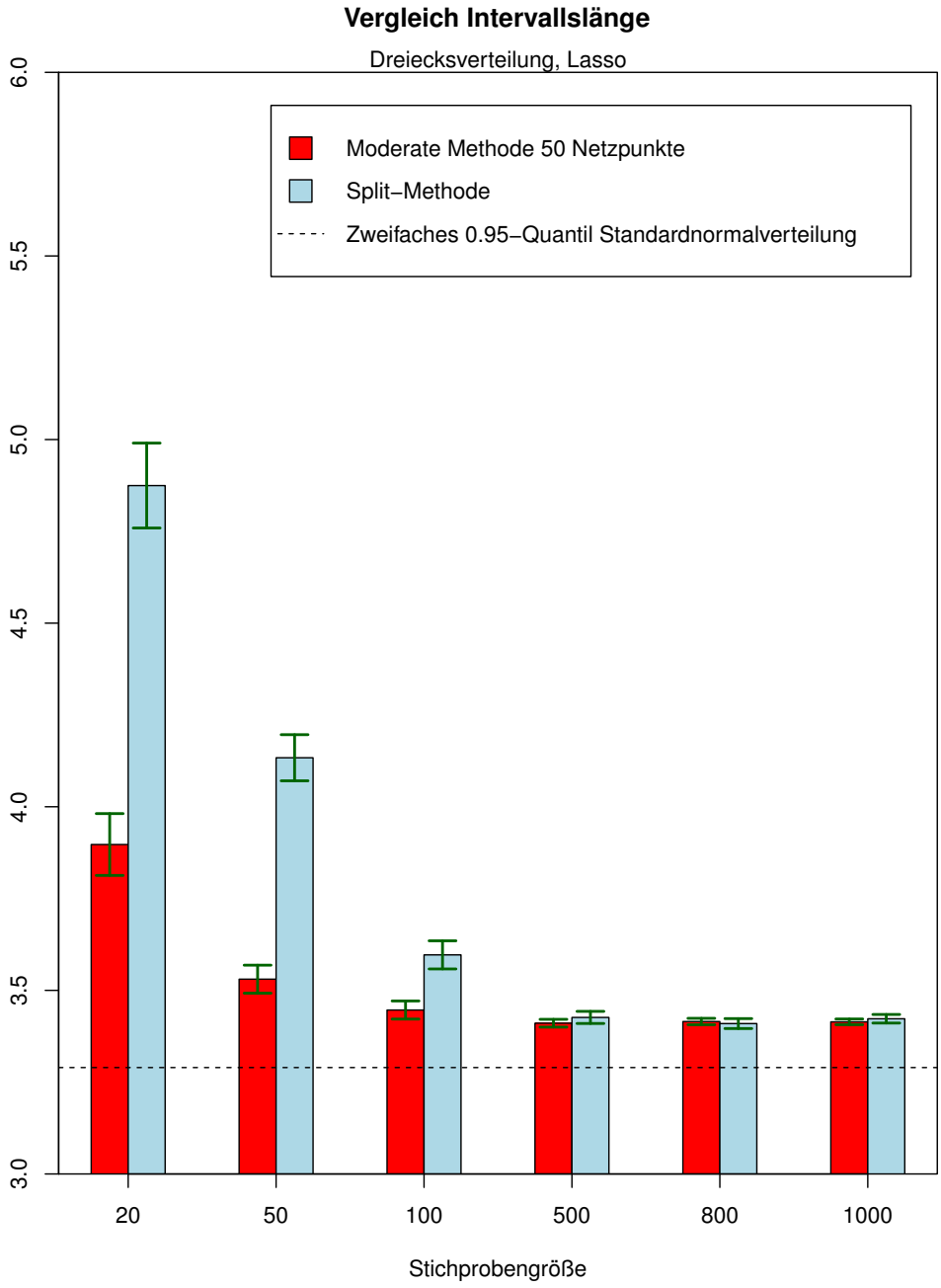

Abbildung 37: *Vergleich zur Entwicklung der gemittelten Intervallslänge sowie der Länge der dazugehörigen asymptotischen* $(1-\alpha)$*-Konfidenzintervalle zwischen der moderaten Full-Methode unter* $m = 50$ *und der Split-Methode, welche beide die Lasso-Methode verwenden, für die Stichprobengrößen* 20, 50, 100, 500, 800 *und* 1000.

B Quellcodes zur Berechnung der Konfidenzbereiche

Wir wollen in diesem Anhang eine kurze Erläuterung über die *Implementierung* der Methoden, welche wir in Kapitel 5 zur numerischen Bestimmung der konformalen Konfidenzbereiche verwendet haben, in der Skriptsprache R (die wir außerdem zur Erzeugung der in dieser Arbeit enthaltenen Grafiken verwendet haben) geben. Die hier vorgestellten *Quellcodes* bzw. *Funktionen* basieren auf das R-Paket `conformalInference` und sind für unsere Zwecke entsprechend vereinfacht, da in dieser Arbeit die Bestimmung der Konfidenzbereiche immer auf einer Kleinste-Quadrate-Schätzung oder einer Lasso-Schätzung beruht. Die Dreiecksverteilung zur Realisierung der Regressoren haben wir durch Ausführung der Funktion `rtriangle` aus dem R-Paket `triangle` simuliert.

