Ein funktionaler Kernschätzer mit verallgemeinerter Bandbreite – Asymptotik und Bandbreitenwahl –

Dissertation

zur Erlangung des Grades

eines Doktors der Naturwissenschaften

der Universität Dortmund

Dem Fachbereich Statistik der Universität Dortmund

vorgelegt von

Rafael Pflüger

aus Haan/Rheinland

Erlangen 2001

Prüfungskommission:	Prof. Dr. W. Urfer	(Vorsitzender)
	Prof. Dr. O. Gefeller	(Gutachter)
	Prof. Dr. S. Schach	(Gutachter)
	Prof. Dr. C. Weihs	(Gutachter)
	Dr. P. Sibbertsen	(wissenschaftlicher Mitarbeiter)

Tag der mündlichen Prüfung: 15. Februar 2001

Danksagung

An dieser Stelle möchte ich allen danken, die mir bei dieser Arbeit geholfen haben.

Mein primärer Dank gilt Herrn Prof. Dr. Olaf Gefeller für die geeignete Themenstellung, die inspirierende Betreuung, sowie die zur Verfügung gestellten Computer- und Zeitressourcen.

Für wertvolle Diskussionen danke ich – in chronologischer Reihenfolge – Herrn Prof. Dr. Holger Dette, Frau Dr. Annette Pfahlberg, Herrn Dr. Matthias Land, Herrn Prof. Dr. Nils Lid Hjort, Herrn Ricercatore Andrea Ongaro, Herrn Prof. Dr. Ørnulf Borgan, Frau Dipl.-Math. Christine Vogel, Frau Dr. Daniela Schneider, Frau Prof. Dr. Claudia Czado und Frau Dipl.-Hum. Biol. Stefanie Thomas.

Für die Freigabe der Daten der Blasenkarzinom Studie des Ontario Cancer Institute, Toronto, und Anregungen zur praktischen Anwendung danke ich Frau Prof. Lillian Siu.

Die Arbeit wäre ohne die finanzielle Unterstützung der Deutschen Forschungsgemeinschaft im Rahmen des Projekts „Epidemiologische Risikokonzepte“ nicht möglich gewesen, wofür ich auch ihr danken möchte.

Für die Begeisterung danke ich außerordentlich meinen Eltern und für ihre Unterstützung meiner Freundin Dipl.-Math. Silke Weißbach.

Erlangen, März 2001

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Als ein frühes verbürgtes Zeugnis schreibt René Descartes (1637) über „...[die] Gesundheit, die ohne Zweifel das erste Gut ist und der Grund aller übrigen Güter dieses Lebens.“ So sind wir bestrebt, Unheil von diesem Gut abzuwenden. Das setzt voraus, dass wir uns ein Bild von den möglichen Risiken machen. Wir wollen eine Bewertung, besser eine Quantifizierung, der Risiken, bevor wir diese Information zielorientiert nutzen können. Am Beispiel der Krebserkrankung wird dieses Bestreben insbesondere deutlich in der Frage nach der Rekurrenz von Tumoren. Statistisch kann man diese Frage umformulieren in die Frage nach der Wahrscheinlichkeit der Rekurrenz, oder exakter nach der Wahrscheinlichkeit einer Rekurrenz zu einem bestimmten Zeitpunkt nach einer Therapie. Diese Wahrscheinlichkeit eines dichotomen Zielereignisses gilt es einzuschätzen.

Auf der Suche nach derartigen Risikoquantifizierungen sind wir von einer einheitlichen Definition der Lebensqualität weit entfernt. Das valideste und reliabelste Surogat zur Beurteilung ist die Lebensspanne, und so belegen auch aktuelle Zitate, dass „death [...] the primary (and ultimate) endpoint“ ist (Chuang-Stein & DeMasi (1998)). Die Wahrscheinlichkeit für dieses wiederum dichotome Ereignis gilt es zu quantifizieren und so ist insbesondere die momentane Sterbewahrscheinlichkeit – in ihrer statistischen Formulierung als *Hazardrate* – von vordringlicher Bedeutung. Die

Hazardrate ist das stetige Analogon zur momentanen Ausfallwahrscheinlichkeit und als Grenzwert der Wahrscheinlichkeit, dass ein Ereignis X zum Zeitpunkt t eintritt, definiert:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\left(X \in [t, t+\Delta t] \mid X \geq t\right).$$

Der Schätzung dieser in der Zeit variierenden Grenzwahrscheinlichkeit ist diese Arbeit gewidmet.

Grundsätzlich gibt es bei der statistischen Auswertung einer klinischen Studie nun zwei Möglichkeiten, die Hazardrate für das interessierenden Zielereignis aus Daten zu schätzen. Wenn man Vorwissen über das Ausfallsverhalten des Zielereignisses hat, kann man im Rahmen der parametrischen Modellierung der Hazardrate eine Verteilungsfamilie wählen, die zu den Daten gut passt. Dann können zum Beispiel mittels Maximum Likelihood Verfahren die Parameter der Verteilung geschätzt werden. Da es häufig aber keine Annahmen über den Verlauf der Hazardrate gibt, will ich nichtparametrisch schätzen. Es gibt in der nichtparametrischen Funktionalschätzung eine Vielzahl von Verfahren, eine frühe Monographie hierzu stellt Prakasa Rao (1983) dar. Insbesondere in den letzten beiden Jahrzehnten hat sich die Realisierbarkeit rechenintensiver Verfahren bemerkbar gemacht. Viele dieser Verfahren fallen unter den Begriff der Glättungsmethoden, zu denen Simonoff (1996) eine aktuelle Referenz darstellt. Eine etablierte Methode stellt die Kernglättung dar, deren Stand Wand & Jones (1995) in einer Monographie darstellen. Da diese Methode schon frühe Wurzeln hat, haben sich auch Varianten für die Hazardratenschätzung etabliert. Gefeller & Michels (1992a, 1992b) geben hierüber eine Übersicht und beschäftigen sich auch mit dem Problem der fehlerhaften Information durch Zensierung, das insbesondere in klinischen Studien zu Überlebenszeiten ein immanentes Problem darstellt. Wegen des Anwendungsbezugs werde auch ich dieses Problem nicht ausgrenzen. Somit stellt sich mir die Kernglättung als ein geeignetes Instrumentarium dar.

Die primäre methodische Aufmerksamkeit jeder Schätzmethodik gilt der Konsistenz also der Frage, ob und unter welchen Bedingungen der Schätzer, hier der Kern-

hazardratenschätzer, für größer werdende Fallzahlen gegen die „wahre“ Hazardrate konvergiert. Trotz des nichtparametrischen Anspruchs kommen wir hier allerdings nicht ohne Glattheitsannahmen für die zu betrachtenden Hazardrate aus, die in der medizinischen Anwendung als gegeben angesehen werden können. Zur Konsistenz der Hazardratenschätzung im zensierten Kontext gibt es einige jüngere Arbeiten mit verschiedener Beweismethodik bei unterschiedlichem Konvergenzverständnis. So erzielen Andersen et al. (1993) in der Martingaltheorie mit Zählprozessmodellierung punktweise Konvergenz in Wahrscheinlichkeit und erhält Zhang (1996) mit dem Gesetz vom iterierten Logarithmus eine punktweise fast sichere Konvergenz. Gefeller & Hjort (1998) belegen die Konsistenz bezüglich des „Visual Error Criterium“ . Ohne auf die Zensierungsproblematik einzugehen, aber für eine gewissen Abhängigkeitsstruktur der Beoachtungen beweisen Estévez-Pérez & Quintela-del-Río (1999) die gleichmäßige Konvergenz und asymptotische Normalität.

Die Hazardratenschätzung hat in ihrer Genese eine starke Beziehung zur Dichteschätzung (Singpurwalla & Wong (1983)), wobei im Kontext der Dichteschätzung allerdings zensierte Daten nur selten thematisiert werden. Historisch schließen sich in der nichtparametrischen Funktionalschätzung an das einfache Histogramm Analysen für die Kernglättungsidee zur Dichteschätzung an. Für die Dichte gibt es viele Konsistenzaussagen, zum Beispiel zeigt Parzen (1962) für einen Schätzer der Dichte, den Rosenblatt (1956) eingeführt hat, die punktweise Konsistenz für den zu erwartenden quadratischen Fehler (MSE), die gleichmäßige Konvergenz in Wahrscheinlichkeit und die (punktweise) asymptotische Normalität. Außerdem gibt er Schranken für den Bias und die Varianz an. Nadaraya (1965) greift diese Entwicklung auf und erweitert sie für die Regressionsschätzung. Silverman (1978) zeigt die schwache und die starke gleichmäßige Konvergenz des Parzen-Schätzers der Dichte und deren Rate $\sqrt{\frac{\log \frac{1}{b}}{nb}}$ bei fixer Bandbreite b. Marron & Tsybakov (1995) stellen eine neue Verlustfunktion, namentlich das „Visual Error Criterium“ für diese vor. Erste Übersichtsarbeiten zu dem Themenkomplex stellen Silverman (1986), Hall et al. (1987) und Izenman (1991) dar.

Trotz der zahlreichen asymptotischen Ergebnisse seit 1962 hat die nichtparame-

trische Funktionalschätzung kaum Eingang in die angewandte Statistik oder gar die Analyse konkreter Studien gefunden. Das liegt neben der Rechenintensität der Methoden, die ein Hinderungsgrund bis in die 80er Jahre gewesen sein mag, entscheidend an der Schwierigkeit, eine Bandbreite im Falle einer finiten Stichprobe zu wählen. Es gibt für Dichteschätzer im unzensierten Szenario und mit fixer Bandbreite eine Vielzahl an Arbeiten zur Bandbreitenwahl. Parzen (1962) bestimmt die theoretisch – unter Kenntnis der zu schätzenden Dichte – asymptotisch MISE-optimale Bandbreite. Deheuvels & Hominal (1980) etablieren die Äquivarianz der Bandbreite bezüglich affiner Transformation der Daten als notwendige Bedingung. Die daten-adaptiven Bandbreitenwahlen, die den MISE unverzerrt und mit Kreuz-Validierung schätzen, werden in Marron (1987) diskutiert. Den fließenden Übergang dieser Bandbreitenwahlen, die sich als zu variabel für den praktischen Gebrauch erwiesen, zu den Plug-in Bandbreitenwahlen stellt Scott & Terell (1987) dar. Hierbei wird der asymptotisch minimale MISE mit einem Pilot-Schätzer, dem Plug-in, geschätzt. Park & Marron (1990) bestimmen eine Bandbreitenwahl per Plug-in Schätzung der unbekannten Dichte mit minimaler asymptotischer Varianz. Hall et al. (1991) bestimmen eine Plug-in Bandbreite, die die maximale Konvergenzrate von $n^{-\frac{1}{2}}$ gegen die asymptotisch MISE-optimale realisiert. Park, Kim & Marron (1991) geben eine Übersicht über die Bandbreitenwahlen bei der Dichteschätzung mit fixer Bandbreite.

Auch für Schätzer der Hazardrate im zensierten Szenario und mit fixer Bandbreite existieren zahlreiche Arbeiten zur Bandbreitenwahl, derer ich hier aber nur zwei nennen möchte. Patil (1993) bestimmt die Bandbreite mit dem Ziel der Minimierung einer unverzerrten Schätzung des MISE durch Kreuz-Validierung und González-Manteiga et al. (1996) bestimmen rechenintensiv die Bandbreite per Bootstrap bezüglich eines zu erwartenden integrierten gewichteten quadratischen Fehlers (MIWSE).

Parallel zur fixen Bandbreite, die zum Beispiel von Parzen (1962) zur Dichteschätzung verwandt wurde und in dem überwiegenden Teil theoretischer Arbeiten wegen ihrer Einfachheit favorisiert wird, entwickelten sich Modifikationen, für die später auch die Konsistenz der Schätzer betrachtet wurde. Fix & Hodges (1951)

führten die nächste-Nachbarn Bandbreite ein. Loftsgaarden & Quesenberry (1965) verwenden sie als erste für die Dichteschätzung. Ralescu (1995) zeigt die punktweise fast sichere Konvergenz des nächste-Nachbarn Dichteschätzers. Hall & Marron (1995) erzielen eine Verbesserung der Bias-Konvergenz durch eine quasi-fixe Bandbreite. Die selben Raten erzielen Jones, Marron & Sheather (1995) mit ihrer Idee der Bias-Korrektur mittels Multiplikation mit einem Schätzer der Identität und durch wiederholte Schätzung der Dichte mit identischer Bandbreite. Für alternative Bandbreitendefinitionen bei der Hazardratenschätzung wenden Tanner & Wong (1984) die Idee der variablen Bandbreite auf die Bandbreitenselektion an.

Die Ergebnisse der Konsistenz- und Bandbreitenwahlarbeiten nutzend und – in Teilen – zusammenfassend, möchte ich in dieser Arbeit eine verallgemeinerte Darstellung für einen Funktionalschätzer erzielen, die die Konvergenz der Schätzung einer große Klasse von Funktionen und insbesondere einer großen Klasse von Bandbreitendefinitionen gewährleistet.

Was die Bandbreitenwahl für die Hazardratenschätzung im Kontext zensierter Daten anbetrifft, muss aber noch immer ein Mangel an praktischen und theoretisch fundierten Bandbreitenwahlen bemerkt werden. Die Verallgemeinerung nutzend, werde ich versuchen, diese Lücke zu schließen. Es wird sich herausstellen, dass die nächste-Nachbarn Bandbreite, als variable Bandbreite verstanden, hier viele Probleme behebt, so dass ich für sie Bandbreitenwahlen entwickeln werden.

1.2 Struktur der Arbeit

Die Arbeit gliedert sich in drei Bereiche. In den Kapiteln 2 und 3 wird ein verallgemeinerter Kernschätzer entwickelt und analysiert. In Kapitel 4 werden konkrete Spezifikationen dieses allgemeinen Schätzers aufgezeigt. Eine dieser Spezifikationen, ein Kernschätzer für die Hazardrate bei zensierten Beobachtungen, wird in den beiden folgenden Kapiteln 5 und 6 auf die Praxisrelevanz in einem biometrischen Beispiel und einer umfangreichen Simulationsstudie untersucht.

In Kapitel 2 wird eine neue, allgemeine Formulierung hergeleitet, die auf der einen Seite Dichte- und Hazardratenschätzung gemeinsam umfasst und auf der anderen Seite fixe und variable Bandbreiten zulässt. Außerdem berücksichtigt diese Verallgemeinerung auch die (Rechts-)Zensierung von Beobachtungen wie sie in der Überlebenszeitanalyse häufig anzutreffen ist.

Abschnitt 2.4 stellt die Konsistenz der Schätzung sicher und erzielt Konvergenzraten für die gleichmäßige Konvergenz.

Um die Fehler der Schätzung analytisch zu bewerten, werden im Abschnitt 2.5 der Bias, die Varianz und die asymptotische Verteilung des Schätzers ermittelt.

Da die Wahl der Kernfunktion ein bereits abgeschlossenes Thema in der nichtparametrischen Funktionalschätzung darstellt, werden im Abschnitt 2.6 nur die Ergebnisse für die im Kontext dieser Arbeit wichtigen Fragen wiedergegeben.

Wegen der Allgemeinheit der Schätzung müssen innovative Ideen der Bandbreitenwahl in Kapitel 3 verfolgt werden. Die Ergebnisse der Konvergenz aus Abschnitt 2.4 führen zu einer asymptotisch optimalen Bandbreitenwahl. In einem zweiten Versuch werden bestehende Bandbreitenwahlen durch eine additive Glättung verbessert, die sich in einer weiteren neuen Bandbreitenwahl niederschlägt.

In Kapitel 4 nähere ich mich zwei konkreten Spezifikationen der allgemeinen Formulierung auf unterschiedliche Weise. Die Dichteschätzung mit fixer Bandbreite ohne die Zensierungproblematik kann für weitere mathematische Analysen genutzt werden. So kann bei ihr die gängige Bandbreitenwahl angewandt werden, die das quadratische Risiko minimieren soll. Wegen der Einfachheit dieser Spezifikation können die zentralen Argumente der von mir verwandten Bandbreitenwahl, die den gleichmäßigen Fehler zu minimieren sucht, wiederholt werden. So lassen sich Unterschiede der Bandbreitenwahlen erklären. Es wird auf eine Anwendung der in Abschnitt 2.5 ausgeführten asymptotischen Verteilung eingegangen.

Bei der Hazardratenschätzung wird als variable Bandbreite die nächste-Nachbarn Bandbreite gewählt, da sie in ihrer Definition die Möglichkeit der Adaption an die beobachteten Daten bietet. Es werden die entwickelten Bandbreitenwahlen aus Kapitel 3 konkretisiert und eine weitere Bandbreitenwahl eingeführt, die explizit für

dieses Szenario von GEFELLER, PFLÜGER & BREGENZER (1996) aufbauend auf TANNER & WONG (1984) entwickelt wurde.

Im Kapitel 5 wird die Anwendbarkeit der Methodik in einer klinischen Überlebenszeitstudie anhand der Hazardratenschätzung illustriert und der Vergleich zur gängigen Methodik der KAPLAN-MEIER Analyse geführt. Die verschiedenen Bandbreitenwahlen werden angewandt und Unterschiede interpretiert.

In einer Simulationsstudie werden in Kapitel 6 Vor- und Nachteile der Bandbreitenwahlen bei der Hazardratenschätzung für zensierte Daten systematisch für die Familie der exponentiellen Weibull Verteilungen kontrolliert evaluiert.

Das abschließende Kapitel 7 fasst die wesentlichen Resultate dieser Arbeit knapp zusammen.

Der Anhang enthält die Ergebnisse der Simulationsstudie als Erwartungswertgraphiken und Verlustmaßzahlen sowie den Quelltext zu einer Hazardratenschätzung in SAS/IML und das Literaturverzeichnis.

Kapitel 2

Der allgemeine Kernschätzer

2.1 Historische Herleitung der Kernschätzung – die Kernglättung

Um Kernschätzverfahren allgemein zu motivieren, betrachten wir zunächst den Fall der Dichteschätzung in einem unzensierten Szenario, das heißt wir beobachten

$$X_1, \ldots, X_n \quad \text{u.i.v. stetige Zufallsvariablen jeweils mit Dichte } f(\cdot).$$

Als ersten „naiven“ Schätzer der Dichte kann man das Histogramm betrachten (siehe Abbildung 2.1). Hierbei stellen sich die folgenden praktischen und theoretischen Fragen.

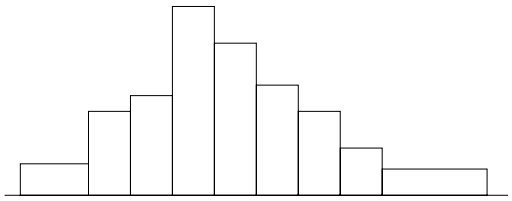


Abbildung 2.1: Skizze eines Histogramms

- Wie sind die Klassenbreiten zu wählen?
- Wie sind die Klassenmitten zu wählen?
- Wie „sicher“ ist die Schätzung?
- Wie sind die Konvergenzraten?

- Warum ist der Schätzer einer Dichte nicht „glatt“, wenn man vielleicht „Glattheit“ der Dichte vermutet, oder auch nur stetig, wie von der Dichte vorausgesetzt?

Um die Mängel zu beheben, die die zweite und letzte Frage andeuten, ist es naheliegend, einen „glatten“ Dichteschätzer zu suchen. Dazu stellen wir zunächst fest, dass die Punkte $X_i, i = 1, \dots n$, auch für ihre nähere Umgebung eine Wahrscheinlichkeitsmasse suggerieren. So kann auch ihre Masse – $1/n$ im vorliegenden Kontext – in eine Umgebung „verschmiert“ werden. Realisieren kann man das, indem man an jedes X_i ein von X_i unabhängiges Y_i addiert, das eine zufällige Schwankung um Null beschreibt. Dieses Y_i habe dann eine Dichte $f_Y(\cdot) = \frac{1}{b}K(\frac{\cdot}{b})$ mit Skalierungsparameter $b \in \mathbb{R}^+$ und Kernfunktion $K(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$ mit $\int_{\mathbb{R}} K(t)dt = 1$, was f_Y zu einer Dichte macht, und $\int_{\mathbb{R}} tK(t)dt = 0$, was äquivalent zu $EY = 0$ ist.

Die Dichte von $Z := X + Y$, wobei X Dichte $f_X := f$ habe, ist dann

$$f_Z(z) = \int_{\mathbb{R}} f_Y(z - y)f_X(y)dy = \int_{\mathbb{R}} \frac{1}{b}K(\frac{z - y}{b})dF_X(y).$$

Wenn wir nun die Dichte von Z aus den Beobachtungen $(X_i)_{i=1,\dots,n}$ schätzen wollen – bemerke, dass die Dichte von Y bekannt ist –, bietet sich das Lebesgue-Stieltjes-Integral

$$\hat{f}_Z(z) := \int_{\mathbb{R}} \frac{1}{b}K(\frac{z - y}{b})dF_n(y)$$

an, wobei die empirische Verteilungsfunktion

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}}(x)$$

die Verteilungsfunktion von X – wegen der Isomorphie zu den Daten trivialerweise suffizient – schätzt (Nussbaum (2000)). Mit dem Argument des „Verschmierens“ stellt dieser Schätzer auch einen Schätzer für die Dichte von X dar. Also halten wir als Kernschätzer für $f(x)$ fest:

$$f_n(x) := \int_{\mathbb{R}} \frac{1}{b}K(\frac{x - y}{b})dF_n(y). \tag{2.1}$$

Dieser PARZEN-Schätzer (PARZEN (1962)) löst zunächst auf jeden Fall das eine Problem des Histogramms, nämlich das der Wahl der Klassenmitten. Dass der Schätzer jetzt auch wie die wahre Dichte $f(\cdot)$ stetig ist, kann man sich leicht überlegen, wenn man an den Kern $K(\cdot)$ noch die Bedingung stellt, dass dieser zu den Rändern stetig auf Null abfällt. Bemerke, dass diese Eigenschaft für den Rechtecks-Kern als einzigen geläufigen (siehe GASSER, MÜLLER & MAMMITZSCH (1985)) nicht erfüllt ist. In dieser Form, mit Rechtecks-Kern, wurde die Dichteschätzung von ROSENBLATT (1956) vorgestellt, weswegen der Schätzer auch PARZEN-ROSENBLATT-Schätzer genannt wird.

Zur Veranschaulichung wollen wir uns das Lebesgue-Stieltjes-Integral $f_n(x)$, mit der Erkenntnis, dass $F_n(x)$ eine Treppenfunktion ist, als Summe anschauen:

$$f_n(x) = \sum_{i=1}^{n} \frac{1}{b} K(\frac{x - X_i}{b})(F_n(X_i) - F_n(X_i - 0)) = \frac{1}{bn} \sum_{i=1}^{n} K(\frac{x - X_i}{b}).$$

Beispielhaft sei nun der Dreiecks-Kern $K(\cdot) = 1_{[-\frac{1}{2},\frac{1}{2}]}(\cdot)(2 - |\cdot|)$ eingesetzt. Dann ist

$$f_n(x) = \frac{1}{bn} \sum_{i=1}^{n} 1_{[-\frac{1}{2},\frac{1}{2}]}(\frac{x - X_i}{b})(2 - |\frac{x - X_i}{b}|). \tag{2.2}$$

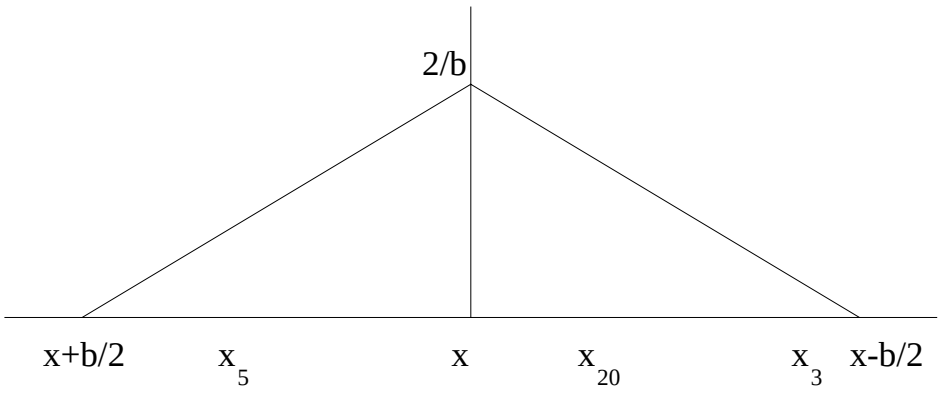


Abbildung 2.2: Dreiecks-Kern-Schätzung

An Formel (2.2) und der Abbildung 2.2 sieht man, dass es sich um das Zusammensammeln von „Massen“ aus der $b/2$-Umgebung um x handelt. Die Beobachtungen werden mit linear (bezüglich der Entfernung zu x) abfallenden Gewichten aufsummiert.

2.2 Eine allgemeine Bandbreite

Die **fixe** Bandbreite b, die unabhängig von der Stelle der Schätzung x und den Daten $X_1, \ldots, X_n$ ist, hat den Nachteil, dass in Regionen dichter Daten, also dort wo die Dichte $f(\cdot)$ groß ist, bei Benutzung einer großen Bandbreite b viele Daten zur Schätzung herangezogen werden, das heißt lokale Eigenschaften der Dichte werden durch die Faltung mit einem breiten Kern relativiert. In diesem Sinne wird hier „überglättet“ . Wählt man die Bandbreite aber klein, so werden in spärlich besetzten Regionen, also dort wo die Dichte $f(\cdot)$ klein ist, nur wenige Beobachtungen zur Schätzung an der Stelle x verwandt. Die Schätzung hängt dann stark an der Ausprägung dieser wenigen Daten und wird somit sehr variabel. Man kann, um dieser Diskrepanz Rechnung zu tragen, auf die Idee verfallen zu fordern, dass nicht in *konstant breiten* Fenstern Daten verwandt werden, sondern *konstant viele*. Zum Beispiel erscheint es plausibel, eine konstante Anzahl nächster Nachbarn (im anschaulichen Sinne oder axiomatisch wie in Minkowski (1907)) für die Bandbreitendefinition zu verwenden. Im vorliegenden Dichteschätzungskontext unzensierter Beobachtungen lässt sich diese Bandbreitendefinition offensichtlich über Ordnungsstatistiken realisieren. Wegen der einfacheren Verallgemeinerungsfähigkeit wähle ich hier aber eine Formalisierung, die in Gefeller & Dette (1992) verfolgt wird und auf der Erkenntnis beruht, dass die empirische Verteilungsfunktion $F_n(\cdot)$ in jeder Beobachtung um $1/n$ springt. So muss man zur Bestimmung einer k-nächste-Nachbarn Bandbreite in t eine Umgebung um t finden, in der $F_n(\cdot)$ um k/n springt, also k/n empirische Masse enthält. Sei also die k-nächste-Nachbarn Bandbreite definiert als

$$b^{NN}(t) = R^{NN}(t) = R_n^{NN}(t) := \inf\{r > 0 \mid |F_n(t - \frac{r}{2}) - F_n(t + \frac{r}{2})| \geq \frac{k}{n}\}.$$

Man veranschauliche das für $k = 4$ an der skizzenhaften Abbildung 2.3. Man be-

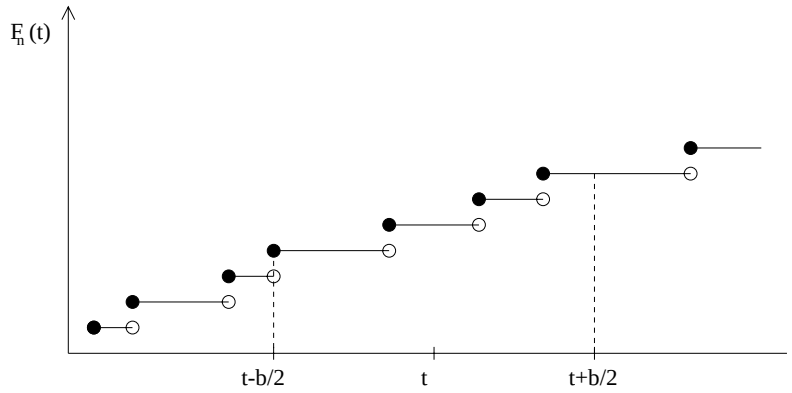


Abbildung 2.3: Die 4-nächste-Nachbarn Bandbreite von t

merke, dass b, beziehungsweise $R(t)$ jetzt nicht nur lokal-variiert, sondern auch von $F_n(\cdot)$ abhängig, also zufällig ist. Damit wird dem Balancierungsproblem zwischen vergrößertem Bias bei vielen zur Schätzung verwandten Beobachtungen und Varianz bei wenigen zur Schätzung verwandten Beobachtungen Rechnung getragen. Da $R_n(t) \approx \frac{k_n}{n} f^{-1}(x)$ ist (Dette & Gefeller (1995)) und damit in Regionen geringer Dichte eine größere Bandbreite verwandt wird als für dichtere Regionen, wird die Varianz verringert. Das ist natürlich nur eine relative Aussage zur fixen Bandbreite. Absolut kommt es auf die Anzahl nächster Nachbarn an, die die Bandbreite parametrisiert.

Man kann jetzt diese Definition von Bandbreiten verallgemeinern, so dass auch gemeinsam die fixe Bandbreite modelliert wird. Definiere

$$R_n(t) := \inf\{r > 0 \mid |\tilde{\Psi}_n(t - \frac{r}{2}) - \tilde{\Psi}_n(t + \frac{r}{2})| \geq p_n\}, \tag{2.3}$$

mit monotonem stochastischem „Glättungsprozess“ $\tilde{\Psi}_n(\cdot)$ und Bandbreitenparameter p_n. Dann ist für

- $\tilde{\Psi}_n(\cdot) = F_n(\cdot)$ mit $p_n = \frac{k}{n}$ die Bandbreite $R_n(\cdot)$ die k-nächste-Nachbarn Bandbreite und

- $\tilde{\Psi}_n(\cdot) = c \cdot id(\cdot) + d$ mit $p_n = |c| \cdot b$ die Bandbreite $R_n(\cdot) \equiv b$ die fixe Bandbreite.

Auch lässt sich in dieser Verallgemeinerung eine nächste-Nachbarn Bandbreite für das zensierte Modell, das ich im folgenden Abschnitt 2.4 genau definieren möchte, definieren. Man setze $\tilde{\Psi}_n(\cdot) = S_n(\cdot)$, den KAPLAN-MEIER-Schätzer der Überlebenszeitfunktion der X_i (siehe KAPLAN & MEIER (1958)), beziehungsweise äquivalent $\tilde{\Psi}_n(\cdot) = 1 - S_n(\cdot)$, was die Analogie noch deutlicher macht. Dass dieses spezielle Beispiel sinnvoll ist, haben GEFELLER & DETTE (1992) und DETTE & GEFELLER (1995) gezeigt.

Nun ist es anschaulich klar, dass wir, wenn wir $n \longrightarrow \infty$ gehen, aber p_n konstant (unabhängig von n) lassen, bestenfalls eine Konvergenz von $f_n(\cdot)$ gegen die Dichte von Z, $f_Z(\cdot)$, erwarten können. Vielmehr ist es so, dass man annehmen sollte, dass es für große n ausreichen sollte, kleinere Fenster zu benutzen. Allerdings ist es anschaulich nicht klar, wie schnell die Fenster (Bandbreite) kleiner werden dürfen. So werden wir im Folgenden sehen, dass es beispielsweise nicht ausreicht, eine feste Anzahl von nächsten Nachbarn zu wählen, was auch schon mit einer Verkleinerung der Fenster (Bandbreite) für große n einhergehen würde. Vielmehr wird sich herausstellen, dass die Anzahl der nächsten Nachbarn immer noch (und schneller als $\log n$) gegen ∞ gehen muss.

2.3 Der allgemeine Funktionalschätzer

Nun wollen wir zunächst einen weiteren wichtigen Schritt zur Verallgemeinerung machen. So ist es evident, dass man mit der „Faltungsmethode“ (= Kern-Glättung) nicht nur die Ableitung einer Verteilungsfunktion schätzen kann. Auch kann man zum Beispiel die Ableitung der kumulativen Hazardrate, die Hazardrate, auch Hazardfunktion genannt, selbst, schätzen. Man benötigt nur wieder einen Schätzer für die kumulative Hazardrate. Allgemein kann man jede Funktion $\psi(\cdot)$ schätzen, von deren Stammfunktion $\Psi(\cdot)$ man eine Schätzung $\Psi_n(\cdot)$ hat. Die Integration bezüglich $\Psi_n(\cdot)$ kann dann eventuell nicht mehr eine Lebesgue-Stieltjes Integration sein, sondern allgemeinere Formen der stochastischen Integration wie das ITÔ-Kalkül bedeuten (siehe ØKSENDAL (1995)). Wir wollen im Folgenden Bedingungen an die Konvergenz dieser stochastischen Prozesse $\Psi_n(\cdot)$ gegen $\Psi(\cdot)$, sowie an die Konver-

	Unzensiertes Szenario	Zensiertes Szenario
Verteilungs-funktion	Empirische Verteilungs-funktion	1-Kaplan-Meier-Schätzer
kumulative Ha-zardfunktion	vereinfachter Nelson-Aalen-Schätzer	Nelson-Aalen-Schätzer

Tabelle 2.1: Schätzer der kumulativen Funktionen

genz der Bandbreite $R_n(\cdot)$ (vielmehr der Prozesse $\tilde{\Psi}_n(\cdot)$) formulieren, unter denen die Konvergenz von $\psi_n(\cdot)$ gegen $\psi(\cdot)$, also insbesondere die Konsistenz des allgemeinen variablen Kernschätzers

$$\psi_n(x) := \int_{\mathbb{R}} \frac{1}{R_n(t)} K(\frac{x-t}{R_n(t)}) d\Psi_n(t) \tag{2.4}$$

bewiesen wird. Insbesondere werde ich mich auf monotone Prozesse $\Psi_n(\cdot) \to [0, \infty)$ beschränken, da diese Maße induzieren. Die Prozesse und zugehörigen Situationen, die auch schon im jetzigen Verlauf implizit aufgetaucht sind, sind in der Tabelle 2.1 angemerkt. Sowohl für unzensierte Daten als auch für zensierte (Überlebenszeit-)Daten können jeweils die Dichte und die Hazardrate geschätzt werden. Die Schätzung der Dichte wird ausführlich im Abschnitt 4.1 für unzensierte Beobachtungen besprochen. Im Abschnitt 4.2 wird die Hazardratenschätzung im zensierten Szenario diskutiert. Hierbei wird insbesondere auf die Bandbreitenwahl bei Nutzung der verallgemeinerten Bandbreite eingegangen. Damit werden zwei Felder der Tabelle 2.1 nicht explizit diskutiert. Bei der Diskussion der Fälle der Dichteschätzung bei Zensierung und der Hazardratenschätzung ohne Zensierung können aber keinen neuen Aspekte erschlossen werden. Die Schätzmethoden können analog angewandt werden. Entscheidend ist das Problem der Zensierung. Die Schätzmethodik im Fall vollständiger Daten wird anhand der Dichteschätzung demonstriert, wohingegen die Methodik im Fall von Zensierung für die Hazardrate untersucht wird. Diese Aufteilung ist deswegen natürlich, da die anderen beiden Fälle in der Praxis selten sind.

Es sei bemerkt, dass sich der nächste-Nachbarn Schätzer zum Beispiel der Dichte

$$f_n(x) := \int_{\mathbb{R}} \frac{1}{R_n^{NN}(x)} K(\frac{x-t}{R_n^{NN}(x)}) dF_n(t),$$

der die Bandbreite in Abhängigkeit der Stelle der Schätzung x und nicht der Beobachtung X_i definiert, nicht in die Verallgemeinerung einbetten lässt. Das ist aber auch nicht wünschenswert, da dieser die Pathologie aufweist, selbst keine Dichte mehr zu sein, also $\int f_n = 1$ nicht erfüllt (siehe BREIMAN, MEISEL & PURCELL (1977)). Er verschwindet nämlich nirgendwo. Das liegt an der Eigenschaft des Schätzers, in Regionen, in denen die Dichte tatsächlich verschwindet, noch Masse zu vermuten, die er von „weit entfernten" Regionen sammelt, da er für jedes x Masse von k nächsten Nachbarn kumuliert. Das ist in dieser Verallgemeinerung ausgeschlossen, denn

$$\begin{aligned}
\int_{\mathbb{R}} f_n(x)dx &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{R_n(t)} K(\frac{x-t}{R_n(t)}) dF_n(t)dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} K(x)dx dF_n(t) \\
&= \int_{\mathbb{R}} dF_n(t) = 1.
\end{aligned}$$

Für weitere Motivation siehe PATIL ET AL. (1994).

2.4 Asymptotik – stark gleichmäßig

Für den eingeführten Schätzer $\psi_n(\cdot)$ (2.4) gibt es keine Darstellung des mittleren integrierten quadratischen Fehlers (MISE - von „mean integrated squared error“) zwischen wahrer Funktion $\psi(\cdot)$ und Schätzer, da die Verzerrung (Bias) und die Varianz nicht mit Hilfe einer Taylorapproximation von $\psi(\cdot)$ eine einfache Darstellung finden. (Vergleiche Jones & Wand (1995) für die Darstellung des MISE für die fixe Kerndichteschätzung.) Da es allerdings auch keinen statistischen Grund gibt, warum der MISE als Diskrepanzkriterium prädestiniert ist, wähle ich den gleichmäßig absoluten Abstand (UAE – von „uniform absolute error“) zur Evaluation der Asymptotik, der dann die Asymptotik bezüglich des integrierten quadratischen Fehlers (ISE – von „integrated squared error“) impliziert. An dieser Stelle möchte ich anfügen, dass ich die Asymptotik des Verlusts evaluieren werde und nicht des Risikos, also des zu erwartenden Verlusts. Das hängt zum einen mit technischen Schwierigkeiten bei der Erwartungswertbildung zusammen, folgt aber auch den Ergebnissen von Devroye (1991), der für den Fall der Dichteschätzung den Risikoanteil einer additiven Abschätzung des Verlusts mit Risiko und zufälligem Rest als überwiegend herausstellt. Somit können technische Schwierigkeiten über die Wahl des Kriteriums entscheiden, ohne die Interpretation zu gefährden.

2.4.1 Das Modell

Für die Funktionen $\Psi : \mathbb{R} \to \mathbb{R}_0^+$ und $\tilde{\Psi} : \mathbb{R} \to \mathbb{R}_0^+$ gelte:

$$|\Psi(x) - \Psi(y)| \leq M|x-y|, \quad |\Psi(x) - \Psi(y)| \geq m|x-y| \qquad \text{und} \tag{2.5}$$

$$|\tilde{\Psi}(x) - \tilde{\Psi}(y)| \leq \tilde{M}|x-y|, \quad |\tilde{\Psi}(x) - \tilde{\Psi}(y)| \geq \tilde{m}|x-y| \tag{2.6}$$

$\forall x, y \in [A, B] \subset \mathbb{R}$ $(A < B)$ mit $0 < m \leq M < \infty$ und $0 < \tilde{m} \leq \tilde{M} < \infty$, woraus die Existenz der Ableitungen $\psi(\cdot)$ und $\tilde{\psi}(\cdot)$ mit $m \leq |\psi(x)| \leq M$ und $\tilde{m} \leq |\tilde{\psi}(x)| \leq \tilde{M}$ $\forall x \in [A, B]$ folgt. Weiter seien $\psi(\cdot)$ und $\tilde{\psi}(\cdot)$ Lipschitz-stetig auf $[A, B]$ mit Lipschitz-Konstanten L_ψ und $L_{\tilde{\psi}}$.

Nun gebe es Folgen rechtsstetiger monotoner stochastischer Prozesse $\Psi_n(x)$ und $\tilde{\Psi}_n(x)$ für die es $0 < D < \infty$ und $0 < \tilde{D} < \infty$ gibt, so dass

$$P\left(\lim_{n\longrightarrow\infty}\sup_{I\subset[A,B],\Psi(I)\leq p_n}\frac{|\Psi_n(I)-\Psi(I)|}{\sqrt{\frac{\log(n)p_n}{n}}}=D\right)=1 \tag{2.7}$$

$$P\left(\lim_{n\longrightarrow\infty}\sup_{I\in[A,B],\Psi(I)\leq p_n}\frac{|\tilde{\Psi}_n(I)-\tilde{\Psi}(I)|}{\sqrt{\frac{\log(n)p_n}{n}}}=\tilde{D}\right)=1 \tag{2.8}$$

mit

$$\begin{aligned} 0 < p_n \in \mathbb{R} < 1 \\ p_n &\longrightarrow 0 \\ \frac{np_n}{\log(n)} &\longrightarrow \infty. \end{aligned} \tag{2.9}$$

Die erste Konvergenzaussage der stochastischen Prozesse (2.7) ist äquivalent zu: für alle $C > D$ gilt

$$P\left(\sup_{I\subset[A,B]\Psi(I)\leq p_n}|\Psi_n(I)-\Psi(I)|\leq C\sqrt{\frac{\log(n)p_n}{n}}\text{ für grosse n}\right)=1. \tag{2.10}$$

Das gleiche gilt natürlich für die zweite Konvergenzaussage (2.8).

Diese Äquivalenz ist sichergestellt durch das folgende noch mehrmals gebrauchte

Theorem 2.4.1 *$(Y_i)_{i\in\mathbb{N}}$ sei eine Folge unabhängiger identisch verteilter Zufallsvektoren $Y_i : \Omega \to \mathbb{R}^d$.*
$S_n : (\mathbb{R}^d)^n \to \mathbb{R}^+$ sei eine Folge Borel-messbarer symmetrischer Abbildungen, das heißt $S_n(y_1,\ldots,y_n) = S_n(\sigma(y_1),\ldots,\sigma(y_n))$ für alle $y_1,\ldots,y_n \in \mathbb{R}^d$ und alle Permutationen $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$.
Sei (a_n) eine Nullfolge positiver reeller Zahlen.

Dann sind folgende Aussagen äquivalent:

- *Für alle $\alpha > 1$ ist $P\{\omega|\exists N \in \mathbb{N}\,\forall n > N :\ S_n(Y_1(\omega),\ldots,Y_n(\omega)) \leq \alpha a_n\} = 1$*

- *Es existiert eine Konstante* $D \leq 1$ *mit* $P\left\{\frac{\limsup_{n\to\infty} S_n(Y_1(\omega),\ldots,Y_n(\omega))}{a_n} = D\right\} = 1$

Für einen Beweis siehe SCHÄFER (1986A). Bemerke hier, dass $R_n(\cdot)$ auf $\mathbb{R}$ stetig ist, was (im Gegensatz zu $r_n(\cdot)$) nicht trivial ist, da $\tilde{\Psi}_n(\cdot)$ nicht stetig ist. Damit folgt die Borel-Messbarkeit von $\psi_n(\cdot)$, und damit ist sie an allen Stellen gegeben, wo sie gebraucht wird. Es wird jeweils auf einen eingehenden Nachweis verzichtet.

Der Ausdruck $\Psi_n(I)$ heißt „das von Ψ_n induzierte Maß von I“ und analog für $\tilde{\Psi}$. Konkret ist dann $\Psi_n(I) := \int_I |d\Psi_n(\xi)|$ beziehungsweise $\Psi(I) := \int_I |\psi(\xi)|d\xi$.

Um zu Beispielen zu kommen, wollen wir an dieser Stelle das Studiendesign und die Notation der in der Überlebenszeitanalyse häufigen *zensierten* Daten einführen.

Seien $T_1, \ldots, T_n$ u.i.v. mit Verteilungsfunktion $F(\cdot)$ und davon unabhängig $C_1, \ldots, C_n$ u.i.v. mit Verteilungsfunktion $G(\cdot)$. Tatsächlich beobachtet man aber $X_i := \min\{T_i, C_i\}$, $i = 1, \ldots, n$ und die *Zensierungsindikatoren* $\delta_i := 1_{\{X_i = T_i\}}$, $i = 1, \ldots, n$. Interessierend ist aber an dieser Stelle die Verteilungsfunktion $F(\cdot)$, die Überlebenszeitfunktion $S(\cdot) = 1 - F(\cdot)$ oder die kumulative Hazardrate $H(\cdot) = -\log S(\cdot)$. Schätzer hierfür weisen die geforderte lokale Konvergenzeigenschaft auf, so

- die empirische Verteilungsfunktion F_n im unzensierten unabhängigen Szenario mit Verteilungsfunktion F, mit $D = 9$,
- der KAPLAN-MEIER-Schätzer S_n der Überlebenszeitfunktion S (im zensierten Szenario) (siehe KAPLAN & MEIER (1958)) mit $D = 9(1 - G(B))^{-\frac{1}{2}}$,
- der NELSON-AALEN-Schätzer H_n der kumulativen Hazardfunktion H (im zensierten Szenario) (siehe ANDERSEN ET AL. 1993) mit $D = 9(1 - F^{obs}(B))^{-\frac{1}{2}}$ mit F^{obs} Verteilungsfunktion der beobachteten Zeiten (X_i) und
- die Funktion $J_n(t) = J(t) = c \cdot t + d$,

mit B als oberer Schranke eines Intervalls, auf das die Schätzung begrenzt werden muss. Für die Beweise, die auf Verschärfungen der Tschebyscheff-Ungleichung von BERNSTEIN (1924), BENNETT (1962)und HOEFFDING (1962) zurückgreifen, sei

für die Verteilungsfunktion F_n auf STUTE (1982A), für die Überlebenszeitfunktion S_n auf SCHÄFER (1986B) und für die kumulative Hazardrate H_n auf SCHÄFER (1986A) verwiesen. Für die Funktion J_n ist die Konvergenz trivial.

Weiter halten wir uns noch einmal die allgemeine Bandbreite

$$R_n(t) = \inf\{r > 0 || \tilde{\Psi}_n(t + \frac{r}{2} - 0) - \tilde{\Psi}_n(t - \frac{r}{2})| \geq p_n\} \tag{2.11}$$

vor Augen und definieren das deterministische Analogon

$$r_n(t) := \inf\{r > 0 || \tilde{\Psi}(t + \frac{r}{2}) - \tilde{\Psi}(t - \frac{r}{2})| \geq p_n\}. \tag{2.12}$$

In der Definition des allgemeinen Kernschätzers von $\psi(x)$ (2.4)

$$\psi_n(x) = \int_{\mathbb{R}} \frac{1}{R_n(t)} K(\frac{x-t}{R_n(t)}) d\Psi_n(t)$$

sei die Kernfunktion $K(\cdot)$, auch nur Kern genannt, eine stetige und stückweise – mit Konstante L_K – Lipschitz-stetige Dichtefunktion von endlicher Totalvariation $V(K)(:= \int_{\mathbb{R}} |dK|)$ und mit Träger $supp(K) \subset (-\frac{1}{2}, \frac{1}{2})$.

Falls $\Psi_n(\cdot)$ keine Sprungprozesse sind, bei denen es sich für $\psi_n(\cdot)$ um pfadweise Lebesgue-Stieltjes Integration handelt, verwende man das ITÔ–Kalkül zur Integrationsdefinition. In meinen Anwendungen aber kommen derartige Prozesse nicht vor.

2.4.2 Die Dreiecksungleichung

Schätze den L_∞-Abstand von $\psi_n(x)$ und $\psi(x)$ auf dem Intervall $[a, b] \subset (A, B)$ mit $a < b$ mit der Dreiecksungleichung ab:

$$\sup_{x \in [a,b]} |\psi_n(x) - \psi(x)| \leq \sup_{x \in [a,b]} |\psi_n(x) - \breve{\psi}_n(x)| \tag{2.13}$$

$$+ \sup_{x \in [a,b]} |\breve{\psi}_n(x) - \bar{\psi}_n(x)| + \sup_{x \in [a,b]} |\bar{\psi}_n(x) - \psi(x)|, \tag{2.14}$$

mit

$$\breve{\psi}_n(x) := \int_{\mathbb{R}} \frac{1}{r_n(t)} K(\frac{x-t}{r_n(t)}) d\Psi_n(t) \quad \text{und zu erwartendem Schätzer}$$
$$\bar{\psi}_n(x) := \int_{\mathbb{R}} \frac{1}{r_n(t)} K(\frac{x-t}{r_n(t)}) d\Psi(t).$$

Bemerkung: Nun läßt sich für den ersten Term mit Zählprozess-Methoden keine Konsistenz beweisen, da der sich ergebende Integrand nicht vorhersagbar ist. Wohl ließe sich diese Methodik aber für den zweiten Term anwenden. Als Literatur sei auf Andersen et al. (1993) und Fleming & Harrington (1991) verwiesen. Es sei aber darauf hingewiesen, dass die Konvergenzaussagen von Andersen et al (1993), insbesondere Theorem IV.2.2, nur gleichmäßig in Wahrscheinlichkeit gelten und keine Konvergenzgeschwindigkeitsaussagen machen. Zudem stellen sie stärkere Konvergenzanforderungen an die Bandbreitenfolge. Der dritte Term ist deterministisch und kann einfacher behandelt werden.

2.4.3 Der Term $|\psi_n(x) - \breve{\psi}_n(x)|$

Zunächst zeigen wir, dass nur in einem mit Rate p_n kleiner werdenden Interval um x Beobachtungen einen Beitrag zu $\psi_n(x)$ und $\tilde{\psi}_n(x)$ leisten.