Die Funktion, welche basierend auf der Kleinste-Quadrate-Schätzung die drei *Full-Methoden* implementiert, hat folgende Gestalt:

```
full.ols <- function(x, y, x.new, conf = 0.9, num.grid = 50,
                     grid.factor = 0.25){
    X <- rbind(cbind(1, x), c(1, x.new))
    n <- dim(X)[1]
    sw <- max(y) - min(y)
    Y <- seq(min(y) - grid.factor * sw, max(y) + grid.factor * sw,
             length = num.grid)
    y.g <- rbind(matrix(rep(y, length(Y)), length(y)), Y)
    beta.hat <- solve(crossprod(X)) %*% crossprod(X, y.g)
    U <- abs(y.g - X %*% beta.hat)
    ränge <- colSums(U <= matrix(rep(U[n, ], each = n), nrow = n))
    h <- ceiling(n * conf)
    intervalle <- grid.interval(netz = Y, r = ränge, grenze = h)
    return(intervalle)
}
```

Die Funktion `full.ols` wird durch die Parameter `x`, `y`, `x.new`, `conf`, `num.grid` und `grid.factor` charakterisiert. Dem Parameter `x` wird eine Designmatrix übergeben, welche die realisierten Werte der verwendeten Regressoren enthält. Die Zahl der Spalten von `x` entspricht der Zahl der Regressoren, während die Zahl der Zeilen die Anzahl der Beobachtungen pro Regressor repräsentiert. Dem Parameter `y` werden die realisierten Werte des Regressanden als Vektor

übergeben. Die Anzahl der Einträge von `y` entspricht dabei der Anzahl der Zeilen von `x`. Dem Parameter `x.new` wird ein Vektor zusätzlicher Beobachtungen für die einzelnen Regressoren übergeben, wobei die Anzahl der Einträge dieses Vektors die Zahl der Spalten von `x` entspricht. Das nominelle Konfidenzniveau wird der Funktion über dem Parameter `conf` (welcher defaultmäßig auf den Wert 0.9 eingestellt ist) übergeben. Die Parameter `num.grid` und `grid.factor` steuern die Anzahl der Punkte und die „Breite“ eines *endlichen Netzes*, welches aus dem Dantenvektor `y` erzeugt wird und dem Netz äquidistanter Punkte aus Unterkapitel 5.2 entspricht. Dabei repräsentiert `num.grid` (defaultmäßig auf 50 eingestellt) die Größe m und `grid.factor` (defaultmäßig auf den Wert 0.25 gesetzt) die Größe g. Die Netzpunkte werden dabei lokal in der Variable `Y` gespeichert. Für jeden Netzpunkt bestimmen wir anschließend Kleinste-Quadrate-Schätzer im Rahmen einer *multivariaten* Schätzung. Das ermöglicht uns, ausschließlich auf Vektorarithmetik zuzugreifen und auf eine (langsamere) Zählschleife zu verzichten. Durch die Anwendung der Hilfsfunktion `grid.interval` werden die Intervallsgrenzen gleichzeitig nach der *hyperkonservativen*, der *hypokonservativen* und der *moderaten* Methode bestimmt. Diese werden in der Matrix `intervalle`, die aus 3 Zeilen und 2 Spalten besteht, gespeichert. Die erste Zeile enthält die Intervallsgrenzen zur hyperkonservativen Methode, die zweite Zeile die Intervallsgrenzen zur hypokonservativen Methode und die dritte Zeile enthält die Grenzen der moderaten Methode, während die erste Spalte die unteren Intervallsgrenzen und die zweite Spalte die oberen Intervallsgrenzen enthält. Das Objekt `intervalle` wird schließlich von der Funktion `full.ols` zurückgegeben. Der Funktion `grid.interval` werden folgende Parameter übergeben: den in der Variable `Y` gespeicherten Netzpunkten, den in `ränge` gespeicherten Rängen der Residuen, welche jeweils mit $U_{n+1,(X_{n+1},y)}$ zum entsprechenden Netzpunkt y aus Unterkapitel 5.2 zusammenfallen und in der letzten Zeile der Matrix `U` gespeichert sind, und den Wert `ceiling(n*conf)`, welcher in der Variable h aus Unterkapitel 3.1 bzw. 5.2 seine Entsprechung findet. Der in dieser Arbeit verwendete Quellcode für diese Funktion lautet:

```
grid.interval <- function(netz, r, grenze) {
  m <- length(netz)
  below <- (r <= grenze)
  ii <- which(below)
  if (length(ii) == 0) {
    return(matrix(rep(0, each = 6), ncol = 2))
```

```
  }
  i <- min(ii)
  j <- max(ii)
  hyper <- c(netz[max(i-1, 1)], netz[min(j+1, m)])
  hypo <- c(netz[i], netz[j])
  if (i == 1) {
    m.u <- netz[i]
  }
  else {
    sl <- (r[i-1] - grenze) / (r[i-1] - r[i])
    m.u <- netz[i-1] + sl * (netz[i] - netz[i-1])
  }
  if (j == m) {
     m.o <- netz[j]
  }
  else {
     sr <- (r[j] - grenze) / (r[j] - r[j+1])
     m.o <- netz[j] + sr * (netz[j+1] - netz[j])
  }
  moderat <- c(m.u, m.o)
  intervalle <- matrix(c(hyper, hypo, moderat), ncol = 2, byrow = TRUE)
  return(intervalle)
}
```

Die auf die Kleinste-Quadrate-Schätzung beruhende *Split-Methode* ist über folgende Funktion implementiert:

```
split.ols <- function(x, y, x.new, delta = 0.5, conf = 0.9){
   X <- cbind(1, x)
   X.new <- c(1, x.new)
   n <- dim(X)[1]
   m <- floor(delta * n)
   id <- sample(1:n, size = m)
   X.s <- X[id, ]
   y.s <- y[id]
   beta.hat <- solve(crossprod(X.s)) %*% crossprod(X.s, y.s)
   U <- abs(y[-id] - X[-id, ] %*% beta.hat)
   U <- sort(U)
   h.tilde <- ceiling((n - m + 1) * conf)
   if (h.tilde <= n - m) {
```

```
    q <- U[h.tilde]
  }
  else {
    q <- Inf
  }
  y.new <- X.new %*% beta.hat
  s.u <- y.new - q
  s.o <- y.new + q
  return(c(s.u, s.o))
}
```

Die Funktion `split.ols` besitzt die Parameter `x`, `y`, `x.new`, `conf` und `delta`. Dabei haben `x`, `y`, `x.new` und `conf` die gleiche Bedeutung wie bei `full.ols`. Der Parameter `delta` (defaultmäßig auf 0.5 eingestellt) steuert die Größe der Trainingsmenge, die beim „Split“ der Daten gewonnen wird. Dieser entspricht der Variable δ aus Unterkapitel 3.2. Die konkrete Trainingsmenge, repräsentiert durch die Indexmenge `ind`, wird durch zufälliges Ziehen aus der verfügbaren Stichprobe bestimmt. Im Vektor `U` werden die durch die Split-Methode bestimmten Residuen gespeichert. Aus diesem wird dann der `h.tilde`-größte Eintrag gewählt (`h.tilde` entspricht der Größe $\tilde{h}$ aus Unterkapitel 3.2), der lokal in der Variable `q` gespeichert und zur Berechnung der Intervallsgrenzen herangezogen wird. Die untere Intervallsgrenze wird im Objekt `s.u` und die obere Intervallsgrenze in `s.o` gespeichert. Diese werden schließlich als Vektor von `split.ols` zurückgegeben.

Zur Umsetzung der Lasso-Methode verwenden wir die Funktion `glmnet` aus dem R-Paket `glmnet`. Diese binden wir in folgende Funktion zur Anwendung der Full-Methoden, welche auf der Lasso-Schätzung beruhen, ein:

```
full.lasso <- function(x, y, x.new, conf = 0.9, num.grid = 50,
                       grid.factor = 0.25, lambda){
    X <- rbind(x, x.new)
    n <- dim(X)[1]
    sw <- max(y) - min(y)
    Y <- seq(min(y) - grid.factor * sw, max(y) + grid.factor * sw,
             length = num.grid)
    y.g <- rbind(matrix(rep(y, length(Y)), length(y)), Y)
    U <- matrix(numeric(nrow(y.g) * ncol(y.g)), ncol = ncol(y.g))
    for (i in 1:ncol(y.g)) {
      qm <- glmnet(x = X, y = y.g[ , i], family = "gaussian",
```

```
                    lambda = lambda)
      pm <- predict(object = qm, newx = X)
      U[ , i] <- abs(y.g[ , i] - pm)
    }
    ränge <- colSums(U <= matrix(rep(U[n, ], each = n), nrow = n)
    h <- ceiling(n * conf)
    intervalle <- grid.interval(netz = Y, r = ränge, grenze = h)
    return(intervalle)
}
```

Die Funktion `full.lasso` ist also nach dem gleichen Schema aufgebaut wie `full.ols`, wobei `lambda` für den Tuningparameter λ der Lasso-Schätzung steht. Im Gegensatz zu `full.ols` kommt hier allerdings eine Zählschleife zum Einsatz, mittels derer für jeden Netzpunkt eine Lasso-Schätzung durch `glmnet` durchgeführt wird. Analog zu `split.ols` ist die folgende Funktion zur Umsetzung der Split-Methode unter Einbindung einer Lasso-Schätzung konstruiert:

```
split.lasso <- function(x, y, x.new, delta = 0.5, conf = 0.9, lambda){
   n <- dim(x)[1]
   m <- floor(delta*n)
   id <- sample(1:n, size = m)
   X.s <- x[id, ]
   y.s <- y[id]
   qs <- glmnet(x = X.s, y = y.s, family = "gaussian", lambda = lambda)
   ps <- predict(object = qs, newx = x[-id, ])
   U <- abs(y[-id] - ps)
   U <- sort(U)
   h.tilde <- ceiling((n - m + 1) * conf)
   if (h.tilde <= n - m) {
     q <- U[h.tilde]
   }
   else {
     q <- Inf
   }
   y.new <- predict(object = qs, newx = t(x.new))
   s.u <- y.new - q
   s.o <- y.new + q
   return(c(s.u, s.o))
}
```

Auch hier entspricht `lambda` wieder dem Tuningparameter λ.