Theorem 2.4.2 *Seien (2.6), (2.8) und (2.9) erfüllt, dann gelten für $\alpha > 1$,*

$$I_n(x) := \left[x - \frac{1}{2}\frac{p_n}{\tilde{m}}, x + \frac{1}{2}\frac{p_n}{\tilde{m}}\right] \quad \textit{und} \quad I_n^\alpha(x) := \left[x - \alpha\frac{p_n}{\tilde{m}}, x + \alpha\frac{p_n}{\tilde{m}}\right] :$$

$$\exists N \in I\!N \, \forall n > N : \sup_{x\in[a,b]} \sup_{t \in I\!R \setminus I_n(x)} K\left(\frac{t-x}{r_n(t)}\right) = 0 \tag{2.15}$$

$$P\left\{\exists N \in I\!N \, \forall n > N : \sup_{x\in[a,b]} \sup_{t \in I\!R \setminus I_n^\alpha(x)} K\left(\frac{t-x}{R_n(t)}\right) = 0\right\} = 1. \tag{2.16}$$

Beweis: Sei $A < A' < a$ und $b < B' < B$. Man wähle $N \in \mathbb{N}$ so, dass für alle $n > N$

$$\frac{\alpha p_n}{\tilde{m}} < \min\{|a - A'|, |b - B'|, |A' - A|, |B' - B|\} \tag{2.17}$$

und fast sicher

$$\sup_{x\in[a,b]} \left|\tilde{\Psi}_n \left[x, x + \alpha\frac{p_n}{\tilde{m}}\right] - \tilde{\Psi} \left[x, x + \alpha\frac{p_n}{\tilde{m}}\right]\right| \leq \tilde{D}\alpha^{\frac{3}{2}}\tilde{M}^{\frac{1}{2}}\tilde{m}^{-\frac{1}{2}}\sqrt{\frac{\log(n)p_n}{n}} \tag{2.18}$$

gilt, was wegen (2.10) ($C := \tilde{D}\alpha$) und $\sup_{x\in[a,b]} \tilde{\Psi}[x, x + \alpha\frac{p_n}{\tilde{m}}] \leq \alpha\frac{\tilde{M}}{\tilde{m}}p_n$ möglich ist und

$$\tilde{D}\alpha^{\frac{3}{2}}\tilde{M}^{\frac{1}{2}}\tilde{m}^{-\frac{1}{2}}\sqrt{\frac{\log(n)}{np_n}} < \alpha - 1, \tag{2.19}$$

was durch (2.9) garantiert wird.

zu (2.15): Sei $n > N$. Sei $x \in [a, b]$ und $t \notin I_n(x)$. Ohne Beschränkung der Allgemeinheit sei $t > x + \frac{1}{2}\frac{p_n}{\tilde{m}}$.

Fall 1: $t > B'$
Dann ist

$$\tilde{\Psi}[t - |t - x|, t + |t - x|] \geq \tilde{\Psi}[x, t] \geq \tilde{\Psi}[b, B'] \geq |b - B'|\tilde{m} \geq \alpha p_n > p_n$$

wegen (2.17) und (2.6). Gemäß der Definition von $r_n(t)$ folgt somit $r_n(t) \leq 2|t - x|$ also, dass $\frac{t-x}{r_n(t)} \geq \frac{1}{2}$ und somit nicht mehr im Träger von $K(\cdot)$ ist.

Fall 2: $t \leq B'$
Dann ist $[t-\frac{1}{2}\frac{p_n}{\tilde{m}}, t+\frac{1}{2}\frac{p_n}{\tilde{m}}] \subset [A, B]$ nach (2.17) und dann gilt nach (2.6) $\tilde{\Psi}[t-\frac{1}{2}\frac{p_n}{\tilde{m}}, t+\frac{1}{2}\frac{p_n}{\tilde{m}}] \geq p_n$. Dann folgt mit der Definition $r_n(t) \leq \frac{p_n}{\tilde{m}}$ und andererseits ist $t > x + \frac{1}{2}\frac{p_n}{\tilde{m}}$ vorausgesetzt und somit

$$2|t - x| > \frac{p_n}{\tilde{m}} \geq r_n(t),$$

woraus wieder

$$\frac{t-x}{r_n(t)} \notin Supp(K)$$

folgt.

zu (2.16): Sei $n > N$. Sei $x \in [a,b]$ und $t \notin I_n^\alpha$. Ohne Beschränkung der Allgemeinheit nehmen wir an $t > x + \alpha\frac{p_n}{\tilde{m}}$. Nach (2.17) und (2.6) ist $\tilde{\Psi}[x, x+\alpha\frac{p_n}{\tilde{m}}] \geq \alpha p_n$ und folglich

$$\begin{aligned}
&\tilde{\Psi}_n[x,t] \\
&\geq \tilde{\Psi}_n[x, x+\alpha\frac{p_n}{\tilde{m}}] \\
&= \tilde{\Psi}[x, x+\alpha\frac{p_n}{\tilde{m}}] + \tilde{\Psi}_n[x, x+\alpha\frac{p_n}{\tilde{m}}] - \tilde{\Psi}[x, x+\alpha\frac{p_n}{\tilde{m}}] \\
&\geq \alpha p_n - |\tilde{\Psi}_n[x, x+\alpha\frac{p_n}{\tilde{m}}] - \tilde{\Psi}[x, x+\alpha\frac{p_n}{\tilde{m}}]|.
\end{aligned}$$

Nun gilt mit (2.18) und (2.19)

$$\begin{aligned}
\tilde{\Psi}_n[x,t] &\geq \alpha p_n - \tilde{D}\alpha^{\frac{3}{2}}\tilde{M}^{\frac{1}{2}}\tilde{m}^{-\frac{1}{2}}\sqrt{\frac{\log(n)p_n}{n}} \\
&> \alpha p_n - (\alpha-1)p_n = p_n \qquad f.s.,
\end{aligned}$$

was wegen der Definition von $R_n(t)$ bedeutet, dass $R_n(t) \leq 2|t-x|$. Somit erhalten wir wieder $K(\frac{t-x}{R_n(t)}) = 0$. □

Nun gilt fast sicher für hinreichend große n (siehe (2.4.2)):

$$\begin{aligned}
|\psi_n(x) - \breve{\psi}_n(x)| &= \left| \int_{\mathbb{R}} \frac{1}{R_n(t)} K\left(\frac{x-t}{R_n(t)}\right) - \frac{1}{r_n(t)} K\left(\frac{x-t}{r_n(t)}\right) d\Psi_n(t) \right| \\
&= \left| \int_{I_n^\alpha(x)} \frac{1}{R_n(t)} K\left(\frac{x-t}{R_n(t)}\right) - \frac{1}{r_n(t)} K\left(\frac{x-t}{r_n(t)}\right) d\Psi_n(t) \right| \\
&\leq \Psi_n(I_n^\alpha(x)) \sup_{t \in I_n^\alpha(x)} \left| \frac{1}{R_n(t)} K\left(\frac{x-t}{R_n(t)}\right) - \frac{1}{r_n(t)} K\left(\frac{x-t}{r_n(t)}\right) \right| \qquad (2.20)
\end{aligned}$$

Für die Abschätzung des Integranden sei vorab noch ein Hilfssatz formuliert.

Theorem 2.4.3 *Die Voraussetzungen (2.6), (2.8) und (2.9) seien erfüllt und es sei* $[a,b] \subset (A,B)$.
Dann existiert eine Konstante $E \leq \frac{\tilde{D}}{\tilde{m}}$ *mit*

$$P\left\{\lim\sup_{n\longrightarrow\infty} \frac{\sup_{t\in[a,b]} |R_n(t) - r_n(t)|}{\sqrt{\frac{\log(n)p_n}{n}}} = E\right\} = 1 \tag{2.21}$$

Beweis: Gemäß 2.4.1 reicht es zu zeigen, dass für jedes $C > \tilde{D}$ und für große n fast sicher gilt

$$\sup_{t\in[a,b]} |R_n(t) - r_n(t)| \leq \frac{C}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}}.$$

Setze abkürzend $\varepsilon_n := \frac{C}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}}$. Sei C' eine Konstante mit $C > C' > \tilde{D}$. Betrachte für jedes $t \in [a,b]$

$$\begin{aligned} I_n^+(t) &:= \left[t - \frac{1}{2}(r_n(t) + \varepsilon_n), t + \frac{1}{2}(r_n(t) + \varepsilon_n)\right] \text{ und} \\ I_n^-(t) &:= \left[t - \frac{1}{2}(r_n(t) - \varepsilon_n), t + \frac{1}{2}(r_n(t) - \varepsilon_n)\right]. \end{aligned}$$

Bemerkung: Da

$$\begin{aligned} p_n &= \int_{t-\frac{r_n(t)}{2}}^{t+\frac{r_n(t)}{2}} |\tilde{\psi}(\xi)| d\xi \\ \Rightarrow \inf_{t\in[a,b]} r_n(t) \geq \frac{p_n}{\tilde{M}} &\quad \text{und } \sup_{t\in[a,b]} r_n(t) \leq \frac{p_n}{\tilde{m}}, \end{aligned} \tag{2.22}$$

für n hinreichend groß, da dann (wegen $r_n(t) \longrightarrow 0$) im gesamten Integrationsinterval gilt $\tilde{m} \leq |\tilde{\psi}(\cdot)| \leq \tilde{M}$.

Zusätzlich mit $\frac{\varepsilon_n}{p_n} \to 0$ (nach (2.9)) gilt, dass $r_n(t) - \varepsilon_n$ für hinreichend große n positiv ist.)

Es ist $\tilde{\Psi}(I_n^-(t)) \leq \tilde{\Psi}(I_n^+(t)) \leq p_n + \tilde{M}\varepsilon_n$ für n hinreichend groß, so dass $I_n^-(t) \subset I_n^+(t) \subset [A,B]$ $(r_n, \varepsilon_n \to 0)$. Und somit gilt wegen (2.10) und $\acute{p_n} := p_n + \tilde{M}\varepsilon_n$, welches wieder (2.9) erfüllt,

$$\tilde{\Psi}_n(I_n^+(t)) \geq \tilde{\Psi}(I_n^+(t)) - C'\sqrt{\frac{\log(n)(p_n + \tilde{M}\varepsilon_n)}{n}}$$
$$\tilde{\Psi}_n(I_n^-(t)) \leq \tilde{\Psi}(I_n^-(t)) + C'\sqrt{\frac{\log(n)(p_n + \tilde{M}\varepsilon_n)}{n}}$$

fast sicher für hinreichend große n und gleichmäßig für alle $t \in [a,b]$. Außerdem gilt

$$\begin{aligned}
\inf_{t\in[a,b]} \tilde{\Psi}(I_n^+(t)) &\geq \tilde{\Psi}\left[t - \frac{1}{2}r_n(t), t + \frac{1}{2}r_n(t)\right] + \tilde{m}\varepsilon_n = p_n + \tilde{m}\varepsilon_n \text{ und} \\
\sup_{t\in[a,b]} \tilde{\Psi}(I_n^-(t)) &\leq \tilde{\Psi}\left[t - \frac{1}{2}r_n(t), t + \frac{1}{2}r_n(t)\right] - \tilde{m}\varepsilon_n = p_n - \tilde{m}\varepsilon_n
\end{aligned}$$

und zusammen gilt fast sicher

$$\begin{aligned}
\inf_{t\in[a,b]} \tilde{\Psi}_n(I_n^+(t)) &\geq p_n + \tilde{m}\varepsilon_n - C'\sqrt{\frac{\log(n)(p_n + \tilde{M}\varepsilon_n)}{n}} \text{ und} \\
\sup_{t\in[a,b]} \tilde{\Psi}_n(I_n^-(t)) &\leq p_n - \tilde{m}\varepsilon_n + C'\sqrt{\frac{\log(n)(p_n + \tilde{M}\varepsilon_n)}{n}}
\end{aligned}$$

für große n. Nun ist auf der rechten Seite dieser Ungleichungen

$$\tilde{m}\varepsilon_n - C'\sqrt{\frac{\log(n)(p_n + \tilde{M}\varepsilon_n)}{n}} = \sqrt{\frac{\log(n)p_n}{n}}\left(C - C'\sqrt{1 + \frac{\tilde{M}\varepsilon_n}{p_n}}\right) > 0$$

für große n da $\frac{\varepsilon_n}{p_n} \longrightarrow 0$ nach (2.9) und da $C > C'$ gewählt wurde. Somit gelten

$$\begin{aligned}
\inf_{t\in[a,b]} &\quad \tilde{\Psi}_n\left[t - \tfrac{1}{2}(r_n(t) + \varepsilon_n), t + \tfrac{1}{2}(r_n(t) + \varepsilon_n)\right] > p_n \text{ und} \\
\sup_{t\in[a,b]} &\quad \tilde{\Psi}_n\left[t - \tfrac{1}{2}(r_n(t) - \varepsilon_n), t + \tfrac{1}{2}(r_n(t) - \varepsilon_n)\right] < p_n
\end{aligned}$$

was mit der Definition von $R_n(t)$ zu beweisen war.

Nun gilt für den zweiten Faktor in (2.20)

$$\begin{aligned}
&\left|\frac{1}{R_n(t)}K\left(\frac{x-t}{R_n(t)}\right)-\frac{1}{r_n(t)}K\left(\frac{x-t}{r_n(t)}\right)\right| \\
&\leq \left|\frac{1}{R_n(t)}K\left(\frac{x-t}{R_n(t)}\right)-\frac{1}{r_n(t)}K\left(\frac{x-t}{R_n(t)}\right)\right| \\
&\quad+\left|\frac{1}{r_n(t)}K\left(\frac{x-t}{R_n(t)}\right)-\frac{1}{r_n(t)}K\left(\frac{x-t}{r_n(t)}\right)\right| \\
&\leq \sup(K)\left|\frac{1}{R_n(t)}-\frac{1}{r_n(t)}\right|+\frac{\tilde{M}}{p_n}\left|K\left(\frac{x-t}{R_n(t)}\right)-K\left(\frac{x-t}{r_n(t)}\right)\right| \qquad (2.23)
\end{aligned}$$

wegen (2.22).

Nun gilt für den ersten Betrag-Term

$$\begin{aligned}
\sup_{t\in[a,b]}\left|\frac{1}{R_n(t)}-\frac{1}{r_n(t)}\right| &= \sup_{t\in[a,b]}\left|\frac{r_n(t)-R_n(t)}{R_n(t)r_n(t)}\right| \\
&\leq \tilde{D}\alpha^2\frac{\tilde{M}^2}{\tilde{m}}\sqrt{\frac{\log(n)}{(np_n^3)}} \qquad (2.24)
\end{aligned}$$

wegen

$$\inf_{t\in[a,b]} r_n(t)\geq\frac{p_n}{\tilde{M}},\quad \sup_{t\in[a,b]}|R_n(t)-r_n(t)|\leq\frac{C}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}}\text{ und}$$
$$\inf_{t\in[a,b]} R_n(t)\geq\frac{1}{\alpha}\frac{p_n}{\tilde{M}}.$$

$\forall C>\tilde{D}$ und somit auch für $C=\tilde{D}\alpha$. Die ersten beiden Ungleichungen ergeben sich aus (2.22) und 2.4.3. Die letzte Ungleichung begründet sich wie folgt:
Wegen 2.4.3 ist

$$R_n(t)\geq r_n(t)-\tilde{D}\frac{\alpha}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}}$$

für hinreichend großes n gleichmäßig auf $[a,b]$. Also ist insbesondere

$$\inf_{t\in[a,b]} R_n(t) \geq \inf_{t\in[a,b]} r_n(t) - \tilde{D}\frac{\alpha}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}}$$

und mit (2.22)

$$\begin{aligned}\inf_{t\in[a,b]} R_n(t) &\geq \frac{p_n}{\tilde{M}} - \tilde{D}\frac{\alpha}{\tilde{m}}\sqrt{\frac{\log(n)p_n}{n}} \\ &= \frac{p_n}{\tilde{M}}(1 - \tilde{D}\frac{\alpha\tilde{M}}{\tilde{m}}\sqrt{\frac{\log(n)}{p_n n}}) \\ &\longrightarrow \frac{p_n}{\tilde{M}}\end{aligned}$$

wegen (2.9). Für hinreichend großes n gilt somit, da $\alpha > 1$ ist, die Ungleichung. $\square$
Für den zweiten Betrag-Term gilt

$$\left|K\left(\frac{x-t}{R_n(t)}\right) - K\left(\frac{x-t}{r_n(t)}\right)\right| \leq k\left(\sup_{x\in[a,b],t\in I_n^\alpha(x)}\left|\frac{x-t}{R_n(t)} - \frac{x-t}{r_n(t)}\right|\right) \tag{2.25}$$

mit dem Stetigkeitsmodul $k(\cdot)$ vom Kern K

$$k(\delta) := \sup\{|K(x) - K(y)| : \ |x-y| \leq \delta\}.$$

Weiter gilt dann

$$\begin{aligned}\left|\frac{x-t}{R_n(t)} - \frac{x-t}{r_n(t)}\right| &= |x-t|\left|\frac{1}{R_n(t)} - \frac{1}{r_n(t)}\right| \\ &\leq \alpha\frac{p_n}{\tilde{m}}\tilde{D}\alpha^2\frac{\tilde{M}^2}{\tilde{m}}\sqrt{\frac{\log(n)}{(np_n^3)}}\end{aligned}$$

wegen $t \in I_n^\alpha(x)$ und obigen Überlegungen. Somit lässt sich das Stetigkeitsmodul wie folgt abschätzen,

$$\left| K\left(\frac{x-t}{R_n(t)}\right) - K\left(\frac{x-t}{r_n(t)}\right)\right| \leq \alpha^3 L_K \tilde{D} \frac{\tilde{M}^2}{\tilde{m}^2}\sqrt{\frac{\log(n)}{(np_n)}} \tag{2.26}$$

wenn man nachfolgende Bemerkung bedenkt.

Bemerkung:

Da es im Träger von K keinen Häufungspunkt von Lipschitz-Unstetigkeitsstellen gibt, gibt es für hinreichend kleines $\delta > 0$ in jedem Intervall der Länge δ nur eine Lipschitz-Unstetigkeitsstelle. Aus der Definition von $k(\cdot)$ folgt dann für diese δ

$$\begin{aligned} k(\delta) &= \sup\{|K(x) - K(y)| : \quad |x-y| \leq \delta\} \qquad (2.27)\\ &\leq \sup\{|K(x) - K(z_{LU})| + |K(z_{LU}) - K(y)| : \quad |x-y| \leq \delta\} \\ &\quad z_{LU} \text{ ist die potentielle Unstetigkeitsstelle zwischen } x \text{ und } y \\ &\leq \sup\{L_K|x - z_{LU}| + L_K|z_{LU} - y| : \quad |x-y| \leq \delta\} \\ &\quad \text{weil } K \text{ zwischen } x \text{ und } z_{LU} \text{ und } y \text{ und } z_{LU} \text{ lipschitz-stetig ist} \\ &= L_K\delta. \end{aligned}$$

Und wegen der Konvergenz des Prozesses

$$\sup_{I\subset[A,B]:\,\Psi(I)\leq \acute{p}_n} |\Psi_n(I) - \Psi(I)| \leq C\sqrt{\frac{\log(n)\acute{p}_n}{n}} \; f.s.$$

für $\acute{p}_n = Konstante \cdot p_n$ und alle $C > \tilde{D}$ und hinreichend großes n, gilt, da

$$\Psi(I_n^\alpha(x)) \leq 2\alpha\frac{p_n}{\tilde{m}}M =: \acute{p}_n \text{ (berücksichtige } |\psi| \leq M),$$

$$\begin{aligned}\Psi_n(I_n^\alpha(x)) - \acute{p}_n &\leq \Psi_n(I_n^\alpha(x)) - \Psi(I_n^\alpha(x)) \\ &\leq |\Psi_n(I_n^\alpha(x)) - \Psi(I_n^\alpha(x))| \\ &\leq C\sqrt{\frac{\log(n)2\alpha p_n M}{n\tilde{m}}}\end{aligned}$$

fast sicher für hinreichend großes n. Das heißt mit $C = \tilde{D}\alpha$

$$\Psi_n(I_n^\alpha(x)) \leq 2\alpha^2\frac{p_n}{\tilde{m}}M\left(\frac{1}{\alpha} + \tilde{D}\sqrt{\frac{\log(n)\tilde{m}}{np_n 2M\alpha}}\right) \leq 2\alpha^2\frac{M}{\tilde{m}}p_n, \tag{2.28}$$

weil die Klammer < 1 ist für hinreichend großes n, da der zweite Summand $\longrightarrow 0$ geht (2.9).

Wenn man nun (2.26) zusammen mit (2.24) in (2.23) eingesetzt, dann ergeben (2.23) und (2.28) in (2.20) eingesetzt:

$$\begin{aligned}|\psi_n(x) - \breve{\psi}_n(x)| &\leq 2\alpha^2\frac{M}{\tilde{m}}p_n\left[\sup(K)\tilde{D}\alpha^2\frac{\tilde{M}^2}{\tilde{m}}\sqrt{\frac{\log(n)}{np_n^3}} + \frac{\tilde{M}}{p_n}L_K\alpha^3\tilde{D}\frac{\tilde{M}^2}{\tilde{m}^2}\sqrt{\frac{\log(n)}{(np_n)}}\right] \\ &= \alpha^5 2\frac{M\tilde{M}^2}{\tilde{m}^2}\tilde{D}\left[\sup(K)\frac{1}{\alpha}\sqrt{\frac{\log(n)}{(np_n)}} + L_K\frac{\tilde{M}}{\tilde{m}}\sqrt{\frac{\log(n)}{(np_n)}}\right] \\ &\leq \alpha^5 2\tilde{D}\frac{M\tilde{M}^2}{\tilde{m}^2}\left(\sup(K)\tilde{m}^{-1} + L_K\frac{\tilde{M}}{\tilde{m}}\right)\sqrt{\frac{\log(n)}{(np_n)}}\end{aligned}$$

Unter Ausnutzung von 2.4.1 gilt somit, da $x \in [a,b]$ und $\alpha > 1$ (und somit auch $\alpha^5 > 1$) beliebig sind,

Theorem 2.4.4 *Für $\tilde{m}$ und $\tilde{M}$ definiert in (2.6) und $\breve{\psi}_n(\cdot)$ definiert in Abschnitt 2.4.2 gilt*

$$P\left\{\limsup_{n\longrightarrow\infty}\frac{\sup_{x\in[a,b]}|\psi_n(x) - \breve{\psi}_n(x)|}{\sqrt{\frac{\log(n)}{(np_n)}}} = D_1'\right\} = 1 \tag{2.29}$$

für

$$D_1' \leq 2\tilde{D}\frac{M\tilde{M}^2}{\tilde{m}^2}(\sup(K)\tilde{m}^{-1} + L_K\frac{\tilde{M}}{\tilde{m}})$$

mit $0 < \tilde{D} < \infty$ *definiert in Abschnitt 2.4.1.*

2.4.4 Der Term $|\breve{\psi}_n(x) - \bar{\psi}_n(x)|$

Theorem 2.4.5 *Mit* $\breve{\psi}_n(x)$ *und* $\bar{\psi}_n(x)$ *definiert in Kapitel 2.4.2 existiert eine Konstante*

$$D_2' \leq \frac{D\tilde{M}M^{\frac{1}{2}}V(K)}{\tilde{m}^{\frac{1}{2}}}$$

mit $0 < D < \infty$ *definiert in 2.4.1, so dass*

$$P\left\{\lim\sup_{n\longrightarrow\infty}\frac{\sup_{x\in[a,b]}|\breve{\psi}_n(x) - \bar{\psi}_n(x)|}{\sqrt{\frac{\log(n)}{(np_n)}}} = D_2'\right\} = 1. \tag{2.30}$$

Beweis: Sei nun $\alpha > 1$. Für hinreichend großes n gilt wegen 2.4.2

$$\begin{aligned}\left|\breve{\psi}_n(x) - \bar{\psi}_n(x)\right| &= \left|\int_{\mathbb{R}}\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right)d\Psi_n(t) - \int_{\mathbb{R}}\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right)d\Psi(t)\right| \\ &= \left|\int_{I_n(x)}\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right)d\Psi_n(t) - \int_{I_n(x)}\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right)d\Psi(t)\right|\end{aligned}$$

mit $I_n(x) = [x - \frac{p_n}{2\tilde{m}}, x + \frac{p_n}{2\tilde{m}}]$ gleichmäßig für alle $x \in [a, b]$.
Partielle Integration ergibt weiter mit SIRJAEV (1988) (Satz 11) (oder SCHÄFER (1986A))

$$\left|\breve{\psi}_n(x) - \bar{\psi}_n(x)\right| \leq V_{I_n(x)}\left(\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right)\right)\sup_{I Intervall\subset I_n(x)}|\Psi(I) - \Psi_n(I)| \tag{2.31}$$

da $\frac{1}{r_n(t)}K(\frac{t-x}{r_n(t)})$ stetig in t auf $I_n(x)$ für hinreichend große n ist, wegen des folgenden Satzes.

Theorem 2.4.6 *Mit $\tilde{\Psi}(\cdot)$ (2.6) erfüllend gilt für $s,t \in [A,B]$ die Lipschitzbedingung*

$$|r_n(t) - r_n(s)| \leq 2|s-t|.$$

Beweis:
Sei ohne Einschränkung $t > s$, somit gilt

$$\left[s - \frac{r_n(s)}{2}, t + \frac{r_n(s)}{2}\right] \supset \left[s - \frac{r_n(s)}{2}, s + \frac{r_n(s)}{2}\right],$$

und also

$$\tilde{\Psi}\left[s - \frac{r_n(s)}{2}, t + \frac{r_n(s)}{2}\right] \geq p_n$$

nach Definition von $r_n(s)$ und damit

$$\begin{aligned} \frac{r_n(t)}{2} &\leq \sup\left\{\left|t - (s - \frac{r_n(s)}{2})\right|, \left|t - (t + \frac{r_n(s)}{2})\right|\right\} \\ &= |t-s| + \frac{r_n(s)}{2}. \end{aligned}$$

Mittels $\left[s - \frac{r_n(t)}{2}, t + \frac{r_n(t)}{2}\right]$ folgt analog

$$r_n(s) \leq 2|t-s| + r_n(t),$$

was den Beweis beendet.

Wir wollen nun zunächst einige Aussagen treffen, die für die Abschätzung der Variation von $\frac{1}{r_n(t)}K(\frac{t-x}{r_n(t)})$ auf $I_n(x)$ nötig sind.

Theorem 2.4.7 *Die Voraussetzungen für $\tilde{\Psi}$ und $\tilde{\psi}$ aus Kapitel 2.4.1 gelten mit den dort definierten Konstanten $A, B, \tilde{M}, \tilde{m}$ und $L_{\tilde{\psi}}$. Sei $[a,b] \subset (A,B)$, dann gilt*

$$|r_n(s) - r_n(t)| \leq \frac{L_{\tilde{\psi}}}{\tilde{m}^2} p_n |s-t| \quad \forall s,t, \in [a,b].$$

Beweis:

$$\begin{aligned}
\left|\tilde{\Psi}\left[t-\frac{r}{2},t+\frac{r}{2}\right]-p_n\right| &= \left|\tilde{\Psi}\left[t-\frac{r}{2},t+\frac{r}{2}\right]-\tilde{\Psi}\left[t-\frac{r_n(t)}{2},t+\frac{r_n(t)}{2}\right]\right| \\
&= \left|\int_{t-\frac{r}{2}}^{t-\frac{r_n(t)}{2}}|\tilde{\psi}(\xi)|d\xi+\int_{t+\frac{r_n(t)}{2}}^{t+\frac{r}{2}}|\tilde{\psi}(\xi)|d\xi\right| \\
&\geq \tilde{m}|r_n(t)-r| \\
\Longrightarrow |r_n(t)-r_n(s)| &\leq \frac{1}{\tilde{m}}\left|\int_{t-\frac{r_n(s)}{2}}^{t+\frac{r_n(s)}{2}}|\tilde{\psi}(\xi)|d\xi-\int_{s-\frac{r_n(s)}{2}}^{s+\frac{r_n(s)}{2}}|\tilde{\psi}(\xi)|d\xi\right| \\
&\leq \frac{1}{\tilde{m}}\left|\int_{s-\frac{r_n(s)}{2}}^{s+\frac{r_n(s)}{2}}(|\tilde{\psi}(\xi-s+t)|-|\tilde{\psi}(\xi)|)d\xi\right| \\
&\leq \frac{r_n(s)}{\tilde{m}}L_{\tilde{\psi}}|t-s| \\
&\leq \frac{p_n}{\tilde{m}^2}L_{\tilde{\psi}}|t-s|,
\end{aligned}$$

da $\tilde{\psi}(\cdot)$ von 0 getrennt ist und somit auch $|\tilde{\psi}(\cdot)|$ lipschitzstetig mit identischer Konstante $L_{\tilde{\psi}}$ ist, und (2.22). □

Theorem 2.4.8 *Unter den Voraussetzungen des vorangegangenen Satzes gilt für großes n, dass die Funktion*

$$\frac{\cdot - x}{r_n(\cdot)} : t \to \frac{t-x}{r_n(t)}$$

monoton wachsend auf $I_n(x)$ ist.

Zum Beweis siehe SCHÄFER (1986A).

Theorem 2.4.9

- *Für $g, h : [a,b] \longrightarrow \mathbb{R}$ ist*

$$V_a^b(gh) \leq \sup|g| V_a^b(h) + \sup|h| V_a^b(g).$$

- *Für monoton nicht fallendes $h : [a,b] \longrightarrow [c,d]$ und beliebiges $g : [c,d] \longrightarrow \mathbb{R}$*

ist

$$V_a^b(g(h)) \leq V_c^d(g).$$

Für einen Beweis siehe SCHÄFER (1986A).

Wegen der vorangegangenen Sätze und (2.22) gilt nun für $s, t \in I_n(t)$ und n hinreichend groß

$$\left| \frac{1}{r_n(t)} - \frac{1}{r_n(s)} \right| \leq L_{\tilde{\psi}} \tilde{m}^{-2} \tilde{M}^2 p_n^{-1} |s - t|$$

und daher

$$\sup_{x \in [a,b]} V_{I_n(x)} \left(\frac{1}{r_n} \right) \leq L_{\tilde{\psi}} \tilde{m}^{-3} \tilde{M}^2.$$

Wegen des letzten Satzes kann nun die Variation der Funktion

$$\frac{1}{r_n(\cdot)} K \left(\frac{\cdot - x}{r_n(\cdot)} \right) : t \to \frac{1}{r_n(t)} K \left(\frac{t - x}{r_n(t)} \right)$$

für hinreichend großes n mit Satz 2.4.8 und 2.4.2 abgeschätzt werden durch

$$\begin{aligned} V_{I_n(x)} \left(\frac{1}{r_n(\cdot)} K \left(\frac{\cdot - x}{r_n(\cdot)} \right) \right) &\leq \frac{\tilde{M}}{p_n} V_{(-1,1)}(K) + \sup(K) L_{\tilde{\psi}} \frac{\tilde{M}^2}{\tilde{m}^3} \\ &\leq \alpha \tilde{M} V_{(-1,1)}(K) p_n^{-1} \end{aligned}$$

für große n wegen $p_n \longrightarrow 0$. Außerdem gilt wegen (2.7) fast sicher für hinreichend große n

$$\sup_{I Intervall \subset I_n(x)} |\Psi(I) - \Psi_n(I)| \leq C \sqrt{\frac{\log(n) M p_n}{n \tilde{m}}}$$

für alle $C > D$ wegen $\sup_{x\in[a,b]} \Psi(I_n(x)) \leq Mp_n/\tilde{m}$ und (2.10). Mit $C := D\alpha$ folgt dann für beliebiges $x \in [a,b]$ wegen (2.31)

$$\begin{aligned}\left|\breve{\psi}_n(x) - \bar{\psi}_n(x)\right| &\leq \alpha\tilde{M}V(K)p_n^{-1}D\alpha\sqrt{\frac{\log(n)Mp_n}{n\tilde{m}}} \\ &= \alpha^2\frac{D\tilde{M}M^{\frac{1}{2}}V(K)}{\tilde{m}^{\frac{1}{2}}}\sqrt{\frac{\log(n)}{np_n}}\end{aligned}$$

und mit Satz 2.4.1 wegen $\alpha > 1$ beliebig folgt dann Satz 2.4.5.

2.4.5 Der Term $|\bar{\psi}_n(x) - \psi(x)|$

$$\begin{aligned}|\bar{\psi}_n(x) - \psi(x)| &= |\int_{\mathbb{R}} \frac{1}{r_n(t)}K(\frac{t-x}{r_n(t)})d\Psi(t) - \int_{\mathbb{R}} \frac{1}{r_n(x)}K(\frac{t-x}{r_n(x)})\psi(x)dt| \\ &\leq \int_{I_n^1(x)} |\frac{1}{r_n(t)}K(\frac{t-x}{r_n(t)}) - \frac{1}{r_n(x)}K(\frac{t-x}{r_n(x)})||d\Psi(t)| \qquad (2.32) \\ &+ \int_{I_n^1(x)} \frac{1}{r_n(x)}K(\frac{t-x}{r_n(x)})|\psi(t) - \psi(x)|dt \qquad (2.33)\end{aligned}$$

für $I_n(x) \subset I_n^1(x) := [x - \frac{p_n}{\tilde{m}}, x + \frac{p_n}{\tilde{m}}]$ und n hinreichend groß wegen Satz 2.4.2.

Nun ist (2.33) kleiner als

$$\int_{I_n^1(x)} \frac{1}{r_n(x)}K\left(\frac{t-x}{r_n(x)}\right)dt \sup_{t\in I_n^1(x)} |\psi(t) - \psi(x)| \leq 1 \cdot L_\psi \frac{p_n}{\tilde{m}}$$

wegen der Lipschitz-Stetigkeit von $\psi(\cdot)$.

Und für den Integranden von Term (2.32) gilt für $t \in I_n^1(x)$ und gleichmäßig in $x \in [a,b] \in (A,B)$

$$\begin{aligned}
\left|\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right) \right. &- \left. \frac{1}{r_n(x)}K\left(\frac{t-x}{r_n(x)}\right)\right| \\
&\leq \left|\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(t)}\right) - \frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(x)}\right)\right| \\
&+ \left|\frac{1}{r_n(t)}K\left(\frac{t-x}{r_n(x)}\right) - \frac{1}{r_n(x)}K\left(\frac{t-x}{r_n(x)}\right)\right| \\
&= \frac{1}{r_n(t)}\left|K\left(\frac{t-x}{r_n(t)}\right) - K\left(\frac{t-x}{r_n(x)}\right)\right| \\
&+ K\left(\frac{t-x}{r_n(x)}\right)\left|\frac{1}{r_n(t)} - \frac{1}{r_n(x)}\right| \\
&\leq \frac{\tilde{M}}{p_n}k\left(\sup_{x\in[a,b],t\in I_n^1(x)}\left|\frac{t-x}{r_n(t)} - \frac{t-x}{r_n(x)}\right|\right) \quad \text{(wegen (2.22))} \\
&+ \sup(K)\frac{\tilde{M}^2}{p_n^2}L_{\tilde{\psi}}p_n\tilde{m}^{-2}|x-t| \quad \text{(wegen (2.22) und Satz 2.4.7)} \\
&\leq \frac{\tilde{M}}{p_n}k\left(\frac{p_n}{\tilde{m}}L_{\tilde{\psi}}\frac{\tilde{M}^2}{\tilde{m}^3}\right) \quad \text{(wieder wegen (2.22) und Satz 2.4.7)} \\
&+ \sup(K)L_{\tilde{\psi}}\frac{\tilde{M}^2}{\tilde{m}^3} \\
&= \frac{\tilde{M}}{p_n}k(L_{\tilde{\psi}}\frac{\tilde{M}^2}{\tilde{m}^4}p_n) + \sup(K)L_{\tilde{\psi}}\frac{\tilde{M}^2}{\tilde{m}^3}
\end{aligned}$$

Bemerke, dass es $A < \acute{a} < a$ und $b < \acute{b} < B$ gibt, so dass wegen $p_n \longrightarrow 0$ t und x in $[\acute{a}, \acute{b}]$ liegen und (2.22) bzw. Satz 2.4.7 für hinreichend große n auch wirklich angewandt werden können. Außerdem gilt wegen (2.5)

$$\Psi(I_n^1(x)) \leq 2\frac{M}{\tilde{m}}p_n.$$

Somit ist

$$|\bar{\psi}_n(x) - \psi(x)| \leq 2\frac{\tilde{M}M}{\tilde{m}}L_K L_{\tilde{\psi}}\frac{\tilde{M}^2}{\tilde{m}^4}p_n + 2\sup(K)L_{\tilde{\psi}}\frac{\tilde{M}^2 M}{\tilde{m}^4}p_n + L_\psi\tilde{m}^{-1}p_n$$

wegen 2.27, woraus folgt

$$\lim \sup_{n \longrightarrow \infty} \frac{\sup_{x \in [a,b]} |\bar{\psi}_n(x) - \psi(x)|}{p_n} \leq D_3, \tag{2.34}$$

mit

$$D_3 = \frac{2\tilde{M}^3 M L_K L_{\tilde{\psi}}}{\tilde{m}^5} + \frac{2 \sup(K) L_{\tilde{\psi}} \tilde{M}^2 M}{\tilde{m}^4} + L_\psi \tilde{m}^{-1}.$$

2.4.6 Resultierende Konvergenzrate für $|\psi_n(x) - \psi(x)|$

Mittels der Dreiecksungleichung und der vorangegangenen drei Abschnitte folgt die Konvergenzrate von $|\psi_n(x) - \psi(x)|$ als Summe von Konvergenzraten.

Theorem 2.4.10 *Für stetige und stückweise Lipschitz-stetige Kernfunktion (mit Konstante L_K) von beschränker Totalvariation existiert eine Konstante $D_0 \leq \max\{D_1 + D_2, D_3\}$ mit*

$$P\left\{ \lim \sup_{n \longrightarrow \infty} \frac{\sup_{x \in [a,b]} |\psi_n(x) - \psi(x)|}{\sqrt{\frac{\log(n)}{(np_n)}} + p_n} = D_0 \right\} = 1 \qquad \forall\, [a,b] \in (A,B), \tag{2.35}$$

wobei

- $D_1 := 2\tilde{D}\frac{M\tilde{M}^2}{\tilde{m}^2}(\sup(K)\tilde{m}^{-1} + L_K \frac{\tilde{M}}{\tilde{m}})$,
- $D_2 := \frac{D\tilde{M}M^{\frac{1}{2}}V(K)}{\tilde{m}^{\frac{1}{2}}}$ *und*
- $D_3 := \frac{2\tilde{M}^3 M L_K L_{\tilde{\psi}}}{\tilde{m}^5} + \frac{2\sup(K) L_{\tilde{\psi}} \tilde{M}^2 M}{\tilde{m}^4} + L_\psi \tilde{m}^{-1}$

sind.

Beweis:

Fast sicher gilt für hinreichend große n wegen Satz 2.4.4, Satz 2.4.5 und (2.34)

$$\begin{aligned}
\sup_{x\in[a,b]} |\psi_n(x) - \psi(x)| &\leq \sup_{x\in[a,b]} |\psi_n(x) - \breve{\psi}_n(x)| \\
&+ \sup_{x\in[a,b]} |\breve{\psi}_n(x) - \bar{\psi}_n(x)| \\
&+ \sup_{x\in[a,b]} |\bar{\psi}_n(x) - \psi(x)| \\
&\leq D_1 \left(\sqrt{\frac{\log(n)}{(np_n)}}\right) + D_2 \left(\sqrt{\frac{\log(n)}{(np_n)}}\right) + D_3 p_n \\
\leq \quad + \quad & \max\{D_1 + D_2, D_3\} \left(\sqrt{\frac{\log(n)}{(np_n)}} + p_n\right).
\end{aligned}$$

□

Es sei darauf hingewiesen, dass die Wahl von $[A, B]$ Teilmenge des Schnitts beider Träger, für Schätzfunktion $\psi(\cdot)$ und Glättungsfunktion $\tilde{\psi}(\cdot)$, sein muss. Praktisch ist das unproblematisch. Wenn nicht der Träger der Glättungsfunktion Obermenge dessen der Schätzfunktion ist, wie bei der fixen Bandbreite, sind sie in der Regel, aber insbesondere in dem mich interessierenden Beispiel bei Hazardrate und Dichte, gleich.

Für die Konvergenz des MISE betrachte zunächst den MIWSE mit Gewichtsfunktion $g(\cdot) = id_{[a,b]}(\cdot)$. Dann konvergiert der MIWSE, weil er kleiner ist als der maximale Abstand auf dem Intervall $[a, b]$ mit $b - a$ multipliziert. Dieses Produkt geht gegen Null. Da wir die Intervallgrenzen a und b beliebig nahe an die Trägergrenzen setzen können, konvergiert dann auch der MISE, falls $b \neq \infty$ - bemerke, dass $a > 0$ ist. Für unendlichen Träger macht schon das Beispiel der Hazardratenschätzung mit fixer Bandbreite für die Exponentialverteilung wegen des unbeschränken MISE – siehe Abschnitt 4.2.1.1 klar, dass das nicht allgemein zu erwarten ist.

2.5 Bias, Varianz und asymptotische Verteilung

Da die gleichmäßige Konvergenz als Kriterium für das Abweichen von Funktion und Schätzer gewählt wurde, haben Bias und Varianz nicht dieselbe vordringliche Bedeutung wie im Fall der MISE-orientierten Analyse. Dennoch möchte ich kurz auf asymptotische Darstellungen derselben eingehen. Da auch der Nutzen der asymptotischen Verteilung als von praktisch geringer Relevanz angesehen werden muss, soll hier auch nur kurz auf mögliche Entwicklungen eingegangen werden.

2.5.1 Bias des Schätzers

Der Bias (die Verzerrung) ist bei Kernschätzungen von besonderem Interesse, da er ein immanenter Fehler der Kernidee ist. Für den vorliegenden Schätzer (2.4) ist nicht klar, was der Erwartungswert von $\psi_n(x)$ ist. Bei der Annahme der deterministischen Bandbreite $r_n(t)$ (2.12) ist er

$$E(\psi_n(x)) = \int_{\mathbb{R}} \frac{1}{r_n(t)} K(\frac{x-t}{r_n(t)}) \psi(t) dt,$$

für erwartungstreues $\Psi_n(\cdot)$. Der approximative Ersatz von $r_n(t) \approx \frac{p_n}{\tilde{\psi}(t)} \approx \frac{p_n}{\tilde{\psi}(x)}$ wegen (2.22) gibt eine Vorstellung von dem Erwartungswert,

$$Bias \approx \frac{1}{2} \left(\frac{p_n}{\tilde{\psi}(x)} \right)^2 f''(x) \mu_2(K),$$

wobei $\mu_2(K) := \int_{\mathbb{R}} K(z) z^2 dz$ ist. Er ähnelt dem des fixen Kernschätzers – siehe Abschnitt 4.1.1–, nur dass dort b statt $\frac{p_n}{\tilde{\psi}(x)}$ steht. Wenn man nun die Proportionalität von b und p_n berücksichtigt – siehe Beispiel zu (2.3) –, bedeutet die Streckung mit dem Faktor $(\tilde{\psi}(x))^{-1}$ eine Vergrößerung der Bandbreite, falls $0 < \tilde{\psi}(x) < 1$ ist, wie im Fall $\tilde{\psi}(x) = f(x)$ gegeben. Dieses führt bekanntermaßen zu einer Vergrößerung des Bias. Da hier allerdings noch keine Bandbreitenwahl p_n getroffen wurde, ist diese Überlegung gegenstandslos, sobald die Bandbreitenwahl für den verallgemeinerten Schätzer selbst vorgenommen wird und nicht eine fixe Bandbreite als p_n eingesetzt

wird. Wenn man sich vorstellt, dass $\tilde{\psi}(x)$ variiert, wie zum Beispiel die Dichte zwischen 0 und 1, dann wird bei größeren Werten von $\tilde{\psi}(x)$ der Bias tendenziell kleiner sein als für Werte x mit kleineren $\tilde{\psi}(x)$. Im Fall $\tilde{\psi}(x) = f(x)$ bedeutet das eine Reduktion des Bias in dichten Regionen und eine Vergrößerung (also Reduktion der Varianz) in schwach besetzten Regionen. Diese Aussagen sind relativ zu einer fixen Bandbreite zu sehen.

2.5.2 Varianz des Schätzers

Abgesehen von einer Einschätzung des Bias ist insbesondere die Varianz des Schätzers von Bedeutung, da sie eine Vergleichbarkeit zu anderen Methoden herstellt.

Für die Berechung der (punktweisen) Varianz des Funktionalschätzer wollen wir zunächst deren asmptotische Darstellung für den PARZEN-Schätzer der Dichte (2.1) (siehe zum Beispiel JONES & WAND (1995)) vor Augen halten:

$$Var(f_n(x)) \approx \frac{1}{nb} f(x) R(K),$$

wobei $R(K) := \int K^2$ ist. Unsere verallgemeinerte Bandbreite (2.3) $R_n(t)$ konvergiert gegen ihr deterministisches Analogon $r_n(t)$, welches ungefähr $r_n(x)$ ist. Wie in DETTE & GEFELLER (1995) sieht man (wegen der Taylorapproximation) ein, dass $r_n(x) \to \frac{p_n}{\tilde{\psi}(x)}$.
Somit haben wir für den PARZEN-Schätzer mit verallgemeinerter Bandbreite,

$$f_n^R(x) := \int_{\mathbb{R}} \frac{1}{R_n(t)} K(\frac{x-t}{R_n(t)}) dF_n(t),$$

$$Var(f_n^R(x)) \approx \frac{1}{np_n} \tilde{\psi}(x) f(x) R(K).$$

Hier tritt also der inverse Faktor $\tilde{\psi}(x)$ zum Bias auf. Das hieße also eine (asymptotische) Reduktion der Varianz um den Faktor $f(x)$, wenn man die nächste-Nachbarn Bandbreite im unzensierten Szenario, also $\tilde{\Psi}_n(\cdot) \equiv F_n(\cdot)$ mit $\tilde{\psi}(x) = f(x)$ und $p_n = b$, annimmt. Aber auch hier gilt wieder, dass der Bandbreitenparameter noch

nicht gewählt wurde. Und so können wir nur eine Aussage relativ zur fixen Bandbreite machen. Im Fall $\tilde{\psi}(x) = f(x)$ ergibt sich eine tendenzielle Vergrößerung der Varianz in dichten Regionen und eine Reduktion in schwach besetzten Regionen. Wenn wir Bias^2 und Varianz zum MSE an einer Stelle addieren, sehen wir, dass kein absoluter Vorteil der Methode vorliegt, sondern „nur“ der Bias-Variance Trade-off anders balanciert ist.

Tatsächlich ist die Verallgemeinerung auf allgemeinere Prozesse $\Psi_n(\cdot)$ nicht direkt, da die Darstellung von $\psi_n(\cdot)$ als Summe unabhängiger Zufallsvariablen verloren geht. Man muss sich indes im zensierten Szenario mit der Sub-Verteilung der unzensierten Ereignisse aushelfen (siehe Diehl & Stute (1988)). Zusammen mit der Arbeit von Yandell (1983) ergibt sich folgendes Bild:

$$np_n Var(\psi_n(x)) \approx \tilde{\psi}(x)\psi(x)\Xi_\Psi(x)R(K),$$

für

$$\begin{aligned} \Xi_F(x) &\equiv 1 \\ \Xi_{F^{KM}}(x) &= \frac{1}{1-G(x)} \\ \Xi_H(x) &= \frac{1}{(1-F(x))G(x)} = \frac{1}{y(x)}, \end{aligned}$$

wobei F Dichteschätzung im unzensierten, F^{KM} im zensierten Szenario, und H die Schätzung der Hazardrate (mittels Nelson-Aalen-Schätzers) symbolisieren. Weiter ist $y(x) := \lim_{n\to\infty} \frac{Y_n(x)}{n}$, wobei $Y_n(x)$ die „population under risk“ darstellt.

Konsistente Schätzung der Komponenten ergibt dann einen konsistenten Schätzer der Varianz. So sollte $\Xi_H(x)$ durch die empirische „population under risk“ $Y_n(x) = \Sigma_{i=1}^n I_{X_i \geq x}(x)$ geschätzt werden und $1 - G(x)$ durch den modifizierten Kaplan-Meier-Schätzer

$$1 - G_n(t) = \prod_{\{j: X_{(j)} \leq t, \delta_{(j)}=0\}} \frac{n-j+1}{n-j+2},\ t < X_{(n)} \quad (0 \ sonst).$$

Bei einer Interpretation der einzelnen Konstellationen stellt sich die Analogie zu der

schon erörterten Situation der Dichteschätzung mit nächste-Nachbarn Bandbreite heraus. Insbesondere gilt dieses für die praktisch relevante Situationen der Hazardratenschätzung.

2.5.3 Asymptotische Verteilung und Anwendungen

Zunächst sei angemerkt, dass für den PARZEN-Schätzer der Dichte (2.1)

$$(nb)^{\frac{1}{2}}(f_n(x) - f(x)) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

gilt, wobei σ^2 die Varianz des Abschnitts 2.5.2 $f(x)\int K^2$ darstellt, wenn b schneller gegen 0 geht als $n^{-\frac{1}{5}}$, das heißt $bn^{\frac{1}{5}} = O(1)$.

Beweisskizze: Sei b zunächst fest, dann ist

$$\begin{aligned}
F_n &\longrightarrow F \quad \textit{fast sicher} \\
\Rightarrow Z_n := \sqrt{nb}(F_n(t) - F(t)) &\xrightarrow{\mathcal{D}} W^0(F(t)) = Z(t) \text{ Brown'sche Brücke} \\
\sqrt{n}(f_n(x) - \bar{f}_n(x)) &= \int \frac{1}{b} K\left(\frac{t-x}{b}\right)[\sqrt{n}dF_n(t) - f(t)dt] \\
= \int \frac{1}{b} K\left(\frac{x-t}{b}\right) dZ_n(t) &\xrightarrow{\mathcal{D}} \int \frac{1}{b} K\left(\frac{t-x}{b}\right) dZ(t) \sim N(0, f(x)\int K^2).
\end{aligned}$$

Wenn nun der Bias $\bar{f}_n(x) - f(x)$ wegen der Konvergenzannahme für b gegen eine deterministische Null geht, kommt keine zufällige Schwankung durch den Bias mehr hinzu und die Asymptotik gilt schon für $f_n(x) - f(x)$.

Es sei angemerkt, dass für $b = O(n^{-\frac{1}{5}})$ noch ein Erwartungswert von $f''(x)\frac{1}{2}\int z^2 K(z)dz$ additiv hinzukommt.

Wenn wir diese Aussage für den allgemeinen Schätzer $\psi_n(x)$ adaptieren, erhalten wir die Aussage:

$$(np_n)^{\frac{1}{2}}(\psi_n(x) - \psi(x)) \xrightarrow{\mathcal{D}} N(0, \sigma^2),$$

wobei σ^2 die Varianz des Abschnitts 2.5.2 darstellt, wenn p_n schneller gegen 0 geht

als $n^{-\frac{1}{5}}$, das heißt $p_n n^{\frac{1}{5}} = O(1)$. Für $p_n = O(n^{-\frac{1}{5}})$ kommt hierbei noch ein Erwartungswert von $\psi''(x)\frac{1}{2}\int z^2 K(z)dz\tilde{\psi}^2(x)$ additiv hinzu.

Allerdings ist diese punktweise Aussage für die Testtheorie oder die Ermittlung von simultanen Konfidenzbändern nicht zu verwenden. Hierfür betrachte man die asymptotische Verteilung der maximalen Abweichung in leicht abgeänderter, standardisierter Form. Die konkreten Anforderungen an die beinhalteten Funktionen und Folgen nenne ich nur, wenn sie von den in Abschnitt 2.4.1 angenommenen abweichen.

So gilt

$$P\left(l_n(M_n - d_n) > x\right) \longrightarrow \exp(-2\exp(-x)) \tag{2.36}$$

für

$$M_n = \sup_{x\in[a,b]} \left|\left| \left(\frac{np_n}{Var(\psi_n(x))}\right)^{\frac{1}{2}} (\psi_n(x) - \psi(x)) \right|\right|,$$
$$l_n = \sqrt{2\log(\frac{b-a}{p_n})} \quad \text{und} \quad d_n = l_n + \frac{1}{l_n}\frac{[\frac{\int K'^2}{\int K^2}]^{\frac{1}{2}}}{2\pi}.$$

Hierbei kann $[a, b]$ auch durch einen Schätzer von $[A, B]$ ersetzt werden, also zum Beispiel $[X_{(1)}, X_{(n)}]$, wobei beide unzensiert zu wählen sind. Außerdem ist der Kern differenzierbar und zu den Grenzen des Trägers ($[\frac{1}{2}, -\frac{1}{2}]$) hin verschwindend (also stetig) zu wählen. Diese Aussage ist eine Adaption des Theorems 4.1. aus Yandell (1983) und des Corollars 4 aus Diehl & Stute (1988).
Mit dieser Konvergenzaussage können nun (asymptotische) Konfidenzbänder, Tests gegen eine hypothetische Funktion sowie Tests auf Unterschied zwischen verschiedenen Gruppen berechnet werden. Ich möchte hier aber nur ein Beispiel geben und darauf verzichten, diese Anwendung in der Simulation zu untersuchen, da die Konvergenz gegen diese aus der Extremwert-Statistik bekannte Verteilung leider sehr langsam ist. Dies beschreiben Bickel & Rosenblatt bereits 1973. Praktische Anwendung sind deshalb kaum möglich. Ein symmetrisches Konfidenzband für die

Hazardraten bei fixer Bandbreite ist

$$h_n(x) \pm k \cdot (\frac{h_n(x)}{Y_n(x)})^{\frac{1}{2}}$$

mit

$$k = (d_n + \frac{z}{l_n})(\frac{\int K^2}{nb})^{\frac{1}{2}} \quad \text{und} \quad z = \log(\frac{-2}{\log(1-\alpha)}).$$

Auf die Möglichkeit die asymptotische Verteilungaussage zur Evaluierung einer optimalen Bandbreite, oder besser eines optimalen Bandbreitenparameters, zu verwenden, möchte ich in diesem Kontext noch hinweisen. Dieser Möglichkeit soll aber erst im Abschnitt 4.1.1.4 für die Dichteschätzung nachgegangen werden.

2.6 Wahl der Kernfunktion

Das Thema der optimalen Kernfunktion und Kerne höherer Ordnung ist in der Literatur ausführlich behandelt worden, siehe zum Beispiel GASSER, MÜLLER & MAMMITZSCH (1985) und WAND & JONES (1995) für eine Überblick. Für die Optimalität in Bezug auf die Minimierung des asymptotischen integrierten mittleren quadratischen Fehlers (AMISE) (siehe auch Abschnitt 4.1.1) hat sich zwar der Epanechnikov- Kern (oder auch Bartlett-Kern)

$$\frac{3}{4}(1-x^2), \quad |x| \leq 1$$

herausgestellt (Epanechnikov (1969)), aber die Effizienz der übrigen gebräuchlichen Kernfunktionen wie dem Normalverteilungskern, dem bi-quadratischen Kern und sogar dem Dreieckskern liegt weit über 0.9. Eine Effizienz von über 0.9 bedeutet, dass man einen identischen AMISE mit dem Epanechnikovkern mit über 90% der Beobachtungsanzahl erreicht (WAND & JONES (1995)). Das hat dazu geführt, dass die Optimalität des Kerns (bezüglich des AMISE) im Vergleich zur Bandbreite als nachgeordnet relevant betrachtet wird, und ich deshalb mein Augenmerk hiervon abwenden möchte. Hingegen spielt die Ordnung des Kerns für die asymptotische Betrachtung eine wichtigere Rolle. Formal gilt für die Momente

$$\mu_j(K) := \int_{\mathbb{R}} x^j K(x) dx$$

eines Kerns k-ter Ordnung

$$\mu_0(K) = 1,\ \mu_j(K) = 0,\ j = 1, \ldots, k-1, \text{ und } \mu_k(K) \neq 0.$$

Die Ordnung, die auch in der historischen Herleitung 2.1 als natürlich anzusehen ist, ist wegen der Symmetrie die 2. Kerne mit höherer Ordnung, also mit erstmalig nichtverschwindenden höheren Momenten, haben Aufmerksamkeit erregt. Diese Kerne, die ab Ordnung 3 keine Wahrscheinlichkeitsdichten mehr sind, führen zu schnellerer

Konvergenz des Bias. So hat der Bias bei der Dichte- oder Hazardratenschätzung mit fixer Bandbreite b für einen Kern der Ordnung 2 Konvergenz $\sim b^{-2}$ (gegen Null) (und der MISE die Ordnung $\sim b^{-\frac{4}{5}}$ für seine Konvergenz gegen Null). Für einen Kern der Ordnung 4 ist die Konvergenz schon $\sim b^{-4}$ (und die des MISE $\sim b^{-\frac{8}{9}}$). Mit weiterer Steigerung der Ordnung kann man die Ordnung des MISE beliebig an die in der parametrischen Statistik üblichen b^{-1} annähern (Literatur siehe WAND & JONES (1995) und GLAD, HJORT & USHTCHAKOV (1998)). Kerne höherer Ordnung beschleunigen also die Konvergenz des MISE. Ihr Nachteil ist allerdings, dass sich finite Konstanten für die Varianz verschlechtern. Weil sie zum Rand hin negativ werden, implizieren höhere Kerne eine geringere Bandbreite. Das bedeutet letztendlich – wenn auch nur finit – Biasreduktion und Varianzvergrößerung. Diese Balanzierung des MISE möchte ich aber auf die Bandbreitenwahl verschieben. Außerdem möchte ich von Kernen höherer Ordnung im Folgenden auch Abstand nehmen, da die durch sie implizierten möglicherweise negativen Dichteschätzer schwer zu interpretieren sind. Es sei an dieser Stelle darauf hingewiesen, dass Kerne ungerader Ordnung auftreten, wenn Asymmetrie vorliegt, was gewöhnlich dann vorkommt, wenn Randkerne verwandt werden. Ich will von der Betrachtung der Randkerne in dieser Arbeit absehen, da sie konzeptionell mit der variablen Bandbreitendefinition überlappt und so Effekte schlecht zu trennen wären. Nichts desto trotz sei darauf hingewiesen, dass die notwendige Ordnung für die asymptotischen Überlegungen, die zu Satz 2.4.10 führen, 1 ist, was eine sehr schwache Annahme an den Kern darstellt. Wir werden am Beispiel der Dichteschätzung mit fixer Bandbreite noch sehen, wohin der Unterschied in den Ordnungen 1 und 2 der Kerns für die Konvergenzraten der Verzerrung führt.

Der exakten MISE ist nur für einzelne Kernfunktion-Dichte-Konstellationen berechenbar, die in der Anwendung unrealistisch sind. Ich werde deshalb von dieser Möglichkeit keinen Gebrauch machen.

Kapitel 3

Bandbreitenwahl

Die Literatur zur Bandbreitenwahl ist äußerst umfangreich (siehe Wand & Jones (1995)). Dies deutet einerseits auf die vielen Unklarheiten zum Thema, andererseits auch auf ein ausgeprägtes Interesse am Gebiet hin. Die Bandbreitenwahl bei Funktionalkernschätzern hat seit Parzen (1962) eine lange Tradition, der eine asymptotische Darstellung des mean integrated squared errors (MISE) für die fixe Kerndichteschätzung nutzt, um eine optimale und insbesondere objektive Bandbreite zu ermitteln (siehe Absatz 4.1.1). Ich will aber zunächst auf eine Bandbreitenwahl abzielen, die sich an dem bisher Betrachteten orientiert, das heißt eine Bandbreite, die der gleichmäßigen Konvergenz Rechnung trägt. Gemeinsam haben diese beiden Ansätze, wie auch weitere Ansätze wie zum Beispiel der L_1-Abstands Ansatz (siehe Devroye & Györfi (1985)), den Versuch, eine Bandbreite zu wählen, die den Abstand zwischen Schätzer und wahrer Funktion minimiert oder zumindest asymptotisch minimiert. Da dieser Abstand von der unbekannten zu schätzenden Funktion abhängt, über die man in der Regel gerade keine (restriktiven) Annahmen machen möchte, werden Schätzer für den Abstand verwandt.

3.1 Das Supremums-Abstandsmaß

Aus dem Beweis zum Satz 2.4.10 geht hervor, dass der zweite Summand der Rate, nämlich p_n, als Bias zu betrachten ist, da er den Abstand zwischen dem wahren

$\psi(\cdot)$ und dem mit dem Kern $K(\cdot)$ gefalteten $\psi(\cdot)$ ausdrückt (siehe WAND & JONES (1995)). Wenn nur dieser Term die Konvergenz beschreiben würde, könnte man letztere durch eine schnelle Konvergenz von p_n, also der Bandbreite, gegen 0 beschleunigen. Der ersten Summand kann insbesondere wegen der Implikation einer größeren Bandbreite der Varianz zugeschrieben werden. Wir haben also auch für das Supremums-Abstandsmaß ein Balancierungsproblem vorliegen.

3.1.1 Dateninvariant

Mit Satz 2.4.10 haben wir eine fast sichere asymptotische Schranke für den Supremumsabstand gegeben. Diese Schranke $D(\sqrt{\frac{\log(n)}{(np_n)}}+p_n)$ ist eine konkave Funktion im Bandbreitenparameter p_n, da sie der Bias- und der Varianzkomponente Rechnung tragen muss. Wie auch für die fixe Bandbreite b bei der Kerndichteschätzung gilt also für ihn, je größer der Bandbreitenparameter, desto geringer die Varianz (und größer der Bias) des Schätzers und vice versa je kleiner der Bandbreitenparameter, desto geringer der Bias (und größer die Varianz). Wir können also einen stabilisierenden Bandbreitenparameter ermitteln, der die Schranke minimiert:

$$p_n^{opt} = (\frac{1}{2})^{\frac{2}{3}}(\frac{\log(n)}{n})^{\frac{1}{3}}. \tag{3.1}$$

Diese Wahl erfüllt die Bedingungen (2.9). Dass die Wahl auch die Ordnung der Konvergenz maximiert, lässt sich aus Folgendem schließen. Wegen der Restriktionen (2.9) hat p_n die Gestalt

$$p_n \sim (\frac{\log(n)}{n})^{\xi}, \quad 0 < \xi < 1, \tag{3.2}$$

und somit hat die Schranke die Gestalt

$$OS(p_n) \sim \sqrt{\frac{\log(n)}{n}}(\frac{\log(n)}{n})^{-\frac{\xi}{2}} + (\frac{\log(n)}{n})^{\xi}. \tag{3.3}$$

Nun steigt die Rate des zweiten Summanden monoton in ξ auf dem Interval $(0, 1)$,

wogegen die des ersten monoton fällt. Da die minimale Rate der beiden Summanden die Konvergenz bestimmt, ist das ξ optimal, bei der die Summanden gleiche Rate haben, also $\xi_{opt} = \frac{1}{3}$. Formaler kann man also das Maximum von $\min_{\xi \in (0,1)}(\frac{1-\xi}{2}, \xi)$ suchen. Die Lösung veranschaulicht sich anhand der Grafik 3.1.

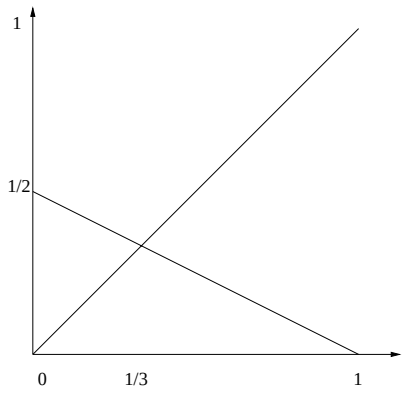


Abbildung 3.1: Maximale Ordnung der Konvergenz

Diese Bandbreite erfüllt aber nicht die Äquivarianzforderung bezüglich linearer und affiner Transformationen von Deheuvels & Hominal (1980). Diese Forderung beruht auf der Überlegung, das die Skalierung der Daten auch zu einer analogen Skalierung der (optimalen) Bandbreite führen muss, da sie sonst willkürlich ist. Um die Skalenäquivarianz zu gewährleisten, multipliziere man die Bandbreite mit der (geschätzten) Varianz der Daten. Man geht dann davon aus, dass die Fehlerrechnung sich auf eine Stichprobe der Varianz 1 bezieht, und erreicht mit Multiplikation dieser Stichprobe die aktuelle. Hier muss dann aber die Varianz 1 in der Fehlerrechnung ausgenutzt werden, was nicht der Fall ist. Also ist nicht klar, wie die Varianz einzugehen hat. Zur Motivation siehe das Beispiel des Parzen-Schätzers in Abschnitt 4.1.1.2.

3.1.2 Allgemein

Wenn man nun den letzten Teil des Beweises zum Satz 2.4.10 betrachtet, so findet man die (im Vergleich zum Satz 2.4.10) feinere Abschätzung des gleichmäßigen

Fehlers

$$\sup_{x\in[a,b]} |\psi_n(x) - \psi(x)| \leq D_1 \left(\sqrt{\frac{\log(n)}{(np_n)}}\right) + D_2 \left(\sqrt{\frac{\log(n)}{(np_n)}}\right) + D_3 p_n,$$

welches minimal wird für

$$p_{data}^{opt} = \left(\frac{D_1 + D_2}{2D_3}\right)^{\frac{2}{3}} \left(\frac{\log n}{n}\right)^{\frac{1}{3}}. \tag{3.4}$$

Dass hierfür die Skalenäquivarianz gegeben ist, kann angenommen werden. Für die fixe Bandbreitenwahl bei der Dichteschätzung wird dies in Abschnitt 4.1.1.2 gezeigt. Allgemein steht der analytische Beweis noch aus.

3.2 Eine Daumenregel

Häufig ist es auch in der nichtparametrischen Funktionalschätzung nützlich, initial eine Verteilung der Beobachtung für Teilaspekte anzunehmen. Hjort & Glad (1995) benutzen zum Beispiel einen parametrischen Start-Schätzer für die nichtparametrische Dichteschätzung. In einem der früheren Verfahren für die Bandbreitenwahl der fixen Bandbreite für den Parzen-Schätzer der Dichte (2.1) wird für die sogenannte „normal-scale-rule“ – auch „Rule-of-Thumb“ oder „Daumenregel“ genannt – eine Normalverteilung der Daten angenommen, um den asymptotischen MISE minimierend eine Bandbreite zu ermitteln (siehe Parzen (1962) und Fryer (1976)). Es stellt sich heraus, dass in die Schätzung der somit optimalen Bandbreite nur die Schätzung der Varianz eingeht:

$$b^{\mathrm{RoT}} = \left[\frac{8\pi^{\frac{1}{2}} R(K)}{3\mu_2(K)^2 n}\right]^{\frac{1}{5}} \hat{\sigma}. \tag{3.5}$$

Hierbei sind $R(K) := \int K(z)^2 dz$ und $\mu_2(K) := \int z^2 K(z) dz$ für die verschiedenen Kerne auch Jones & Wand (1995) zu entnehmen. $\hat{\sigma}$ ist ein Schätzer für die Standardabweichung. Es ist wünschenswert, dieses einfache, schnelle und objektive

Verfahren auch für eine verallgemeinerte Bandbreite 2.3 zu verallgemeinern. Nun kann man die Einbettung der fixen Bandbreite in die verallgemeinerte Modellierung aus Abschnitt 2.4.1 nutzen und die Funktion $\tilde{\Psi}(\cdot)$ durch eine lineare Funktion approximieren. Die sich ergebende Steigung der Gerade β lässt dann eine fixe Bandbreite b^{RoT}, die nach einer beliebigen etablierten Daumenregel erzielt wurde, in einen Bandbreitenparameter

$$p_n^{RoT} = |\beta| b^{\mathrm{RoT}} \tag{3.6}$$

uminterpretieren.

Beweis:
Allgemein gilt für die Umrechnung

$$\begin{aligned} R_n(t) &= \sup\{r > 0 | |\tilde{\Psi}_n(t - \frac{r}{2}) - \tilde{\Psi}_n(t + \frac{r}{2})| \leq p_n\} \\ \text{speziell} \quad b &= \sup\{r > 0 | |c \cdot (t - \frac{r}{2}) + d - (c \cdot (t + \frac{r}{2}) + d)| \leq p_n\} \\ &= \sup\{r > 0 | | - 2c \cdot \frac{r}{2}| \leq p_n\} \\ &= \sup\{r > 0 | |c| \cdot r = p_n\} \\ &= \frac{p_n}{|c|}. \end{aligned}$$

□

Die Wahl der linearen Funktion hängt von der Funktion $\tilde{\Psi}(\cdot)$ ab, so dass ich erst am Beispiel der Hazardratenschätzung mit nächste-Nachbarn Bandbreite exemplarisch darauf eingehen werde. Es sei angemerkt, dass p_n die Konvergenzeigenschaften von b^{RoT} „erbt“ und somit die gleichmäßige Konvergenzaussage des Satzes 2.4.10 gilt, wenn b^{RoT} die Konvergenzeigenschaften (2.9) erfüllt.

Kapitel 4

Anwendung der allgemeinen Konvergenzaussage

4.1 Der PARZEN-Schätzer der Dichte – unzensiertes Szenario

Zunächst will ich eine gleichmäßig absolute Konvergenzaussage für den PARZEN-Schätzer der Dichte mit fixer Bandbreite (2.1) ableiten. Man wähle $\tilde{\Psi}_n(t) = c \cdot t + d = \tilde{\Psi}(t)$ und $\Psi_n(t) = F_n(t)$ als empirische Verteilungsfunktion von n u.i.v. Zufallsvariablen X_i mit Verteilungsfunktion F, wobei dann $\Psi(t) = F(t)$ ist. Dann gilt für den Bandbreitenparameter $p_n := |c| \cdot b$ wegen der Definition (2.11) $R_n(t) = r_n(t) \equiv b$, und der Schätzer vereinfacht sich zu

$$\begin{aligned} \psi_n(x) &= \int_{\mathbb{R}} \frac{1}{R_n(t)} K(\frac{x-t}{R_n(t)}) d\Psi_n(t) \\ &= \int_{\mathbb{R}} \frac{1}{b} K(\frac{x-t}{b}) dF_n(t) \qquad (4.1) \\ &= \frac{1}{nb} \sum_{i=1}^{n} K(\frac{x-X_i}{b}). \qquad (4.2) \end{aligned}$$

Bemerke, dass wir hier eine ganze Schar von Glättungsprozessen $\tilde{\Psi}_n(\cdot)$ sowie von Glättungsparametern p_n zulassen können. So haben wir also bei gegebener Kon-

vergenz von p_n nach den Kriterien (2.9) für jeden Bereich, wo die Dichte f von F von Null weg beschränkt ist, mit Hilfe von Satz 2.4.10, eine Konsistenzaussage mit Konvergenzrate $\sqrt{\frac{\log(n)}{(nb)}} + b$ bezüglich des gleichmäßigen Fehlers. Zu weiteren Konsistenzaussagen für diesen Schätzer bezüglich des MISE siehe NADARAYA (1965) und bezüglich des gleichmäßigen Fehlers STUTE (1982B).

4.1.1 Bandbreitenwahl

4.1.1.1 Der integrierte mittlere quadratische Fehler

Zunächst will ich hier die gewöhnliche Methode zur Bandbreitenwahl für fixe Bandbreiten rekapitulieren, die bei der verallgemeinerten Bandbreite in Ermangelung einer expliziten Darstellung des asymptotischen MISE (AMISE) nicht möglich ist, um später Unterschiede zwischen den Bandbreitenwahlen aufzeigen zu können. Als etablierte Verlustfunktion zur Bewertung asymptotischen Verhaltens (siehe WAND & JONES (1995)), und somit zur Wahl der Bandbreite, stellt sich der quadratische Verlust dar. Das heißt, wir „messen“ die Diskrepanz zwischen zu schätzender Funktion $f(\cdot)$ und Schätzer $f_n(\cdot)$ mittels

$$ISE := \int_{\mathbb{R}} (f(x) - f_n(x))^2 dx.$$

Da deren mittlere Ausprägung

$$MISE := E \int_{\mathbb{R}} (f(x) - f_n(x))^2 dx,$$

wie wir weiter unten sehen werden, einfacher zu approximieren ist, werde ich mich auf deren Betrachtung als goodness-of-fit Kriterium beschränken, anhand dessen wir Entscheidungen über die Wahl der Bandbreite treffen wollen.

Im Dichteschätzkontext mit fixer Bandbreite ist wegen der Zerlegung des MSE in Bias^2 + Varianz eine asymptotische Darstellung des MISE gegeben durch

$$AMISE = \frac{1}{4} b^4 \mu_2(K) R(f'') + \frac{R(K)}{nb}, \tag{4.3}$$

wobei ein Kern zweiter Ordnung verwandt wird. Man rufe sich noch einmal ins Gedächtnis, dass $R(g(\cdot)) = \int g(z)dz$ und $\mu_j(g(\cdot)) = \int z^j g(z)dz$ bezeichnen. Der erste Summand, der dem Bias zuzuschreiben ist, ergibt sich aus folgender Überlegung:

$$\begin{aligned} Ef_n(x) &= E\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}K(\frac{x-X_i}{b}) \\ &= E\frac{1}{b}K(\frac{x-X_i}{b}) \\ &= \int_{\mathbb{R}} K(z)f(x-bz)dz \\ &= \int_{\mathbb{R}} K(z)\sum_{\nu=0}^{\infty}\frac{f^{(\nu)}(x)}{\nu!}(bz)^\nu dz \\ &\approx (O(b^3)) \quad f(x) + \frac{1}{2}b^2 f''(x)\int_{\mathbb{R}} K(z)z^2 dz, \end{aligned} \tag{4.4}$$

wegen der Symmetrie (und der Dichte-Eigenschaft) des Kerns $K(\cdot)$, $\int K(z)zdz = 0$. Der erste Summand in (4.3) ist dann

$$\begin{aligned} \int_{\mathbb{R}} \text{Bias}^2 &= \int_{\mathbb{R}} (f(x) - f_n(x))^2 dx \\ &= \frac{1}{4}b^4\mu_2^2(K)\int_{\mathbb{R}} (f''(x))^2 dx. \end{aligned}$$

4.1.1.2 Der gleichmäßig absolute Fehler

Aufgrund des vereinfachten Ausdrucks (4.1) können wir hier eine Verschärfung der Ungleichungen aus Abschnitt 2.4.6 erzielen und somit die Bandbreite (3.4) verbessern. Zusätzlich stellt diese Wiederholung der Konvergenzüberlegungen eine Verdeutlichung einiger zentraler Argumente dar. Wenn man eine Reanalyse der Konvergenzgeschwindigkeit für die Kernschätzung der Dichte mit fixer Bandbreite, das heißt $R_n(t) \equiv b$, vornimmt, erhält man Folgendes. (Zu Illustrationzwecken wird hier auf den Dreieckskern zurückgegriffen, da dessen Totalvariation leicht zu berechnen ist. Das schlägt sich aber nicht verändernd in den Konvergenzraten nieder.) Für hinreichend großes n ist die Integration zur Berechnung des Schätzers beschänkt auf das Interval $I_n(x) = [x - \frac{b}{2}, x + \frac{b}{2}]$, vergleiche Satz 2.4.2 wobei $\tilde{m} = c \equiv \frac{dc \cdot x + d}{dx}$ und

$p_n = |c| \cdot b$ sind, das heißt

$$\begin{aligned}
|f_n(x) - \bar{f}(x)| &= |\int_{I_n(x)} \frac{1}{b} K(\frac{x-t}{b}) dF_n(t)) - \int_{I_n(x)} \frac{1}{b} K(\frac{x-t}{b}) dF(t))| \\
&\leq \sup_{I \subset I_n(x)} |F_n(I) - F(I)| V_{I_n(x)}(\frac{1}{b} K(\frac{x-t}{b})) \\
&\leq 9\sqrt{\frac{\log(n) M b}{n}} \int_{-\frac{b}{2}}^{\frac{b}{2}} \frac{4}{b^2} dt \quad \text{Schäfer (1986b)} \\
&= 36\sqrt{M}\sqrt{\frac{\log n}{n}} b^{-\frac{1}{2}} \quad \text{wegen des Dreieckskerns.}
\end{aligned}$$

Zur Totalvariationsberechung des Dreieckskerns betrachte dessen Bild 4.1.

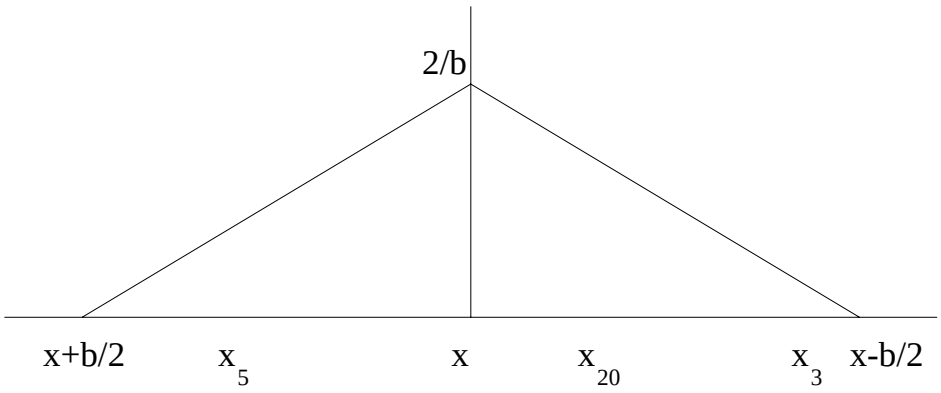


Abbildung 4.1: Dreiecks-Kern

Ebenso gilt für hinreichend großes n

$$\begin{aligned}
|\bar{f}(x) - f(x)| &\leq |\int_{I_n(x)} \frac{1}{b} K(\frac{x-t}{b}) |f(t) - f(x)| dt \\
&\leq \sup_{t \in I_n(x)} |f(t) - f(x)| \leq \sup_{t \in I_n(x)} L_f |x - t| \\
&\leq \frac{1}{2} L_f b \qquad (4.5)
\end{aligned}$$

(mit Lipschitz-Konstanter L_f). Und es folgt

$$|f_n(x) - f(x)| \leq 36\sqrt{M}\sqrt{\frac{\log n}{n}} b^{-\frac{1}{2}} + \frac{1}{2} L_f b \qquad (4.6)$$

unabhängig von x auf dem Intervall $[a,b]$ und minimal für

$$b_{UE} = M^{\frac{1}{3}} L_f^{-\frac{2}{3}} 36^{\frac{2}{3}} (\frac{\log n}{n})^{\frac{1}{3}}, \tag{4.7}$$

welches für hinreichend großes n auch (2.9) erfüllt. Eben diese Konstante gilt dann auch für $\sup_{x\in[a,b]} |f_n(x) - f(x)|$. Der Defekt, dass b_{UE} nicht p_{data}^{opt} (3.4) mit $\tilde{\Psi}_n(\cdot) \equiv c\cdot id(\cdot)+d$ und $\Psi_n(\cdot) = F_n(\cdot)$ darstellt, liegt daran, dass die Totalvariation des Kerns hier exakt ermittelt wird, im allgemeinen Fall dagegen abgeschätzt wird.

An dieser Stelle sei auf die Darstellung des Bias (4.4) in Vergleich mit der Darstellung (4.5) hingewiesen, die eine Rate von $O(b^2)$ belegt. Die höhere Rate ist einzig dem symmetrischen Kern zuzuschreiben. Verzichtet man auf die Symmetrie – und den Erwartungswert gleich Null –, wie in den Voraussetzungen zu Satz (2.4.10), können auch Randkerne verwendet werden und man erhält mit der Taylorapproximation dieselbe Rate $O(b)$.

Interessant ist es an dieser Stelle auch die Skalenäquivarianz der optimalen Bandbreite zu untersuchen. Wenn wir von unserer Stichprobe $X_1, \ldots, X_n$ zu einer Stichprobe $X_i^c := cX_i,\ i+1, \ldots, n$ übergehen, so ist in der Bandbreite (4.7) nur der Faktor $M^{\frac{1}{3}} L_f^{-\frac{2}{3}}$ betroffen. Wegen $f_c(x) = \frac{1}{c} f(cx)$ ist $M_c = \frac{1}{c} M$ und $L_{f_c} = \frac{1}{c^2} L_f$ (wegen $|f_c(y) - f_c(x)| = |\frac{1}{c} f(\frac{y}{c}) - \frac{1}{c} f(\frac{x}{c})| \leq \frac{1}{c} L_f |\frac{y}{c} - \frac{x}{c}| = \frac{1}{c^2} L_f |y - x|$). Es folgt, dass $b_{opt}^{f_c} = c b_{opt}^{f}$ ist, was die Äquivarianz gegenüber linearer Transformation belegt, die Invarianz bezüglich Addition einer Konstanten ist offensichtlich, da kein Faktor der Darstellung (4.7) betroffen ist. Somit haben wir die von Deheuvels & Hominal (1980) geforderte Äquivarianz bezüglich affiner Transformationen für die Bandbreitenwahl bewiesen. Interessant übrigens auch, dass bei linearer Streckung der Daten ($c > 1$) sich beide Summanden des gleichmäßigen Fehlers (4.6) (also Varianz und Bias) verkleinern. Das Phänomen ist noch nicht begriffen.

Für die Anwendung des Donsker'schen Invarianzprinzips auf den gleichmäßigen Fehler des Rosenblatt-Parzen-Schätzers siehe Stute (1986a,b). Er erzielt ähnliche Konvergenzraten für andere Glattheitsannahmen an die Dichte $f(\cdot)$. So belegt er die gleichmäßige fast sichere Rate von $(\frac{\log n}{n})^{\frac{2}{5}}$ für optimale Bandbreite $(\frac{\log n}{n})^{\frac{1}{5}}$.

Auswirkung der Bandbreitenwahl auf Bias und Varianz Bei Wahl der Bandbreiten $\sim (\frac{\log n}{n})^{\frac{1}{3}}$ ergibt sich für Bias und Varianz folgendes Bild.

$$\begin{aligned} Bias(x) &\sim \frac{1}{2}(\frac{\log n}{n})^{\frac{2}{3}}\mu_2(K)f''(x) \quad \text{und} \\ Varianz(x) &\sim (log(n)n^2)^{-\frac{1}{3}}R(K)f(x), \end{aligned}$$

so dass also der MISE $\sim (\log(n)n^2)^{-\frac{1}{3}}$ ist und somit langsamer als $n^{-\frac{4}{5}}$ gegen Null konvergiert. Wenn wir uns also das Verhältnis der beiden zueinander anschauen, so ist

$$\begin{aligned} \frac{Bias(x)^2}{Varianz(x)} &\sim \frac{\log(n)^{\frac{5}{3}}}{n^{\frac{2}{3}}} \longrightarrow 0 \\ \frac{Bias(x)}{Varianz(x)} &\sim \log n \longrightarrow \infty. \end{aligned}$$

Es ist weiter leicht einzusehen, dass $\frac{Bias(x)^\epsilon}{Varianz(x)} \to 0 \; \forall \; \epsilon > 1$ und $\frac{Bias(x)^\epsilon}{Varianz(x)} \to \infty \; \forall \; \epsilon \leq 1$. Dies kann als Bestreben der Bandbreite interpretiert werden, die Summe aus Bias und Varianz möglichst schnell gegen Null streben zu lassen. Dies ist im Unterschied zu der MISE-Minimierung zu sehen, wobei die Summe aus Bias^2 und Varianz schnell gegen Null konvergieren soll, was ein konstantes Verhältnis (ungleich Null) der Raten von Bias^2 und Varianz, zur Folge hat. Das heißt, die Bandbreitenwahl legt in Bezug auf den gleichmäßigen Fehler mehr Gewicht auf den Bias als die Methode der MISE-Minimierung in dem Sinne, dass sie eine schnellere Bias-Konvergenz gegen Null zur Folge hat (im Verhältnis zur Varianz-Konvergenz) als die MISE-Minimierung. Dass der Exponent des Bias 1 bei optimaler Bandbreitenwahl bezüglich des gleichmäßigen Fehlers ist, kann ich nur so deuten, dass es eine Zerlegung des gleichmäßigen Fehlers der Art $|f_n(x) - f(x)| \approx Bias(x) + Varianz(x)$ oder allgemeiner $|\hat{\theta} - \theta| \approx E(\hat{\theta} - \theta) + E((\hat{\theta} - \theta)^2)$ gibt.

4.1.1.3 Normalverteilungsapproximation

Für eine Bandbreite, die sich als $C(\frac{\log n}{n})^{\frac{1}{3}}$ – C sei eine Konstante – schreiben lässt, definiere

$$W_n(x) := n^{\frac{1}{3}}(\log n)^{\frac{6}{3}}(f_n(x) - f(x)).$$

Dann ist $Var(W_n(x)) = \frac{1}{C}R(K)f(x)$ und

$$\begin{aligned} E(W_n(x)) &= n^{\frac{1}{3}}(\log n)^{\frac{6}{3}}\frac{1}{2}C^2(\frac{\log n}{n})^{\frac{2}{3}}\mu_2(K)f''(x) \\ &= \frac{1}{2}C^2\mu_2(K)f''(x)(\frac{\log(n)^{\frac{1}{2}}}{n^{\frac{1}{3}}} \\ &\longrightarrow 0. \end{aligned}$$

Wie in Abschnitt 2.5.3 gilt

$$W_n(x) \xrightarrow{\mathcal{D}} N(0, \frac{1}{C}R(K)f(x)),$$

was wieder die Bias-Bezogenheit der Bandbreitenwahl belegt.

4.1.1.4 Asymptotische Verteilung des maximalen Abstandes zur Bandbreitenwahl

Man kann auf die Idee verfallen, die asymptotische Verteilungaussage (2.36) zur Evaluierung einer optimalen Bandbreite zu verwenden. Die Minimierung des Erwartungwertes oder des (einfacher zu berechnenden Medians) liegen dann nahe. Allerdings stellt man bei der genaueren Betrachung der Darstellung des asymptotischen Erwartungswertes von M_n fest, dass dieser *linear* (fallend) in b ist. Das hängt damit zusammen, dass eigentlich nur die asymptotische Verteilung von $f_n(x) - \bar{f}(x)$ ermittelt wird. Der Bias-Term wird (zunächst) vernachlässigt. Er wird erst später mit Glattheitsaussagen motiviert miteinbezogen, indem die Konvergenz der Bandbreitenfolge schnell (schneller als $n^{-\frac{1}{5}}$) vorausgesetzt wird (siehe Bickel & Rosenblatt (1973), Korollar 1). Der erwähnte Term ist also der Varianz zuzuschrei-

ben, von der wir schon wussten, dass sie für größere Bandbreiten kleiner wird.

Zerlegt man aber von vorherein wie folgt, kann man sinnvolle Berechnungen anstellen.

$$
\begin{aligned}
& E \lim_{n\to\infty} \sup_{[a,b]} \left| \frac{f_n(x) - f(x)}{\sqrt{f(x)}} \right| \\
\leq\ & E \lim_{n\to\infty} \sup_{[a,b]} \left| \frac{f_n(x) - \bar{f}(x)}{\sqrt{f(x)}} \right| + \lim_{n\to\infty} \sup_{[a,b]} \left| \frac{\bar{f}(x) - f(x)}{\sqrt{f(x)}} \right| \\
\approx\ & E \lim_{n\to\infty} \sup_{[a,b]} \left| \frac{f_n(x) - f(x)}{\sqrt{f(x)}} \right| + \sup_{[a,b]} \left| \frac{f''(X)}{\sqrt{f(x)}} \right| b^2 \frac{\mu_2(K)}{2} \\
=\ & (E(e^{-2e^{-x}}(2\log b)^{-\frac{1}{2}} + d_n)\sqrt{R(K)}(nb)^{-\frac{1}{2}} + \sup_{[a,b]} \left| \frac{f''(X)}{\sqrt{f(x)}} \right| b^2 \frac{\mu_2(K)}{2}
\end{aligned}
$$

So stellt wieder $\frac{1}{5}$ die optimalen Bandbreitenordnung einer Bandbreite der Form $n^{-\gamma}$ dar. Gegen weitere Berechungen spricht die bereits erwähnte langsame Verteilungskonvergenz sowie die Unhandlichkeit des Ausdrucks. Man muss numerisch minimieren.

4.2 Ein variabler Hazardratenschätzer – zensiertes Szenario

Wenn wir von der Anwendung in Überlebenszeitstudien ausgehen, treten nicht-negative Zufallsvariablen natürlich auf. Wenn man zum Beispiel Patienten in einer klinischen Studie bis zum Auftreten eines interessierend Ereignisses verfolgt, ist diese Zeitspanne, Überlebenszeit oder Ausfallzeit genannt, eine nicht-negative Zufallsvariable. Man kann sich viele Situationen vorstellen, bei denen das Beobachten dieses interessierenden Ereignisses durch ein potentielles zweites Ereignis während der Beobachtungszeit verhindert wird. Wir haben dann nur noch die Information, dass das interessierende Ereignis nicht bis zum Eintreten des zweiten, zensierenden, Ereignisses aufgetreten ist. Wir wollen diese Information nutzen, auch weil einsichtig ist, dass, bei Vernachlässigung dieser Information, Verzerrungen, nämlich systematische Risikoüberschätzugen die Folge sind. Wiederholend nennen wir ein „(rechts)-zensiertes Szenario“ die Beobachtung von unabhängigen (bivariaten) Zufallsvariablen

$$(X_i, \delta_i) := (\min\{T_i, C_i\}, 1_{\{T_i=C_i\}}) \quad i = 1, \ldots, n$$

für Überlebenszeiten T_i mit Dichte (auf der positiven Halbachse) $f(\cdot)$ unabhängig von den Zensierungszeiten C_i mit Dichte $g(\cdot)$ (und kumulativer Verteilungsfunktion $G(\cdot)$).

Natürlich ist auch hier die Schätzung der Dichte $f(\cdot)$ von Interesse. Allerdings hat sich in der Überlebenszeitanalyse die Funktion der Hazardrate

$$h(x) := \frac{f(x)}{S(x)}$$

– $S(x) := 1 - \int_0^x f(x)$ ist die Überlebenszeitfunktion (survival function) – als interpretierbarer herausgestellt, da es das stetige Analogon zur momentanen Sterbewahrscheinlichkeit

$$h(t) \quad = \quad \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\left(X \in [t, t + \Delta t] \mid X \geq t\right)$$

darstellt. Da das hier verfolgte Prinzip der Funktionalschätzung die Kern-Glättung von Schätzern der Stammfunktion der interessierenden Funktion ist, sei hier die kumulative Hazardrate

$$H(t) := \int_0^t h(s)ds$$

eingeführt. Für ihre Schätzung ist der NELSON-AALEN-Schätzer

$$\begin{aligned} H_n(t) &:= \int_0^t \frac{J(s)}{Y(s)} dN(s) \\ &= \sum_{i:X_{(i)} \leq t} \frac{\delta_{(i)}}{n-i+1} \end{aligned}$$

der seit langem verwendete nichtparametrische Standardschätzer. Hier stellt $Y(s)$ die „population at risk“ bis zum Zeitpunkt s, $J(s) = I(Y(s) > 0)$ und $N(s)$ den Zählprozess der unzensiert bis zum Zeitpunkt s „ausgefallenen“ Beobachtungen dar. Die Notation $\delta_{(i)}$ bezieht sich auf die Ordnung der jeweils dazu gehörenden $X_{(i)}$. Wegen der additiven Zerlegung dieses Schätzers in einen deterministischen und einen Martingal Term findet er im Rahmen der Martingaltheorie, oder spezieller in der Theorie der Zählprozesse, eine Fundierung, die ihn als Analogon bei der Schätzung der kumulativen Hazardrate bei zensierten Daten zur empirischen Verteilungsfunktion bei der Schätzung der Verteilungsfunktion bei unzensierten Daten ausweist. Eingeführt hat diesen Schätzer NELSON (1972) in Kontext grafischer Überprüfung von parametrischen Modellannahmen, AALEN (1978) hat ihn in die Martingaltheorie eingebettet. ANDERSEN ET AL. (1993) und BORGAN (1997) fassen die Entwicklungen zusammen.

Theorem 4.2.1 *Sei der variable Kern-Schätzer*

$$\begin{aligned} \hat{h}_{R_3}(x) &:= \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-i+1} \frac{1}{R_3(k_n, X_i)} K\left(\frac{X_i - x}{R_3(k_n, X_i)}\right) \\ &= \int_{\mathbb{R}} \frac{1}{R_3(k_n, t)} K\left(\frac{t-x}{R_3(k_n, t)}\right) dH_n(t) \end{aligned} \tag{4.8}$$

mit Definition der nächste-Nachbarn Bandbreite von GEFELLER & DETTE (1992)

$$R_3(k_n, x) = \{\inf r > 0 || S_n(t - \frac{r}{2}) - S_n(t + \frac{r}{2})| \geq \frac{k_n}{n}\},$$

definiert. Unter den Voraussetzungen von Satz 2.4.10 konvergiert er fast sicher gleichmäßig gegen die wahre Hazardfunktion $h(\cdot)$ auf dem Intervall $[a, b]$ mit Geschwindigkeit $\sqrt{\frac{\log(n)}{(k_n)}} + \frac{k_n}{n}$, wenn

$$\begin{aligned} 0 < \frac{k_n}{n} \in I\!R < 1 \\ \frac{k_n}{n} \longrightarrow 0 \\ \frac{k_n}{\log(n)} \longrightarrow \infty. \end{aligned} \tag{4.9}$$

Beweis:
Anwendung von Satz 2.4.10: mit dem KAPLAN-MEIER-Schätzer S_n für die Überlebenszeitfunktion S in der Rolle von $\tilde{\Psi}_n$, dem NELSON-AALEN-Schätzer H_n für die kumulative Hazardrate H in der Rolle von Ψ_n und mit $p_n = \frac{k_n}{n}$ folgt die Konvergenzaussage. □

4.2.1 Bandbreitenwahl

4.2.1.1 Die Daumenregel

Es wurde bereits angemerkt, dass die Darstellung des MISE im Falle der Hazardraten Schätzung mit fixer Bandbreite b

$$MISE = \int \frac{h(x)}{1 - G(x)} dx \int K^2(z) dz (nb)^{-1} + \frac{1}{4} b^4 (\int z^2 K(z) dz)^2 \int (h''(x))^2 dx$$

zur Unbeschränktheit führen kann. Sei zum Beispiel die Hazardrate konstant, wie im Falle der Exponential-Verteilung, so ist der Term $\int h(x) dx$ (unter Verneinung von Zensierung) unbeschränkt. Das heißt, dass die Varianz unbeschänkt ist. Eine

Möglichkeit, ein endliches Maß zu erzwingen, ist, den Träger der Integration zu restringieren (siehe DETTE & GEFELLER (1995)), also eine gekappte Exponential-Verteilung anzunehmen. Das führt aber zu Entscheidungsnöten bezüglich der Endpunkte und insbesondere zu einer starken Sensitivität bezüglich deren Fixierung. Alternativ kann man eine Gewichtsfunktion $w(x)$ einfügen, für die als geeignete Wahl sich die Funktion herausstellt, die den MIWSE der Hazardrate in den MISE der korrespondieren Dichte überführt (siehe HJORT (1991)). Diese Wahl der Gewichtsfunktion begründet, die Wahl der Bandbreite zur Benutzung für die Hazardratenschätzung auf den Dichteschätzkontext (Dichte-MISE) zu beziehen. Anschaulich kann man auch argumentieren, dass die Bandbreite eine „gleichmäßige“ Einbeziehung der Beobachtungen in die Schätzung realisieren soll. Die Verteilung der Beobachtungen ist aber durch die Dichte und nicht durch die Hazardrate beschrieben. Ein bedeutender praktischer Vorteil, der sich aus diesem Vorgehen ergibt, ist, dass wir die intensive Arbeit vieler auf dem Gebiet der optimalen Bandbreitenwahl im Kern-Dichteschätzkontext mit fixer Bandbreite für viele Anwendungen nutzbar machen.
Sei b eine solche fixe Bandbreite. Nun ist aber nicht klar, wie sich die Bandbreite b für eine Dichte in eine Bandbreite für die Hazardratenschätzung transformiert, wenn man nächste-Nachbarn Bandbreiten betrachtet. Hier stellt sich unsere verallgemeinerte Darstellung (2.3) als hilfreich heraus. Man nutzt die durch die verallgemeinerte Darstellung (2.3) gegebene Brücke zwischen nächste-Nachbarn und fixer Bandbreite und approximiert die Verteilungsfunktion durch eine lineare Funktion:

$$R_n^{opt}(t) := \{r > 0 | |S_n^{approx}(t - \frac{r}{2}) - S_n^{approx}(t + \frac{r}{2})| = p_n\}.$$

Exemplarisch sei hier die Daumenregel, auch „normal scale rule“ oder kurz „sRoT“ genannt, adaptiert.

Man stelle sich vor, b^{RoT} – nach der „normal scale rule“ (Annahme der Normalverteilung) – sei die Bandbreite $R_n^{opt}(t)$ mit unbekanntem Glättungsprozess $\tilde{\Psi}_n(\cdot) = c \cdot id(\cdot) + d$ und Bandbreitenparameter $p_n = |c| \cdot b$ (siehe Abschnitt 3.2). Nun wollen wir zum KAPLAN-MEIER-Schätzer der Überlebenszeitfunktion als Glättungsprozess

übergehen, müssen also $S_n(\cdot) \approx \beta \cdot id(\cdot) + d$ erreichen. Die sich ergebende Steigung einer approximierenden Geraden β lässt dann die fixe Bandbreite b^{RoT} mit einem Bandbreitenparameter von $p_n = |\beta| \cdot b^{RoT}$ (siehe 3.6) und sukzessive mit einer Anzahl nächster Nachbarn von $k_n = p_n \cdot n = n \cdot |\beta| \cdot b^{RoT}$ assoziieren und in eine Anzahl von $[n|\beta|b^{RoT}]$ nächsten Nachbarn uminterpretieren. Die Approximation erfolgt über eine lineare Regression der (Beobachtungs-)Punkte der Kaplan-Meier-Kurve $(X_i, \frac{1}{2}(S_n(X_i-) + S_n(X_i))), i = 1, \ldots, n$ mit deren Steigungskoeffizient $\hat{\beta}$.

Zusammenfassend ergibt sich als Anzahl nächster Nachbarn folgende Darstellung.

Theorem 4.2.2 *Eine fixe Bandbreite für die Kern-Dichteschätzung b^{RoT} lässt sich für die variable Kern-Schätzung der Hazardrate mit nächste-Nachbarn Bandbreite in eine Anzahl*

$$k_n^{RoT} = \left[n \cdot |\hat{\beta}| \cdot b^{RoT} \right] \tag{4.10}$$

nächster Nachbarn uminterpretieren, wobei die Gaussklammer $[\cdot]$ nur eine Ganzzahligkeit garantieren soll.

Bemerke, das diese Bandbreitenwahl die Bedingung (2.9) erfüllt, da $b^{RoT} \sim n^{-\frac{1}{5}}$ diese Bedingung erfüllt.

Bemerkung: Man kann in der Überlebenszeitanalyse die Normalverteilung als hypothetische Verteilung ablehnen, da sie nicht nur positiv ist. Als Alternative eignet sich die log-Normalverteilung für diese Idee allerdings nicht, da für sie keine geschlossene Form der Hazardfunktion existiert (Lee (1992)). Aber auch die Maximum Likelihood Schätzung der Parameter der Weibull Verteilung ist nicht explizit gegeben (Lee (1992)). Ich will deswegen hier die Daumenregel aus der Dichteschätztheorie unter Annahme der Normalverteilung verwenden.