Anmerkungen

1 Im Allgemeinen werden im Bereich des sogenannten *überwachten maschinellen Lernens* (*supervised learning*), welches nicht nur Regressionsanalysen sonder auch Klassifikationsprobleme umfasst, für den Ausdruck Regressor sehr oft die Termini *Input* oder *Attribut* (in der englischsprachigen Literatur auch *independent variable*, *predictor variable* oder *feature* genannt) verwendet, während im Zusammenhang mit dem Regressanden sehr oft von *Output* oder *Zielvariable* (*response variable*, *dependent variable* oder *label* in der englischsprachigen Literatur) die Rede ist; siehe zum Beispiel Friedman et al. (2010), Mohri et al. (2012) oder Hastie et al. (2013).

2 Dies kann in praktischer Hinsicht dann sinnvoll sein, wenn zum Zeitpunkt der Prognose mehrere unbekannte Instanzen vorliegen oder sich die jedesmalige Anwendung des On-Line Settings beim Auftauchen neuer Instanzen als zu aufwändig erweist; siehe auch Gammerman et al. (2005, S.7).

3 In der Literatur wird für Funktionen, die wir hier Algorithmen nennen, oft die Bezeichnung *Prediktor* (*predictor*) oder *Prognosefunktion* (*prediction function*) verwendet; siehe z.B. Gammerman et al. (2005) oder Bottou und Bousquet (2008).

4 Für beliebige n betrachte die Funktionenmenge

$$\bar{M} := \{\sigma \,|\, \sigma : \{1, ..., n\} \to \{1, ..., n\}\}.$$

Aus der Kombinatorik wissen wir, dass die Menge $\bar{M}$ die Mächtigkeit n^n besitzt. Gleichzeitig nimmt bei gegebenen Realisationen $r_1, ..., r_n$ der Zufallsvariablen $R_1 - \bar{R}, ..., R_n - \bar{R}$ der (diskret verteilte) Zufallsvektor $\mathbf{E}^*$ ausschließlich Werte aus der Menge

$$\{(r_{\sigma(1)}, ..., r_{\sigma(n)}) \,|\, \sigma \in \bar{M}\}$$

und damit bis zu n^n verschiedene Werte an.

5 Es wird häufig in der Literatur eine wohldefinierte *Verlustfunktion* (*loss function*) herangezogen. Diese soll die „Kosten“ bemessen, die aus einer Diskrepanz zwischen dem realisierten Wert einer als Regressand verwendeten Variable und dem durch ein Modell prognostizierten Wert für die gleiche Variable resultieren. Eine solche Diskrepanz kann z.B. als Absolutbetrag oder Quadrat der Differenz zwischen realisiertem und prognostiziertem Wert definiert werden. Man setzt dann die Existenz eines Modells (das kann z.B. eine Funktion sein, die einen bestimmten Zusammenhang zwischen Regressoren und Regressand beschreibt) voraus, welches die Verlustfunktion über eine vorher festgelegte Menge an möglichen Modellen minimiert. Dieses soll anschließend von einem Algorithmus bzw. von einem Prediktor möglichst gut approximiert werden; siehe auch Bottou und Bousquet (2008).

6 Für gegebene Stichprobengröße n und einer beliebigen Regressionsfunktion $\hat{r}_n$, die höchstens von $Z_1, ..., Z_n, X_{n+1}$ abhänge und für welche der Erwartungswert des Schätzfehlers $\mathbb{E}(a(X_{n+1}) - \hat{r}_n(X_{n+1}))$ bezüglich des Maßes P existiere, betrachten wir den Erwartungswert des *quadratischen Prognosefehlers* $\mathbb{E}((Y_{n+1} - \hat{r}_n(X_{n+1}))^2)$ von $\hat{r}_n$ für Y_{n+1}, also der Erwartungswert des Quadrates des Residuums, das sich zwischen Y_{n+1} und $\hat{r}_n(X_{n+1})$ ergibt. Nach algebraischer Umformung und unter Ausnutzung der Linearität des Erwartungswertes erhalten wir unter den Bedingungen des Referenzmodells