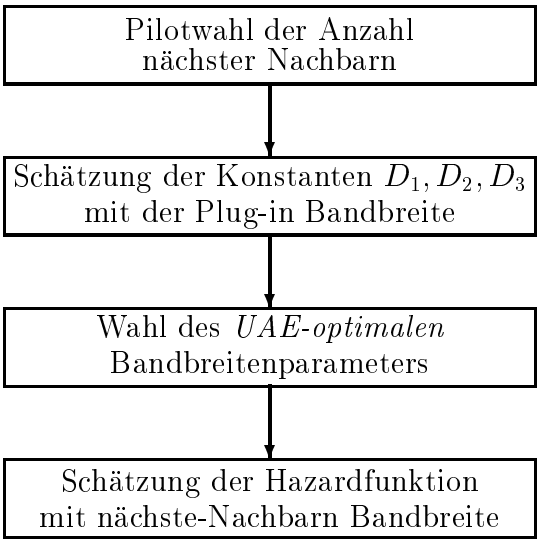
4.2.1.2 Der gleichmäßig absolute Fehler

Wie bei der Dichteschätzung kann man die Minimierung der oberen Schranke des gleichmäßig absoluten Fehlers anstreben. Die Varianten (3.1) und „(3.1) mit Varianzmuliplikator“ des optimalen Bandbreitenparameters sind nicht sinnvoll, da die

erste nicht die zwingende Bedingung der Skaleninvarianz (siehe DEHEUVELS & HOMINAL (1980)) erfüllt und auch die zweite sich in praktischen Tests als zu instabil erwiesen hat. So muss man die Variante (3.4)

$$p_{data}^{opt} = (\frac{D_1 + D_2}{2D_3})^{\frac{2}{3}} (\frac{\log n}{n})^{\frac{1}{3}} \tag{4.11}$$

wählen, mit n multiplizieren und ganzzahlig wählen, um die Anzahl der nächsten Nachbarn zu optimieren. Nun sind die Größen zu schätzen, die zu deren Berechnung notwendig sind. Das ist kein unbekanntes Problem: alle Bandbreitenwahlen neueren Datums, die in der Einleitung erwähnt werden, haben dieses Problem, einen „Vorab-Schätzer“ , „Plug-in“ genannt, benutzen zu müssen. Es sei angemerkt, dass für die Konsistenz, die in Abschnitt 2.4 gezeigt wurde, dann weitere Regularitätsannahmen an die Vorab-Schätzung gemacht werden müssen. Aufgrund der Komplexität sei auf eine Analyse hier verzichtet. Man mache sich das Vorgehen an folgendem Flussdiagramm klar.



Hierbei wurde als Plug-in Schätzer der Hazardrate wieder der variable Kernschätzer mit nächste-Nachbarn Bandbreite verwendet, wobei die Anzahl nächster Nachbarn k so gewählt werden kann, dass zum Beispiel die von k abhängige modified Likelihood (in k) maximiert wurde. Dies ist ein bekanntes Verfahren der Bandbreitenwahl (siehe PATIL (1993)) für die fixe Bandbreite, bei dem die modifizierte Likelihood maximiert wird. Für die Adaption an die nächste-Nachbarn Bandbreitendefinition sind allerdings Modifikationen nötig. Die Konstruktion wird im folgenden Abschnitt angedeutet.

Die Frage, die sich stellt ist dann, wie stabil, soll heißen invariant, die Bandbreitenwahl in Bezug auf diesen Plug-in Schätzer ist. Um das zu überprüfen, werden in der Simulationsstudie zwei Plug-in Schätzer, oder genauer zwei Plug-in Schätzer mit unterschiedlicher Anzahl nächster Nachbarn, verglichen. Als zweite Bandbreitenwahl für den Plug-in Schätzer neben der modified Likelihood Methode bietet sich die oben entwickelte Daumenregel (4.10) an.

Über verbesserte Möglichkeiten, die Charakteristika der zu schätzenden Funktionen zu approximieren, finden sich Informationen im Abschnitt 6.3 „Technische Umsetzung der Bandbreitenwahlen“ .

4.2.1.3 Die modifizierte Likelihood

Die nichtparametrische Likelihood ist als Produkt der Dichteschätzungen an den Beobachtungsstellen definiert:

$$ML_x(k_n) = \prod_{i=1}^{n} f_n(X_i).$$

Damit nicht die Bandbreite, beziehungsweise die Anzahl nächster Nachbarn, „0“ die nichtparametrische Likelihood maximiert – dort ist die Anpassung der Daten trivialerweise optimal – wird bei jeder Beobachtung die Dichte ohne Nutzung eben der Beobachtung an deren Stelle man gerade schätzt geschätzt, in dem Sinne modifiziert. Dann werden die Schätzungen an allen Stellen zur Gesamt-Likelihood aufmultipliziert und maximiert. Alternativ können natürlich auch die logarithmierten Dichteschätzungen aufsummiert und maximiert werden. Das kann man auch zur Einbeziehung des hier betrachteten variablen Hazardratenschätzers adaptieren, indem man die Dichte als Produkt aus Hazardrate und Überlebenzeitfunktion dekomponiert und die Komponenten nichtparametrisch schätzt. Die Likelihood sieht dann wie folgt aus:

$$mL_x(k_n) = \prod_{i=1}^{n} h_n^{-i}(X_i)^{\delta_i} S_n^{-i}(X_i) \tag{4.12}$$

Asymptotisch wird damit auch das Kullback-Leibler Risiko der entsprechenden Dichte minimiert (siehe Chow, Geman, Wu (1983)). Das Auslassen der Beobachtung, an deren Stelle geschätzt wird, nennt man deswegen auch Kreuz-Validierung (Cross Validation). Für Theorie und Implementierung in der Matrix-orientierten Programmierumgebung SAS/IML siehe Gefeller, Pflüger, Bregenzer (1996). Historisch gesehen wurde die Kreuz-Validierung zunächst genutzt, um einen erwartungstreuen Schätzer des MISE zu erzielen und nachfolgend - in der fixen Bandbreite – zu minimieren; Literaturangaben dazu finden sich in der Einleitung.

Kapitel 5

Ein biometrisches Anwendungsbeispiel

In einer retrospektiven Studie wurde am Princess Margaret Krankenhaus, Toronto, Ontario 1997 bei 114 Blasenkrebsfällen in metastasiertem Zustand die Konzentration von Metallothionein im Tumorgewebe untersucht. Die physiologische Funktion dieses Proteins wird unter anderem im Zusammenhang mit Entgiftungsprozessen gesehen, das heißt es bindet toxische Substanzen und „transportiert“ sie letztendlich aus der Zelle. Es wurde vermutet, dass es dadurch zu einer verringerten Wirkung der Chemotherapeutika kommen kann. Für Details siehe Siu et al. (1997). In der Studie ging es um die prognostische Rolle der Metallothioneinkonzentration auf das Überleben von Blasenkrebspatienten. Tumorgewebe wurde den Patienten durch Biopsien entnommen. In der Pathologie derselben Klinik wurde durch immunohistologische Färbung histologischer Schnittpäparate die Konzentration von Metallothionein bestimmt. Aufgrund der bisherigen Kenntnisse zu effektiven Konzentrationen von Metallothionein wurde das Patientenkollektiv in zwei Straten mit jeweils $>$ oder $\leq 10\%$ Färbung auf Metallothionein zerlegt. Es ergaben sich Gruppengrößen von von 45 und 69 Patienten. Es wurden die Überlebenszeitfunktionsschätzer für das Überleben nach der Entfernung des Primärtumors in diesen zwei Straten nach Kaplan-Meier geschätzt (siehe Abbildung 5.1).

Bei der Betrachtung der gefundenen Überlebenszeitfunktionen muss als erstes be-

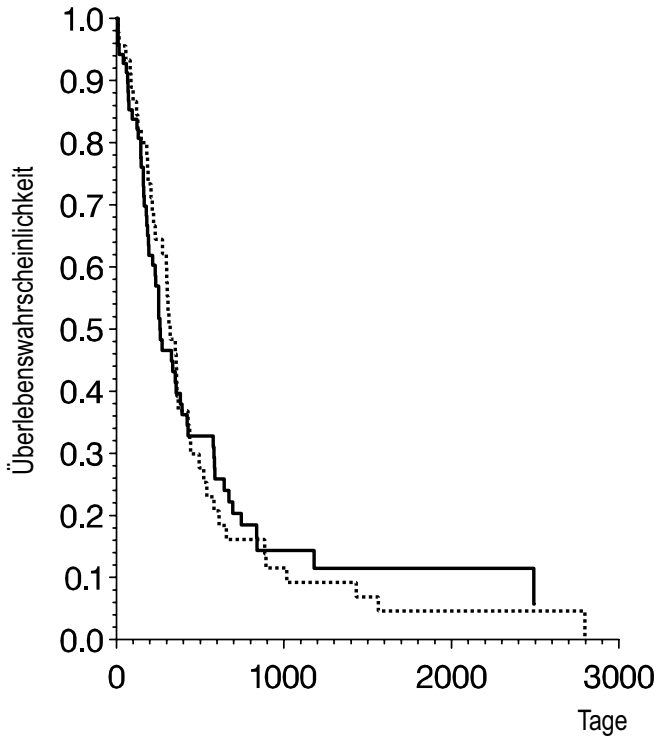


Abbildung 5.1: Schätzung der Überlebenszeitfunktion nach KAPLAN-MEIER für die Patientengruppe mit geringerer Metallothioneinkonzentration ($\leq 10\%$) (durchgezogen) und die mit höherer Konzentration ($> 10\%$) (gestrichelt)

merkt werden, dass nach circa 1000 Tagen insgesamt nur noch sieben Personen unzensiert beobachtet wurden. Daher sollte man sich also bei der Interpretation auf die Zeitspanne bis 1000 beschränken. Es fällt auf, dass die geschätzte Überlebenszeitfunktion des Kollektives mit geringerer Konzentration von Metallothionein im Gewebe – die durchgezogene Kurve – die andere – bis circa 400 Tage – unterschreitet, dann kreuzt und danach überschreitet. Nach 1000 Tagen fällt sie schließlich wieder auf das Niveau der zweiten. Dieser Verlauf ist allerdings nur sehr schwach sichtbar. Da es für das Beispiel interessant ist, soll hier auch die schließende Statistik erwähnt werden. Der log-rank Tests – durchgeführt mit der Prozedur „lifetest“ in SAS – wird nicht signifikant (p-Wert= 0.6757). Interpretierend kann man so nicht trennen, ob nicht der Unterschied zu schwach war, um mit vorliegender Stichprobenstärke aufgedeckt zu werden, ob also die Macht des Tests nicht ausreichte, oder

ob tatsächlich kein relevanter Unterschied existiert. Bekannt ist, dass der log-rank Test geringe Macht hat, da er auf der asymptotischen Verteilung des gleichmäßigen Abstand beruht, auf deren (finite) Unsicherheit ich in Abschnitt 2.5.3 hingewiesen habe.

Ich will hier diesen Aspekt nicht weiter verfolgen, sondern die Punktschätzung der Kurve betrachten. Natürlich bleibt die Ungewissheit, ob es sich dann um die Überinterpretation von Zufallsschwankungen handelt, oder die gefundenen Aspekte substantiell sind.

Inhaltlich interpretierend kann man aber mutmaßen, dass im Intervall 0 bis circa ein Jahr nach der Operation die Patienten mit weniger Metallothionein ein höheres Risiko tragen und erst ab dann ein geringeres Risiko für sich beanspruchen können – immer im Vergleich zur anderen Gruppe. Wir können also eine Umkehr des Risikoverhaltens annehmen. Nur wann hat sich das Risikoverhalten wirklich umgekehrt? Bei der 400-Tage Grenze hat sich die Umkehr kumulativ bemerkbar gemacht, die eigentliche Risikoumkehr muss aber früher eingesetzt sein.

Das momentane Riskoverhalten, das für die medizinische Nachsorge der Patienten viel wichtiger ist als das kumulative, wird mittels der Hazardrate gemessen. Diese soll nun mit dem variablen Kernschätzer mit nächste-Nachbarn Bandbreite (4.8) geschätzt werden. Dafür sollen die beiden entwickelten Wahlen für die Anzahl nächster Nachbarn, die den gleichmäßigen Fehler minimierende (4.11) und die Daumenregel (4.10) verwendet werden.

In der Abbildung 5.2 wird die erste Hazardratenschätzung grafisch dargestellt. Als Anzahlen der nächsten Nachbarn zur Minimierung des gleichmäßig-absoluten Fehlers ergeben sich nach Optimierung (3.4) 23 beziehungsweise 19. Die Plug-in Anzahl nächster Nachbarn zur Schätzung der Kenngrößen waren 38 beziehungsweise 46 aus der modified Likelihood Maximierung für die zwei Gruppen.

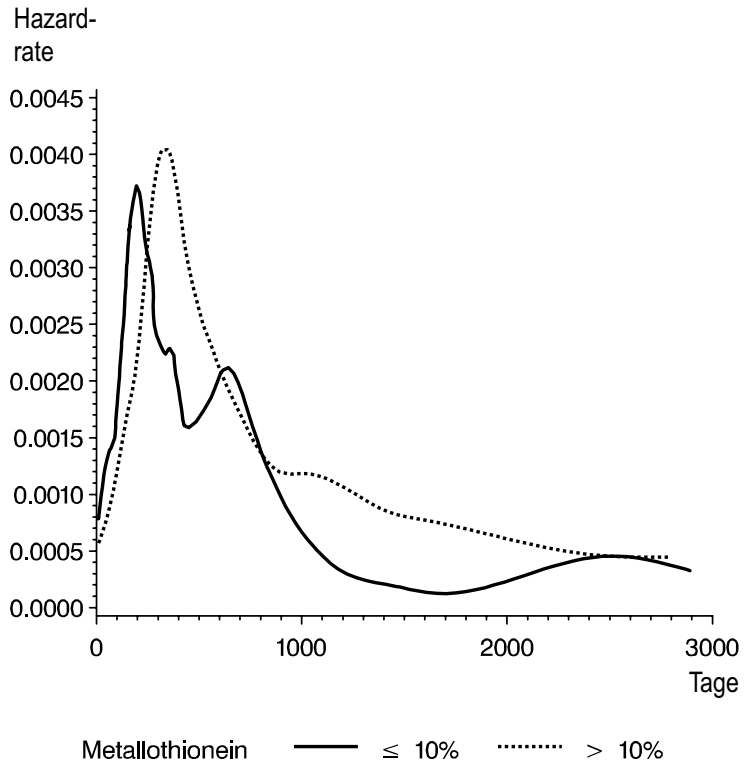


Abbildung 5.2: Schätzung der Hazardfunktion mit UAE-optimaler Wahl der nächste-Nachbarn Bandbreite für die Patientengruppe mit geringerer Metallothioneinkonzentration ($\leq 10\%$) (durchgezogen) und die mit höherer Konzentration ($> 10\%$) (gestrichelt)

Die Abbildung 5.3 stellt die Hazardratenschätzung mit der im Abschnitt 4.2.1.1 vorgestellten Daumenregel dar. Für die Metallothionein Strata ergeben sich als Anzahlen nächste Nachbarn $k_n^{RoT} = 12$ für $\leq 10\%$ und $k_n^{RoT} = 17$ für $> 10\%$.

Man sieht an den beiden Abbildungen, dass beide Bandbreitenwahlen zu ähnlichen Formen der Hazardrate führen, selber aber erheblich von einander abweichen. Die Bandbreite, beziehungsweise die Anzahl nächster Nachbarn, die die modified Likelihood Maximierung empfiehlt, ist indes wesentlich größer als die veranschaulichten. Es sei angemerkt, dass die Daumenregel innerhalb von Sekunden vom Computer (Pentium II Prozessor, 450 MHz, 256 MB RAM) berechnet wird, während die UAE-optimale Anzahl nächster Nachbarn einige Minuten benötigt. Das liegt auch an der rechenintensiven Enumeration der Kreuz-Validierung und gilt deswegen auch für die

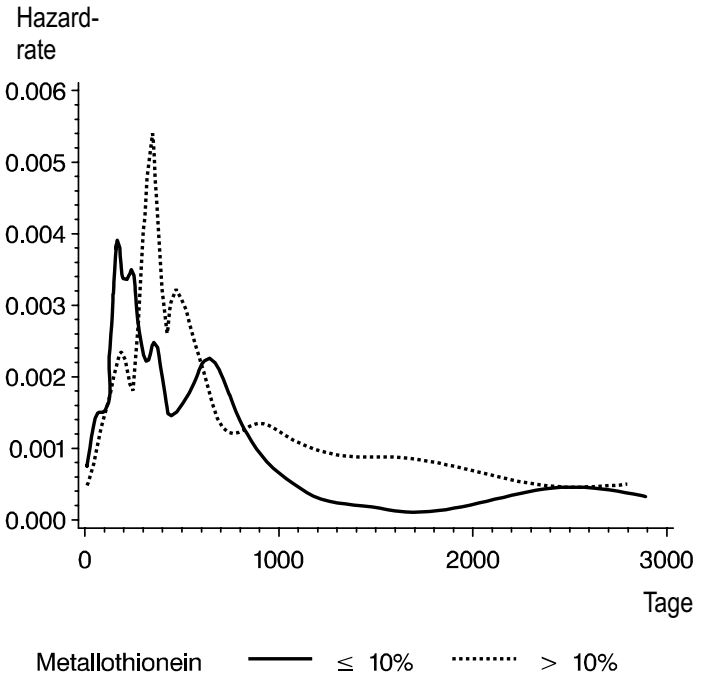


Abbildung 5.3: Schätzung der Hazardfunktion mit Daumenregel zur Wahl der nächste-Nachbarn Bandbreite für die Patientengruppe mit geringerer Metallothioneinkonzentration ($\leq$ 10%) (durchgezogen) und die mit höherer Konzentration ($>$ 10%) (gestrichelt)

hier nur als Plug-in gewählte Bandbreite nach der modified Likelihood Maximierung. Weitere Hinweise auf die Rechenintensität der verschiedenen Bandbreitenwahlen liefert die Simulationsstudie, die im folgenden Kapitel ausgeführt wird.

Inhaltlich interpretierend kann man nun den gewünschten Wendepunkt des Risikoverhaltens, der im ersten Schnittpunkt der Hazardraten – bei circa 300 Tagen – liegt ablesen. Zudem können die Moden der Hazardraten Aufschluss über inhomogene Subgruppen geben. Die Bimodalität der Hazardrate innerhalb der Patientengruppe mit geringerer Konzentration deutet auf eine solche Inhomogenität hin. In der Tat kann durch den Ausschluss der Patienten, die keinen Tumor des Stadium III haben, ein sogar in der schließenden Statistik signifikanter Unterschied in den Überlebenszeitfunktionen gefunden werden (p-Wert= 0.04). Diese Signifikanz ist umso erstaunlicher, da sich durch den Ausschluss das Patientenkollektiv auf 59 Patienten

reduziert, (siehe hierzu wieder Siu et al. (1997)).

Kapitel 6

Simulationsstudie

Zweck einer Simulationsstudie für Schätzer ist es, das Verhalten der Schätzer bei vorgegebener Struktur der Daten, das heißt im allgemeinen bei vorgegebener Wahrscheinlichkeitsverteilung der Beobachtungen, zu evaluieren. In der vorliegenden Simulationsstudie soll das Verhalten des allgemeinen Funkionalschätzers (2.4) anhand der Hazardratenschätzung mit nächste-Nachbarn Bandbreite (4.8) beurteilt werden.

Zunächst soll über die Wahl der Verteilungsfamilie entschieden werden, aus der wir unsere Daten generieren wollen.

6.1 Datenerzeugung

6.1.1 Die Exponentielle Weibull Familie

Um die Güte einer Schätzung per Simulation zu evaluieren, muss man sich für eine (oder mehrere) Verteilung(en) entscheiden. Diese Verteilung soll den praktischen Problemen angepasst sein. Das heißt, sie soll keine für die behandelte Problematik unspezifischen Charakeristika enthalten und dennoch flexibel die unterschiedlichen — zum Beispiel biologischen oder biomedizinischen — Prozesse abbilden. Ersteres bewog mich, von der in der Kern-Dichteschätzung ausgezeichneten Verteilungsfamilie der gemischten Normalverteilungen von Marron & Wand (1992) abzusehen. Mit ihr werden für die Humanbiologie untypische, stark variierende Dichten wie die

fünf-modale Claw(Klauen)-Verteilung modelliert. Diese Familie scheint einerseits „zu reich“ — zum Beispiel, was die Anzahl der Moden angeht — und andererseits „zu arm“ zu sein, da insbesondere die Modellierung schiefer Verteilungen, die in der Überlebenszeitanalyse aufgrund der Positivität der Daten üblich sind, schwierig ist.

Eine Verteilungsfamilie, die die Modellierung schiefer und explizit positiver Verteilungen zulässt, ist die Familie der HJORTH-Verteilungen (HJORTH (1980)), die fallende, steigende, konstante und badewannenförmige Hazardraten modelliert. Diese Familie wurde zum Beispiel von GEFELLER (1986) zur Evaluation von Bandbreitenwahlen bei der Hazardratenschätzung verwendet.

Vor kurzem haben MUDHOLKAR, SRIVASTAVA & FREIMER (1995) die exponentielle Weibull Familie vorgestellt und diskutiert, die zwar keine analytische Verallgemeinerung der HJORTH-Verteilungen ist, aber dieselben Hazardratenformen mit der zusätzlichen und für die Anwendung wichtigen unimodalen Hazardrate darstellt. Deshalb beschränke ich mich in meiner Simulationstudie auf diese Verteilungen, die sich analytisch als Überlebenszeitfunktionen folgenden Typs darstellen lassen:

$$S(x) = 1 - (1 - \exp(-(x/\sigma)^\alpha))^\theta, \tag{6.1}$$

mit $0 < x < \infty$, $\alpha > 0$, $\theta > 0$ und $\sigma > 0$. Somit stellt die Familie eine Verallgemeinerung zur Weibullverteilung dar, die man für $\theta = 1$ erhält (siehe zum Beispiel KALBFLEISCH & PRENTICE (1980)). Die Dichte ist

$$f(x) = \frac{\alpha\theta}{\sigma}(1 - \exp(-(x/\sigma)^\alpha))^{\theta-1} \exp(-(x/\sigma)^\alpha)(\frac{x}{\sigma})^{\alpha-1}. \tag{6.2}$$

Als Hazardrate stellt sich die Verteilungsfamilie wie folgt dar:

$$h(x) = \frac{\alpha\theta(1 - \exp(-(x/\sigma)^\alpha))^{\theta-1} \exp(-(x/\sigma)^\alpha)(x/\sigma)^{\alpha-1}}{\sigma(1 - (1 - \exp(-(x/\sigma)^\alpha))^\theta)}.$$

Die vier zu modellierenden Formen einer Hazardrate segmentieren den Parameterraum explizit. Die Grenzlinien sind $\alpha = 1$ und $\alpha \cdot \theta = 1$. Die Entsprechungen sind in der Tabelle 6.1 dargelegt.

Tabelle 6.1: Segmente des Parameterraums und zugehörige Hazardfunktionsformen für die exponentielle Weibull Familie

Verteilung	Hazardrate	Parameterraum
Typ I	Badewannenform	$\alpha > 1$ und $\alpha \cdot \theta < 1$
Typ II	unimodal	$\alpha < 1$ und $\alpha \cdot \theta > 1$
Typ III	monoton fallend	$\alpha \leq 1$ und $\alpha \cdot \theta \leq 1$
Typ IV	monoton steigend	$\alpha \geq 1$ und $\alpha \cdot \theta \geq 1$

Hier sind die Monotonien strikt bis auf den Punkt $\alpha = \theta = 1$, der mit der Exponentialverteilung korrespondiert, also eine konstante Hazardrate hat.

Grafen der vier Typen sind im Folgenden aufgelistet. Da für die Ergebnisse der simulierten Hazardraten die Dichte der Daten in den verschiedenen Bereichen der x-Achse wichtig ist, werden Dichte und Überlebenszeitfunktion zusätzlich dargestellt. Es ist der „Abszissenausschnitt“ $\in (0, 100]$ gewählt, weil er in den dargestellten Beispielen von Mudholkar, Srivastava & Freimer (1995) die entscheidende Charakteristik der Hazardraten enthält.

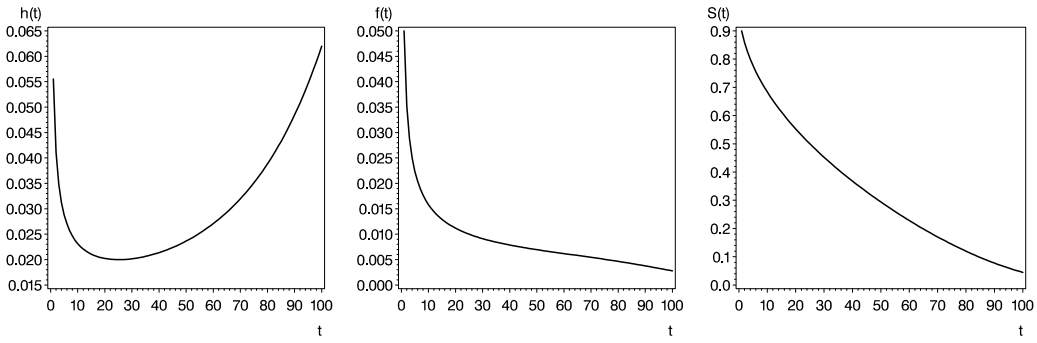


Abbildung 6.1: Hazardfunktion, Dichtefunktion und Überlebenszeitfunktion einer Typ I-Verteilung mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$

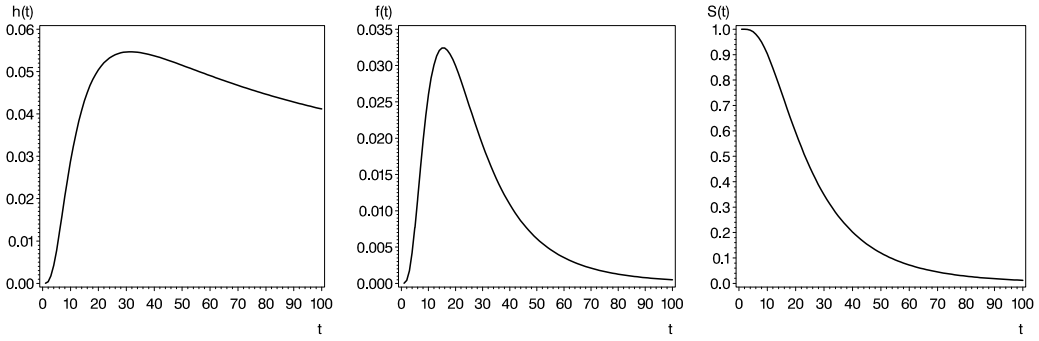


Abbildung 6.2: Hazardfunktion, Dichtefunktion und Überlebenszeitfunktion einer Typ II-Verteilung mit Parametern $\alpha = 0.6$, $\theta = 12$, $\sigma = 4$

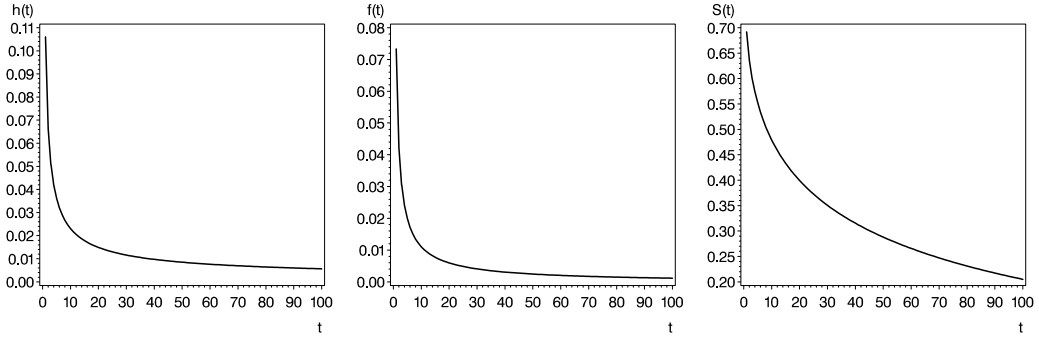


Abbildung 6.3: Hazardfunktion, Dichtefunktion und Überlebenszeitfunktion einer Typ III-Verteilung mit Parametern $\alpha = 0.5$, $\theta = 0.5$, $\sigma = 100$

6.1.2 Erzeugung der zensierten Zufallsvariablen

Wenn wir die Verteilung, das heißt in unserem Fall die Hazardfunktion, aus Daten schätzen wollen, müssen wir letztere im Rahmen einer Simulation zunächst erzeugen. Das ist im allgemeinen für eine Zufallsvariable X mit Verteilungsfunktion $F(\cdot)$ bei Kenntnis deren Umkehrfunktion $F^{-1}(\cdot)$ möglich. Mit der Erkenntnis, dass

$$P(X \leq x) = F(x) = P(U \leq F(x)) = P(F^{-1}(U) \leq x)$$

für eine gleichverteilte Zufallsvariable U auf $[0, 1]$ gilt, simuliert man U und erzeugt mit $F^{-1}(U)$ eine entsprechende Zufallsvariable. Für die exponentielle Weibull Fami-

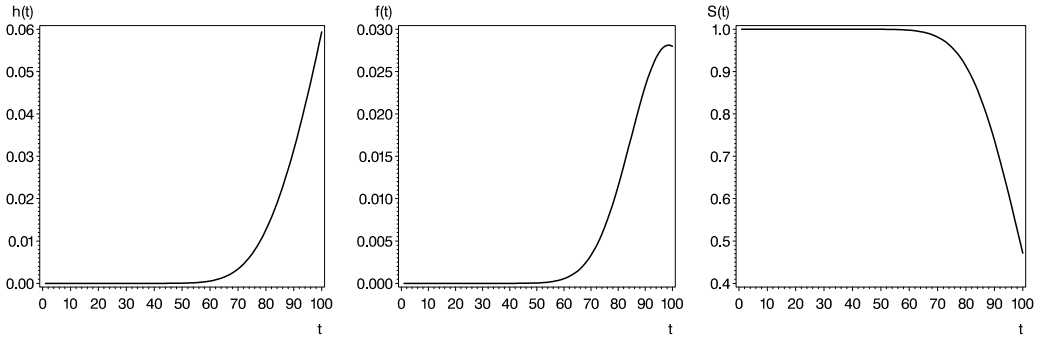


Abbildung 6.4: Hazardfunktion, Dichtefunktion und Überlebenszeitfunktion einer Typ IV-Verteilung mit Parametern $\alpha = 4$, $\theta = 4$, $\sigma = 85$

lie ist die Quantilfunktion gegeben durch

$$F^{-1}(u) = \sigma[-\log(1 - u^{\frac{1}{\theta}})]^{\frac{1}{\alpha}}, \quad 0 < u < 1.$$

Wenn wir zensierte Daten gemäß der exponentiellen Weibull Familie generieren wollen, müssen wir sowohl n unabhängige Ereignisse T_i, $i = 1, \ldots, n$ als auch – unabhängig davon – n unabhängige „Zensierungszeiten“ C_i, $i = 1, \ldots, n$ erzeugen. Im klassischen rechts-zensierten Modell sind die beobachteten Zeiten dann $X_i := \min\{T_i, C_i\}$ und die Zensierungsindikatoren $\delta_i := 1_{\{X_i = T_i\}}$, $i = 1, \ldots, n$. Der Zensierungsgrad, also der Anteil der zensierten ($\delta_i = 0$) Beobachtungen an allen Beobachtungen, ist ein Maß für den Informationsgehalt unseres Datensatzes. Ein höherer Grad bedeutet, dass bei mehr Beobachtungen das Zielereignis zum Beobachtungszeitpunkt (noch) nicht eingetreten ist, also weniger Infomation vorhanden ist. Somit ist es wünschenswert, den Zensierungsgrad in eine Simulation einzubeziehen, beziehungsweise verschiedene Zensierungsgrade vorzugeben. Hierfür ist es zunächst notwendig, den Zensierungsgrad in einem Datensatz zu bestimmen. Wenn wir also davon ausgehen, dass sowohl die „Todeszeiten“ als auch die „Zensierungszeiten“ einer exponentiellen Weibull Verteilung – mit Dichten $f_T(\cdot)$ und $f_C(\cdot)$ – folgen, so ist

der zu erwartende Zensierungsgrad

$$\begin{aligned} P(C<T) &= \int_{\{(x,y):x>y\}} f_T(x)f_C(y)dydx \\ &= \int_0^\infty \int_0^x f_T(x)f_C(y)dydx \\ &= \int_0^\infty f_T(x)\int_0^x f_C(y)dydx \quad \text{siehe nun (6.2) und (6.1)} \\ &= \int_0^\infty \frac{\alpha_T\theta_T}{\sigma_T}(1-\exp(-(x/\sigma_T)^{\alpha_T}))^{\theta_T-1}\exp(-(x/\sigma_T)^{\alpha_T})(\frac{x}{\sigma_T})^{\alpha_T-1} \\ &\quad (1-\exp(-(x/\sigma_C)^{\alpha_C}))^{\theta_C}dx. \end{aligned} \tag{6.3}$$

Es sei nun bemerkt, dass ein Zensierungsgrad von 50% am leichtesten dadurch erreicht wird, das Ereignis- und Zensierungszeitenverteilung gleich gewählt werden. Da der Integrand – und damit auch das Integral – in (6.3) in σ_C, $\frac{1}{\alpha_C}$ und θ_C monoton (und gleichmäßig auf 0) fällt, wähle man einen der drei Parameter per Simulation so, dass der gewünschte Zensierungsgrad erreicht wird.

Bemerkung: Nun ist 6.3 aber auch

$$\begin{aligned} &= \frac{\alpha_T\theta_T\sigma_T}{\sigma_T(\theta_T-1)\alpha_T}\int_0^\infty (1-\exp(-(x/\sigma_T)^{\alpha_T}))^{\theta_T-1}(\theta_T-1) \\ &\quad (1-\exp(-(x/\sigma_T)^{\alpha_T}))^{\theta_T-2}\exp(-(x/\sigma_T)^{\alpha_T})\alpha_T(\frac{x}{\sigma_T})^{\alpha_T-1}\sigma_T^{-1}dx \\ &\quad \text{für} \quad \theta_C=\theta_T-2, \sigma_C=\sigma_T, \alpha_C=\alpha_T \\ &= \frac{\theta_T}{2(\theta_T-1)}, \end{aligned}$$

wobei der letzte Schritt durch die Substitution $z=(1-\exp(-(x/\sigma_T)^{\alpha_T}))^{\theta_T-1}$ einsichtig ist. Die Integrationsgrenzen transformieren sich dann in $[0,1]$, da $\theta_T > 2$ durch die Definition $\theta_C = \theta_T - 2 > 0$ notwendig ist. Nun kann man in jeder Simulationssituation $\frac{\theta_T}{2(\theta_T-1)}$ ausrechnen, so dass man für die Fälle, in denen θ nicht größer als 2 ist, auch von dieser Konstellation aus, die Monotonie ausnutzend, die gewünschte Zensierungsrate erzielen kann.

Wir wollen die Zensierungsraten 0% und 40% als Simulationsszenarien wählen, um erstens in einem unzensierten Szenario die Relevanz des Schätzverfahrens jenseits der

Tabelle 6.2: Parameterwahl für die Simulationsverteilungen der exponentiellen Weibull Familie und resultierende innere 80%-Bereich

Typ	Todeszeit	Zensierungszeit	$[F^{-1}(0.1); F^{-1}(0.9)]$
I (Badewannenform)	$(5, 0.1, 100)$	$(5, 0.15, 100)$	$[1; 84.419192]$
II (unimodal)	$(0.5, 9.5, 5)$	$(0.5, 14.3, 5)$	$[11.796502; 101.57423]$
III (monoton fallend)	$(0.5, 1.8, 10)$	$(0.5, 2.75, 10)$	$[1.0633123; 82.212834]$
IV (monoton steigend)	$(1.5, 1.0, 33)$	$(1.5, 1.51, 33)$	$[7.3614923; 57.54281]$

Zensierungsproblematik zu eruieren und zweitens ein typisches Szenario im Kontext klinischer (Überlebenszeit-)Studien zu generieren. Erwartet wird hierfür eine Reproduktion der Ergebnisse von 0% Zensierung, allerdings wegen des Infomationsverlusts mit geringerer Präzision. Als Parameter (α, θ, σ) für die Ereignis- und Zensierungszeitenverteilung der vier Typen der exponentiellen Weibull Familie sind dann die in Tabelle 6.2 aufgeführten geeignet, wobei die Zensierungzeiten für den unzensierten Fall entfallen.

Die Simulation der Zensierungsrate wurde mit Stichprobenumfängen von 100.000 durchgeführt, so dass von einer exakten Zensierungrate in jenseits der Nachkommastellen ausgegangen werden kann, da die Standardabweichung maximal 0.00158 beträgt, was für unsere Zwecke ausreicht.

Das Verhalten der Schätzfunktionen ist an den Rändern, das heißt in den Außenbereichen der Verteilung, wegen der spärlichen Beobachtungen besonders instabil. Man sollte also, wie im vorangegangen Beispiel, auf die Interpretation des Verhaltens des Schätzers dort verzichten. In der Simulation kann die Variabilität die Beurteilung der verschiedenen Bandbreitenwahlen stören und sogar unbrauchbar machen. Ich will mich deshalb auf den inneren symmetrischen 80%-Bereich $[F^{-1}(0.1); F^{-1}(0.9)]$ der Verteilungen beschränken. Die Intervalllängen, innerhalb derer sich 80% der zu erwartenden Beobachtungen befinden, sind trotz des Versuchs, sie über die vier zu untersuchenden Typen gleich zu halten, für die obige Parameterwahl nicht bei allen Typen genau gleich. Man kann das bei der Interpretation gegebenenfalls berücksichtigen, aber es wird sich als unproblematisch herausstellen, da nicht über die Verteilungen hinweg absolute Maßzahlen verglichen werden

müssen. Es sei bemerkt, dass selbst eine algebraische Gleichheit der Grenzen der 80%-Träger ($[F^{-1}(0.1); F^{-1}(0.9)]$) für die vier Verteilungsrepräsentanten nicht zu gewährleisten ist. Es ergibt sich ein Gleichungssystem mit 6 Gleichungen und 4 beliebig zu wählenden σ_i. Wenn man über die anderen, an Restriktionen gebundenen, Parameter keine Annahmen machen möchte, ist dieses nicht lösbar.

Die Parameterkonstellationen, die beispielhaft in Mudholkar, Srivastav & Freimer (1995) genannt werden und deren Form im Text schon dargestellt wurde, werden nicht verwandt, da bei diesen zwar der Wertebereich ($[0; 100]$), in denen die Charakeristik der Hazardraten zu erkennen ist, gleich ist, aber dieses nicht der Bereich ist, in denen der Großteil der Daten liegt. So liegen zum Beispiel 80% der Daten für die angegebene Verteilung des Typs IV mit Parametervektor $(4, 4, 85)$ im Interval $[81.041125; 117.48604]$.

Da die Zensierung sich nicht gleichmäßig (proportional) über die Achse verteilen muss, stimmt der Zensierungsgrad nicht mehr genau auf dem betrachteten Intervall. Streng genommen müssten wir somit in obigen Überlegungen von Ereignissen T_i und nicht von Beobachtungen X_i sprechen, da der mittlere 80%-Bereich nicht vor und nach der Zensierung für beide gleich sein muss. Ich werde also die tatsächlichen Zensierunggrade im 80%-Quantil mitprotokollieren und bei Abweichungen von der nominalen Vorgabe auf sie eingehen.

Man bemerke, dass die Parameterwahl der exponentiellen Weibull Verteilung vom Typ IV $(1.5, 1.0, 33)$ einer Weibullverteilung mit den Parametern $p = 1.5$ und $\lambda = \frac{1}{33}$ (gemäß Kalbfleisch & Prentice (1980)) entspricht.

6.2 Ziele und Design

Das Ziel dieser Simulation ist die Bandbreitenevaluation. Zum einen soll der Nutzen der fast sicher gleichmäßig absoluten Asymptotik für die Bandbreitenwahl betrachtet werden. So soll ermittelt werden, ob die UAE-optimale Bandbreite (3.4) einen positiven Nutzen hat, das heißt, ob sie bezüglich der noch zu definierenden Zielkriterien eine Verbesserung im Vergleich mit der Plug-in Anzahl nächster Nachbarn erwirtschaftet. Wir wollen hier als Referenz Plug-in denjenigen nehmen, der per

Kreuz-Validierung (cross-validation) den Kullback-Leibler Verlust (asymptotisch) minimiert (siehe 4.2.1.3). Diese Bandbreitenwahl, die man auch als modifizierte Likelihood Maximierung erhält, und die wir deswegen mit „ml-Bandbreite“ abkürzen wollen, ist ein typischer Vertreter der Kreuz-Validierungsbandbreitenwahlen und wurde in Abschnitt 4.2.1.3 vorgestellt. Darüber hinaus ist interessant, wie sich die Wahl der Plug-in Bandbreite auf diese UAE-optimale Bandbreite auswirkt. Dieser Effekt ist insbesondere deshalb interessant, da die in der UAE-optimalen Bandbreite enthaltenen Minima und Maxima schwierig zu schätzen sind. Die Variabilität dieser Schätzung könnte sich auf die Plug-in Bandbreite übertragen.

Zum anderen soll die Frage, ob die neuentwickelte Daumenregel („Rule of Thumb“ (RoT) 4.2.1.1) sich gegen die asymptotisch motivierte UAE-optimale Bandbreite behaupten kann und ob der einfache Kernschätzer mit fixer Bandbreite und herkömmlicher Daumenregel (siehe Wand & Jones (1995)) als Konkurrent klar unterlegen ist, beantwortet werden.

6.2.1 Zielkriterien

Das zentrale qualitive Zielkriterium ist die punktweise Mittelung – über die Simulationsdatensätze – der Hazardratenschätzungen in ihrer grafischen Darstellung und im Vergleich mit der zu schätzenden Hazardrate. Dieses Vorgehen ist pro Situation zu betrachten. Zu erwarten sind sowohl Schlüsse über die Verzerrung, als auch über die Variabilität der Schätzung.

Als quantitative Zielkriterien sind die in der Literatur üblichen

- MISE,
- integrierter Bias (IBias) und
- integrierte Varianz (IVarianz)

sowie die hier interessanten

- gleichmäßig absoluter Fehler (UAE) und

- zu erwartender integrierter Kullback-Leibler Fehler(MIKLE)

vorgesehen. Wie oben motiviert, sollen die Kriterien auf dem Interval $[F^{-1}(0.1); F^{-1}(0.9)]$ ausgewertet werden, um Randeffekte unberücksichtigt zu lassen. Dasselbe gilt auch für die grafische Bewertung.

Für alle Kriterien sollen empirische Varianz und Mittelwert aus den Simulationswiederholungen über die Variabilität und Präzision der Schätzung Auskunft geben.

Zudem soll für alle Fragestellungen die Bandbreite selbst beziehungsweise die Anzahl nächster Nachbarn in ihrer Streuung Auskunft über die Variabilität der Bandbreitenwahl und mittelbar auch der Schätzung an sich geben. Dass es ich hierbei um eine relevante Information handelt, illustriert die Arbeit von Park & Marron (1990), in der die Variabilität der Bandbreite für den fixen Dichteschätzer minimiert wird.

6.2.2 Design

Die Fragestellungen sollen für in der Praxis übliche Studienumfänge und Subgruppenumfänge von 50, 100 und 300 Beobachtungen betrachtet werden. Wie in Hess, Serachitopol & Brown (1999) sollen 500 Simulationsdurchläufe gewählt werden. Dabei können die 500 Simulationsdurchläufe in einigen Fällen aufgrund der Computerlaufzeit auf 250 Durchläufe reduziert werden. Das scheint ausreichend, da es sich um die Evaluation einer Punktschätzung handelt und nicht um die einer Konfidenz- oder Testniveauschätzung. Außerdem ließ sich trotz der guten Computerressourcen (circa zehn Intel Pentium II Prozessoren mit 450 MHz Takt 256 MB RAM und einem Intel Pentium III Prozessor mit 600 MHz Takt und 1024 MB RAM) keine höhere Anzahl realisieren.

Insgesamt werden alle genannten Ziele für alle verschiedenen Bandbreitenwahlen an den *vier* Typen der exponentiellen Weibull Familie, *drei* Stichprobenumfängen und *zwei* Zensierungsgraden evaluiert. Die konkreten Szenarien zur Zielevaluation sind in Tabelle 6.3 aufgelistet. Es wird erwartet, dass aufgrund der empfindlichen Schätzung der UAE-optimalen Bandbreite durch Schätzung von Maxima $(M, \tilde{M})$, Minima $(m, \tilde{m})$ und Lipschitz-Konstanten $(L_\psi, L_{\tilde{\psi}})$ - also Maxima der Ableitung -

Tabelle 6.3: Ziel-orientierter Szenarienentwurf für die Simulation

Ziel	**Szenario 1**	**Szenario 2**
1.Etablierung einer neuen Daumenregel für die nächste-Nachbarn Bandbreite	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und Daumenregel	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und mL-Anzahl
2.Externer Vergleich der Daumenregel	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und Daumenregel	Hazardratenschätzung mit fixer Bandbreite und deren Daumenregel
3.Evaluation der UAE-optimalen Bandbreite	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und ml-Anzahl	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und UAE-optimaler Bandbreite bei mL-Plug-in
4.Evaluation der Plug-in Robustheit	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und UAE-optimaler Anzahl bei mL-Plug-in	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und UAE-optimaler Anzahl bei Daumenregel-Plug-in
5.Externer Vergleich der UAE-Methodik	Hazardratenschätzung mit nächste-Nachbarn Bandbreite und UAE-optimaler Anzahl bei mL-Plug-in	Hazardratenschätzung mit fixer Bandbreite und deren Daumenregel

die Bandbreite eine hohe Variabilität aufweisst. In einer praktischen Situation würde man dann die Bandbreite mit sinnvollen Schranken versehen, um zufällige extreme Unter- und Überglättung zu vermeiden. In einer Simulation ist dies dagegen nicht sinnvoll, da wertvolle Information über das Verhalten der Bandbreitenwahl verloren ginge, das heißt es wäre unklar, ob man die Ergebnisse der Bandbreitenwahl oder den Schranken (oder beiden) zuzuschreiben hätte.

Wiederholtes Sampling von Zufallsvariablen stellt eine unnötige Quelle für Zufallsschwankungen der interessierenden Zielgrößen dar. Um sie zu vermeiden, wurden für alle Simulationssituationen einmalig Datensätze zufallsgeneriert und dauerhaft gespeichert. Die verschiedenen Bandbreitenwahlen wurden dann immer an dem selben Sample angewandt.

Die Simulation wird in SAS/IML, der Matrix-orientierten Programmiersprache im statistischen Programmpacket SAS, Version 6.12, durchgeführt. Die Generierung der Zufallszahlen wird mit RANUNI durchgeführt. Eine sorgfältige Validierung des Quellcodes wurde mit kleinen Testdatensätzen gewährleistet. Ein Auflistung des Quellcodes ist im Anhang zu finden.

6.3 Technische Umsetzung der Bandbreitenwahlen

In allen vorkommenden Fällen wurde die numerische Integration mit der Trapezregel verwandt:

$$\int_A^B g(x)dx \approx \frac{B-A}{n}(\sum_{i=1}^{n} g(x_i) - \frac{1}{2}(g(x_1) + g(x_n))).$$

In allen Fällen wird der bi-quadratische Kern mit Träger $[-\frac{1}{2}; \frac{1}{2}]$

$$K(x) = I_{[-\frac{1}{2};\frac{1}{2}]} \frac{240}{23}(\frac{1}{4} - x^2)^2$$

verwandt. Die Charakteristika des Kerns $R(K) := \int K(z)^2$ und $\mu_2(K) := \int z^2 K(z)$,

Tabelle 6.4: Charakeristika des biquadratischen Kerns

Charakteristikum	numerischer Wert
Supremum $\sup(K)$	0.652173913
Lipschitzkonstante L_K	2.008174849
Totalvariation $V(K)$	0.326086957

die für alle Bandbreitenwahlen benötigt werden, sind, wie auch für andere Kerne, JONES & WAND (1995) zu entnehmen.

Die Charakteristika des bi-quadratischen Kerns, die nur in der UAE-optimale Bandbreitenwahl benötigt werden sind in Tabelle 6.4 aufgelistet. Sie ergeben sich aus einer elementaren Kurvendiskussion.

6.3.1 Daumenregel für nächste-Nachbarn Bandbreite

Bei der Daumenregel (4.10) für die Anzahl nächster Nachbarn k_n^{RoT} muss der Schätzer für die Standardabweichung spezifiziert werden. $\hat{\sigma}$ wird als Wurzel des unverzerrten Schätzers der Varianz $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ über alle Beobachtungen gebildet.

Man könnte auch die SJPI-Methode (nach SHEATHER & JONES, siehe JONES, MARRON & SHEATHER (1996)) zur Schätzung der MISE-optimalen fixen Bandbreite verwenden, um – wie bei der Daumenregel – zu einer optimalen Anzahl nächster Nachbarn bezüglich des MIWSE zu kommen. Leider ist die Implementierung in SAS 8.0 – in der Prozedur „proc kde“ – erst vor kurzem auf den Markt gekommen.

6.3.2 Modified-Likelihood-Anzahl nächster Nachbarn

Die modified likelihood Maximierung, die einer Kreuz-Validierung mit Kullback-Leibler Verlustfunktion entspricht, ist ausführlich in GEFELLER, PFLÜGER, BREGENZER (1996) besprochen. Es wurde die vollständige Enumeration ersetzt durch eine Enumeration, die

- von $\frac{1}{4}n$ bis $\frac{3}{4}n$ läuft, da das Maximum der bekanntermaßen konkaven Likelihood in der „Mitte“ zu erwarten ist,

- abbricht, wenn die Likelihood 3-malig in Folge gefallen ist (auch wegen der Konkavität) und

- in $\frac{nobs}{50}$-er Schritten läuft.