$$\mathbb{E}((Y_{n+1} - \hat{r}_n(X_{n+1}))^2)$$

$$= \mathbb{E}((Y_{n+1} - a(X_{n+1}) + a(X_{n+1}) - \hat{r}_n(X_{n+1}))^2)$$

$$= \mathbb{E}((Y_{n+1}-a(X_{n+1}))^2+2\cdot(Y_{n+1}-a(X_{n+1}))\cdot(a(X_{n+1})-\hat{r}_n(X_{n+1}))+(a(X_{n+1})-\hat{r}_n(X_{n+1}))^2)$$

$$= \mathbb{E}((\epsilon_{n+1})^2 + 2\cdot\epsilon_{n+1}\cdot(a(X_{n+1})-\hat{r}_n(X_{n+1})) + (a(X_{n+1})-\hat{r}_n(X_{n+1}))^2)$$

$$= \mathbb{E}((\epsilon_{n+1})^2) + 2\cdot\mathbb{E}(\epsilon_{n+1})\cdot\mathbb{E}(a(X_{n+1})-\hat{r}_n(X_{n+1})) + \mathbb{E}((a(X_{n+1})-\hat{r}_n(X_{n+1}))^2)$$

$$= \mathbb{E}((\epsilon_{n+1})^2) + \mathbb{E}((a(X_{n+1})-\hat{r}_n(X_{n+1}))^2),$$

wobei im letzten Schritt zu beachten ist, dass $\mathbb{E}(\epsilon_{n+1}) = 0$ im Referenzmodell gilt. Der quadratische Fehler setzt sich also aus der Varianz der Störgröße ϵ_{n+1} und dem quadratischen Fehler, der beim Schätzen von a durch $\hat{r}_n$ entsteht, zusammen. Falls a durch $\hat{r}_n$ *perfekt* geschätzt wird, also wenn $\hat{r}_n = a$ erfüllt ist, dann erhalten wir

$$\mathbb{E}((a(X_{n+1})-\hat{r}_n(X_{n+1}))^2) = 0,$$

womit der quadratische Schätzfehler verschwindet.

7 Beachte, dass die empirische Verteilungsfunktion $\hat{F}_{n,\delta}$ auf der Indexmenge $I_{n,2}$ basiert. Da mit verschiedenen $u, v \in \mathbb{N}$ die Indizes $j(u,s)$ und $j(v,s)$ mit $s \leq \min\{k_u, k_v\}$ nicht zwingend die gleichen Werte annehmen müssen, verhält sich $\hat{F}_{n,\delta}$ mit wachsendem n im Allgemeinen anders als die über die Indexmenge $\{1,...,n\}$ definierte konventionelle empirische Verteilungsfunktion. Andererseits können im Zuge des auf dem Gesetz der großen Zahlen aufbauenden Beweises des Satzes von Glivenko-Cantelli (wie man ihn in Lehrbüchern zur Wahrscheinlichkeitstheorie wie etwa Chow und Teicher (1978) findet) alle Transformationen, die sich der Eigenschaft von Folgen unabhängiger und identisch verteilter Zufallsvariablen bedienen, zu fester Stichprobengröße durchgeführt werden und bei $\hat{F}_{n,\delta}$ handelt es sich für festes n immerhin um eine empirische Verteilungsfunktion im herkömmlichen Sinn. Man erhält dann Aussagen, die unabhängig von der konkreten Form der empirischen Verteilungsfunktion im Bezug auf wachsender Stichprobengröße Grenzprozesse zulassen (wir erinnern uns dabei, dass $k_n \to \infty$ für $n \to \infty$ gilt), die zu jener Schlussfolgerung führt, wie wir sie schon von dem für konventionelle empirische Verteilungsfunktionen gültigen Satz von Glivenko-Cantelli kennen.

8 Beachte, dass es sich dabei um *ad hoc* ausgesuchte Zahlenwerte handelt. Für beliebige i und t betrachte die Größe

$$\pi_i^t := \frac{\mathbb{E}\left((\nu_i^t)^2\right)}{\mathbb{E}\left((\epsilon_i^t)^2\right)}$$

als *Signal-Rausch-Verhältnis* (*signal-to-noise ratio*) für Modell (5.1), wobei die Zufallsvariable

$$\nu_i^t := \sum_{k=1}^{d} \beta_k \cdot X_{i,k}^t$$

den *Signalterm* für den Regressanden Y_i^t darstellt. Aus dem Umstand, dass die Regressoren $X_{i,1}^t, ..., X_{i,d}^t$ im Rahmen des Simulationsmodells unabhängig und identisch verteilt sind mit 0 als gemeinsamer Erwartungswert und 1 als gemeinsame Varianz, folgt

$$\mathbb{E}\left((\nu_i^t)^2\right) = Var(\nu_i^t) = \sum_{k=1}^{d} \beta_k^2 \cdot Var(X_{i,k}^t) = \sum_{k=1}^{d} \beta_k^2.$$

Wegen $Var(\epsilon_i^t) = 1$ ist somit

$$\pi_i^t = \sum_{k=1}^{d} \beta_k^2$$

erfüllt. Mit den von uns gewählten Zahlenwerten für die Koeffizienten $\beta_1, ..., \beta_5$ gilt also näherungsweise $\pi_i^t \approx 0.56$. Wir hätten beispielsweise für die Koeffizienten betragsmäßig größere Werte wählen können, was ein größeres π_i^t nach sich gezogen hätte. Die Störgrößenvarianz wäre dann von der Varianz des Signalterms stärker überlagert worden.

9 Shrinkage-Methoden führen unter anderem auch bei stark korrelierten Regressoren zu einer entsprechenden Schrumpfung der Schätzwerte der dazugehörigen Regressoren, was zur Eliminierung der in den Daten enthaltenen „überflüssigen“ Information führen soll, welche durch die Korrelation verursacht wird. Während die Ridge-Methode dazu neigt, die geschätzten Koeffizienten der korrelierten Regressoren gleichzeitig schrumpfen zu lassen, tendiert die Lasso-Methode eher dazu, unter den korrelierten Regressoren einen auszuwählen und die geschätzten Koeffizienten der anderen Regressoren auf 0 zu setzen; vgl. Friedman et al. (2010, S.3).

10 Wir finden uns also in einer Situation wieder, in der wir es mit einem Lernalgorithmus $a_{\lambda,n}$ zu tun bekommen, welcher ein Algorithmus zur Full-Methode im Sinne von Definition 3.1 ist und darüber hinaus von einem Parameter $\lambda \in \mathbb{R}$ abhängt. Wenn wir nun zu einer gegebenen Realisation $z_1, ..., z_{n+1}$ der Zufallsvariablen $Z_1, ..., Z_{n+1}$ einen datengetriebenen Tuningparameter $\lambda(z_1, ..., z_{n+1})$ betrachten, dann ist allgemein nicht mehr garantiert, dass der vom Tuningparameter abhängige Algorithmus die Eigenschaft der Permutationsinvarianz erfüllt, d.h. es kann Permuationen $\sigma \in S_{n+1}$ und Werte $x \in \mathbb{R}^d$ geben, sodass die Funktionswerte

$$a_{\lambda(z_1,...,z_{n+1}),n}(z_1, ..., z_{n+1})(x)$$

und

$$a_{\lambda\left(z_{\sigma(1)},...,z_{\sigma(n+1)}\right),n}(z_{\sigma(1)}, ..., z_{\sigma(n+1)})(x)$$

nicht übereinstimmen (sogar dann nicht, wenn der Algorithmus $a_{\lambda,n}$ bei festem λ permutationsinvariant ist).

11 Beachte, dass zum Iterationsschritt t für alle $i \in \mathbb{N}$ der Algorithmus a_i^t permutationsinvariant im Sinne von Definition 3.1 ist, da er auf Kleinste-Quadrate-Schätzungen bzw. auf der Lasso-Methode mit festem (bzw. von den Daten unabhängigem) Tuningparameter aufbaut. Falls außerdem im Falle der korrekten Spezifikation $\hat{\beta}_{i,y}^t$, wobei hier y für die Zufallsvariable Y_{i+1}^t stehe, ein konsistenter Schätzer für β ist, d.h. falls

$$\operatorname*{plim}_{n\to\infty} |\hat{\beta}_{n,y,k}^t - \beta_k| = 0$$

für alle $k \in \{0, ..., d\}$ gilt, erfüllt die Funktionenfolge $(a_n^t)_{n\geq d+1}$ im Rahmen des Simulationsmodells die Konsistenzbedingung (KF), weil auf ganz Ω die Beziehung

$$\begin{aligned}
&\sup_{j\in\{1,...,i\}} |a_i^t(Z_1^t, ..., Z_i^t, (X_{i+1}^t, y))(X_j^t) - a(X_j^t)| \\
&= \sup_{j\in\{1,...,i\}} |\hat{\beta}_{i,y,0}^t - \beta_0 + \sum_{k=1}^{d} (\hat{\beta}_{i,y,k}^t - \beta_k) \cdot X_{j,k}^t| \\
&\leq |\hat{\beta}_{i,y,0}^t - \beta_0| + \sum_{k=1}^{d} |\hat{\beta}_{i,y,k}^t - \beta_k| \cdot \sup_{j\in\{1,...,i\}} |X_{j,k}^t|
\end{aligned}$$

und bezüglich des gegebenen Wahrscheinlichkeitsmaßes fast sicher die Ungleichung

$$|\hat{\beta}_{i,y,0}^t - \beta_0| + \sum_{k=1}^{d} |\hat{\beta}_{i,y,k}^t - \beta_k| \cdot \sup_{j\in\{1,...,i\}} |X_{j,k}^t|$$

$$\leq |\hat{\beta}^t_{i,y,0} - \beta_0| + \sqrt{6} \cdot \sum_{k=1}^{d} |\hat{\beta}^t_{i,y,k} - \beta_k|$$

gelten (beachte, dass fast sicher $|X^t_{j,k}| \leq \sqrt{6}$ für alle $k \in \{0, ..., d\}$ erfüllt ist).