Diese Reduktion des Rechenaufwands ist notwenig, um für große Stichprobenumfänge eine praktikable Rechenlaufzeit der Computer zu erzielen.

6.3.3 Gleichmäßig-absoluter-Abstands-Optimalität

Die Minimierung des gleichmäßigen Abstandes resultiert in mehreren Bandbreitenwahlen: (3.1), deren Variante mit Multiplikation der Streuung zur Skalenäquivarianz und (3.4). Bei der Implementierung letzterer müssen Charakteristika der zu schätzenden Funktion - also der Hazardrate - und der glättenden Funktion - also der Dichte - geschätzt werden. Dabei wird der x-Wert der maximalen Dichte dort vermutet, das heißt geschätzt, wo die Intervalllänge der aufeinander folgenden, geordneten, unzensierten, Beobachtungen minimal wird. Maximierung der Intervalllänge deutet auf minimale Dichte hin. Zur Schätzung der Maxima ($\tilde{M}$), beziehungsweise Minima ($\tilde{m}$) wird in der (arithmetischen) Mitte dieses Intervalls die Dichte geschätzt. Hierfür ist die initiale Bandbreite nötig, die im Referenz Plug-in per Kreuz-Validierung gewählt wird. Für die Schätzung der Kenngrößen für die Hazardrate (M und m) wird die Vorstellung gewählt, dass die Hazardrate dann maximal/minimal ist, wenn die Intervalllänge der aufeinander folgenden, geordneten, unzensierten Beobachtungen – in Bezug gesetzt zu der Anzahl schon beobachteter kumulativer empirischer Wahrscheinlichkeitsmasse – maximal/minimal ist. Anders ausgedrückt bedeutet das, dass die Hazardrate, also der Quotient aus Dichte und Überlebenszeitfunktion, geschätzt und maximiert/minimiert wird.

Die Lipschitz-Konstanten ($L_{\tilde{\psi}}$ und L_{ψ}) werden dort vermutet, wo die aufeinander folgenden Intervalle maximal voneinander abweichen, also derer Quotient - oder dessen Logarithmus - maximal wird. Die Konstanten werden dann als Differenzenquotient an den die beiden Intervalle umschließenden Beobachtungen ermittelt.

6.3.4 Daumenregel für fixe Bandbreite

Die Daumenregel unter Normalverteilungssannahme (3.5) ist zum Beispiel FRYER (1976) zu entnehmen. Die Schätzung $\hat{\sigma}$ der Standardabweichung wird wie für die nächste-Nachbarn Daumenregel realisiert (siehe (6.3.1)).

6.4 Zeitmanagement der Simulation

Es werden folgende Bandbreitenwahlen für die Hazardratenschätzung untersucht.

(1) Nächste-Nachbarn Bandbreite mit RoT-Anzahl nächster Nachbarn

(2) Nächste-Nachbarn Bandbreite mit ml-Anzahl nächster Nachbarn

(3) Nächste-Nachbarn Bandbreite mit UAE_{opt} Anzahl nächster Nachbarn bei RoT-Plug-in

(4) Nächste-Nachbarn Bandbreite mit UAE_{opt} Anzahl nächster Nachbarn bei mL-Plug-in

(5) Fixe Bandbreite mit RoT-Bandbreitenwahl

Es sollen bei 500 Simulationschleifen für

- 4 Hazardraten-Typen
- 2 Zensierungsraten 0/40%
- 3 Stichprobenumfäng 50/100/300

die Zielkriterien für die Bandbreitenwahlen simuliert werden. Es soll nur bei den Szenarien mit 300 Beobachtungen von den 100 Stützstellen bei der Schätzung auf 50 abgewichen werden.

Es wurden nach Vorsimulationen 5240 Stunden Simulationsrechenzeit unter Benutzung von SAS Version 8 geschätzt, was bei der simultanen und permanenten Benutzung von vier Computern (Intel Pentium II mit 450 MHz Takt und 256 MB RAM) 54 Tagen entspricht.

Nomenklatur der Simulationsdateibezeichnungen:
Die Bezeichnung der einmalig am Anfang (15.3.2000) gesamplen Daten wurden schreibgeschützt gesichert und haben folgende Notation:

expoijk.sd2 heißt

- i=Type der Hazardrate
- j=Anzahl der Beobachtungen (1=50,2=100,3=300)
- k=Zensierungsgrad (1=0 %, 2=40 %)

Die Simulationoutputs, das heißt die durch die Progamme simzahl$l$.sas, wobei l für die Bandbreitenwahl steht, erzeugten Kriteriumsgrößen und Grafikelemente haben dieselbe Nomenklatur. Es kommt lediglich in der Bezeichnung der Dateien simijkl.sd2 und dgraijkl.sd2 das l zur Bezeichnung für die Bandbreitenwahl hinzu.

6.5 Die Simulation - Ergebnisse

Um die Ergebnisse der Simulation adäquat interpretieren zu können, sollen zunächst einige einzelne Hazardratenschätzungen diskutiert werden, die auf *einzelnen* zufällig erzeugten Datensätzen beruhen. Es wurde versucht, diese Beispiel repräsentativ zu wählen. Die Beispiele stellen eine kontrollierte Situation für einzelne Studien dar und helfen dort Variabilität und Verzerrung für eine Hazardratenschätzung zu erkennen. Anschließend werden die sich als Mittel über die Simulationen ergebenden Erwartungwertschätzer der Hazardratenschätzungen als Gütemaß zum Vergleich der Bandbreitenwahlen betrachtet. Es folgt eine kurze Analyse der Recheneffizienz der verschiedenen Bandbreitenwahlen. Abschließend erfolgt eine Beurteilung der numerischen Verlustkriterien zur Evaluation der Bandbreitenwahlen.

6.5.1 Beispiele

Als erstes Beispiel wollen wir uns eine „wahre“ Hazardfunktion des Typs III der exponentiellen Weibull Familie und deren Schätzung, ohne Restriktion, vor Augen halten und Randeffekte identifizieren.

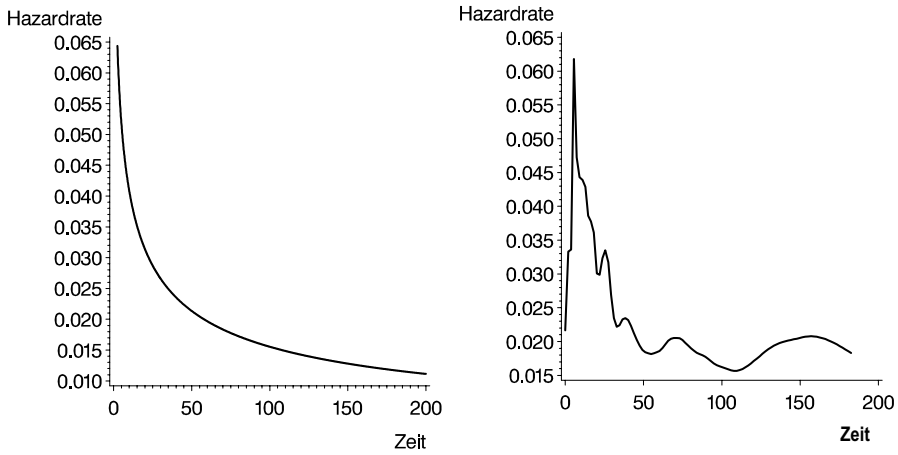


Abbildung 6.5: Hazardrate von Typ III mit Parametern $\alpha = 0.5$, $\theta = 1.8$, $\sigma = 10$ (links) und deren Schätzung bei 100 Beobachtungen mit 10%iger Zensierung und UAE-optimalen 16 nächsten Nachbarn (und modified Likelihood plug-in von 55 nächsten Nachbarn)(rechts)

In Abbildung 6.5 sieht man am linken Rand einen Randeffekt, das heißt das Abfallen der Schätzung fast auf Null, der aus dem Kumulieren von Nullmassen aus der Region unter Null resultiert. Dort ist aber Modell-theoretisch keine Masse. Die Unstetigkeit in Null ist ein eher theoretisches Argument. Man könnte auch sagen, dass empirische Masse ungerechtfertigterweise in die Region unter Null verschmiert wird und dann fehlt. Diese Effekte können mit sogenannten Randkernen (siehe zum Beispiel Müller & Wang (1994), einfach gesprochen durch Rücktransformation der verlorenen Masse, behoben werden. Hier wird von dieser Möglichkeit aber abgesehen, da die Randkorrekturen nur für die fixe Bandbreite direkt verfügbar sind. Wenn man bemerkt, dass der innere 80%-Bereich $[F^{-1}(0.1); F^{-1}(0.9)] = [1.0633123; 82.212834]$ ist, dann kann man am rechten Rand, beim 90%-Quantil von circa 80, das abfallen deutlich unter 0.02 auch als Randeffekt deuten. Den positiven Effekt des Stichprobenumfangs sieht man in Abbildung 6.6 auf Basis von 300 statt 100 Beobachtungen und trotz höherer Zensierung – 40% statt 10%.Der Randeffekt am Ursprung verschwindet fast und der am linken Rand verschiebt sich auf circa 100.

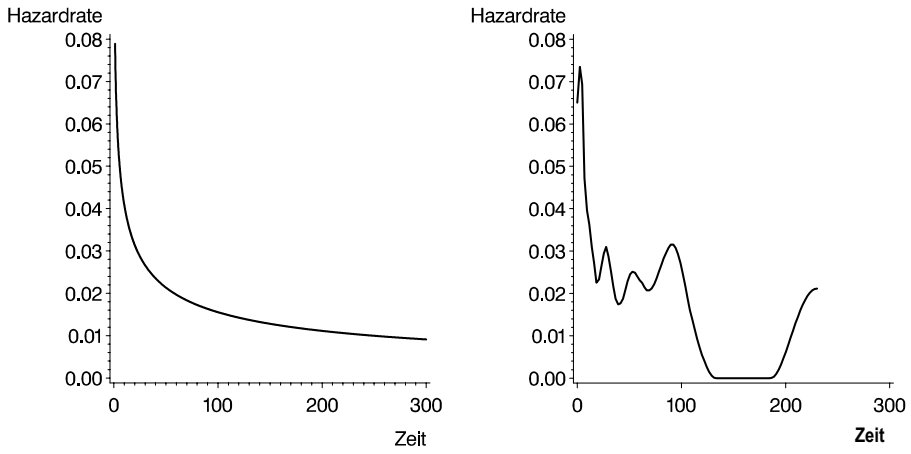


Abbildung 6.6: Hazardfunktion von Typ III mit Parametern $\alpha = 0.5$, $\theta = 1.8$, $\sigma = 10$ (links) und deren Schätzung bei 300 Beobachtungen mit 40%iger Zensierung und Rule-of-Thumb 48 nächsten Nachbarn (rechts)

Die Randkorrekturen sind auch nur am linken Rand, der Null, einfach, wie die Beispiele der Typen I und II exponentieller Weibull Hazardraten (Abbildungen 6.7 und 6.9) zeigen. Schwieriger ist die Beseitigung der Randeffekte am rechten Rand. Die

90%-Quantile der beiden korrespondierenden Dichten liegen bei 84.4 beziehungsweise 101.6 (siehe Tabelle 6.2), das heißt nach diesen Zeiten ereignen sich im Erwartungswert nur noch 30 Ausfälle, abzüglich der erwarteten 12 Zensierten. Anhand der Summendarstellung der Hazardratenschätzung (4.8) sieht man aber ein, dass nur dann entscheidende Beiträge für die Schätzung an der Stelle x erfolgen, wenn in der Umgebung Daten liegen. Das heißt, die Hazardrate mag an der Stelle x hoch sein, wenn aber die Dichte dort klein ist, und deshalb keine Daten in die nähere Umgebung fallen, so kann die Hazardrate kaum geschätzt werde, das heißt man wird sie entscheidend unterschätzen. Diesen Randeffekt, kann man für die Hazardrate nicht beheben. Für die Dichteschätzung wäre so ein Randeffekt, der durch das Ende des Trägers der Dichte entsteht, wenn die Dichte bis dahin nicht auf Null gesunken ist, wieder mittels Randkernen behebbar; dies allerdings nur, wenn die Lokation der Randes bekannt ist. In der Praxis der medizinischen Anwendungen der Hazardrate ist die Problematik der Randverzerrungen, durch die explizite Beschränkung auf – die typischerweise bekannten – Regionen ausreichend hoher Dichte vermeidbar.

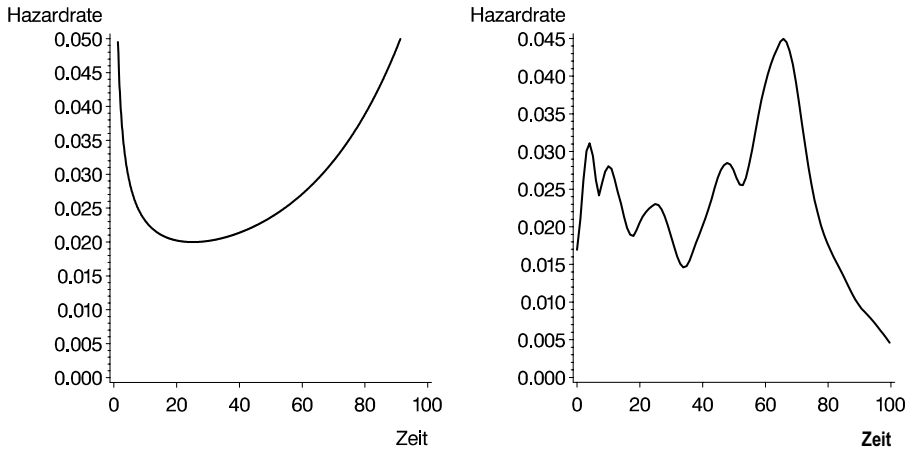


Abbildung 6.7: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (links) und deren Schätzung bei 300 Beobachtungen mit 40%iger Zensierung und Rule-of-Thumb 59 nächsten Nachbarn (rechts)

Man beschränke sich zum Beispiel auf die Region, in die die mittleren 80% der Daten fallen. Das ist auch deshalb sinnvoll, da die vorliegende nichtparametrische Funktionalschätzmethodik eindeutig nicht für die Schätzung von *Tail*-Effekten verwendbar ist. Das ist auch in der Dichteschätzung bekannt.

Das Beispiel mit geschätzter und wahrer Hazardrate des Repräsentanten des Typs I im inneren 80%-Quantil bei 50 Beobachtungen und mit 15 nächsten Nachbarn nach der Daumenregel in Abbildung 6.8 veranschaulicht, dass bei dieser Beschränkung auch schon bei geringem Stichprobenumfang Konturen der wahren Hazardrate erkannt werden können.

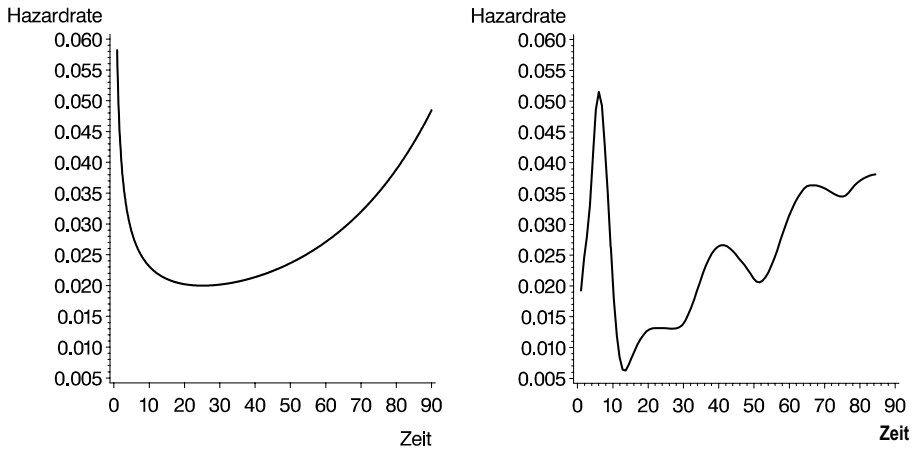


Abbildung 6.8: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (links) und deren Schätzung bei 50 Beobachtungen mit 10%iger Zensierung und Rule-of-Thumb 15 nächsten Nachbarn (rechts)

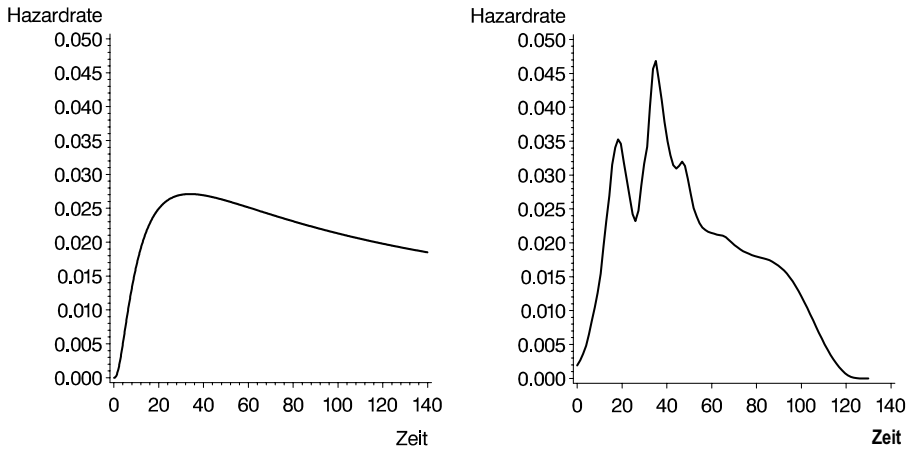


Abbildung 6.9: Hazardfunktion von Typ II mit Parametern $\alpha = 0.5$, $\theta = 9.5$, $\sigma = 5$ (links) und deren Schätzung bei 300 Beobachtungen mit 40%iger Zensierung und Rule-of-Thumb 59 nächsten Nachbarn (rechts)

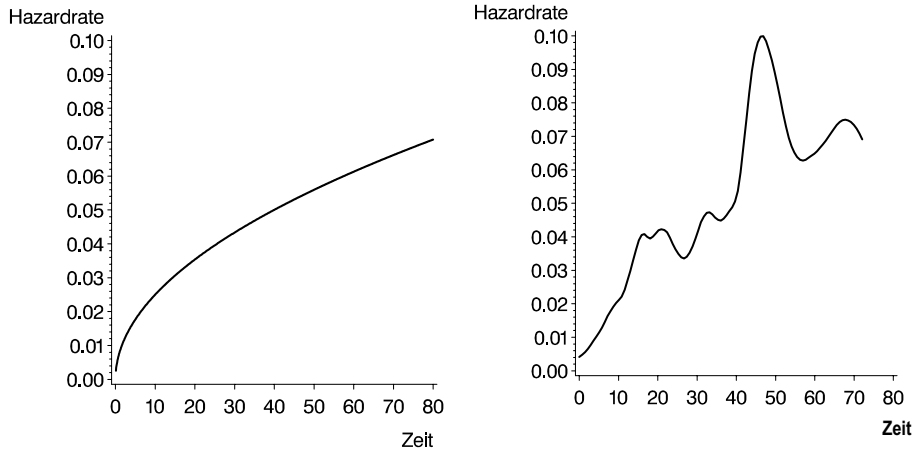


Abbildung 6.10: Hazardfunktion von Typ IV mit Parametern $\alpha = 1.5$, $\theta = 1.0$, $\sigma = 33$ (links) und deren Schätzung bei 300 Beobachtungen mit 40%iger Zensierung und Rule-of-Thumb 64 nächsten Nachbarn (rechts)

6.5.2 Erwartungswertschätzer

6.5.2.1 Daumenregeln und Kreuz-Validierung

In der Simulation werden nun die - als Mittel über die in der Regel 500 Simulationen - geschätzten Erwartungswerte der Hazardratenschätzung untersucht. Somit können wir Aussagen über den Bias der Schätzung machen, sowie über die Varianz, die sich, falls sie groß ist, auch in den Erwartungswert Schätzungen abzeichnet, da die „Stichprobe“ nur 500 Simulationen beträgt. Man bedenke, dass die Varianz, im Vergleich zum einzelnen Schätzer, um den Faktor $\sharp\{Simulationen\}^{-1}$ gedämpft ist. Anhand von zwei Beispielen will ich zunächst die Überlegenheit der Daumenregel für die Anzahl nächster Nachbarn (4.10) über die herkömmliche Daumenregel (3.5) für die fixe Bandbreite demonstrieren.

In allen folgenden Grafiken werden jeweils die „wahre“ und die geschätzte Hazardrate auf dem inneren 80%-Bereich der jeweiligen exponentiellen Weibull-Verteilung dargestellt.

Sowohl für die Hazardrate des Typs I im zensierten Szenario in Abbildung 6.11 als auch für die des Typs IV im zensierten Szenario in Abbildung 6.12 stimmen Schätzung und wahre Hazardrate quantitativ bei der Daumenregel für die nächste-Nachbarn Bandbreite besser überein als bei fixer Bandbreite mit Daumenregel. Vielleicht noch wichtiger für die Praxis ist, dass qualitativ Moden der wahren Haradrate für die erste Bandbreitenwahl (im Erwartungwert) besser erkannt werden. Bei der fixen Bandbreite indes werden diese immer noch durch Randeffekte überschätzt. Die Daumenregel für die fixe Bandbreitenwahl neigt – auch für die betrachteten glatten Hazardraten – zur Überglättung und führt somit teilweise zur Fehlinterpretation von Moden. So kann sie zum Beispiel zur bimodalen Annahme bei wahrer badewannenförmiger Hazardrate führen (Abbildung B.14(rechts)) oder zur unimodalen bei fallender Hazardrate (Abbildung B.43-B.45). Zumindest für die Moden am linken Rand könnte man dieses Phänomen wie angedeutet mit Randkernen beheben. Ich habe auf deren Verwendung hier verzichtet, da man sie sonst für alle Bandbreite anwenden müsste, so aber nicht sichtbar würde, dass die nächste-Nachbarn Bandbreite die Randproblematik am linken Rand schon konzeptionell löst.

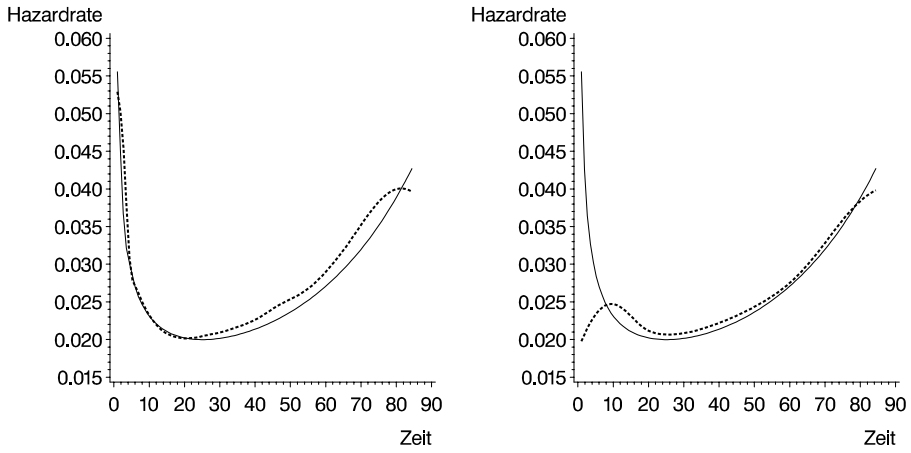


Abbildung 6.11: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 250 Schätzung bei 300 Beobachtungen mit 40%iger Zensierung (gestrichelt) für die Daumenregel für nächste-Nachbarn (links) die Daumenregel für die fixe Bandbreite (rechts)

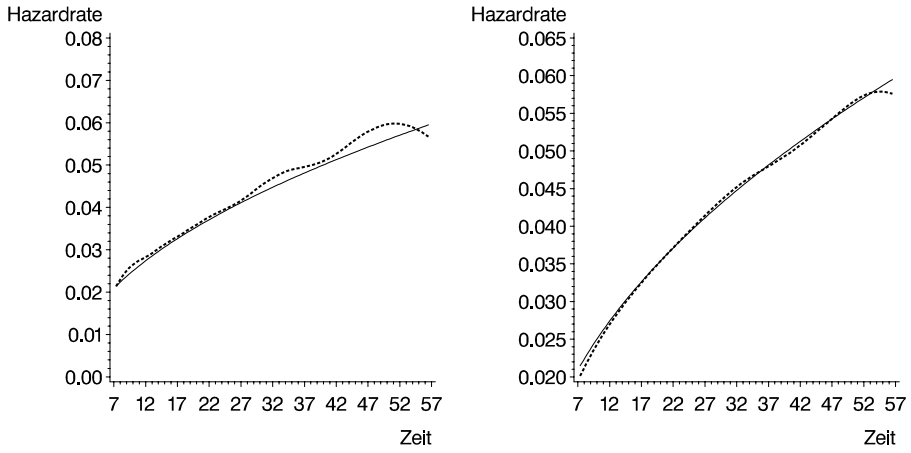


Abbildung 6.12: Hazardfunktion von Typ IV mit Parametern $\alpha = 1.5$, $\theta = 1.0$, $\sigma = 33$ (durchgezogen) und das Mittel von 500 Schätzung bei 300 Beobachtungen mit 40%iger Zensierung (gestrichelt) für die Daumenregel für nächste-Nachbarn (links) die Daumenregel für die fixe Bandbreite (rechts)

Natürlich stellen sich aber auch für die nächste-Nachbarn Bandbreite Verzerrungen ein. So gibt es im unzensierten Szenario für die Bandbreitenwahl gemäß der Daumenregel für die Anzahl nächster Nachbarn als auch nach der modified Likelihood Kreuz-Validierung (4.12) eine positive Verzerrung am rechten Rand des inneren 80%-Bereichs (siehe Abbildung 6.13). Da die Dichte des Typs I der exponentiellen

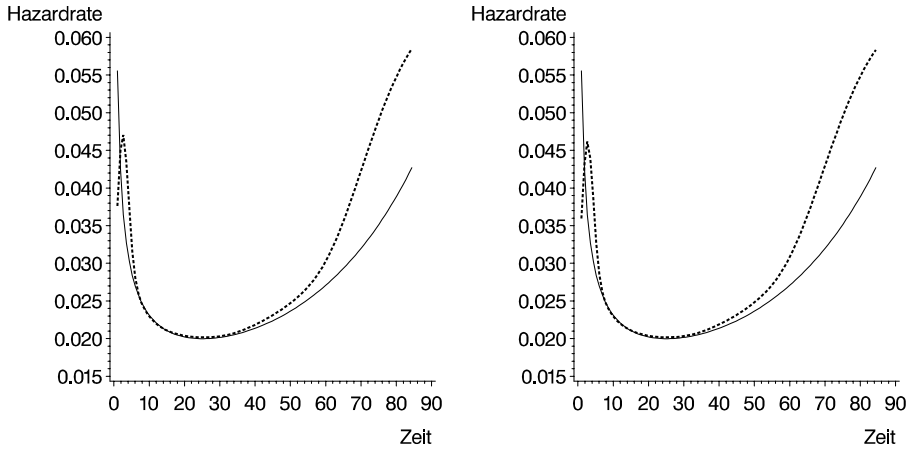


Abbildung 6.13: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 500 Schätzung bei 300 Beobachtungen ohne Zensierung (gestrichelt) für die Daumenregel für nächste-Nachbarn (links) und modified Likelihood Maximierung (rechts)

Weibull Verteilung (siehe Abschnitt 6.1.1) dort gering ist, wird eine große Bandbreite – mit trotzdem „wenigen“ Beobachtungen – gewählt. Wegen der überproportionalen Zunahme $O(\frac{1}{n})$ der Hazardrate wird diese dann überschätzt. Der Rückgang ist der Randeffekt fehlender Beobachtungen nach rechts hin. Dieser Rückgang überkompensiert den positiven Bias für zensierte Beobachtungen gerade bei geringem Stichprobenumfang – wie man in Abbildungen 6.14 und 6.15 sehen kann – erheblich.

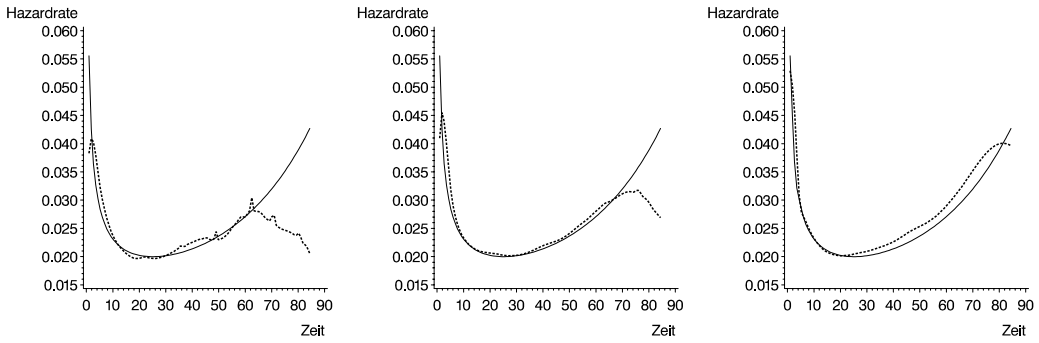


Abbildung 6.14: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von Schätzungen mit der Daumenregel für nächste-Nachbarn mit 40%iger Zensierung (gestrichelt) bei 500 Schätzung mit 50 Beobachtungen (links), 500 Schätzung mit 100 Beobachtungen (Mitte) und 250 Schätzung mit 300 Beobachtungen (rechts)

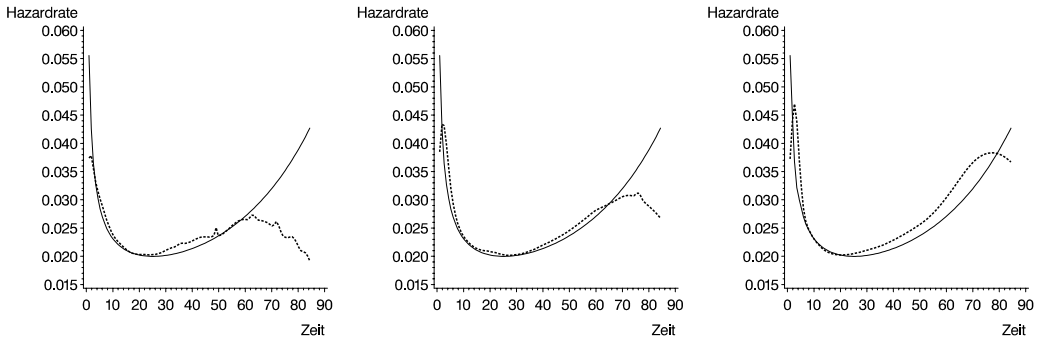


Abbildung 6.15: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von Schätzungen mit modified-Likelihood Wahl nächster Nachbarn mit 40%iger Zensierung (gestrichelt) bei 500 Schätzung mit 50 Beobachtungen (links), 500 Schätzung mit 100 Beobachtungen (Mitte) und 250 Schätzung mit 300 Beobachtungen (rechts)

Generell führt in fast allen Schätzungen, die im Anhang zu finden sind, die Zensierung zum Abfall der Schätzung am rechten Rand – ohne Ansehen des tatsächlichen Verlaufs. Wenn man sich die Zensierung nur als Verlust von Beobachtungen vorstellt, so müssten eine Beobachtungszahl von $n = 100$ mit 40%-iger Zensierung also ähnliche Ergebnisse liefern wie 50 unzensierte Beoachtungen. Als Gegenbeispiel betrachten wir für die fixe Bandbreite Abbildung 6.16

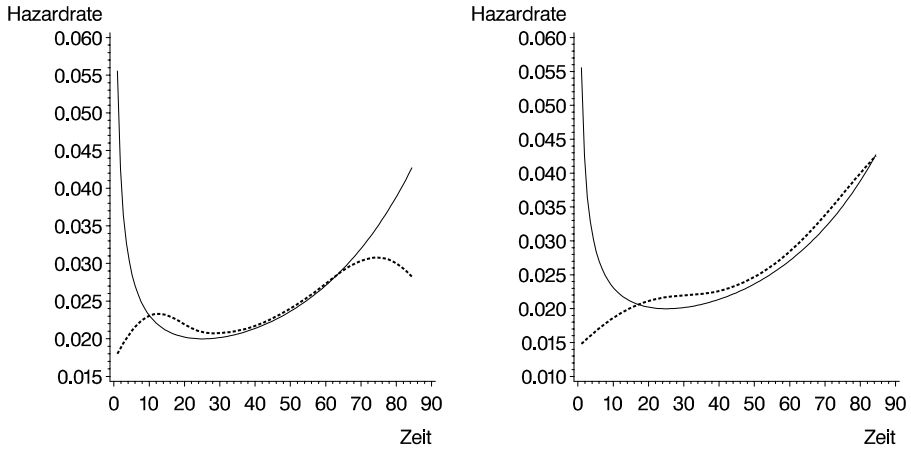


Abbildung 6.16: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 500 Schätzung mit der Daumenregel für die fixe Bandbreitenwahl (gestrichelt) bei 100 Beobachtungen mit 40%iger Zensierung (links) und 50 unzensierten Beoachtungen (rechts)

Zur Erklärung dieses verstärkten Randeffektes müssen wir die Zensierungverteilung betrachten. Hier sehen wir, dass – wie in der Praxis häufig – für lange Überlebenszeiten die Zensierungen häufiger werden. Da aber in der Summe circa 40% zensiert sein sollen, sind im Bereich langer Überlebenszeiten wenig unzensierte Beobachtungen, was zu dem verstärkten Randeffekt führt.

Auffällig ist, dass sich die Schätzungen des Erwartungswerts zur Daumenregel und zur Kreuz-Validierung stark ähneln. Das ist auch der Grund, warum sie hier in einem Abschnitt behandelt werden. Die Ähnlichkeit erstaunt um so mehr, als dass die Bandbreitenwahlen völlig unterschiedliche Motivationen haben.

6.5.2.2 UAE-optimale Bandbreitenwahl

Ein Ziel der Simulation war es zu überprüfen, ob die UAE-optimale Bandbreitenwahl zu weniger Bias in der Hazardratenschätzung im Vergleich mit MISE-optimalen Bandbreitenwahlen führt, wie das die Theorie vermuten läßt. Ein weiteres Ziel war es, die Stabilität der UAE-optimalen Bandbreitenwahl (4.11) bezüglich der Plug-in Bandbreite zu untersuchen. Die Schätzung bei 300 Beobachtungen für die Typ I Hazardrate ohne Zensierung in Abbildung 6.17 lässt letzteres annehmen. Auch die

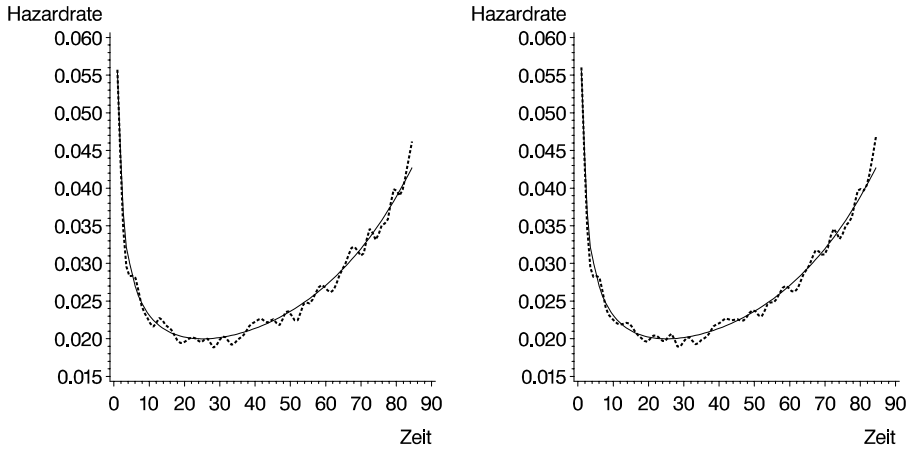


Abbildung 6.17: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 250 Schätzung bei 300 Beobachtungen ohne Zensierung (gestrichelt) für die UAE-optimalen Bandbreitenwahl mit Plug-in nach der Daumenregel für nächste Nachbarn (links) und Plug-in nach der modified Likelihood Maximierung für nächste Nachbarn (rechts)

Durchsicht der weiteren Szenarien lässt, zumindest was die Plug-in Stabilität angeht, keinen Zweifel aufkommen. Der Biasvorteil prägt sich aber erst für große Stichproben aus. An diesen beiden Grafiken sieht man bereits, dass die Variabilität erheblich ist. Es ist also anzunehmen, dass die Bandbreitenwahl eine hohe Variabilität durch die sensible Schätzung der Charakteristika der unterliegenden Funktionen aufweist. Das wird bei der Analyse der Maßzahlen geklärt werden können. Ein weiterer Grund für die hohe Variabilität könnte die Ungenauigkeit der Dreiecksungleichung (2.13) sein. Davon gehe ich aber nicht aus, da Korostelev & Nussbaum (1999) für den Fall der Dichteschätzung die Schärfe der Ungleichung $|f - f_n| \leq |f - Ef_n| + |Ef_n - f_n|$ für

den PARZEN-Schätzer gezeigt haben. Die Zensierung behindert die UAE-Methodik eventuell durch vereinfachte Schätzannahmen bei der Schätzung der Charakertistika. Vorab will ich hier bewerten, ob der Ausschluss der grenzwertigen Bandbreitenwahlergebnisse zu einem anderen Ergebnis führt. So kann man annehmen, dass diese Konstellation bei der Anwendung als Ausfall der Bandbreitenwahl gewertet wird. Im vorliegen Fall der Abbildung 6.18 betrifft das aber nur 6 Simulationen, die mehr als 300 oder weniger als 2 nächste Nachbarn gefordert hätten. Beim selben

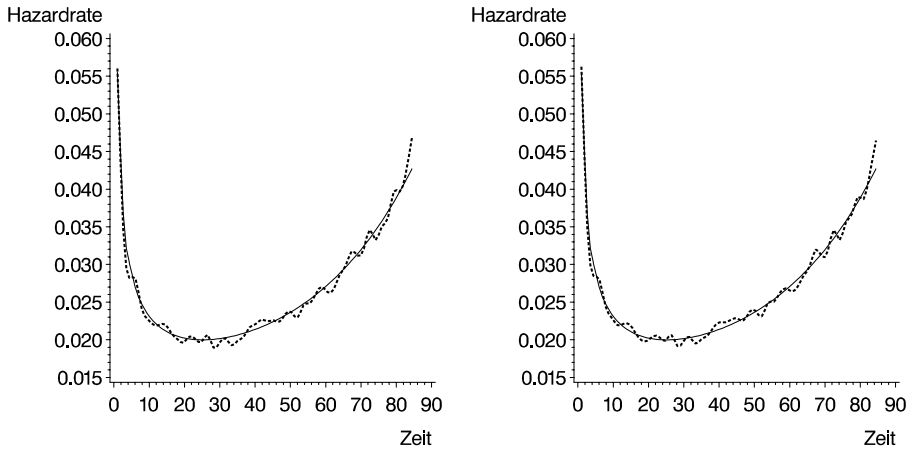


Abbildung 6.18: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 250 Schätzung bei 300 Beobachtungen ohne Zensierung (gestrichelt) für die UAE-optimalen Bandbreitenwahl mit Plug-in nach der modified Likelihood Maximierung für nächste Nachbarn ohne Ausfallabgrenzung (links) und mit Ausfallabgrenzung (rechts)

Szenario allerdings bei 50 Beobachtungen betrifft das schon 210 Simulationen, also einen erheblichen Anteil. Das Ergebnis ist in Abbildung 6.19 dargestellt. Man sieht, dass sich durch die Ausfallabgrenzung die Variabilität leicht reduzieren lässt. Dem systematischen Bias scheint das aber nicht zugute zu kommen.

6.5.3 Recheneffizienz

Interessant sind sicherlich auch die Laufzeiten der einzelnen Simulationen – dargestellt in Tabelle 6.5 – da sie Aufschluss über die Rechenintensität der verschiedenen Bandbreitenwahlen bei den verschiedenen Stichprobenumfängen geben.

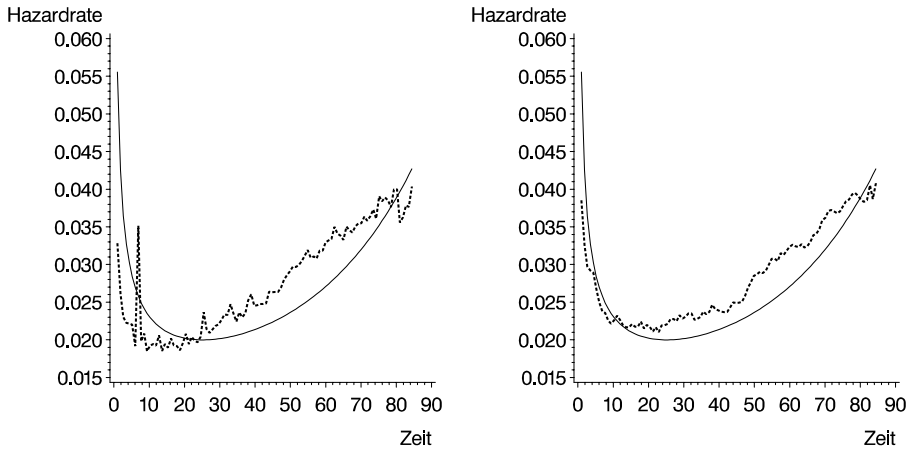


Abbildung 6.19: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 250 Schätzung bei 50 Beobachtungen ohne Zensierung (gestrichelt) für die UAE-optimalen Bandbreitenwahl mit Plug-in nach der modified Likelihood Maximierung für nächste Nachbarn ohne Ausfallabgrenzung (links) und mit Ausfallabgrenzung (rechts)

Für die Mehrheit der nicht gekennzeichneten Situationen wurden Pentium II Prozessoren mit 450 MHz Takt und 256 MB RAM verwendet. Ein relevanter Zeitunterschied zwischen der Version SAS 6.12 und der Version SAS V 8, die im späteren Verlauf teilweise genutzt wurde, konnte nicht ermittelt werden. Deshalb wird auf eine Zuweisung zu den einzelnen Situationen verzichtet. In der Summe wurden circa 6576 Stunden CPU-Zeit simuliert, das entspricht circa ununterbrochenen 9 Monaten auf einem Rechner. Eine Legende der Indizes für die Simulation findet sich im Anhang. Es werden Besonderheiten aufgezeigt, die aber für die Interpretation qualitativ keine große Auswirkung haben.

Tabelle 6.5: Simulationslaufzeiten (in Stunden) für die Bandbreitenwahlen der Daumenregel für die nächste-Nachbarn Bandbreite (Smooth RoT, kurz sRoT), der modified Likelihood Maximierung für die nächste-Nachbarn Bandbreite (mL), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit sRoT Plug-in (UAE(sRoT)), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit mL Plug-in (UAE(mL)) und der Daumenregel für die fixe Bandbreite (fix), die erste Zeit in den Zellen bezieht sich auf unzensierte Beoachtungen, die zwei auf die 40%-tige Zensierung (Die Indices sind in der Simulationslegende aufgelistet)

		Hazardenratentyp			
Bandbreiten	n	Typ I	Typ II	Typ III	Typ IV
Smooth RoT	50	3:46/5:00	3:24/5:24	3:01/4:14	$3:38^{\oplus}/6:38^{\oplus}$
	100	20:00/20:00	20:00/10:00	20:00/20:00	20:00/20:00
	300	$300/40:38^{\dagger}$	315:40/193:35	267:47/135:21	$231:33^{\oplus}/124:44^{\oplus}$
mL	50	2:00/3:00	27:00/27:00	11:59/7:00	$24:49^{\oplus}/28:37^{\oplus}$
	100	29:00/29:00	71/51	39:37/52:00	f/70:42
	300	$398:26^{\xi}/184^{\dagger}$	853:03/664:10	$186:32^{\amalg}/121:25^{\dagger}$	$109:56^{\dagger\ddagger}/555:22^{\dagger}$
UAE(sRoT)	50	2:00/3:00	3:00/4:00	3:00/4:00	2:00/3:00
	100	16:00/12:00	8:40/14:50	4:50/6:40	20:30/18:01
	300	$36:39^{\dagger}/53:58^{\dagger}$	$27:39^{\dagger\oplus}/76:00^{\dagger}$	$17:34^{\dagger}/21:40^{\dagger}$	$41:18^{\dagger}/99:42^{\dagger}$
UAE(mL)	50	3:00/5:00	6:00/6:00	7:00/5:00	6:00/4:00
	100	13:47/13:25	9:40/15:00	5:20/7:00	25:00/19:28
	300	$39:07^{\dagger}/63:37^{\dagger}$	$34:01^{\dagger}/96:56^{\dagger}$	$12:39^{\dagger}/26:07^{\dagger}$	$43:56^{\dagger\ddagger\oplus}/87:21^{\dagger}\oplus$
fix	50	0:07/0:07	0:13/0:13	0:13/0:13	$0:13^{\oplus}/0:13^{\oplus}$
	100	0:30/0:30	0:20/0:20	0:30/0:30	0:30/0:30
	300	18:36/19:05	f/f	18:50/20:01	19:24/f

So ist zum Beispiel die Ähnlichkeit der Schätzungen für die Daumenregel und der modified Likelihood Selektion bei der nächste-Nachbarn Bandbreite für die Erwartungwertschätzungen – siehe zum Beispiel Abbildung 6.20 – im Disproporz zu deren

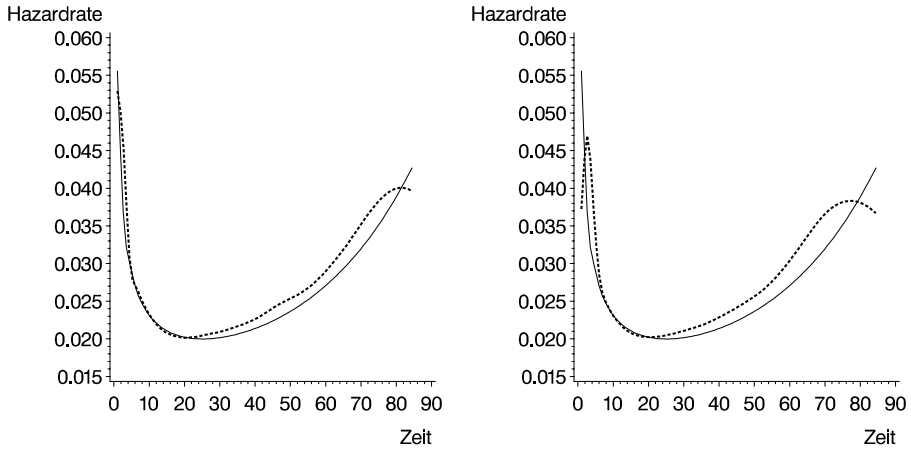


Abbildung 6.20: Hazardfunktion von Typ I mit Parametern $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ (durchgezogen) und das Mittel von 250 Schätzung bei 300 Beobachtungen mit 40%iger Zensierung (gestrichelt) für die Daumenregel zur Wahl nächster Nachbarn (links) und nach der modified Likelihood Maximierung zur Wahl nächster Nachbarn (rechts)

Recheneffizienz. Die Ähnlichkeit belegt die Güte der vergleichsweise einfachen Daumenregel, insbesondere wenn man vorweg nimmt, dass die Bandbreitenvariabilität, der schärfste Kritikpunkt bei der modified Likelihood Methodik, stark verringert ist. Die Wahl der Anzahl nächster Nachbarn nach der Daumenregel benötigt im extremen Fall von Typ I Hazard bei 300 unzensierten Beobachtungen 40 Stunden, wohingegen die der modified Likelihood 184 Stunden erfordert. Natürlich sind dieses Zeitmessungen für 250 Schätzungen. Für die Einzelschätzung bedeutet dies aber einen Unterschied von circa 10 Minuten pro Schätzung bei der Daumenregel oder 44 bei der modified Likelihood Maximierung. Dieser Unterschied ist definitiv von Anwendungsrelevanz.

6.5.4 Verlustkriterien

Die tabellarische Auflistung der geschätzten Verlustparameter stellt eine weitere Möglichkeit dar, Schlüsse über die Qualität und Vergleichbarkeit der einzelnen Schätzer beziehungsweise Bandbreitenwahlen zu ziehen. Eine komplette Auflistung ist im Anhang C zu finden. Hier werde ich nur die Teile detailliert auflisten, die für die Gedankenführung nützlich sind.

Die schon in Abschnitt 6.5.2 angesprochene Ähnlichkeit der Wahlen nächster Nachbarn mit der Daumenregel (4.10) und der modified Likelihood Methode (4.12) soll hier anhand der absoluten Bandbreiten untersucht werden. So fällt auf, dass in allen 24 Simulationssituationen im Mittel die Anzahl der zur Schätzung herangezogenen nächsten Nachbarn von der Daumenregel geringer gewählt werden als von der modified Likelihood Methode ($5-50\%$). Das heißt, der Bias ist geringer und die Varianz höher. Allerdings fällt das im Mittel der Hazardratenschätzungen grafisch nicht auf. Viel wichtiger ist, dass die Variabilität der Anzahl nächster Nachbarn bei der Daumenregel entscheidend reduziert ist. Bis auf einen Fall ist die Standardabweichung der Anzahl für diese geringer, und zwar um das circa drei- bis zehnfache. Das belegt eine gute Anwendbarkeit in der Praxis, da seltener übermäßig unglatte oder glatte Kurvenschätzer auftreten. Beispielhaft soll hier die Situation für Typ I der Weibull Hazard mit 50 unzensierten Beobachtungen betrachtet werden. Hier wählt die Daumenregel im Mittel 17.01 nächste Nachbarn und die modified Likelihood Methode 17.59. Die Standardabweichung von 0.535 ist um mehr als das zehnfache kleiner als die von 6.095 für die modified Likelihood Methode.