12 Beachte, dass für jedes $i \in \{q+1, ..., n\}$ die empirischen Überdeckungsraten

$$\frac{1}{i-q} \sum_{j=q+1}^{i} B^{full}_{j,\alpha,t}$$

über den Laufindex t unabhängig und identisch verteilt sind. Damit ist die Varianz der gemittelten empirischen Überdeckungsrate $A^{full,gem}_{i,\alpha}$ kleiner als die Varianz der empirischen Überdeckungsraten zu den einzelnen Iterationen. Die Intervallslängen $r^t_i - l^t_i$ sind ebenfalls über den Laufindex t unabhängig und identisch verteilt, womit die Varianz der gemittelten Intervallslänge $I^{full}_{i,\alpha}$ kleiner ist als die Varianzen der Intervallslängen zu den einzelnen Iterationsschritten. Mit den gemittelten Größen erhalten wir also Schätzer mit *reduzierter* Varianz.

13 Bei $s = 400$ ist die Länge der Konfidenzintervalle für die empirische Überdeckungsrate nicht größer als

$$\frac{2 \cdot F^{-1}(1-\alpha)}{\sqrt{s-1}} \approx 0.165.$$

Weil für alle $i \in \{q+1, ..., n\}$ und $j \in \{q+1, ..., i\}$ sowie für jedes $t \in \{1, ..., s\}$ auf ganz Ω die Bediengung

$$B^{full}_{j,\alpha,t} \in \{0, 1\}$$

erfüllt ist, gilt

$$\left(\frac{1}{i-q} \sum_{j=q+1}^{i} B^{full}_{j,\alpha,t} - A^{full,gem}_{i,\alpha} \right)^2 \in [0, 1].$$

Daraus folgt

$$sd(A^{full,gem}_{i,\alpha}) \leq \sqrt{\frac{s}{s-1}}$$

und schließlich

$$\frac{2 \cdot F^{-1}(1-\alpha) \cdot sd(A^{full,gem}_{i,\alpha})}{\sqrt{s}} \leq \frac{2 \cdot F^{-1}(1-\alpha)}{\sqrt{s-1}}.$$

Beachte, dass der linke Ausdruck in dieser Ungleichung der Länge des Konfidenzintervalls entspricht. Eine analoge Rechnung lässt sich auch für die Split-Methode anstellen.

14 Sofern für den Iterationsschritt t der Schätzer $\hat{\beta}^t_{i,\delta}$ konsistent für β ist, erfüllt die Funktionenfolge $(a^t_{n,\delta})_{n \geq p}$ im Rahmen des Simulationsmodells die Konsistenzbedingung (KS) (vgl. auch Anmerkung 11).

15 Das wäre auch richtig gewesen, hätten wir für das lineare Modell (5.1) betragsmäßig größere Werte für die Koeffizienten gewählt, was ein größeres Signal-Rausch-Verhältnis bedeutet hätte (siehe auch Anmerkung 8). Dies hätte im Bereich kleinerer Stichproben (bei denen Schätzungen der Funktion a aus dem Referenzmodell durch die betrachteten Algorithmen noch mit entsprechend großen Fehlern behaftet sind) zur Konsequenz gehabt, dass in der Entstehung von Prognosefehlern für Y_{n+1}, die ja von den Konfidenzbereichen abgedeckt werden, die von den Regressoren ausgehende Varianz gegenüber der Störgrößenvarianz eine dominantere Rolle gespielt hätte. Da bei kleinen Stichproben die Split-Methode gröbere Schätzfehler begeht als die Full-Methode, wäre also der von uns beobachtete Unterschied in

der Länge der Konfidenzintervalle lediglich schärfer in Erscheinung getreten, mit wachsenden Stichproben und unter Beibehaltung geeigneter Konsistenzbedingungen aber wieder kleiner geworden.

Literaturverzeichnis

Beran, R. (1992): Designing bootstrap prediction regions, in: *Bootstrapping and Related Techniques. Proceedings, Trier, FRG, 1990*, Hrsg. K.-H. Jöckel, G. Rothe und W. Sendler, Springer: 23-30.

Bottou, L./Bousquet, O. (2008): The tradeoffs of large scale learning, in: *Advances in Neural Information Processing Systems*, **20**, Hrsg. D. Koller, J.C. Platt, S. Roweis und Y. Singer, NIPS Foundation: 161-168.

Carnell, R. (2016): Package 'triangle'. Provides the standard distribution functions for the triangle distribution, in: `https://cran.r-project.org/web/packages/triangle`.

Chow, Y. S./Teicher, H. (1978): *Probability Theory. Independence, Interchangeability, Martingales*, Springer, New York.

Davison, A. C./Hinkley, D. V. (1997): *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

Efron, B. (1979): Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, **7**: 1-26.

Efron, B. (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, Philadelphia.

Efron, B./Tibshirani, R. (1993): *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Friedman, J./Hastie, T./Tibshirani, R. (2001): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer, New York.

Friedman, J./Hastie, T./Tibshirani, R. (2010): Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**: 1-22.