Auch schon im Abschnitt 6.5.2 angesprochen wurde die starke Variabilität der Schätzung, die die UAE-optimale Bandbreitenwahl (4.11) praktisch unattraktiv macht. Das kann man auch an der Bandbreitenwahl nachweisen. Im Fall von 300 Beobachtungen liegt die Standardabweichung zwischen 5% und 30% – bezogen auf das Mittel – und ist im Vergleich zur modified Likelihood Methode groß, bei der sie zwischen 0.1% und 15% liegt. Bei den kleinen Stichproben von 50 und 100 Beobachtungen liegt sie aber im inakzeptablen Bereich von 40% bis 200%(!). Das belegt wieder, dass die Schätzung der Ingredienzen der UAE-optimalen Bandbreite schwierig ist. Die Diskrepanz zwischen großer und kleinerer bis mittlerer Stichpro-

be kann zwei Gründe haben, die hier nicht zu trennen sind. Zum einen kann die Schätzung obiger Ingredienzen für große Stichproben präziser sein. Zum anderen kann die asymptotische Schranke aber auch näher an den wahren absoluten Fehler rücken und damit die Balance erleichtern. Für die Praxis sind diese Fragen aber weniger entscheidend, da auch für große Stichproben die anderen Bandbreitenwahlen geringere Variabilität aufweisen. Interessant sind – neben der Variabilität – auch die Anzahlen nächster Nachbarn selber. Hier folgt die UAE-optimale Methodik auch in Tendenzen nicht der Daumenregel und der modified Likelihood Methode. Häufig schlägt sie für kleine Stichproben deutlich größere Anzahlen vor – zum Teil vom Faktor drei – und für große Stichproben kleinere – zum Teil vom Faktor ein drittel. Allerdings gibt es hierfür Ausnahmen, und somit lässt sich keine allgemein gültige Aussage machen. Eine Interpretation jenseits der hohen Variabilität will mir zudem hierfür nicht gelingen. Der in Abschnitt 6.5.2 vermutete Biasvorteil kann sich auch für große Stichproben an der Quantifizierung des integrierten quadratischen Bias nicht nachvollziehen lassen. Die Mittel sind durchweg höher als zum Beispiel die der Daumenregel für die Anzahl nächster Nachbarn. Die Anschauung aus den Grafiken zur Erwartungswertschätzung zeigt uns aber, dass die Höhe der quantitativen Biasschätzungen auf die Varianz zurückzuführen ist.

Interessant ist auch zu sehen, dass bei der Wahl nächster Nachbarn mit der Daumenregel (4.10) in allen bis auf einen Fall die Zensierung zu einer Verringerung der gewählten Anzahlen führt. Das spricht auch für sie, da ein Informationsverlust, der einem Datenverlust und damit einer kleineren Stichprobe entspricht, entsprechend bewertet wird. Bestärkt wird dieses Urteil, wenn man berücksichtigt, dass die Daumenregel für die fixe Bandbreite (3.5) dieselbe Tendenz hat, was man auch schon wegen deren konstruktiver Bedeutung für die Daumenregel bei nächste-Nachbarn Bandbreite annehmen konnte.

Überhaupt kann man bemerken, dass die Daumenregel zur fixen Bandbreitenwahl sehr gute Ergebnisse liefert, und das ohne eine Randkorrektur. So ist der MISE in circa 80% der Simulationssituationen besser als der der Daumenregel für nächste Nachbarn und im Vergleich mit allen anderen Bandbreitenwahlen sogar in circa 90%. Auch bei anderen Zielkriterien ist die Daumenregel für die fixe Bandbreite

häufig überlegen. Für den KULLBACK-LEIBLER-Verlust ist sie besser als die diesen asymptotisch minimierende modified Likelihood Methode, und sie impliziert auch einen kleineren gleichmäßig absoluten Verlust und Verzerrung im Mittel als die UAE-optimalen Methoden. Allerdings muss man hinzufügen, dass die Glattheit der Normalverteilung, für die diese als optimal konstruiert wurde, auch für die hier verwendeten exponentiellen Weibull Hazardraten gilt.

Um jenseits der Einzelerkenntnisse eine zusammenfassenden Bewertung der Methode zu erreichen, muss man wiederum die Information der Tabellen des Anhangs aggregieren. Dies soll mittels eines Score erreicht werden. Der Score soll den Vergleich der Bandbreitenwahlen über alle Zielkriterien und alle Hazardratentypen pro Stichprobenumfang und Zensierungsgrad ermöglichen. Dabei wird zu den fünf Zielkriteren aus Abschnitt 6.2.1, deren Mittel bewertet werden, das Kriterium der „geringen Variabilität der Bandbreite“ hinzugenommen. Der Score wird über die Ränge der fünf Methoden definiert, die je Hazardratentyp und Zielkriterium ermittelt werden und dann über alle Hazardratentypen und Kriterien summiert werden. Symbolisieren kann man das wie folgt:

$$Score = \sum_{\substack{i=1,\dots 4 \\ j=1,\dots,6}} \text{Rang}(\text{Hazardratentyp}_i, \text{Kriterium}_j).$$

Die Ergebnisse sind in der Tabelle 6.6 aufgelistet. Zunächst fällt die scheinbare Überlegenheit der Daumenregel für die fixe Bandbreite ins Auge. Wie weiter oben angemerkt, ist sie wahrscheinlich insbesondere Resultat der ähnlichen Glattheit zwischen der exponentiellen Weibull Verteilung und der Normalverteilung. Allerdings sieht man auch, dass die Daumenregel nie den theoretisch optimalen Score von $24 = 4 \cdot 6 \cdot 1$ erreicht.

Interessanter ist dann die Tatsache, dass sich die Daumenregel für die nächste-Nachbarn Bandbreite und die modified Likelihood Maximierung für die nächste-Nachbarn Bandbreite kaum unterscheiden. Sie liegen absolut nahe beisammen und in drei von sechs Fällen, nämlich den zensierten, erzielt die modified Likelihood Methode einen besseren Score. Die leichte Majorität der modified Likelihood Me-

Tabelle 6.6: Scorebewertung der Bandbreitenwahlen der Daumenregel für die nächste-Nachbarn Bandbreite (Smooth RoT, kurz sRoT), der modified Likelihood Maximierung für die nächste-Nachbarn Bandbreite (mL), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit sRoT Plug-in (UAE(sRoT)), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit mL Plug-in (UAE(mL)) und der Daumenregel für die fixe Bandbreite (fix)

	Stichprobenumfang					
Bandbreitenwahl	50		100		300	
	unzens.	zens.	unzens.	zens.	unzens.	zens.
smooth RoT	61	65.5	60	67	58	69.5
mL	66	64	66.5	58	70	59
UAE(sRoT)	101.5	110	101.5	101	94.5	85.5
UAE(mL)	89.5	83.5	94.5	84	95.5	88.5
fix	42	36	37.5	50	42	57.5

thode für die zensierten Szenarien lässt vermuten, dass diese durch eine verbesserte Varianzschätzung bei der der Daumenregel explizit für zensierte Daten behoben werden kann. Aber auch bei vorliegendem Kenntnisstand erhärtet die Scorebewertung den Eindruck, dass die Daumenregel für die nächste-Nachbarn Bandbreite vergleichbar mit der optimalen modified Likelihood Maximierung für die nächste-Nachbarn Bandbreite ist.

Klar kann man aber auch an diesen Score die geringe Eignung der UAE-optimalen Bandbreitenwahlen ablesen. Bei Benutzung der Daumenregel als Plug-in ergibt sich aber der Eindruck eines fallenden Scores bezüglich des Stichprobenumfangs, der eine Eignung für große Stichproben vermuten lässt.

6.6 Bewertung

An dieser Stelle soll eine vergleichende Bewertung der fünf Bandbreitenwahlen, die tabellarisch in 6.4 zu finden sind, erfolgen. Die Erkenntnisse der Abschnitte 6.5.2, 6.5.3 und 6.5.4 sollen zusammengefasst werden. Es soll eine Handlungsempfehlung gegeben werden.

Die gleichmäßige absolute Asymptotik enthält in ihren zwei Bandbreitenwahlen un-

terschiedlicher Plug-in Bandbreiten – in meiner Realisation – nur wenig nutzbare Information für die Bandbreitenwahl, auch wenn sie sich durch Robustheit bezüglich der Plug-in Schätzung auszeichnet. Die stärkere Betonung der Biaskonvergenz – im Vergleich zum MISE-Ansatz – scheint erst bei großen Stichprobenumfängen effektiv zu werden. Aber die resultierende Vernachlässigung der Varianz wird durch die empfindlichen Schätzungen der Konstanten D_1, D_2 und D_3 bei der Bandbreitenwahl verstärkt und in der Summe dominant. Diese Bandbreitenwahlen können beim jetzigen Entwicklungsstand für die Analyse von Studiendaten nicht empfohlen werden.

Der Vergleich der Bandbreitenwahlen aus 4.2.1.1, aus 4.2.1.3 und der Daumenregel zur fixen Bandbreite 3.5 fällt hingegen weniger deutlich aus, da hierbei die verschiedenen Bewertungskriterien der drei Abschnitte zu unterschiedlichen Empfehlungen führen.

Der Vergleich der Bandbreitenwahlen aus 4.2.1.1 und 4.2.1.3 ist zunächst noch eindeutig. Neben der theoretischen Fundierung über die verallgemeinerte Darstellung 2.4 und der Vermittelbarkeit empfiehlt die Simulation die Daumenregel (4.2.1.1) für die variable Hazardratenschätzung mit nächste-Nachbarn Bandbreite. Ihre Überlegenheit über die asymptotisch optimale und rechenintensivere Bandbreitenwahl (4.2.1.3), die den zu erwartenden Kullback-Leibler Verlust asymptotisch minimiert, ist in drei Punkten fundiert: erstens in dem simulativ festgestellten äquivalenten Verhalten; zweitens in der geringen Variabilität der Bandbreitenwahl; und drittens schließlich wegen der einfacheren Implementation.

Wenn wir nun abschließend den Vergleich führen zwischen der nächste-Nachbarn Bandbreite und der fixen Bandbreite mit der jeweiligen Daumenregel 4.10 und 3.5 müssen wir zunächst feststellen, dass bezüglich des Abschnitts 6.5.3 sich keine Präferenz ergibt, da beide Bandbreitenwahlen schnell zu berechnen sind. Die graphische Beurteilung in 6.5.2 empfiehlt die nächste-Nachbarn Bandbreite da die Randeffekte korrigiert werden und so die Anzahl der Moden nicht überschätzt wird. Die numerische Beurteilung 6.5.4 empfiehlt aber die fixe Bandbreite in allen Kriterien. Für die Auflösung dieser Diskrepanz konnten nur Vermutungen angestellt werden, die im Rahmen dieser Simulationsstudie nicht zu klären sind. Wegen der subjektiven Bewertung der Mittelwertsgraphiken in 6.5.2 muss – insbesondere wegen der

Möglichkeit der Randkorrektur für bekannte Ränder mittels Randkernen – die fixe Bandbreite vorerst als ausreichend anerkannt werden. Die zusätzliche Datenadaptation, die die nächste-Nachbarn Bandbreite realisiert, führt nicht zu einer verbesserten Schätzung im Hinblick auf die numerischen Verlustkriterien. Für die konkrete Hazardratenschätzung können beide Daumenregeln empfohlen werden, da die Unterschiede in der numerischen Beurteilung nicht vermuten lassen, dass die nächste-Nachbarn Bandbreite mit der Daumenregel zu pathologischen Schätzungen führt.

Die Simulation lehrt aber auch, nur dort Funktionalschätzer zu interpretieren, wo ausreichend viele Daten verfügbar sind. Als praktische Empfehlung kann gesehen werden, dass man seine Interpretation auf ein (geschätztes) inneres $(1-2\alpha)$-Quantil $[F^{\dagger}(\alpha); F^{-1}(1-\alpha)]$ beschränken sollte, wobei die Sicherheit der Interpretation in α zunimmt. Ich habe $\alpha = 0.1$ verwandt. Eine weitere Hilfe ist, durch zusätzliche Schätzung der Dichte auch „innere“ Bereiche spärlicher Daten, also mit geringer Dichte, von der Interpretation auszunehmen. Zudem kann die Dichteschätzung helfen zufällige Variabilität in Regionen geringer Dichte von systematischen Schwankungen in Regionen hoher Dichte zu trennen. Sie liefert somit ein Beurteilungskriterium für die Glaubwürdigkeit des Schätzers.

Kapitel 7

Zusammenfassung und Ausblick

Ausgehend von der Bedeutung der Hazardrate und deren Schätzung in klinischen Überlebenszeitstudien beschäftigte sich die vorliegende Arbeit allgemein mit der nichtparametrischen Kernschätzung von Funktionalen. Es konnte eine Formulierung entwickelt werden, welche die Hazardraten- und die Dichteschätzung für unzensierte und (rechts-)zensierte Beobachtungen umfasst. Die Bandbreite wurde dabei so allgemein formuliert, dass sie die fixe und verschiedene variable daten-adaptive Bandbreitendefinitionen, wie zum Beispiel die der nächste-Nachbarn Bandbreite, enthält.

Für diesen neuen Funktionalschätzer mit verallgemeinerter Bandbreitendefinition wurde die Konsistenz im Sinne der fast sicheren Konvergenz bezüglich der L_∞-Norm gezeigt.

Das Thema der Bandbreitenwahl unterteilte sich nach den Prinzipien der Optimalität bezüglich der L_∞-Norm Konvergenz und der Praktikabilität. Die Konvergenz hatte eine optimale Bandbreitenwahl mit Plug-in Verfahren zur Folge, wohingegen sich als schnell mögliche objektive Bandbreitenwahl eine etablierte Daumenregel für die fixe Bandbreitenwahl bei der Dichteschätzung auf den verallgemeinerten Schätzer sinnvoll übertrug.

Es wurden die Spezifikationen der Dichteschätzung mit fixer Bandbreite ohne Zensierung und die der Hazardratenschätzung mit variabler nächste-Nachbarn Bandbreite bei (Rechts-)Zensierung untersucht.

Als Implikation der asymptotisch optimalen Bandbreitenwahl stellte sich bei der Dichteschätzung heraus, dass die Konvergenzraten für Bias und Varianz eine andere Balancierung erfahren als bei der Bandbreitenwahl, die den zu erwartenden integrierten quadratischen Fehler asymptotisch minimiert. Der Bias muss für die Bandbreitenwahl, die den gleichmäßigen Verlust asymptotisch minimiert, schneller gegen Null gehen als für die das integrierte quadratische Risiko minimierende.

Die Anwendung auf die Hazardratenschätzung bei zensierten Daten wurde anhand eines biometrischen Beispiels und einer umfangreichen Simulationsstudie untersucht. Hierbei stellte sich die als Spezifikation aus der allgemeinen Daumenregel resultierende Daumenregel zur Wahl der Anzahl nächster Nachbarn als praktikable Bandbreitenwahl dar. Die Vorteile im Vergleich zur optimalen Bandbreitenwahl bezüglich des gleichmäßig absoluten Fehlers lagen in der Einfachheit der Darstellung und der Schätzung sowie in der wesentlich schnelleren computer-technischen Ermittelbarkeit, da bei ihr im Gegensatz zur optimalen Bandbreitenwahl keine Plug-in Schätzungen benötigt werden. Die Daumenregel konnte sich auch gegen eine weitere optimale und rechenintensive Bandbreitenwahl, bei der asymptotisch das Kullback-Leibler Risiko minimiert wird, behaupten, in dem Sinne, dass die Ergebnisse bei der Verzerrung gleich gut waren, allerdings die Variabilität, die als Nachteil der Kreuz-Validierungs Bandbreitenwahlen bekannt ist, stark verringert wird.

Eine wichtige Frage, die sich nach der Simulationsstudie stellt, ist, in wieweit die implizite Randkorrektur der nächste-Nachbarn Bandbreite der expliziten Randkorrektur bei fixer Bandbreite überlegen ist. Dem Vorteil, dass der Rand bei der nächste-Nachbarn Bandbreite nicht zu spezifizieren ist, steht die Erkenntnis der Simulationsstudie gegenüber, dass die implizite Randkorrektur keinen Gewinn bezüglich der numerischen Verlustkriterien gegen die fixe Bandbreite ohne Randkorrektur erwirtschaftet. Der Vergleich sollte noch eingeschränkt auf die Umgebung der Ränder durchgeführt werden.

Die Auswirkungen der entwickelten Glättungsmethoden auf die Schätzung der Verteilungsfunktion und der kumulativen Hazardrate wurden nicht untersucht, auch wenn sich in jüngerer Zeit die glatte Schätzung dieser Funktionen als von Interesse herausgestellt hat. Eine weitere Frage, die sich aus der Arbeit ergibt, ist die Auswei-

tung der „Punktschätzung“ zur Bereichsschätzung, die nur kurz thematisiert wurde. So können simultane Konfidenzbänder zur Entscheidung bei Punkt- und Äquivalenzhypothesen verwandt werden, was zum Beispiel bei der Modellvaldierung einsetzbar ist. Allerdings ist hierfür zunächst eine weitere Beurteilung nötig, wann die Asymptotik in der L_∞-Norm greift, da insbesondere die Konvergenz der Verteilung des gleichmäßigen Fehlers als langsam bekannt ist.

Anhang A

Simulationslegende

Legende für alle Grafiken und Tabellen zur Simulation

†	50 Stützstellen und 250 Simulationen
$T1$	„Träger des Typs I ([1; 84.419192]). Das führt im Typ II Fall dazu, dass nur das innerhalb des 10%- und des 85.5%-Quantils eine Evaluation erfolgt. Für die anderen Typen bleibt das innere 80%-Quantil erhalten, es können nur vielleicht (Typ VI) nicht alle Stützstellen genutzt werden.
f	„fertig“ (ohne Zeitnahme)
x:00	exakte Stunden (x:00) sind Schätzungen
II	$nsim = 146$
‡	$nsim = 162$
$\oplus$	auf einen Pentium III mit 1024 MB RAM
ξ	Bei fast allen (75%) mL-Bewertungen wurde das Minimum von $n = 75$ erreicht und deshalb wurde dies bei allen in der Nachsimulation gesetzt

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang B

Erwartungswertgrafiken

In den folgenden Grafiken werden jeweils die „wahre“ (durchgezogene, dünne Linie) und die geschätzte Hazardrate (gepunktete, dicke Linie) auf dem inneren 80%-Bereich der jeweiligen exponentiellen Weibull-Verteilung – Typ I-IV – dargestellt. Links ist die Situation ohne Zensierung zu sehen, wohingegen rechts eine 40%ige Zensierung simuliert wird.

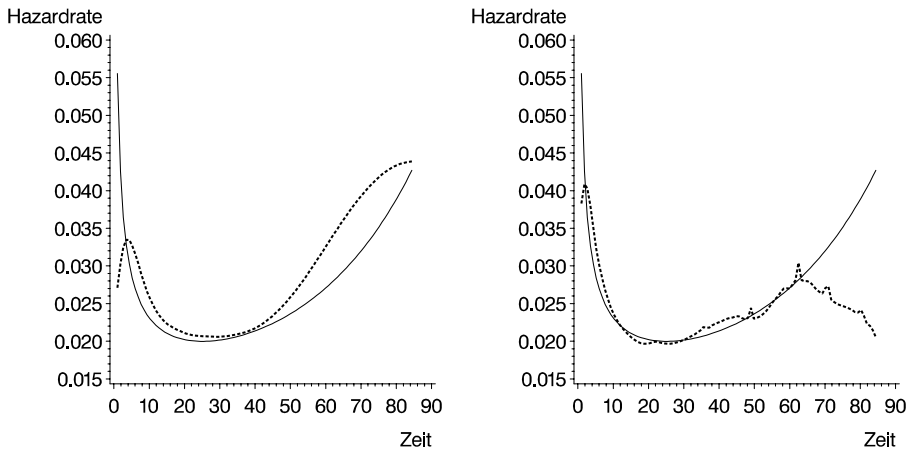


Abbildung B.1: Typ I, $n = 50$, smooth RoT

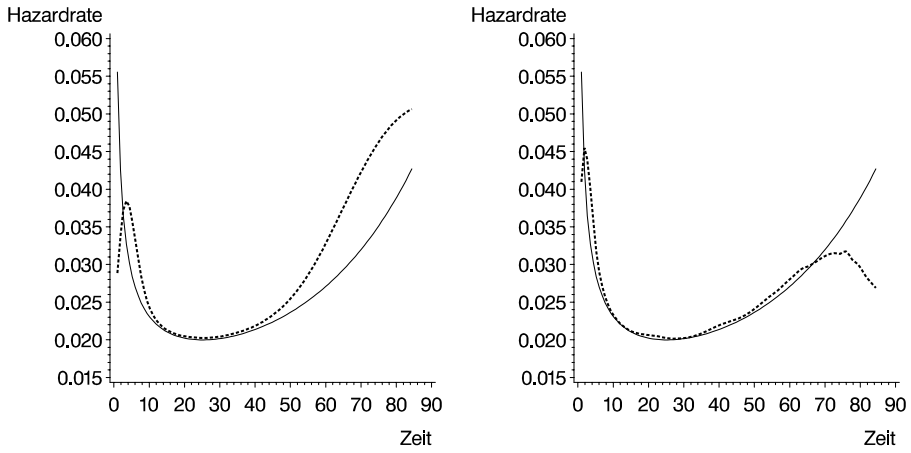


Abbildung B.2: Typ I, $n = 100$, smooth RoT

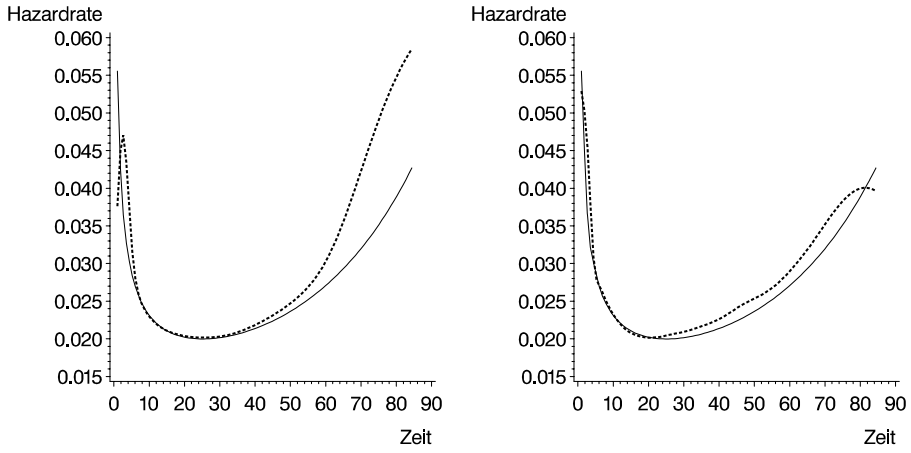


Abbildung B.3: Typ I, $n = 300$, smooth RoT, †

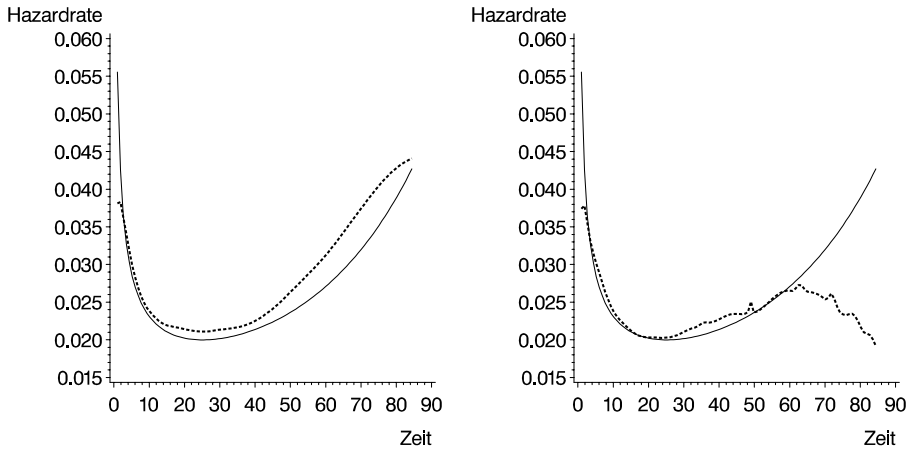


Abbildung B.4: Typ I, $n = 50$, mL

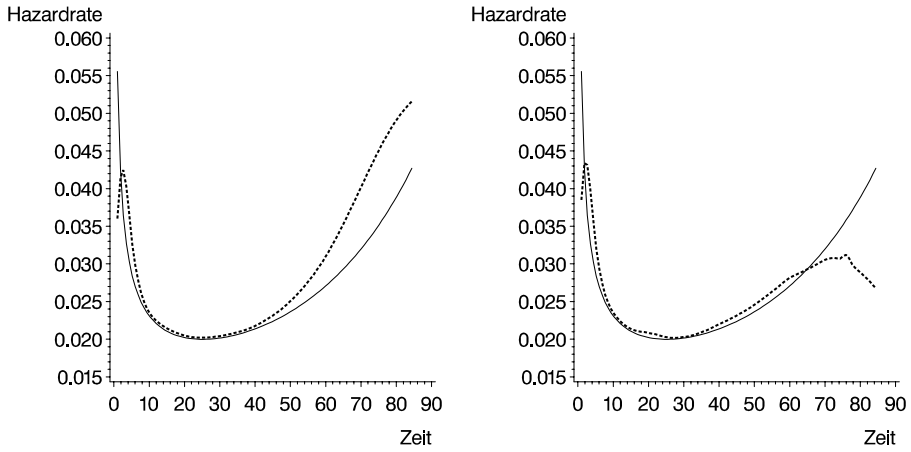


Abbildung B.5: Typ I, $n = 100$, mL

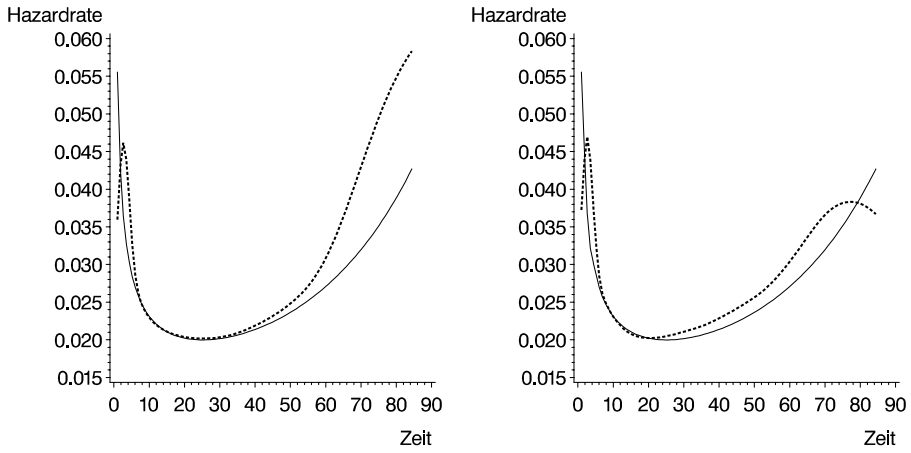


Abbildung B.6: Typ I, $n = 300$, mL, †

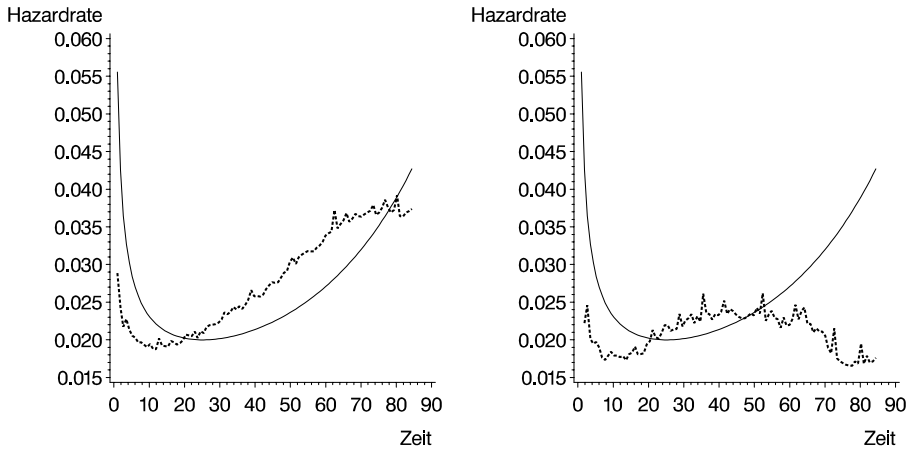


Abbildung B.7: Typ I, $n = 50$, UAE-optimal (mit smooth RoT Plug-in)

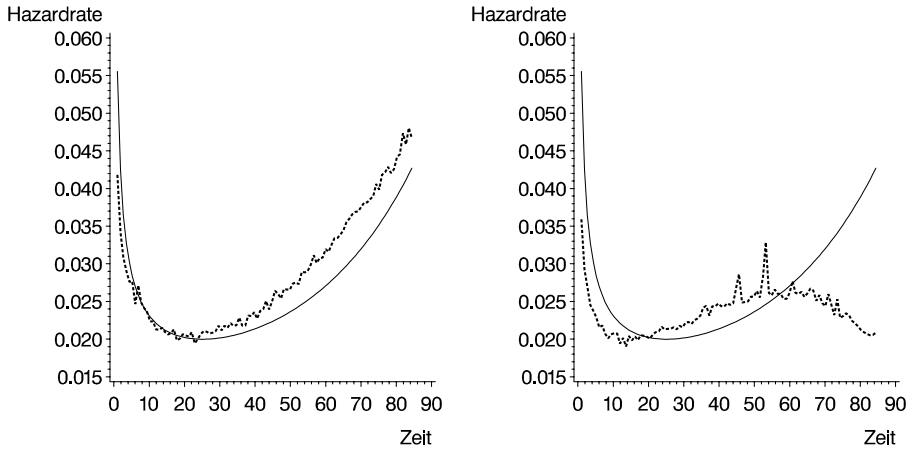


Abbildung B.8: Typ I, $n = 100$, UAE-optimal (mit smooth RoT Plug-in)

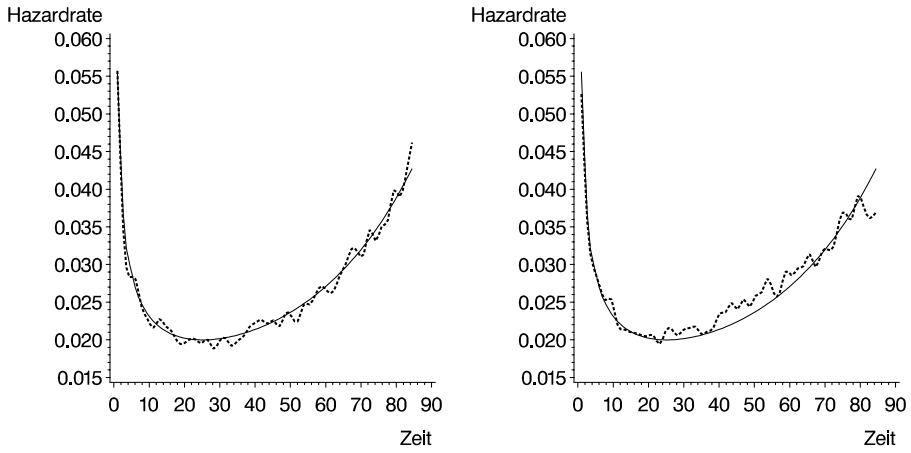


Abbildung B.9: Typ I, $n = 300$, UAE-optimal (mit smooth RoT Plug-in), †

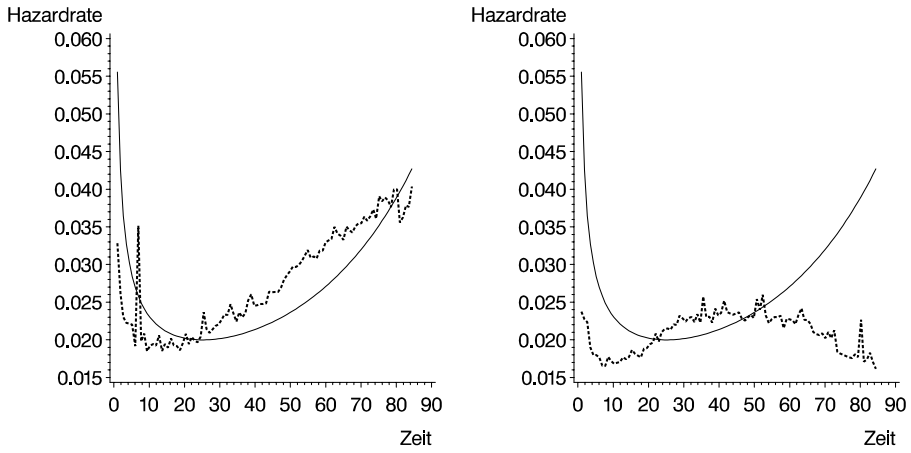


Abbildung B.10: Typ I, $n = 50$, UAE-optimal (mit mL Plug-in)

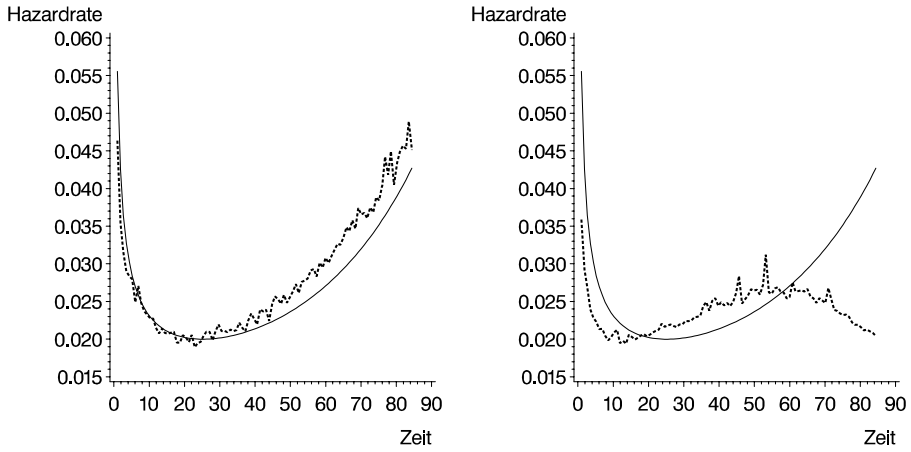


Abbildung B.11: Typ I, $n = 100$, UAE-optimal (mit mL Plug-in)

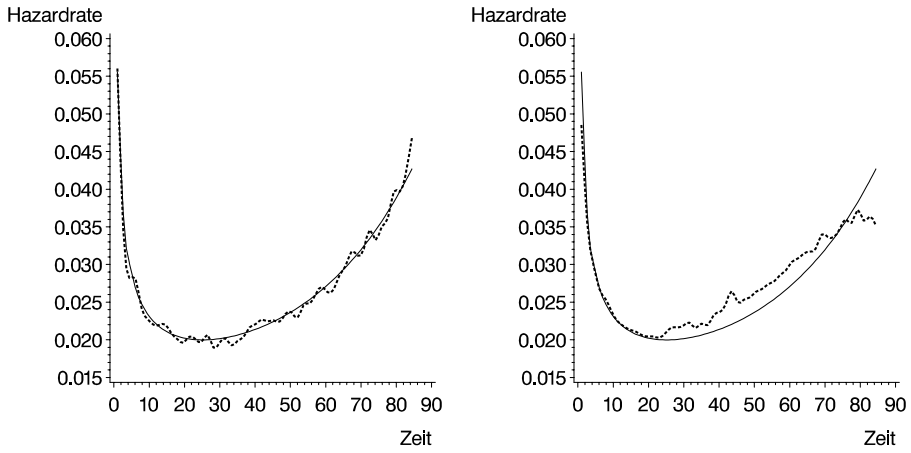


Abbildung B.12: Typ I, $n = 300$, UAE-optimal (mit mL Plug-in), †

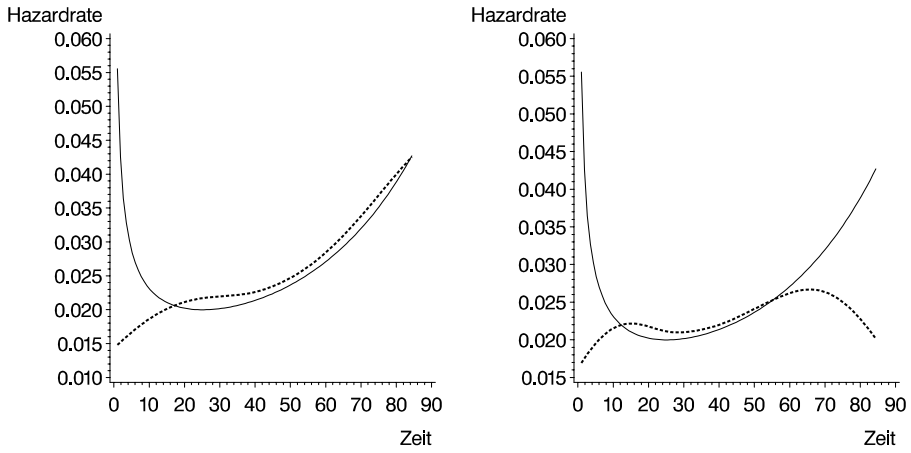


Abbildung B.13: Typ I, $n = 50$, fix

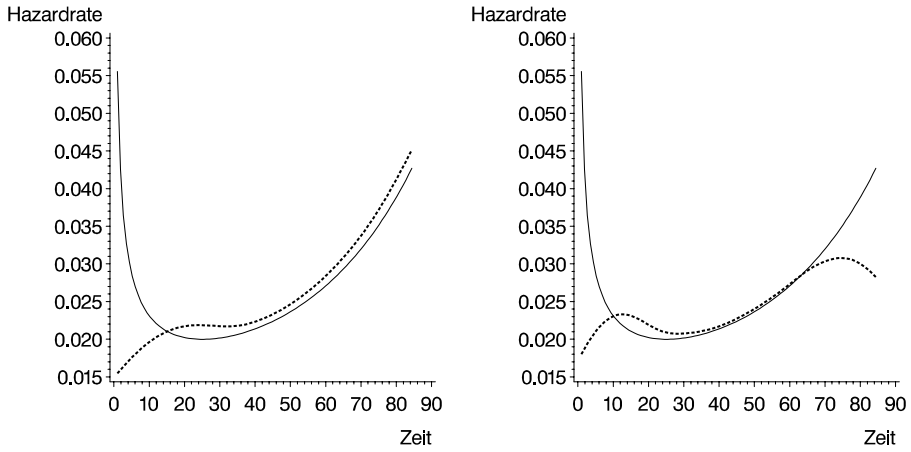


Abbildung B.14: Typ I, $n = 100$, fix

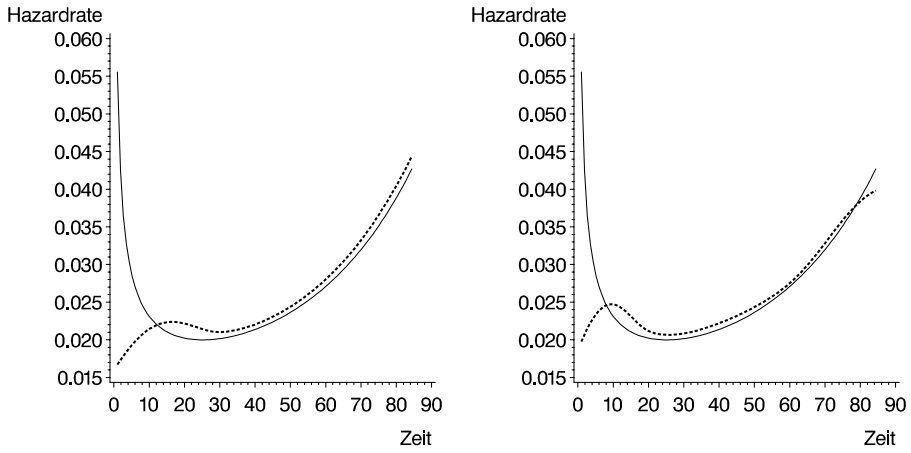


Abbildung B.15: Typ I, $n = 300$, fix

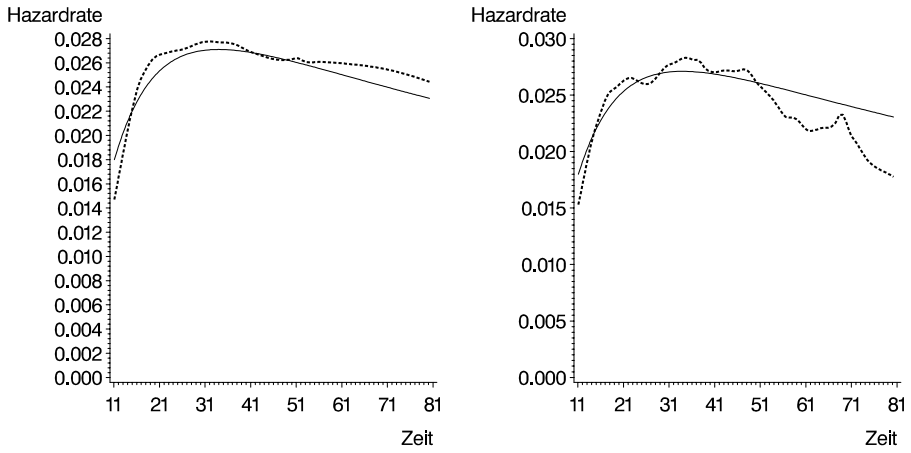


Abbildung B.16: Typ II, $n = 50$, smooth RoT, $T1$

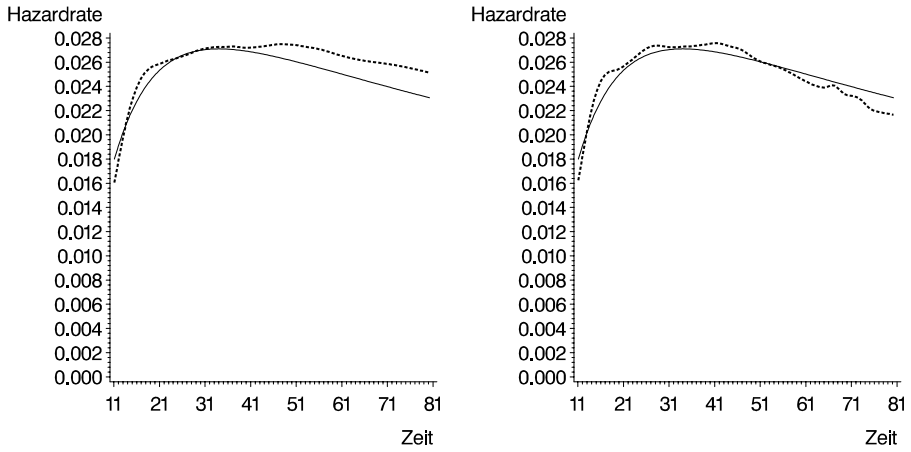


Abbildung B.17: Typ II, $n = 100$, smooth RoT, $T1$

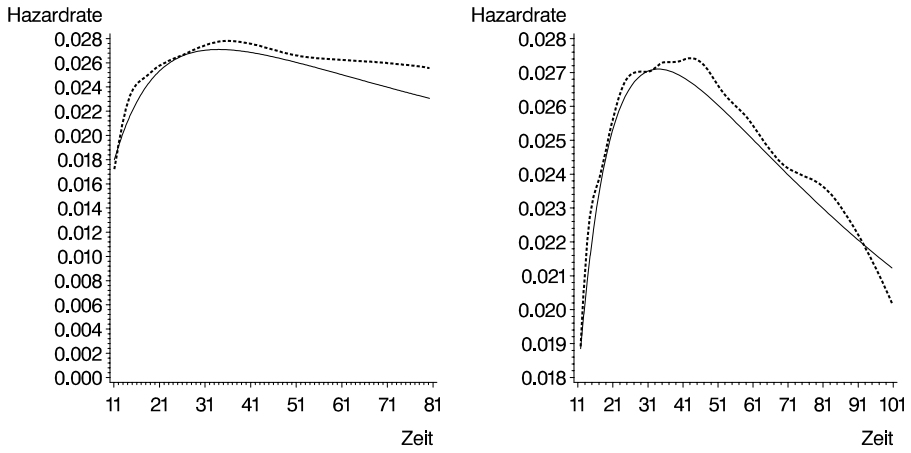


Abbildung B.18: Typ II, $n = 300$, smooth RoT, $T1$, bzw. ohne

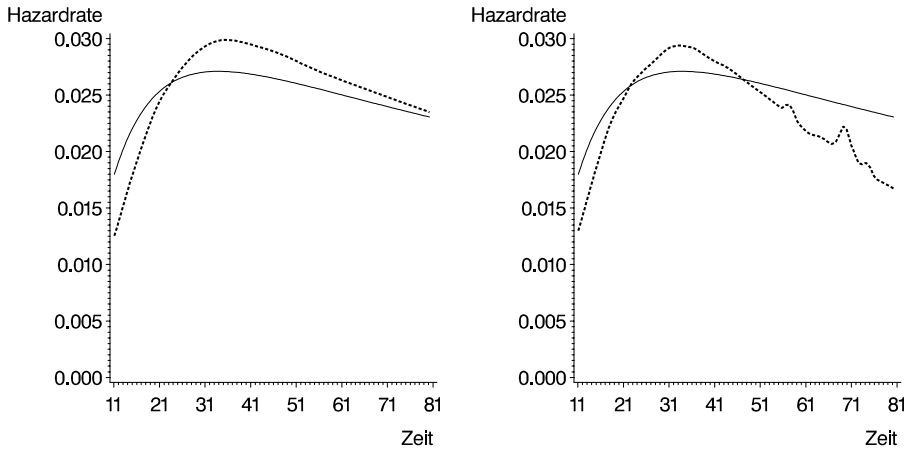


Abbildung B.19: Typ II, $n = 50$, mL, $T1$

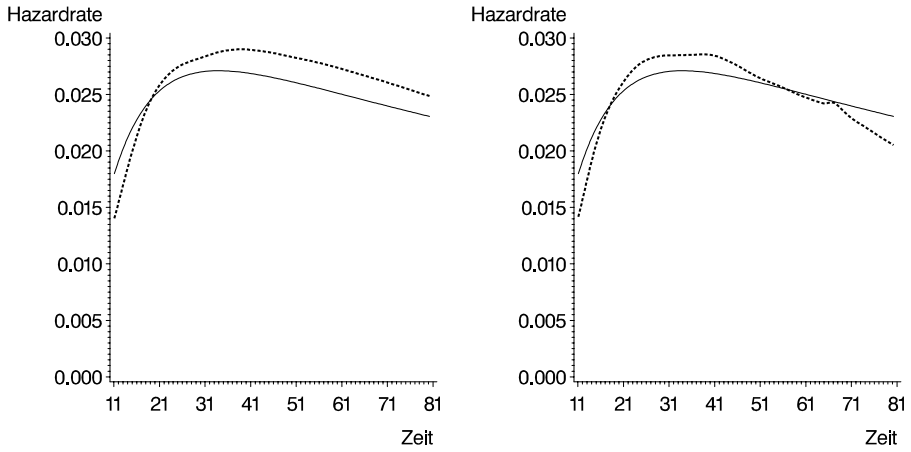


Abbildung B.20: Typ II, $n = 100$, mL, $T1$

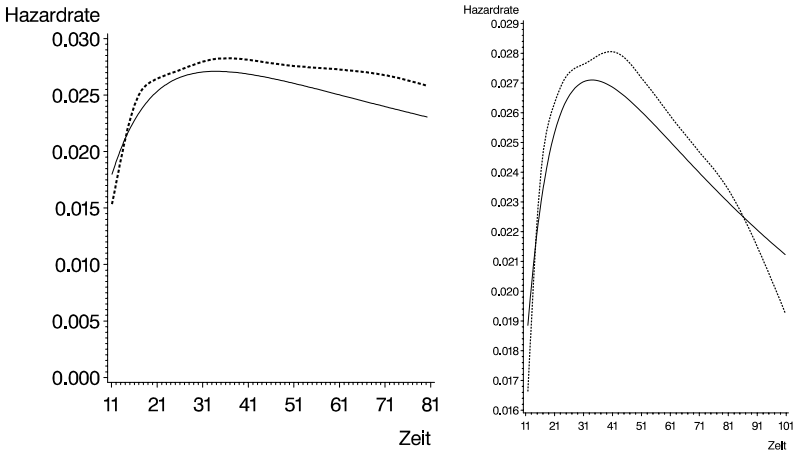


Abbildung B.21: Typ II, $n = 300$, mL, $T1$, bzw. ohne

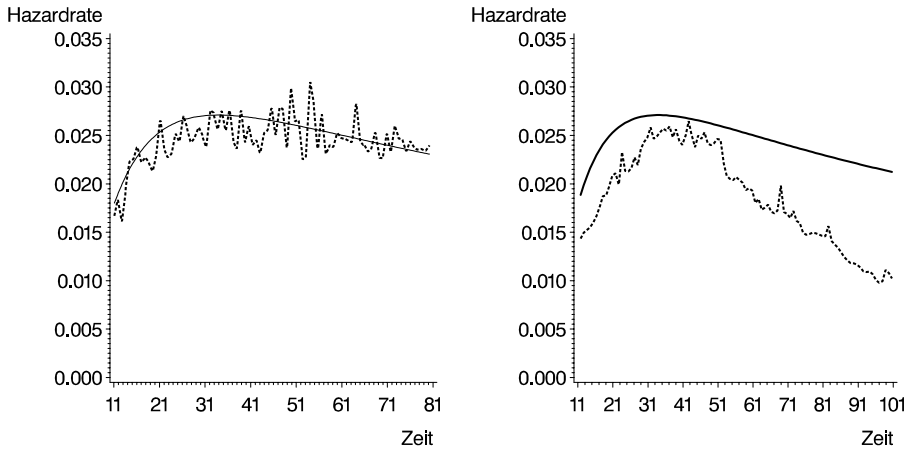


Abbildung B.22: Typ II, $n = 50$, UAE-optimal (mit smooth RoT Plug-in), $T1$, bzw. nicht $T1$

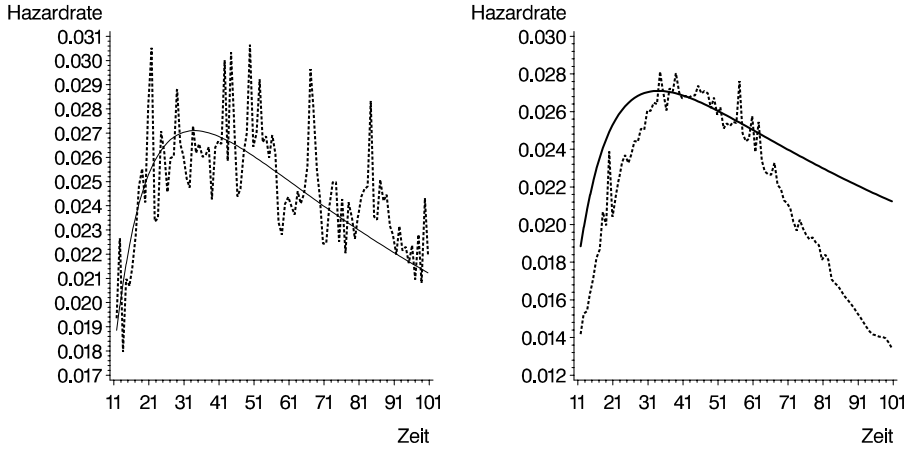


Abbildung B.23: Typ II, $n = 100$, UAE-optimal (mit smooth RoT Plug-in)

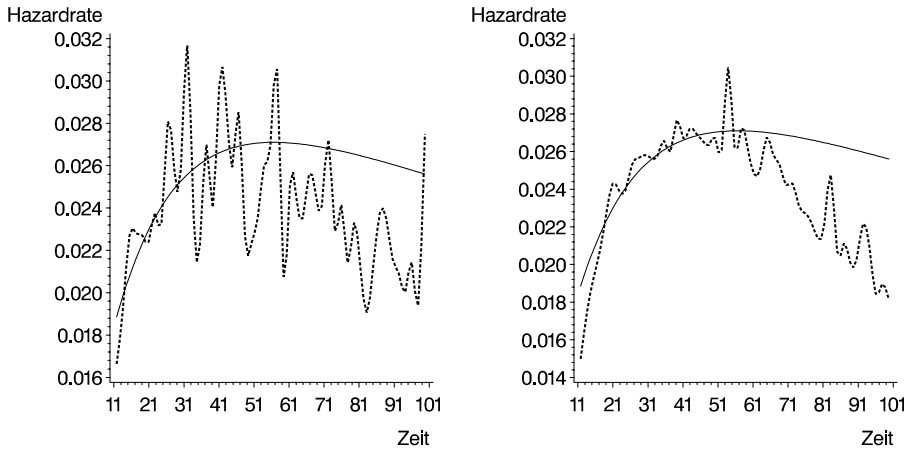


Abbildung B.24: Typ II, $n = 300$, UAE-optimal (mit smooth RoT Plug-in), †

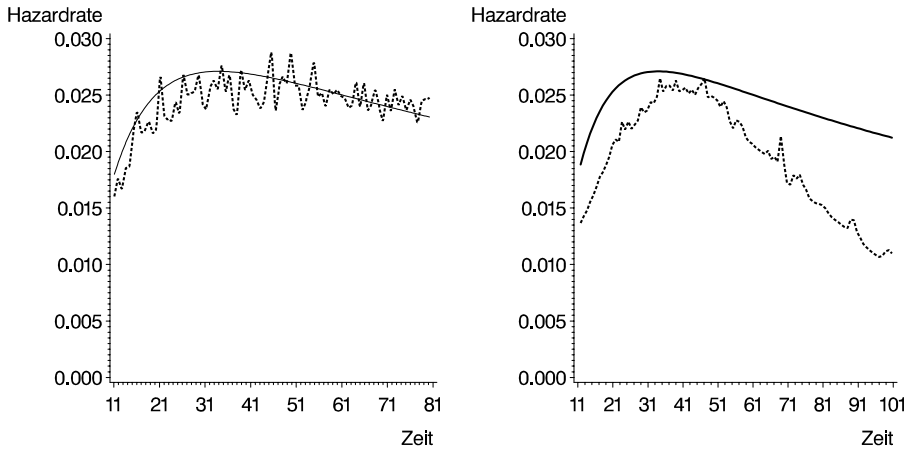


Abbildung B.25: Typ II, $n = 50$, UAE-optimal (mit mL Plug-in), $T1$, bzw. nicht $T1$

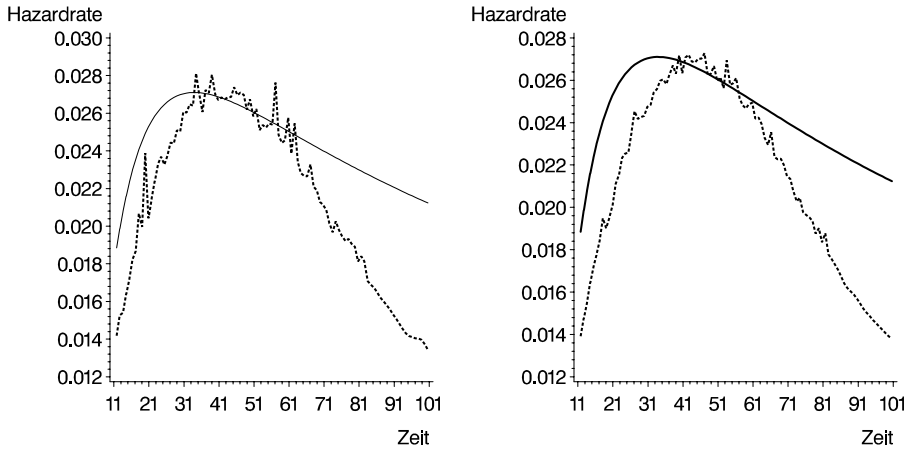


Abbildung B.26: Typ II, $n = 100$, UAE-optimal (mit mL Plug-in)

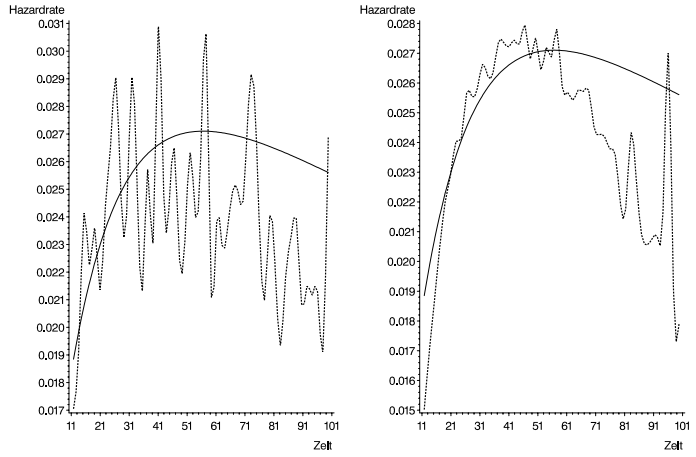


Abbildung B.27: Typ II, $n = 300$, UAE-optimal (mit mL Plug-in), †

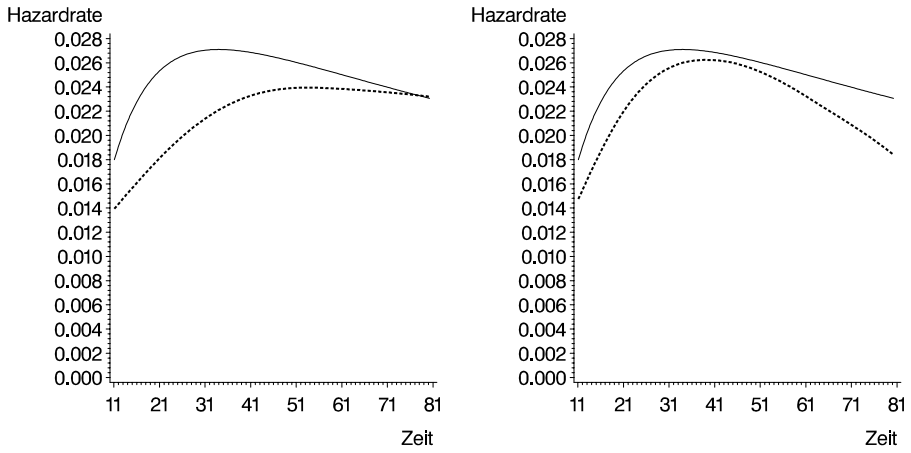


Abbildung B.28: Typ II, $n = 50$, fix, $T1$

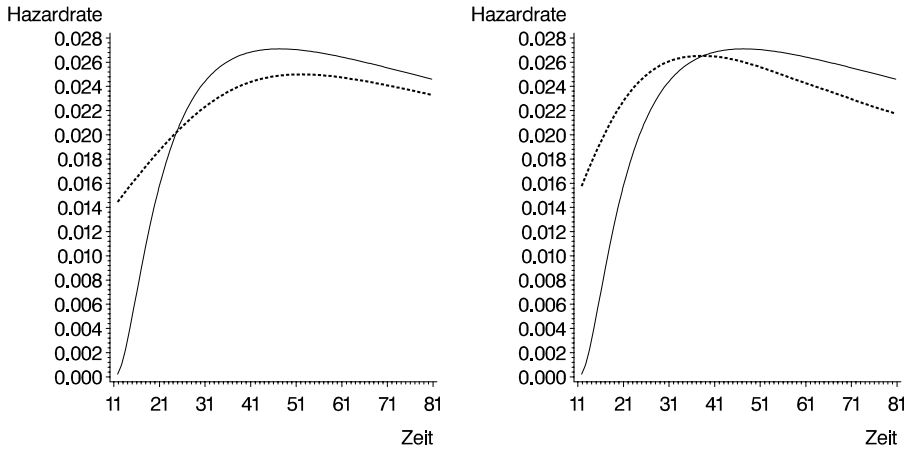


Abbildung B.29: Typ II, $n = 100$, fix

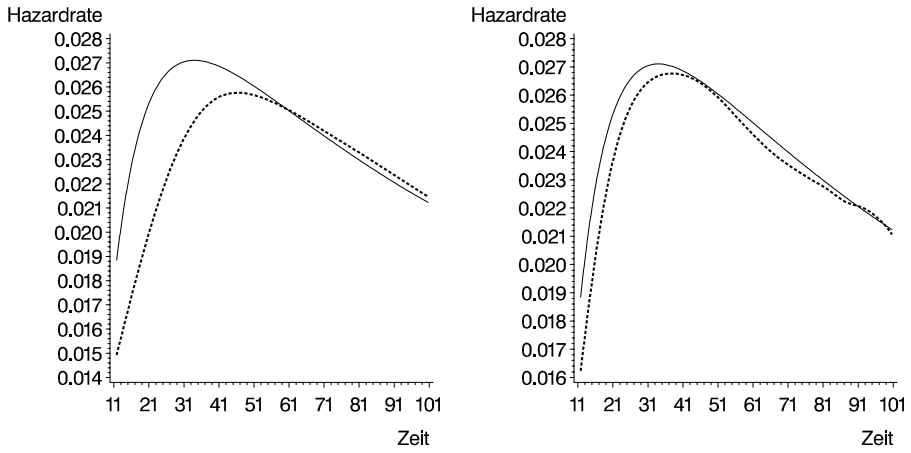


Abbildung B.30: Typ II, $n = 300$, fix

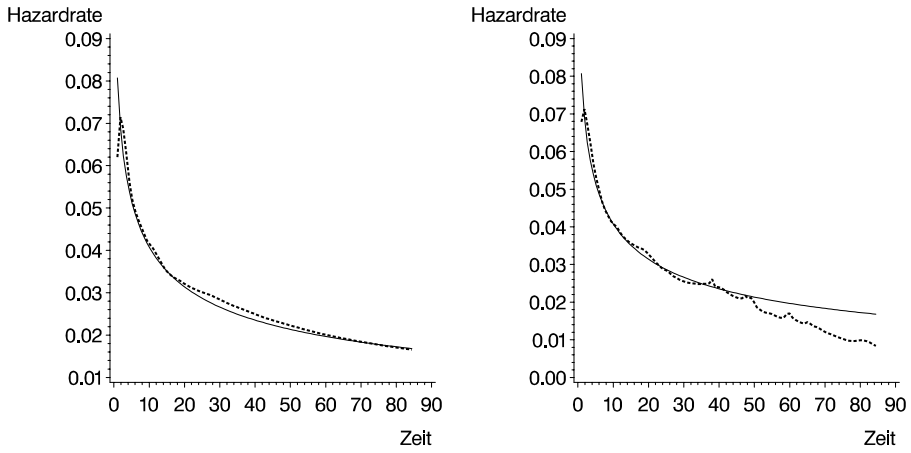


Abbildung B.31: Typ III, $n = 50$, smooth RoT, $T1$

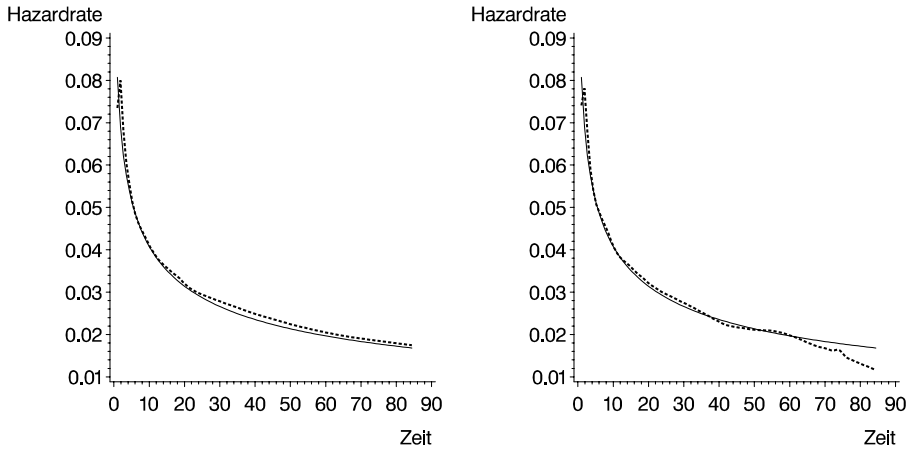


Abbildung B.32: Typ III, $n = 100$, smooth RoT, $T1$

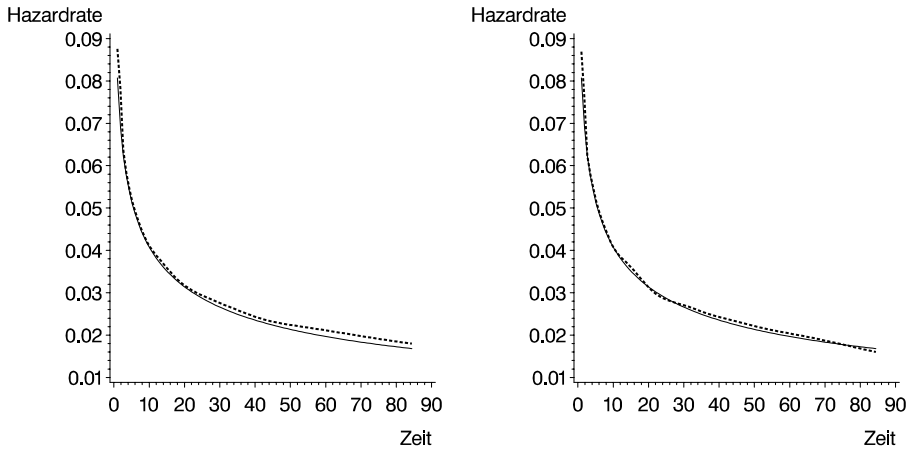


Abbildung B.33: Typ III, $n = 300$, smooth RoT, $T1$

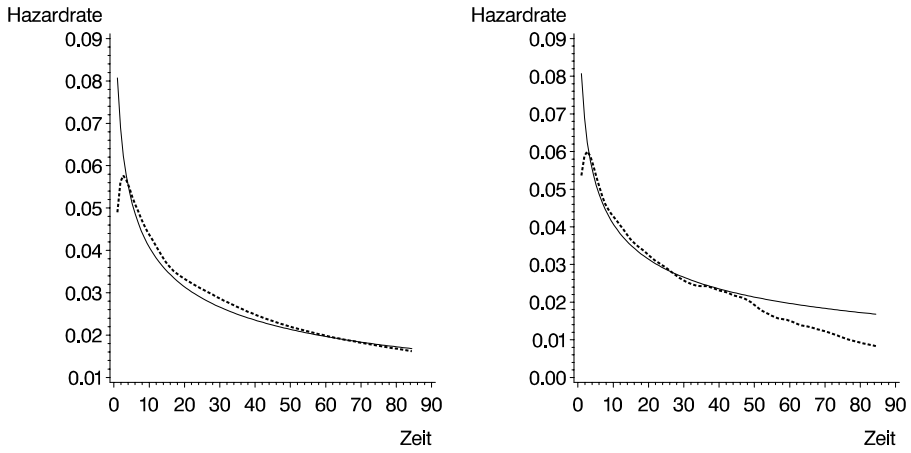


Abbildung B.34: Typ III, $n = 50$, mL, $T1$

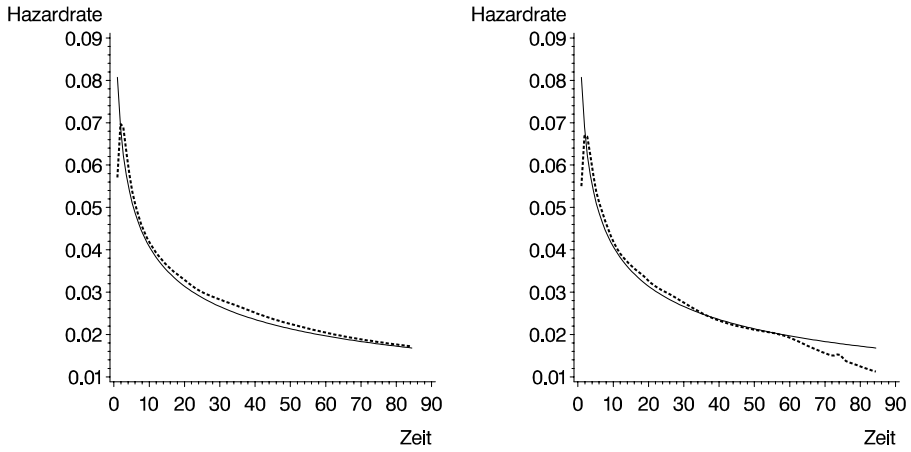


Abbildung B.35: Typ III, $n = 100$, mL, $T1$

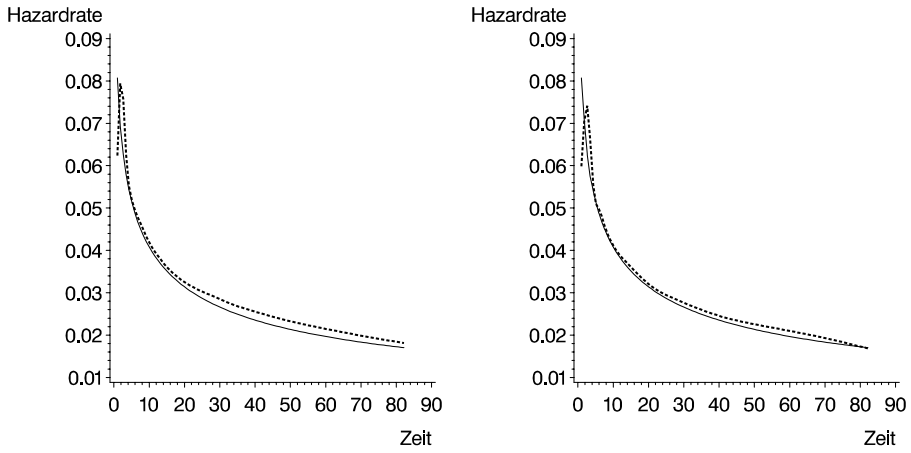


Abbildung B.36: Typ III, $n = 300$, mL, $nsim = 146$, bzw. †

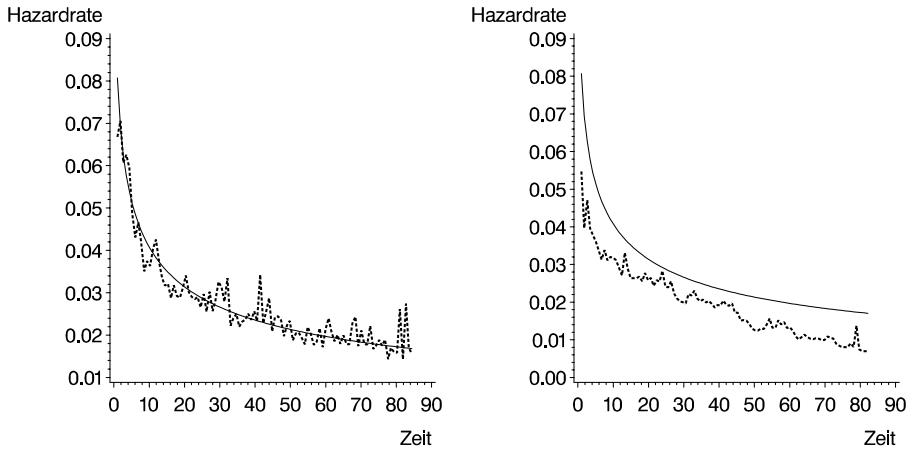


Abbildung B.37: Typ III, $n = 50$, UAE-optimal (mit smooth RoT Plug-in), $T1$, bzw. nicht $T1$

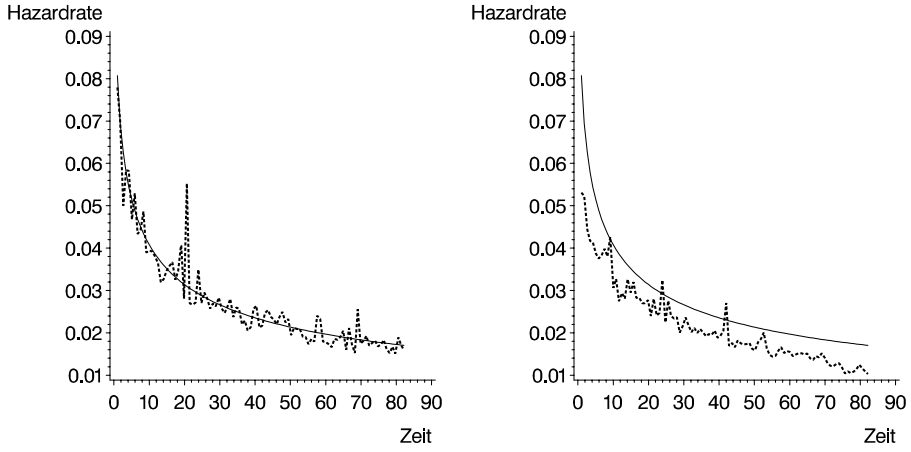


Abbildung B.38: Typ III, $n = 100$, UAE-optimal (mit smooth RoT Plug-in)

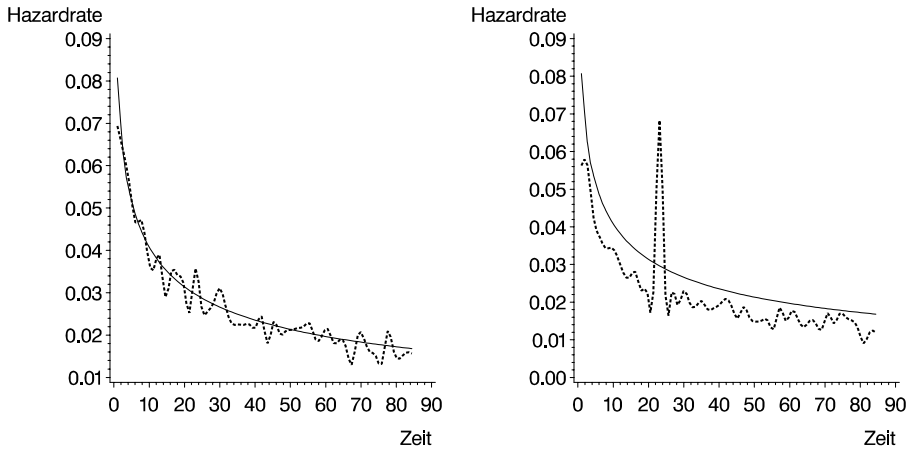


Abbildung B.39: Typ III, $n = 300$, UAE-optimal (mit smooth RoT Plug-in), †, $T1$

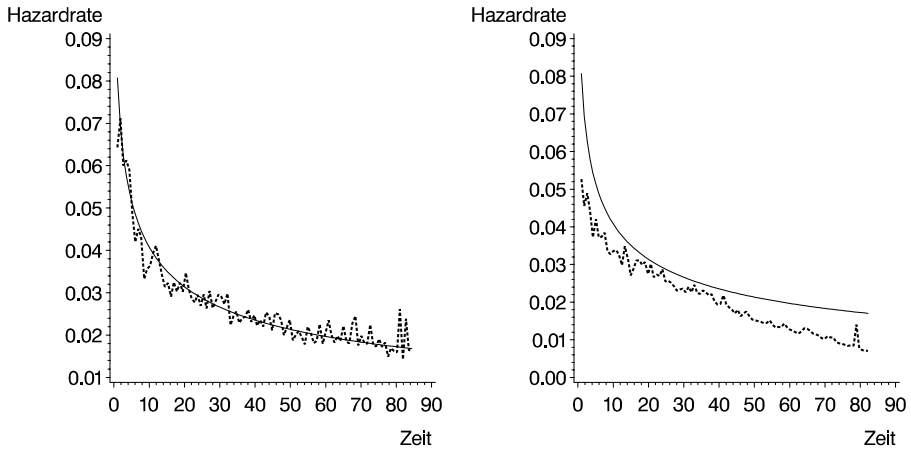


Abbildung B.40: Typ III, $n = 50$, UAE-optimal (mit mL Plug-in), $T1$, bzw. nicht $T1$

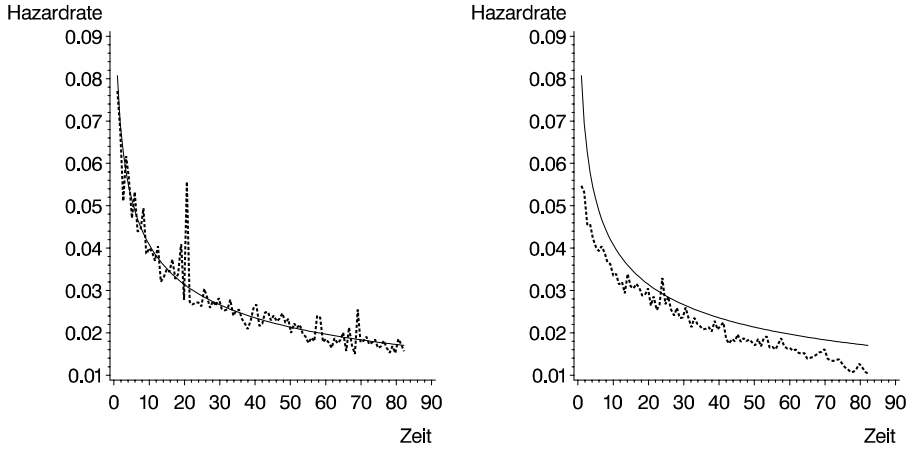


Abbildung B.41: Typ III, $n = 100$, UAE-optimal (mit mL Plug-in)

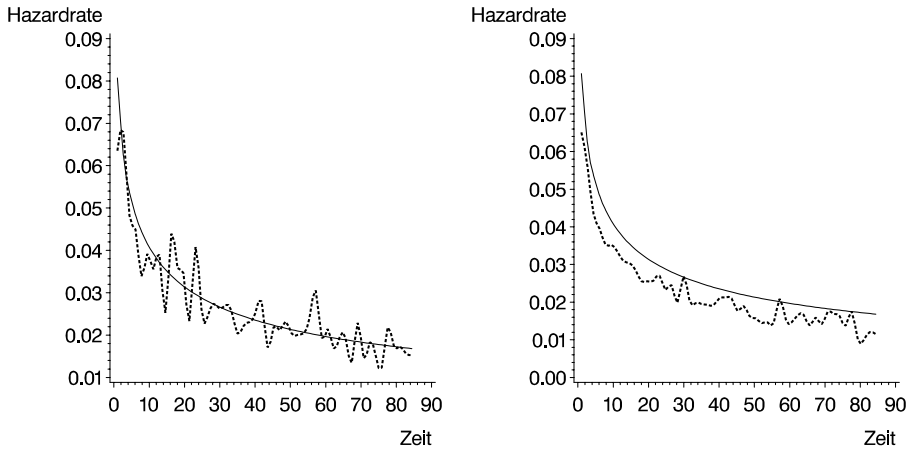


Abbildung B.42: Typ III, $n = 300$, UAE-optimal (mit mL Plug-in), † und $nsim = 146$, bzw. †, $T1$

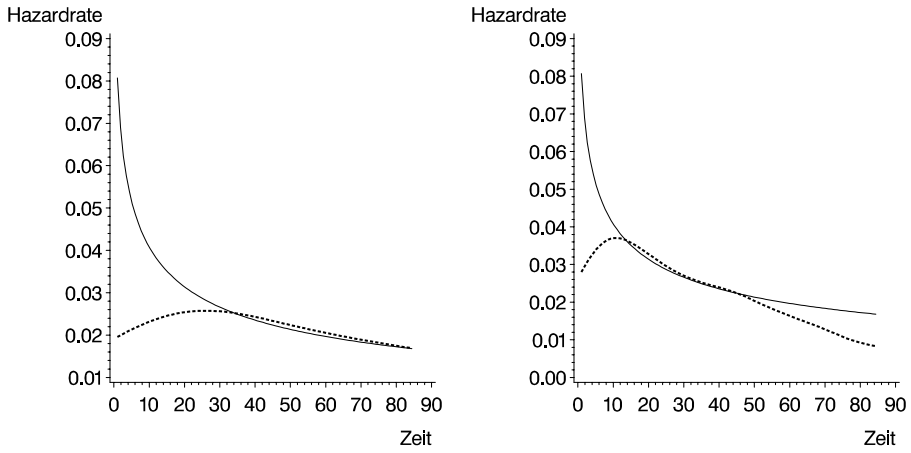


Abbildung B.43: Typ III, $n = 50$, fix, $T1$

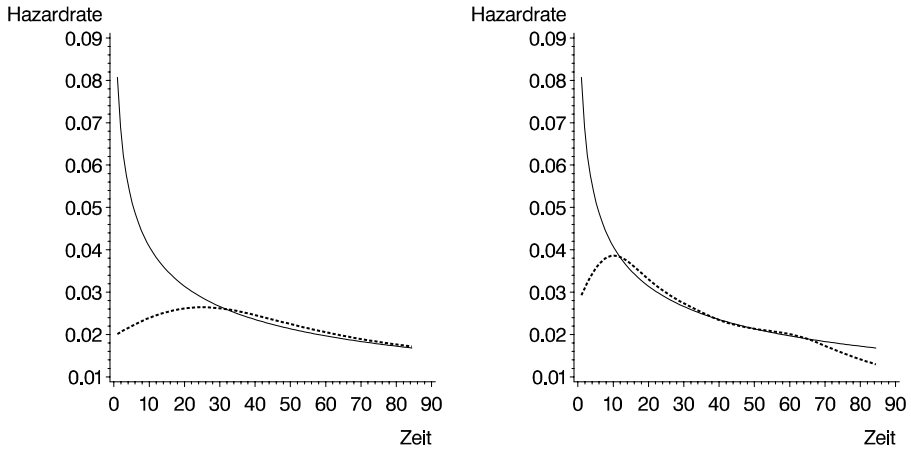


Abbildung B.44: Typ III, $n = 100$, fix, $T1$

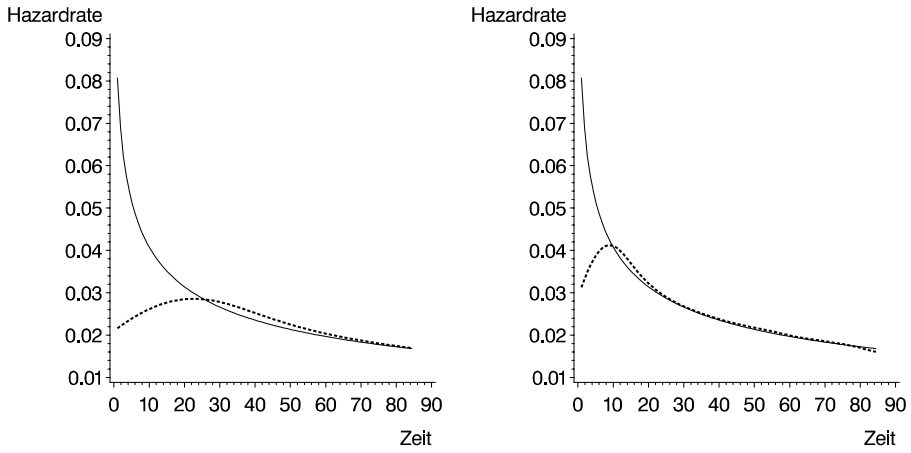


Abbildung B.45: Typ III, $n = 300$, fix, $T1$

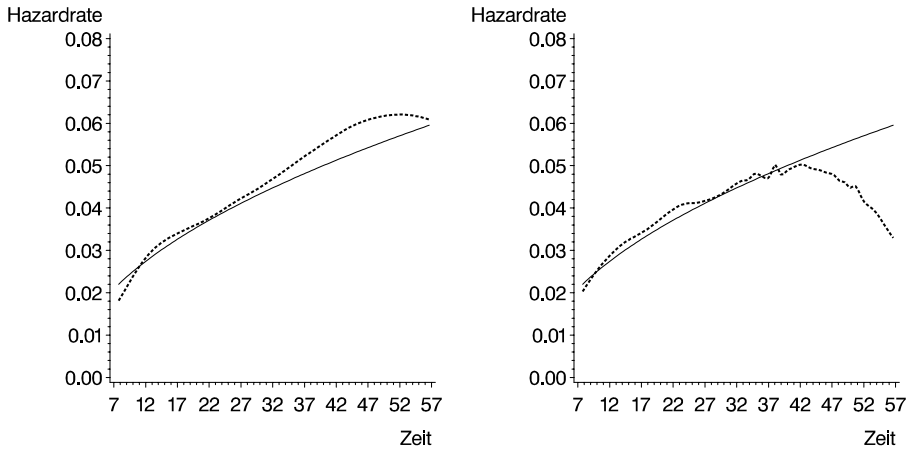


Abbildung B.46: Typ IV, $n = 50$, smooth RoT, $T1$

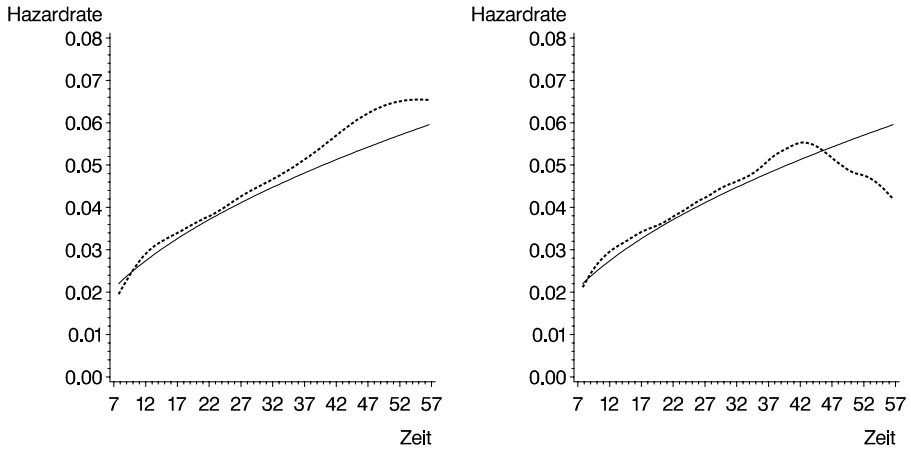


Abbildung B.47: Typ IV, $n = 100$, smooth RoT, $T1$

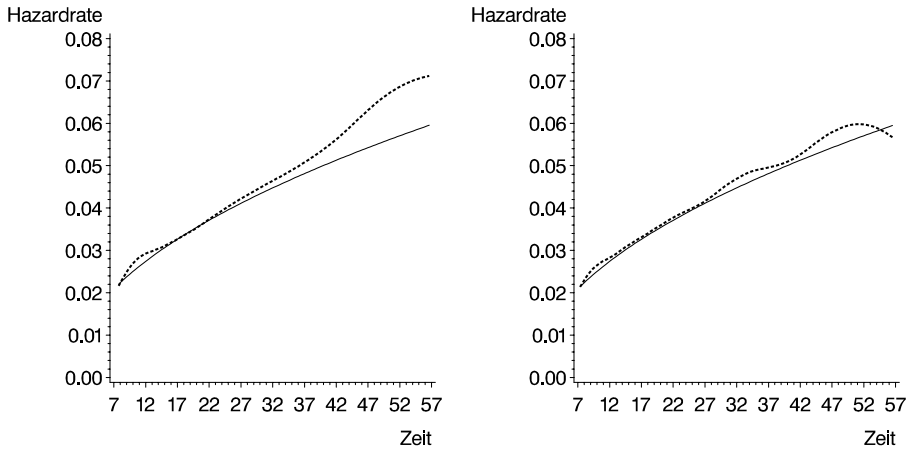


Abbildung B.48: Typ IV, $n = 300$, smooth RoT, $T1$

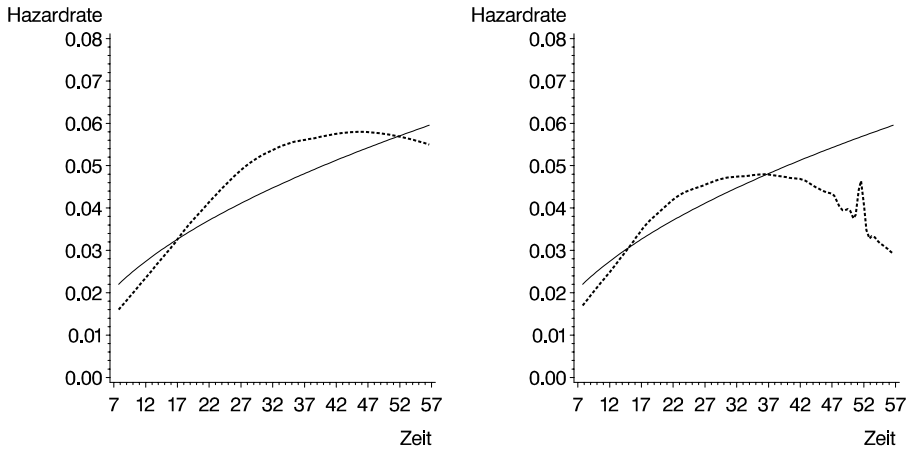


Abbildung B.49: Typ IV, $n = 50$, mL, $T1$

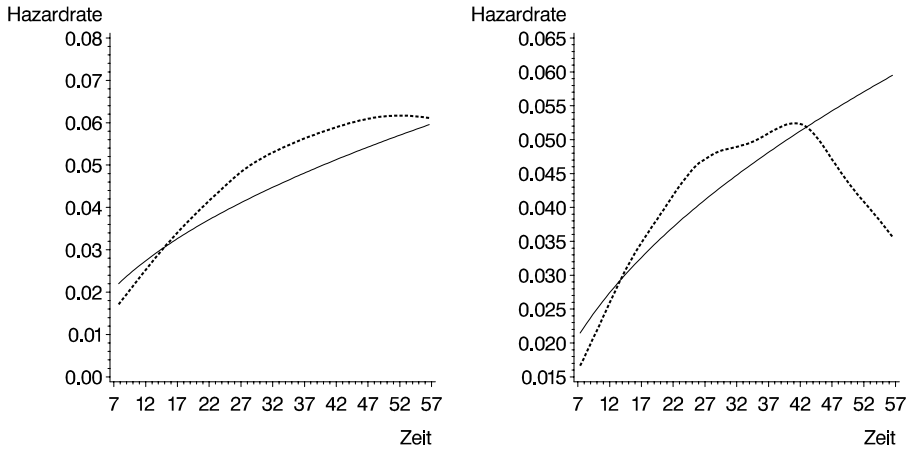


Abbildung B.50: Typ IV, $n = 100$, mL, $T1$, bzw. nicht $T1$

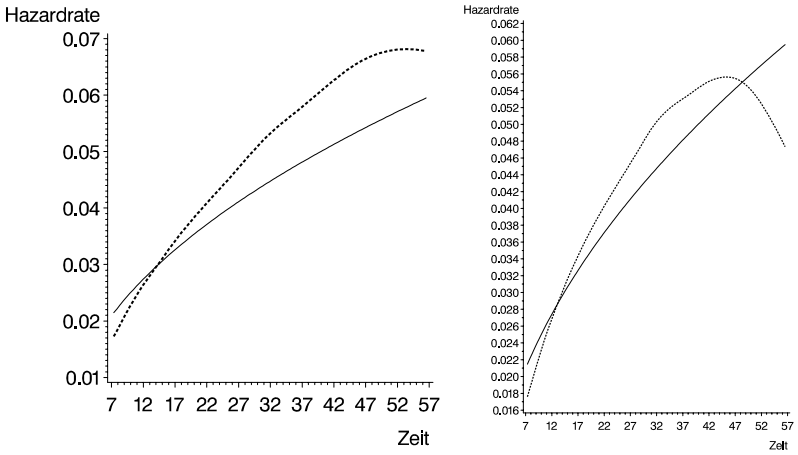


Abbildung B.51: Typ IV, $n = 300$, mL, †‡, bzw. †

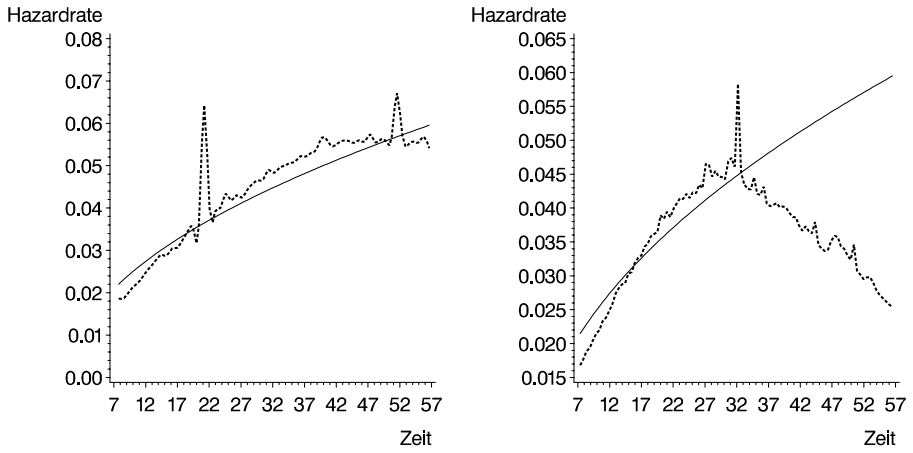


Abbildung B.52: Typ IV, $n = 50$, UAE-optimal (mit smooth RoT Plug-in), $T1$, bzw. nicht $T1$

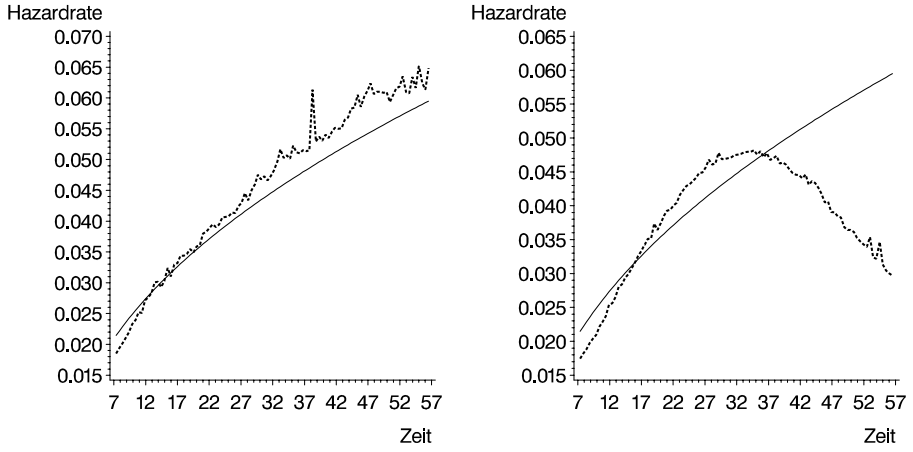


Abbildung B.53: Typ IV, $n = 100$, UAE-optimal (mit smooth RoT Plug-in)

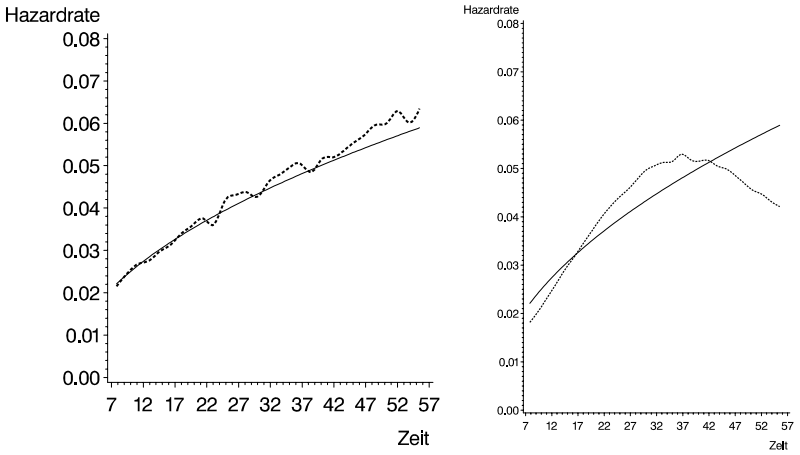


Abbildung B.54: Typ IV, $n = 300$, UAE-optimal (mit smooth RoT Plug-in), †, $T1$

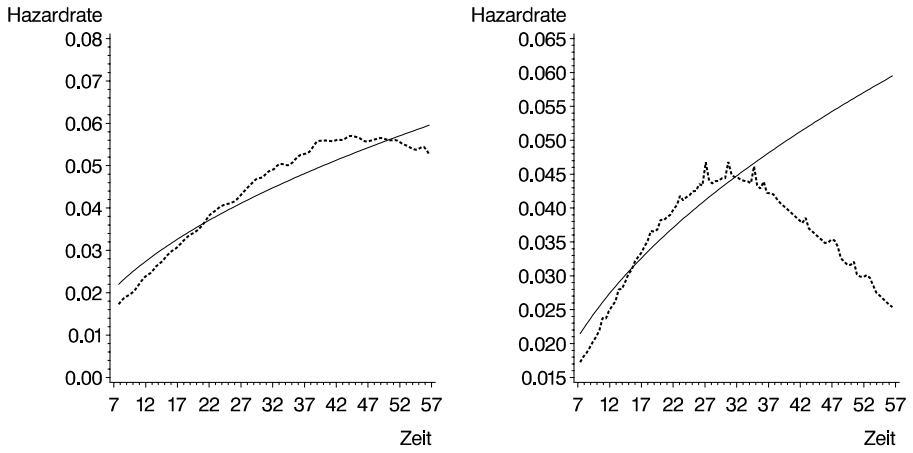


Abbildung B.55: Typ IV, $n = 50$, UAE-optimal (mit mL Plug-in), $T1$, bzw. nicht $T1$

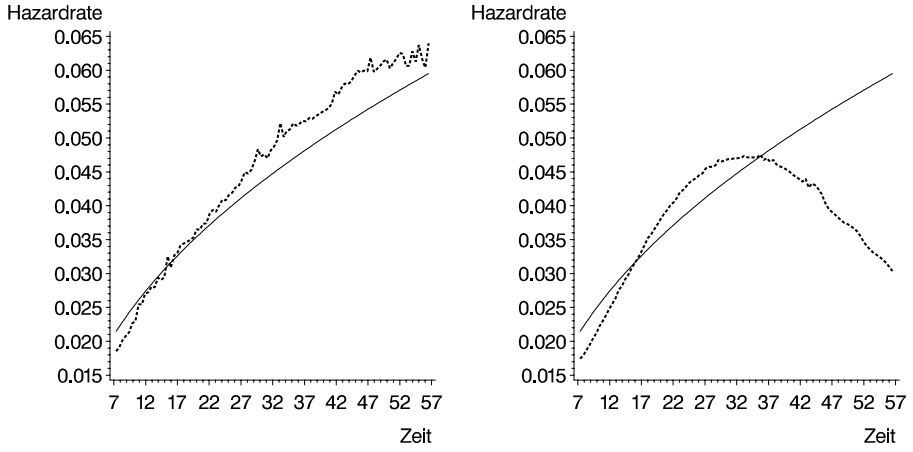


Abbildung B.56: Typ IV, $n = 100$, UAE-optimal (mit mL Plug-in)

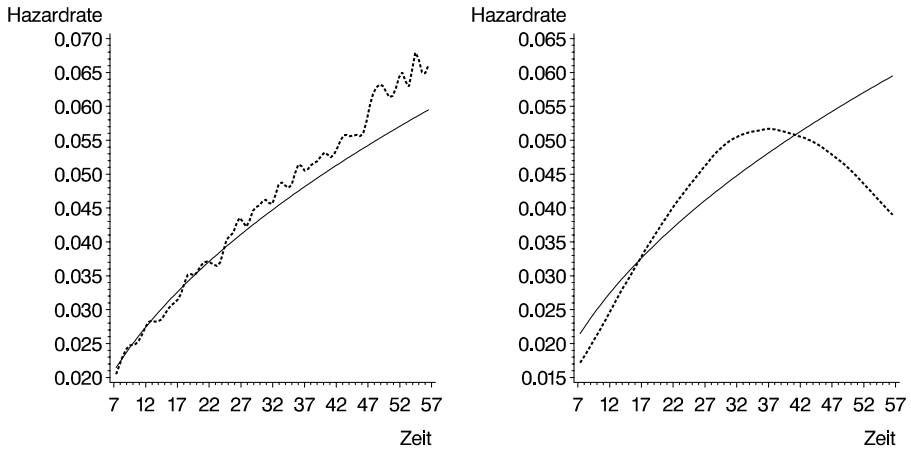


Abbildung B.57: Typ IV, $n = 300$, UAE-optimal (mit mL Plug-in), ††‡ und, bzw. †

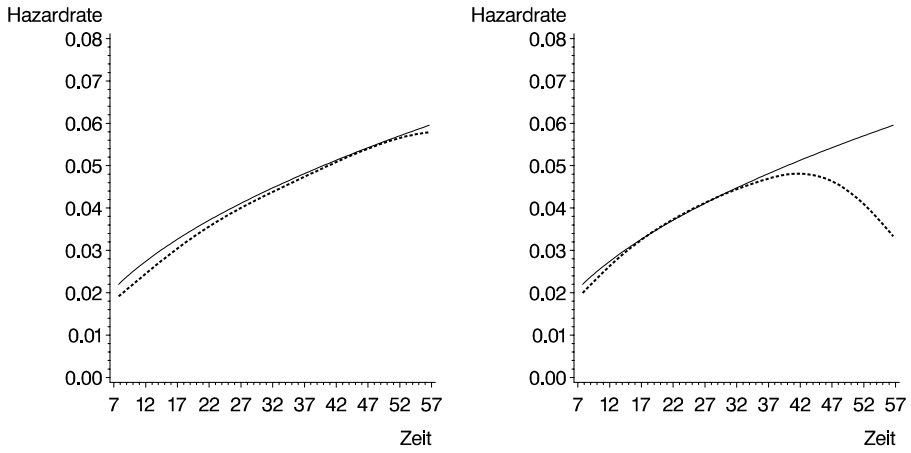


Abbildung B.58: Typ IV, $n = 50$, fix, $T1$

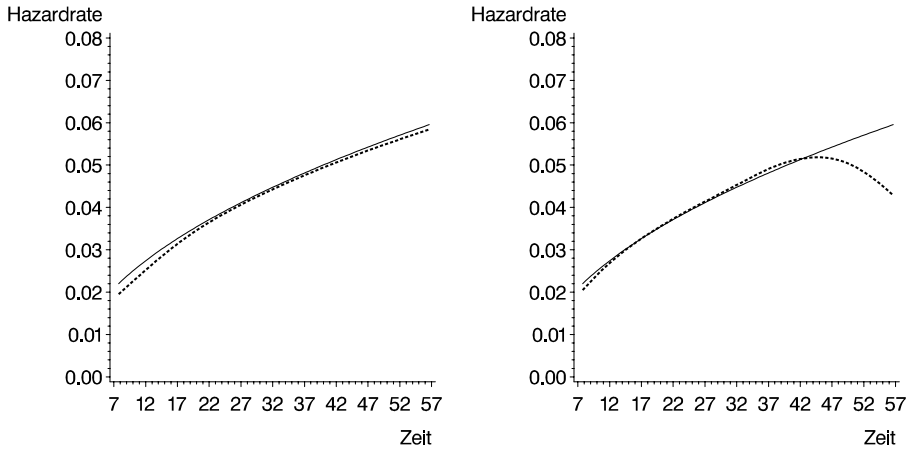


Abbildung B.59: Typ IV, $n = 100$, fix, $T1$

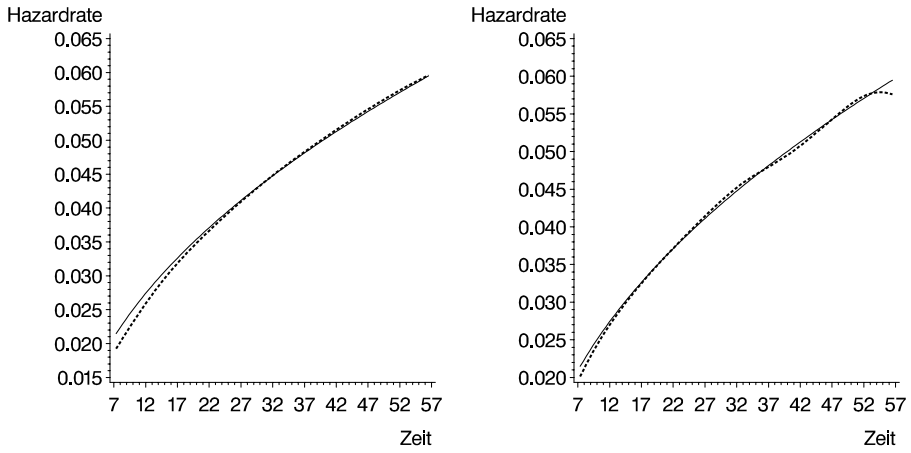


Abbildung B.60: Typ IV, $n = 300$, fix

Anhang C

Verlustmaßzahlen

Die folgenden Tabellen enthalten die Verlustmaßahlen der Bandbreitenwahlen der Daumenregel für die nächste-Nachbarn Bandbreite (Smooth RoT, kurz sRoT), der modified Likelihood Maximierung für die nächste-Nachbarn Bandbreite (mL), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit sRoT Plug-in (UAE(sRoT)), der optimalen nächste-Nachbarn Bandbreitenwahl für die gleichmäßig absoluten Fehler mit mL Plug-in (UAE(mL)) und der Daumenregel für die fixe Bandbreite (fix). Die Maßzahlen sind abgekürzt mit: MISE für zu erwartender integrierter quadratischer Fehler, UAE für gleichmäßig absoluter Fehler, MIKLE für zu erwartender integrierter Kullback-Leibler Fehler, NN für nächste Nachbarn (bei der fixen Bandbreite Bandbreite), CENS80 für den empirischen Zensierungsgrad auf dem inneren 80%-Bereich, IBIAS2 für die integrierte quadratische Verzerrung und IVARIANZ für die integrierte Varianz. Je Weibull Verteilung werden für die drei Stichprobenumfängen von 50, 100 und 300 zunächst die Ergebnisse ohne Zensierung präsentiert und als zweites die Ergebnisse für die 40%-tige Zensierung. Die Tabellen C.1 bis C.6 beziehen sich auf den Hazardratentyp I, die Tabellen C.7 bis C.12 auf den Hazardratentyp II, die Tabellen C.13 bis C.18 auf den Hazardratentyp III und die Tabellen C.19 bis C.24 auf den Hazardratentyp IV.