Friedman, J./Hastie, T./Tibshirani, R. (2016): Package 'glmnet'. Lasso and elastic-net regularized generalized linear models, in: `https://cran.r-project.org/web/packages/glmnet`.

G'Sell, M./Lei, J./Rinaldo, A./Tibshirani, R. J./Wasserman, L. (2016): Distribution-free predictive inference for regression, *ArXiv: 1604.04173*.

G'Sell, M./Lei, J./Rinaldo, A./Tibshirani, R. J./Wasserman, L. (2016): Package 'conformalInference'. Tools for conformal inference in regression, in: `https://github.com/ryantibs/conformal`.

Gammerman, A./Shafer, G./Vovk, V. (2005): *Algorithmic Learning in a Random World*, Springer, New York.

Hastie, T./James, G./Tibshirani, R./Witten, D. (2013): *An Introduction to Statistical Learning. with Applications in R*, Springer, New York.

Heuser, H. (2006): *Lehrbuch der Analysis. Teil 1*, Teubner, Wiesbaden.

Hinderer, K. (1972): *Grundbegriffe der Wahrscheinlichkeitstheorie*, Springer, Berlin.

Horowitz, J. L. (2001): The bootstrap, in: *Handbook of Econometrics*, **5**, Hrsg. J. J. Heckman und E. Leamer, Elsevier Science B. V.: 3160-3228.

Kusolitsch, N. (2011): *Maß- und Wahrscheinlichkeitstheorie. Eine Einführung*, Springer, Wien.

Mohri, M./Rostamizadeh, A./Talwalkar, A. (2012): *Foundations of Machine Learning*, MIT Press, Cambridge.

Shafer, G./Vovk, V. (2008): A tutorial on conformal prediction, in: *Journal of Machine Learning Research*, **9**: 371-421.

Stine, R. A. (1985): Bootstrap prediction intervals for regression, in: *Journal of the American Statistical Association*, **80**: 1026-1031.

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso, in: *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**: 267-288.

von Auer, L. (2003): *Ökonometrie. Eine Einführung*, Springer, Berlin.

Wooldridge, J. M. (2003): *Introductory Econometrics. A Modern Approach*, South-Western, Mason.

Abstract:

Deutsch:

Methoden zur Konstruktion *konformaler Konfidenzbereiche*, welche im Rahmen des sogenannten *On-Line Settings* angewendet werden, haben vor allem durch die Arbeit von Gammerman et al. (2005) sowie von Shafer und Vovk (2008) Bekanntheit erlangt. Die hier vorliegende Arbeit beschäftigt sich mit zwei besonderen Herangehensweisen: die Full- und die Split-Methode, welche beide in G'Sell et al. (2016) vorgestellt werden. Dabei wird der Schwerpunkt auf zwei Aspekte gelegt. Zum einen wird gezeigt, dass beide Methoden unter der Bedingung der *Vertauschbarkeit* der datengenerierenden Folge von Zufallsvariablen bei *fester* Stichprobengröße *gültige* Konfidenzbereiche hervorbringen, d.h. Konfidenzbereiche, deren *Überdeckungswahrscheinlichkeit* das vorgegebene *Konfidenzniveau* nicht unterschreitet und mit *wachsender* Stichprobe sich diesem sogar annähert. Zum anderen wird unter der stärkeren Annahme, dass die datengenerierende Folge aus *unabhängigen* und *identisch verteilten* Zufallsvariablen besteht, für beide Methoden die auf wachsende Stichproben Bezug nehmende *Asymptotik* untersucht, wobei das *Konvergenzverhalten* der *empirischen Überdeckungsrate* sowie der *Breite* der Konfidenzbereiche von Relevanz ist. Diese Untersuchung wird außerdem noch durch eine einfache *Simulation* eines On-Line Settings exemplarisch ergänzt. Im Verlauf dieser Arbeit zeigt sich, dass bezüglich rechentechnischer Umsetzung und Robustheit die Split-Methode gegenüber der Full-Methode einige Vorzüge aufzuweisen hat.

Englisch:

Methods to construct *conformal confidence regions*, which are applied within the framework of the so called *on-line setting*, have been popularised through Gammerman et al. (2005) as well as through Shafer and Vovk (2008). The work in hand deals with two particular approaches: the full- and the split-method, which are both introduced by G'Sell et al. (2016). The focus is laid on two issues. On the one hand, it is shown that under *exchangeability* of the data generating sequence of random variables and *fixed* sample size both methods generate *valid* confidence regions, i.e. confidence regions whose *coverage probability* does not fall short of the given *confidence level* and even converges towards it with *increasing* sample size. On the other hand, the asymptotics of both methods, which refers to increasing sample size, are examined with respect to the *convergence behaviour* of both the *empirical coverage rate* and the width of the relevant *confidence regions* under the stronger assumption that the data generating sequence consists of *independent* and *identically distributed* random variables. In addition, these examinations are supplemented by a simple simulation of an on-line setting framework. Throughout this work, it becomes apparent that the split method exhibits some advantages over the full method regarding computational realization and robustness.