Tabelle C.1: Maßzahlen für Typ I, $n = 50$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0084	0.006	0.0093	0.007	0.0247	0.096	0.1315	1.989	0.0058	0.004
UAE	0.0359	0.012	0.0380	0.016	0.0630	0.088	0.0990	0.349	0.0408	0.003
MIKLE	8.7051	1.242	8.6233	1.296	8.4307	1.184	8.3564	1.665	7.7877	1.199
NN	17.0140	0.535	17.5940	6.095	65.2640	96.393	58.2700	100.794	41.1192	3.823
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0067	0.005	0.0084	0.006	0.0222	0.096	0.1295	1.984	0.0032	0.004
IVARIANZ	0.0022	0.002	0.0015	0.002	0.0037	0.003	0.0031	0.006	0.0026	0.001

Tabelle C.2: Maßzahlen für Typ I, $n = 100$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0072	0.006	0.0073	0.007	0.0281	0.088	0.0584	0.339	0.0047	0.004
UAE	0.0313	0.007	0.0299	0.009	0.0648	0.095	0.0877	0.186	0.0402	0.002
MIKLE	8.9809	1.054	8.8227	1.128	8.6139	1.183	8.4385	1.331	7.8731	0.979
NN	29.9360	0.623	26.9600	4.245	43.2220	85.617	34.6980	79.452	35.7550	2.446
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0042	0.004	0.0052	0.004	0.0269	0.088	0.0576	0.337	0.0022	0.004
IVARIANZ	0.0033	0.003	0.0026	0.003	0.0020	0.002	0.0017	0.003	0.0024	0.001

Tabelle C.3: Maßzahlen für Typ I, $n = 300$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0059	0.004	0.0060	0.004	0.0246	0.049	0.0230	0.048	0.0029	0.001
UAE	0.0250	0.006	0.0252	0.006	0.0544	0.054	0.0531	0.056	0.0388	0.002
MIKLE	9.0269	0.698	9.0799	0.698	7.8753	0.814	7.8983	0.757	7.8750	0.564
NN	72.6960	0.877	75.0000	0.000	12.2520	9.301	13.0560	9.515	28.7072	1.125
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0020	0.001	0.0019	0.001	0.0245	0.049	0.0229	0.048	0.0009	0.001
IVARIANZ	0.0040	0.003	0.0042	0.003	0.0002	0.000	0.0002	0.000	0.0020	0.000

Tabelle C.4: Maßzahlen für Typ I, $n = 50$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0370	0.077	0.0292	0.054	143.076	3198.10	0.0476	0.138	0.0155	0.013
UAE	0.0584	0.081	0.0514	0.055	0.9150	18.421	0.0835	0.127	0.0403	0.004
MIKLE	7.3501	1.931	7.2798	1.888	7.4449	22.353	6.4199	1.999	7.0009	1.868
NN	14.1640	1.193	18.3740	6.526	88.0720	165.245	100.390	154.417	27.7094	3.535
CENS80	0.4225	0.079	0.4225	0.079	0.4225	0.079	0.4225	0.079	0.4225	0.079
IBIAS2	0.0340	0.078	0.0257	0.056	142.803	3185.78	0.0400	0.140	0.0106	0.012
IVARIANZ	0.0051	0.005	0.0055	0.005	0.5504	12.306	0.0090	0.007	0.0059	0.005

Tabelle C.5: Maßzahlen für Typ I, $n = 100$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0195	0.029	0.0172	0.025	0.0485	0.263	0.0416	0.251	0.0120	0.011
UAE	0.0402	0.029	0.0380	0.023	0.0801	0.158	0.0727	0.140	0.0390	0.005
MIKLE	7.7705	1.653	7.7623	1.620	7.1941	1.756	7.2294	1.686	7.4569	1.584
NN	24.9600	1.363	27.5400	4.773	78.3560	127.651	82.3900	126.187	24.2224	2.224
CENS80	0.4206	0.051	0.4206	0.051	0.4206	0.051	0.4206	0.051	0.4206	0.051
IBIAS2	0.0183	0.030	0.0159	0.026	0.0442	0.262	0.0373	0.250	0.0095	0.012
IVARIANZ	0.0023	0.002	0.0025	0.002	0.0055	0.005	0.0055	0.005	0.0032	0.003

Tabelle C.6: Maßzahlen für Typ I, $n = 300$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0078	0.009	0.0069	0.007	0.0326	0.050	0.0145	0.021	0.0066	0.007
UAE	0.0256	0.011	0.0258	0.007	0.0660	0.066	0.0406	0.032	0.0367	0.005
MIKLE	8.3346	1.154	8.4130	1.115	8.1135	1.340	8.2411	1.100	7.9046	1.080
NN	60.6160	1.873	75.0232	0.373	45.4880	54.529	67.0400	57.512	19.2572	1.024
CENS80	0.4228	0.030	0.4234	0.030	0.4228	0.030	0.4228	0.030	0.4242	0.030
IBIAS2	0.0075	0.008	0.0058	0.006	0.0324	0.050	0.0141	0.020	0.0052	0.007
IVARIANZ	0.0007	0.001	0.0012	0.001	0.0006	0.001	0.0009	0.001	0.0014	0.000

Tabelle C.7: Maßzahlen für Typ II, $n = 50$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0066	0.006	0.0045	0.005	0.0947	0.363	0.0565	0.143	0.0033	0.003
UAE	0.0195	0.013	0.0139	0.008	0.1303	0.213	0.0934	0.148	0.0099	0.005
MIKLE	6.9426	1.239	6.9639	1.178	6.5907	1.646	6.5746	1.502	5.9271	1.239
NN	15.0440	1.471	29.3400	6.863	26.9280	84.450	28.8140	59.219	57.7508	16.179
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0065	0.006	0.0042	0.004	0.0946	0.362	0.0564	0.143	0.0021	0.003
IVARIANZ	0.0003	0.000	0.0006	0.001	0.0005	0.001	0.0005	0.001	0.0015	0.001

Tabelle C.8: Maßzahlen für Typ II, $n = 100$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0037	0.004	0.0030	0.003	0.1480	0.459	0.1209	0.430	0.0018	0.002
UAE	0.0144	0.007	0.0115	0.005	0.1782	0.271	0.1494	0.249	0.0072	0.004
MIKLE	7.0380	0.931	7.1366	0.908	8.2568	1.696	8.2705	1.602	6.3326	1.042
NN	25.9320	2.244	45.1920	14.728	15.2140	33.308	22.1820	70.657	50.8817	11.057
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0036	0.003	0.0027	0.003	0.1477	0.457	0.1207	0.428	0.0016	0.003
IVARIANZ	0.0003	0.001	0.0005	0.001	0.0007	0.002	0.0006	0.002	0.0005	0.001

Tabelle C.9: Maßzahlen für Typ II, $n = 300$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0016	0.001	0.0014	0.001	0.1281	0.271	0.1217	0.313	0.0011	0.001
UAE	0.0095	0.004	0.0080	0.003	0.1410	0.162	0.1304	0.161	0.0062	0.002
MIKLE	7.0582	0.517	7.1995	0.508	8.0142	1.736	7.9973	1.729	7.7331	0.648
NN	61.2900	3.964	90.5280	18.432	6.1240	10.455	7.7080	14.477	41.6626	5.873
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0014	0.001	0.0011	0.001	0.1275	0.269	0.1211	0.311	0.0006	0.001
IVARIANZ	0.0002	0.000	0.0004	0.000	0.0012	0.002	0.0011	0.003	0.0005	0.000

Tabelle C.10: Maßzahlen für Typ II, $n = 50$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0241	0.050	0.0175	0.054	9554.45	204475	9135.67	204279	0.0093	0.012
UAE	0.0406	0.048	0.0287	0.054	5.5058	102.600	4.5244	100.369	0.0195	0.011
MIKLE	6.4238	1.803	6.2918	1.722	24.3210	340.604	21.2879	332.753	6.1237	1.765
NN	13.8640	1.349	26.5740	8.397	91.5900	131.993	133.062	164.507	27.6114	5.810
CENS80	0.4023	0.068	0.4023	0.068	0.4300	0.071	0.4300	0.071	0.4023	0.068
IBIAS2	0.0237	0.051	0.0168	0.055	9535.38	203658	9117.41	203462	0.0088	0.013
IVARIANZ	0.0013	0.001	0.0016	0.002	38.1009	816.354	36.4878	815.891	0.0013	0.001

Tabelle C.11: Maßzahlen für Typ II, $n = 100$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0103	0.017	0.0073	0.013	0.0293	0.191	0.0123	0.026	0.0101	0.009
UAE	0.0247	0.021	0.0179	0.017	0.0403	0.124	0.0248	0.046	0.0212	0.010
MIKLE	6.7365	1.386	6.7333	1.356	7.1668	1.854	7.1253	1.714	6.3578	1.743
NN	24.4600	1.615	43.3200	14.491	122.480	158.059	148.920	149.561	24.1426	3.847
CENS80	0.3959	0.051	0.3959	0.051	0.4228	0.055	0.4228	0.055	0.4228	0.055
IBIAS2	0.0103	0.017	0.0071	0.014	0.0279	0.190	0.0109	0.026	0.0096	0.011
IVARIANZ	0.0003	0.000	0.0004	0.000	0.0022	0.002	0.0020	0.002	0.0014	0.001

Tabelle C.12: Maßzahlen für Typ II, $n = 300$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0057	0.006	0.0041	0.004	0.0214	0.100	0.0430	0.471	0.0050	0.006
UAE	0.0160	0.008	0.0124	0.006	0.0309	0.074	0.0323	0.135	0.0157	0.010
MIKLE	8.2522	1.178	8.2961	1.130	7.9207	1.470	8.0305	1.471	7.9639	1.197
NN	59.1040	2.282	90.9960	20.776	166.508	262.319	206.204	295.936	19.6229	1.759
CENS80	0.4236	0.031	0.4236	0.031	0.4239	0.032	0.4239	0.032	0.4236	0.031
IBIAS2	0.0057	0.006	0.0040	0.004	0.0211	0.099	0.0427	0.468	0.0050	0.006
IVARIANZ	0.0001	0.000	0.0002	0.000	0.0005	0.001	0.0005	0.003	0.0001	0.000

Tabelle C.13: Maßzahlen für Typ III, $n = 50$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0128	0.009	0.0101	0.007	0.5311	2.740	0.4481	2.465	0.0150	0.007
UAE	0.0521	0.024	0.0491	0.016	0.3599	0.609	0.3268	0.552	0.0612	0.005
MIKLE	8.1690	1.395	8.1463	1.343	8.0067	2.536	7.9388	2.394	6.7690	1.569
NN	13.0000	1.745	21.7040	6.342	5.7860	18.950	6.7060	19.039	65.9371	25.132
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0125	0.009	0.0093	0.007	0.5301	2.733	0.4473	2.461	0.0034	0.007
IVARIANZ	0.0005	0.001	0.0013	0.001	0.0021	0.008	0.0016	0.005	0.0116	0.005

Tabelle C.14: Maßzahlen für Typ III, $n = 100$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0072	0.005	0.0060	0.004	0.5047	3.756	0.4797	3.710	0.0129	0.004
UAE	0.0392	0.017	0.0393	0.013	0.3569	0.610	0.3490	0.591	0.0606	0.004
MIKLE	8.2375	1.129	8.2766	1.097	7.7720	2.314	7.7861	2.237	6.8684	1.202
NN	22.0440	2.834	32.3120	7.958	5.1480	14.508	5.0980	10.780	60.3321	19.548
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0069	0.005	0.0056	0.004	0.5036	3.738	0.4786	3.692	0.0019	0.003
IVARIANZ	0.0004	0.001	0.0006	0.001	0.0023	0.018	0.0023	0.018	0.0110	0.004

Tabelle C.15: Maßzahlen für Typ III, $n = 300$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0028	0.002	0.0024	0.001	0.2146	0.339	0.2620	0.636	0.0102	0.002
UAE	0.0248	0.011	0.0281	0.008	0.2106	0.199	0.2198	0.242	0.0591	0.002
MIKLE	8.2754	0.640	8.2897	0.633	7.7245	2.056	7.8095	2.299	7.0973	0.652
NN	51.9280	4.654	77.0548	5.151	4.3160	8.193	4.7192	15.870	48.7663	9.358
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0026	0.001	0.0018	0.001	0.2139	0.338	0.2601	0.629	0.0006	0.001
IVARIANZ	0.0003	0.000	0.0006	0.001	0.0016	0.003	0.0035	0.008	0.0096	0.002

Tabelle C.16: Maßzahlen für Typ III, $n = 50$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0420	0.071	0.0268	0.035	0.1649	0.870	0.1403	0.870	0.0251	0.020
UAE	0.0689	0.063	0.0551	0.030	0.1847	0.308	0.1529	0.297	0.0547	0.009
MIKLE	7.2066	2.221	7.1276	2.140	5.5593	2.816	5.9866	2.660	6.8082	2.283
NN	12.7040	1.676	20.7100	6.564	26.3920	63.656	34.0840	64.970	23.9662	8.919
CENS80	0.4160	0.078	0.4160	0.078	0.4159	0.078	0.4159	0.078	0.4160	0.078
IBIAS2	0.0407	0.071	0.0251	0.035	0.1588	0.876	0.1360	0.875	0.0191	0.022
IVARIANZ	0.0027	0.002	0.0030	0.003	0.0089	0.009	0.0064	0.007	0.0064	0.004

Tabelle C.17: Maßzahlen für Typ III, $n = 100$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0203	0.025	0.0148	0.023	0.1728	0.882	0.1068	0.283	0.0188	0.018
UAE	0.0462	0.027	0.0428	0.023	0.1921	0.327	0.1610	0.216	0.0530	0.008
MIKLE	7.8044	1.949	7.7465	1.881	6.1990	2.840	6.5108	2.606	7.4331	2.041
NN	21.7860	2.153	34.1600	8.971	30.0580	76.430	43.3200	113.913	21.1379	5.423
CENS80	0.4204	0.052	0.4204	0.052	0.4204	0.052	0.4204	0.052	0.4204	0.052
IBIAS2	0.0200	0.025	0.0142	0.023	0.1693	0.881	0.1043	0.282	0.0143	0.019
IVARIANZ	0.0007	0.001	0.0011	0.001	0.0058	0.007	0.0042	0.005	0.0045	0.002

Tabelle C.18: Maßzahlen für Typ III, $n = 300$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0074	0.009	0.0062	0.008	1.0124	13.849	0.1027	0.204	0.0104	0.012
UAE	0.0274	0.011	0.0267	0.010	0.1857	0.732	0.1290	0.151	0.0503	0.006
MIKLE	8.0781	1.360	8.0733	1.373	6.4787	5.086	6.5281	2.567	7.6661	1.424
NN	51.9340	3.457	77.9040	7.609	28.6680	80.147	36.7360	105.819	17.2752	2.558
CENS80	0.4198	0.030	0.4208	0.029	0.4208	0.029	0.4208	0.029	0.4198	0.030
IBIAS2	0.0074	0.009	0.0056	0.008	1.0062	13.754	0.1002	0.205	0.0068	0.013
IVARIANZ	0.0002	0.000	0.0008	0.001	0.0102	0.095	0.0044	0.006	0.0036	0.001

Tabelle C.19: Maßzahlen für Typ IV, $n = 50$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0124	0.016	0.0089	0.013	0.3951	7.218	0.0131	0.031	0.0066	0.011
UAE	0.0293	0.016	0.0220	0.013	0.0851	0.664	0.0306	0.051	0.0192	0.013
MIKLE	7.0167	1.179	7.0821	1.071	6.8628	2.128	6.7197	1.152	6.4132	1.170
NN	16.9980	0.653	31.0240	7.480	56.8480	69.527	74.6220	85.744	25.5835	3.225
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0118	0.014	0.0074	0.012	0.3939	7.189	0.0125	0.031	0.0065	0.011
IVARIANZ	0.0017	0.003	0.0023	0.003	0.0022	0.029	0.0011	0.001	0.0003	0.000

Tabelle C.20: Maßzahlen für Typ IV, $n = 100$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0070	0.007	0.0063	0.006	0.0522	0.502	0.0225	0.129	0.0030	0.004
UAE	0.0227	0.010	0.0193	0.010	0.0690	0.245	0.0459	0.116	0.0139	0.009
MIKLE	7.1226	0.808	7.2986	0.770	7.2026	0.952	7.2086	0.913	6.4623	0.764
NN	29.7940	0.770	53.9040	17.859	52.0300	76.906	69.0920	86.962	22.2964	1.927
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0060	0.005	0.0046	0.005	0.0515	0.500	0.0218	0.129	0.0029	0.004
IVARIANZ	0.0017	0.002	0.0023	0.003	0.0014	0.003	0.0013	0.002	0.0002	0.000

Tabelle C.21: Maßzahlen für Typ IV, $n = 300$ und unzensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0041	0.003	0.0058	0.004	0.0206	0.033	0.0155	0.031	0.0013	0.001
UAE	0.0176	0.007	0.0172	0.006	0.0476	0.043	0.0455	0.054	0.0095	0.005
MIKLE	7.1767	0.525	7.7948	0.528	6.9277	0.711	7.1430	0.663	6.7530	0.479
NN	72.1840	1.139	141.370	50.060	40.8960	71.737	75.6296	147.974	17.8956	0.961
CENS80	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000	0.0000	0.000
IBIAS2	0.0026	0.002	0.0024	0.002	0.0203	0.033	0.0149	0.030	0.0013	0.001
IVARIANZ	0.0019	0.002	0.0037	0.003	0.0006	0.001	0.0012	0.001	0.0001	0.000

Tabelle C.22: Maßzahlen für Typ IV, $n = 50$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0561	0.116	0.0750	0.988	0.0933	1.158	0.0267	0.090	0.0208	0.018
UAE	0.0777	0.100	0.0603	0.232	0.0826	0.345	0.0475	0.085	0.0409	0.017
MIKLE	6.2414	1.612	6.0628	1.581	5.6057	1.497	5.5819	1.389	6.0184	1.546
NN	14.6400	1.179	28.9000	8.495	123.728	123.132	137.438	83.775	16.7721	2.097
CENS80	0.3937	0.070	0.3937	0.070	0.4156	0.076	0.4156	0.076	0.3937	0.070
IBIAS2	0.0534	0.117	0.0699	0.992	0.0829	1.152	0.0163	0.089	0.0178	0.018
IVARIANZ	0.0062	0.006	0.0078	0.007	0.0119	0.012	0.0115	0.010	0.0059	0.006

Tabelle C.23: Maßzahlen für Typ IV, $n = 100$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0274	0.040	0.0195	0.037	0.0205	0.052	0.0141	0.024	0.0160	0.020
UAE	0.0504	0.033	0.0378	0.028	0.0447	0.081	0.0345	0.031	0.0373	0.018
MIKLE	6.5586	1.433	6.5366	1.379	6.0070	1.320	5.9868	1.222	6.3714	1.419
NN	25.7360	1.362	54.1560	17.713	185.732	164.365	217.242	137.063	14.6137	1.355
CENS80	0.3976	0.050	0.4200	0.053	0.4200	0.053	0.4200	0.053	0.3976	0.050
IBIAS2	0.0261	0.040	0.0158	0.038	0.0138	0.052	0.0075	0.023	0.0150	0.021
IVARIANZ	0.0031	0.003	0.0054	0.005	0.0075	0.007	0.0074	0.006	0.0026	0.003

Tabelle C.24: Maßzahlen für Typ IV, $n = 300$ und zensiert

	Smooth RoT		mL		UAE(sRoT)		UAE(mL)		fix	
Kriterium	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std	Mittel	Std
MISE	0.0120	0.016	0.0079	0.008	0.0082	0.014	0.0066	0.006	0.0086	0.013
UAE	0.0336	0.020	0.0254	0.013	0.0270	0.020	0.0248	0.012	0.0297	0.019
MIKLE	7.0043	0.988	6.9823	0.921	6.5259	0.903	6.5255	0.893	6.7774	0.973
NN	62.5680	1.833	136.320	47.520	387.008	359.300	396.296	280.257	11.7682	0.573
CENS80	0.4204	0.032	0.4213	0.034	0.3990	0.030	0.4213	0.034	0.4204	0.032
IBIAS2	0.0118	0.015	0.0068	0.009	0.0057	0.016	0.0037	0.006	0.0086	0.013
IVARIANZ	0.0006	0.001	0.0016	0.002	0.0030	0.003	0.0032	0.003	0.0001	0.000

Anhang D

Quelltext für eine Hazardratenschätzung

```
/*************************************************************************/
/*  Programm auf SAS/IML-Basis zur Berechnung des variablen Kernschaetzers    */
/*  der Hazardrate im zensierten (und unzensierten) Modell mit Naechster-     */
/*  Nachbarn-definition durch den Kaplan-Meier-Schaetzer                      */
/*  Ausserdem Implementierung des fixen Kernschaetzers zum Vergleich          */
/*  Fetiggestellt: 22.9.1998                                                  */
/*************************************************************************/



libname a 'u:\promo\sasprog\simdata';



proc iml symsize=1000;              /* Start der IML-Prozedur               */
```

```
/*****************************************************************************/
/*  Modul zur Berechung des Kalpan-Meier-Schaetzers der Ueberlebensszeit im   */
/*  zensierten Szenario (Formel nach E. Lee (1992), p 191)                    */
/*      Eingabeparameter:  xx    : n*2-Matrix mit nach der Zeit geordneten    */
/*                                 Zeiten und zugehoerigen Zensierungs-       */
/*                                 indikatoren                                */
/*      Ausgabeparameter:  s     : (n+1)*2-Matrix mit Schaetzungen an den, in */
/*                                  der 2. Spalte abgelegten Zeiten, wobei die*/
/*                                  erste Zeile einen Randwert enthaelt       */
/*      Nebenwirkungen  : keine                                               */
/*      validiert       : 10.9.98 mit Bsp. Datensatz p 193 (Lee)              */
/*                        a={1 1, 3 1, 5 1, 8 1, 10 0, 15 1, 18 1,            */
/*                             19 1, 22 1, 25 0}; run s_n(a, b); print b;     */
/*****************************************************************************/


 start s_n(xx, s);
  n=nrow(xx);                          /* n=Anzahl der Beobachtungen       */

  uncenobs=ncol(loc(xx[,2]));          /* uncenobs=Anzahl der Unzensierten */

  xx=t(xx);                            /* 2xn-Matrix                       */

s=j(2,n+1,0);
s[1,1]=1;                              /* Der Schaetzer wird sukzessiv be- */
s[2,1]=0;                              /* rechnet und fuer die beobachteten */
                                       /* Zeiten gespeichert               */

do i=1 to n-1; *print i;
 s[1,i+1]=s[1,i]*((n-i)/(n-i+1))**xx[2,i];
 s[2,i+1]=xx[1,i];
end;
if xx[2,n]=0 then s[1,n+1]=s[1,n];
```

```
else s[1,n+1]=0;
s[2,n+1]=xx[1,n];

do j=n+1 to 3 by -1;                        /* anderenfalls wuerden in Bindungen-*/
 if s[2,j]=s[2,j-1] then s[1,j-1]=s[1,j]; /* gruppen verschiedene Schaetzungen */
end;                                        /* fuer gleiche Zeiten angenommen    */
s=t(s);
xx=t(xx);
finish;
```

```
/*****************************************************************************/
/*            Modul zur Berechung des Identitaetsprozesses J_n(t)=t          */
/*      Eingabeparameter:  xx    : n*2-Matrix mit nach der Zeit geordneten    */
/*                                 Zeiten und zugehoerigen Zensierungs-      */
/*                                 indikatoren                               */
/*      Ausgabeparameter:  j     : (n+1)*2-Matrix mit id.Schaetzungen an den,in*/
/*                                 der 2. Spalte abgelegten Zeiten, wobei die */
/*                                 erste Zeile einen Randwert enthaelt       */
/*      Nebenwirkungen  : keine                                              */
/*      validiert       : 11.9.98 mit Bsp. Datensatz p 193 (Lee)             */
/*                        a={1 1, 3 1, 5 1, 8 1, 10 0, 15 1, 18 1,           */
/*                             19 1, 22 1, 25 0}; run j_n(a, b); print b;    */
/*****************************************************************************/


 start j_n(xx, j);
  n=nrow(xx);

j=j(n,2,0);                          /* Berechnung des Identitaetsprozesses */
j[,1]=xx[,1];
j[,2]=xx[,1];
dummy={0 0};
j=dummy//j;
finish;
```

```
/*************************************************************************/
/*  Modul zur Berechung des Nelson-Aalen-Schaetzers der kumulativen Hazardrate */
/*                              im zensierten Szenario                   */
/*      Eingabeparameter:  xx    : n*2-Matrix mit nach der Zeit geordneten     */
/*                                 Zeiten und zugehoerigen Zensierungs-       */
/*                                 indikatoren                               */
/*      Ausgabeparameter:  h     : (n+1)*2-Matrix mit Schaetzungen an den, in  */
/*                                 der 2. Spalte abgelegten Zeiten, wobei die  */
/*                                 erste Zeile einen Randwert enthaelt         */
/*      Nebenwirkungen  : keine                                              */
/*      validiert       : 11.9.98 mit Bsp. Datensatz (Lawless 73/81 anders!)   */
/*                        g={1 1,2 1,3 1,4 0,5 1,6 1,6 0,8 1,9 1,10 1};       */
/*                        run h_n(g, b); print b;                            */
/*************************************************************************/


 start h_n(xx, h);
  n=nrow(xx);      print n;

 b=n - (1:n) +1;
c=t(xx[,2])/b;
mat=j(n,1,1);
do nu=1 to n-1;
mat=mat||(j(nu,1,0)//j(n-nu,1,1));
end;
d=mat*t(c);
h=(0//d)||(0//xx[,1]);

do j=n+1 to 3 by -1;                                          /* s.o.    */
 if h[j,2]=h[j-1,2] then h[j-1,1]=h[j,1];
end;
finish;
```

```
/*****************************************************************************/
/*  Modul zur Berechung der naechste-Nachbarn-Distanzen bzgl. des Prozesses   */
/*                                tildepsi                                    */
/*      Eingabeparameter:  xx    : n*2-Matrix mit nach der Zeit geordneten     */
/*                                 Zeiten und zugehoerigen Zensierungs-       */
/*                                 indikatoren                                */
/*                      tildepsi : (n+1)*2-Matrix mit Schaetzungen an den, in  */
/*                                 der 2. Spalte abgelegten Zeiten, wobei die */
/*                                 erste Zeile einen Randwert enthaelt        */
/*                        k_n    : 1*1 Anzahl naechster-Nachbarn oder fixe BB */
/*                       distart : spezifiziert ob NN-Abstand oder           */
/*                                 fixer Abstand                              */
/*      Ausgabeparameter: r_n_x_i: n*1-Vektor mit NN-Abstaenden an den        */
/*                                 gordneten Zeiten (1. Spalte der xx-Matrix) */
/*      Nebenwirkungen  : keine                                               */
/*      validiert       : 18.9.98 fuer tildepsi=s_n und ohne Bindungen        */
/*                        Validiert durch Vergleich mit nngefk2.sas           */
/*                      a={1 1,3 1,5 1,8 1,10 0,15 1,18 1,19 1,22 1,25 0};    */
/*                        run s_n(a, b1); print b1;                           */
/*                        run h_n(DATA_0, b); print b;                        */
/*                        run r_n(a, b1, 4, 3, vecci1); print vecci1;         */
/*                        run r_n(DATA_0, b, 4, 3, vecci); print vecci;       */
/* VORSICHT: Es wird die NN-Distanz nach Gefeller/Dette bestimmt (r statt r/2) */
/*           und inf statt sup und k_n-1/n statt k_n/n                         */
/*****************************************************************************/


start R_n(xx, tildepsi, k_n, distart, r_n_x_i);
n=nrow(xx);
if distart = 1 then r_n_x_i=k_n*j(n, 1, 1);
else do;
 eps=0.000001;                                  /* Kommentierung siehe Gefeller,*/
 xx=t(xx);                                      /* Pflueger, Bregenzer (1996)   */
```

```
tildepsi=t(tildepsi);
uncenobs=ncol(loc(xx[2,]));
maxind=min(uncenobs+1,k_n+3);
r_n_x_i=j(1,n,1);
intmasse=j(2,maxind,0);



null={0,1};
xx=null||xx;
unzens=xx[1,loc(xx[2,])];
spind = loc(xx[2,]);

do j=1 to n;
  abstand = abs(unzens - xx[1,j]);
  links=xx[1,spind[loc(rank(abstand)=1)]];
  rechts=links;
  intmasse[1,1]=abstand[loc(rank(abstand)=1)];
  intmasse[2,1]=0;



 do l=2 to maxind;               /* Bemerke, dass loc() bei Bindungen Vektoren */
   if xx[1,spind[loc(rank(abstand)=l)]] < links   /* liefen kann bei unzens=.. */
   then do; links=xx[1,spind[loc(rank(abstand)=l)]];
            intmasse[1,l]=abstand[loc(rank(abstand)=l)];
             intmasse[2,l]=abs(tildepsi[1,spind[loc(unzens=links)[1]]]
             - tildepsi[1,spind[loc(unzens=rechts)[1]]]);
        end;
   else do; rechts=xx[1,spind[loc(rank(abstand)=l)]];
            intmasse[1,l]=abstand[loc(rank(abstand)=l)];
            if links > 0
            then intmasse[2,l]=abs(tildepsi[1,spind[loc(unzens=links)[1]-1]]
            - tildepsi[1,spind[loc(unzens=rechts)[1]-1]]);
            else intmasse[2,l]=1 - tildepsi[1,spind[loc(unzens=rechts)[1]-1]];
```

```
        end;

 end;

*print intmasse;

 kindex=maxind;
 do i=1 to maxind while (kindex=maxind);
   if intmasse[2,i] - (k_n-1)/n > eps
   then kindex=i-1;
 end;

 r_n_x_i[j]=intmasse[1,kindex];

end;

hilf=j(2,n,0);
do i=1 to n;
  hilf[,i]=xx[,i+1];
end;
xx=t(hilf);

tildepsi=t(tildepsi);
r_n_x_i=t(r_n_x_i);
end;
finish;
```

```
/*****************************************************************************/
/*   Modul zur Berechnung von Inkrementen von einer Treppenfunktion          */
/*      Eingabeparameter:  psi    : (n+1)*2-Matrix mit Schaetzungen an den, in der*/
/*                                  2. Spalte abgelegten Zeiten, wobei die erste  */
/*                                    Zeile einen Randwert enthaelt               */
/*      Ausgabeparameter:deltavec: n-Vektor mit nach der Zeit geordneten         */
/*                                  Inkrementen der psi-Funktion                */
/*      Nebenwirkungen  : keine                                                  */
/*      validiert       : 18.9.98                                                */
/*                        help={0 2, 0.3 3, 0.4 2, 0.5 4, 0.8 33, 1.3 99};       */
/*                        run deltapsi(help, hapa); print hapa;                  */
/*****************************************************************************/

start deltapsi(psi, deltavec);

n=nrow(psi)-1;

dummy1={1 -1}||j(1, n, 0);
kontrast=shape(dummy1, n, n+1);
deltavec=abs(kontrast*psi[,1]);
finish;
```

```
/*****************************************************************************/
/*     Modul zur Berechnung von Bi-square-Kerngewichten zur Kumulation fuer die  */
/*                              Schaetzung an der Stelle x                       */
/*      Eingabeparameter:  xx    : n*2-Matrix mit nach der Zeit geordneten       */
/*                                 Zeiten und zugehoerigen Zensierungs-          */
/*                                 indikatoren                                   */
/*                          x     : 1*1-Stelle der Schaetzung                    */
/*                         bb     : n-Vektor Bandbreiten an den Auswertungs-     */
/*                                 stellen                                       */
/*      Ausgabeparameter: kernx  : n-Vektor mit nach der Zeit geordneten Gewich- */
/*                                 ten der Schaetzung (bzgl der Bandbreite, bb)  */
/*      Nebenwirkungen  : keine                                                  */
/*      validiert       : 18.9.98                                                */
/*                        stelle=5; abstaend={1, 2, 3, 4};                       */
/*                        daten={3 3, 4 6, 1 5, 6 3};                            */
/*                        run kern1(daten, stelle, abstaend, hallo); print hallo;*/
/*****************************************************************************/

start kern1(xx, x, bb, kernx);

dummy1=(x - xx[,1])/bb;
dummy2=(dummy1#(abs(dummy1)<1))+(abs(dummy1)>1);
kernx=0.9375#(1-dummy2##2)#(1-dummy2##2);

finish;
```

```
/*****************************************************************************/
/*   Hauptprogramm zur Berechnung des variablen Kernschaetzers fuer die Hazard-  */
/*    funktion bei Benutzung der Naechste-Nachbarn-Distanz von Gefeller          */
/*                          oder fixer Bandbreite                                */
/*      Eingabeparameter:  Daten : n*2-Matrix mit nach der Zeit geordneten       */
/*                                 Zeiten                                        */
/*                                 und zugehoerigen Zensierungsindikatoren       */
/*                       bandbrei: Bandbreite bei art=1 und                      */
/*                                 Anzahl naechster Nachbarn bei art=2           */
/*                          art  : Bestimmt die Art der Bandbreite               */
/*                                 NN=2 oder fix=1                               */
/*                        stelle : Stelle der Schaetzung (x)                     */
/*      Ausgabeparameter:h_dach_x: Funktions(Schaetz)wert an der Stelle der      */
/*                                 Schaetzng stelle (h(x))                       */
/*      Nebenwirkungen  : Ausgabe des Breitenvektor im Output                    */
/*      validiert       : 21.9.98                                                */
/*                        breiten={1, 2, 3}; kernvec={0.1, 0.2, 0};              */
/*                        delta_H={0.3, 0.2, 0};                                 */
/*****************************************************************************/


start hazard(Daten, bandbrei, art, stelle, h_dach_x);
run s_n(Daten, surviv);  *print surviv;
run R_n(Daten, surviv, bandbrei, art, breiten); print breiten;

run h_n(Daten, nelson);
run deltapsi(nelson, delta_H); *print delta_H;

run kern1(Daten, stelle, breiten, kernvec); *print kernvec;
h_dach_x=t((1/breiten)#kernvec) * delta_H;

finish;
```

```
/*****************************************************************************/
/*   Programm zur Berechnung des variablen Kernschaetzers fuer die Dichte    */
/*    funktion bei Benutzung der Naechste-Nachbarn-Distanz von Gefeller       */
/*                          oder fixer Bandbreite                            */
/*      Eingabeparameter:  Daten : n*2-Matrix mit nach der Zeit geordneten    */
/*                                 Zeiten                                    */
/*                                 und zugehoerigen Zensierungsindikatoren   */
/*                       bandbrei: Bandbreite bei art=1 und                  */
/*                                 Anzahl naechster Nachbarn bei art=2       */
/*                          art  : Bestimmt die Art der Bandbreite           */
/*                                 NN=2 oder fix=1                           */
/*                        stelle : Stelle der Schaetzung (x)                 */
/*      Ausgabeparameter:f_dach_x: Funktions(Schaetz)wert an der Stelle der   */
/*                                 Schaetzng stelle (h(x))                   */
/*      Nebenwirkungen  : Ausgabe des Breitenvektor im Output                */
/*      validiert       : 6.7.1999 per Analogieschluss zu hazard und Plausibili- */
/*                                 taetskontrolle (Delta_F-richtig,          */
/*                                 S_n abfallend)                            */
/*                        DATA_0={0.1 1, 1 1, 2 0, 3 1, 4 0, 106 1, 109 1,   */
/*                                 109.3 1, 111 1};                          */
/*                        run dichte(DATA_0, 4, 1, 4, fdach); print fdach;   */
/*****************************************************************************/
start dichte(Daten, bandbrei, art, stelle, f_dach_x);

run s_n(Daten, surviv);  *print surviv;
run R_n(Daten, surviv, bandbrei, art, breiten); print breiten;
run deltapsi(surviv, delta_F); *print delta_F;
run kern1(Daten, stelle, breiten, kernvec); *print kernvec;

f_dach_x=t((1/breiten)#kernvec) * delta_F;

finish;
```

```
/*****************************************************************************/
/*****************************************************************************/
/*            BANDBREITENWAHLEN                                              */
/*****************************************************************************/
/*****************************************************************************/

/*****************************************************************************/
/* Berechung der cross-validation idealen Bandbreite                         */
/* *** Kullback-Leibler-loss = modified likelihood                           */
/* Dokumentation siehe Seuge94-paper                                         */
/*       Eingabeparameter: AO      : n*2 Matrix der sortierten Daten         */
/*                                                                           */
/*       Ausgabeparameter: theta   : 1*1 Matrix des optimalen Anzahl naechster */
/*                                   Nachbarn                                */
/*                                                                           */
/*       Nebenwirkungen  : Ausgabe von theta in Outputfenster                */
/*       validiert       : 1994                                              */
/*****************************************************************************/

 start maxtheta(AO, theta);                    /* Start der Enumeration in theta */
  n=nrow(AO);
  uncenobs=ncol(loc(AO[,2]));
  theta=0;

AO=t(AO);

s=j(2,n+1,0);

s[1,1]=1;
s[2,1]=0;

do i=1 to n-1; *print i;
 s[1,i+1]=s[1,i]*((n-i)/(n-i+1))**AO[2,i];
```

```
 s[2,i+1]=A0[1,i];
end;
if A0[2,n]=0 then s[1,n+1]=s[1,n];
else s[1,n+1]=0;
s[2,n+1]=A0[1,n];

do j=n+1 to 3 by -1;                                          /* s.o.    */
 if s[2,j]=s[2,j-1] then s[1,j-1]=s[1,j];
end;


A0=t(A0);                                                /* nx2-Matrix      */

  a=-100000000;                          /* Schleife der Enumeration in theta   */
  do t=3 to n;     print t;       /* um die maximale modified likelihood (moli)  */
  c=moli(n,t,A0,s,uncenobs);   print c;                          /* zu finden  */
   if c>a
    then do;
     theta=t;
     a=c;
    end;
  end;
  print "Die Modified Likelihoodfunktion ist maximal fuer ";
  print theta;
  print "Das ist die Anzahl naechster Nachbarn, die bei einer";
  print "nicht parametrischen Kernschaetzung der Hazardfunktion";
  print "mit nn-Bandbreiten und biquadratischem Kern genutzt werden soll";
 finish;


 start moli(n,theta,A0,s,uncenobs);      /* Berechung des moli Wertes fuer theta */
  run nndef(nndist,nndist_0,n,A0,s,theta,uncenobs,0.000001);/* Subroutinenaufruf */
  run kerneloo(hazard_i,s_int,n,A0,nndist,nndist_0,theta);  /* fuer Hazardrate-, */
```

```
 m=0;                                                    /* Ueberlebenszeit-  */
 do i=1 to n;                                            /* schaetzung        */

  if A0[i,2]=1                          /* Berechnung der log-Version der      */
   then do;                             /* modified Likelihoodfunktion         */
    m=m+log(hazard_i[i]#s_int[i]);
   end;
   else do;
    m=m+log(s_int[i]);                  /* s_int ist die Integral-Version der  */
   end;                                 /* Ueberlebenszeitfunktion, die fuer   */
 end;                                   /* moli benoetigt wird                 */
 return(m);
finish;
```

```
/*****************************************************************************/
/* Modul fuer nndist (Naechste Nachbarn Abstand);                            */
/* erzeugt einen Vektor, der die Abstaende fuer jede Beobachtung enthaelt    */
/* (fuer ein gegebenes theta)                                                */
/*****************************************************************************/
start nndef(nndist,nndist_0,n,A0,s,theta,uncenobs,eps);
A0=t(A0); maxind=min(uncenobs+1,theta+3); nndist=j(1,n,1);
int_valu=j(2,maxind,0); nndist_0=0; help_01={0,1}; A0=help_01||A0;
uncens=A0[1,loc(A0[2,])]; jump_ind = loc(A0[2,]);
j=0;
do j=2 to n+1;
 distance = abs(uncens - A0[1,j]);        /* Erzeugung eines Vektors, der die    */
                                          /* Abstaende zwischen jeder Beobachtung*/
                                          /* und X_j enthaelt                    */

 left=A0[1,jump_ind[loc(rank(distance)=1)]]; /* Nehme das naechste X_k und       */
 right=left;                              /* definiere right=left=X_k            */

 int_valu[1,1]=distance[loc(rank(distance)=1)]; /* Speichern des Abstaende und   */
 int_valu[2,1]=0;                         /* Ueberlebenszeitdifferenzen          */
 do l=2 to maxind;
   if A0[1,jump_ind[loc(rank(distance)=l)]] < left /* fuer die l-naechste Beo.   */
   then do; left=A0[1,jump_ind[loc(rank(distance)=l)]]; /* fahre fort wie oben   */
            int_valu[1,l]=distance[loc(rank(distance)=l)]; /* fuer die naechste  */
            int_valu[2,l]=s[1,jump_ind[min(loc(uncens=left)[1])]]   /*     X_k   */
                          - s[1,jump_ind[min(loc(uncens=right)[1])]];
        end;
   else do; right=A0[1,jump_ind[loc(rank(distance)=l)]];
            int_valu[1,l]=distance[loc(rank(distance)=l)];
            if left > 0
            then int_valu[2,l]=s[1,jump_ind[min(loc(uncens=left)[1])-1]]
                               - s[1,jump_ind[min(loc(uncens=right)[1])-1]];
            else int_valu[2,l]=1 - s[1,jump_ind[min(loc(uncens=right)[1])-1]];
```

```
        end;


end;
kindex=maxind;
do i=1 to maxind while (kindex=maxind);   /* geht durch die i-naechste Beo bis   */
 if int_valu[2,i] - (theta-1)/n > eps     /* die Kumulativen Ueberlebenszeitfkt. */
 then kindex=i-1;                         /* Differenzen das Limit erreichen und */
end;                                      /* nehme die letzte Beo.               */
                                          /* (mit entsprechender Bandbreite) als */
                                          /* theta-naechste Beo.                 */
nndist[j-1]=int_valu[1,kindex];
end;
nndist_0=nndist[1]+A0[1,2] ;              /* Berechnung des NN-Abstandes fuer 0  */
help_00=j(2,n,0);                         /* Wiederherstellung der Matrix A0     */
do i=1 to n;
 help_00[,i]=A0[,i+1];
end;
A0=t(help_00);
finish;
```

```
/*****************************************************************************/
/* Modul zur leave-one-out-Schaetzung der                                    */
/* Hazardrate und der Ueberlebenszeit (Integral-Version)                     */
/* fuer jede Beoachtung (fuer gegebens theta)                                */
/* gespeichert in den Vektoren hazard_i und s_int                            */
/*****************************************************************************/
start kerneloo(hazard_i,s_int,n,A0,nndist,nndist_0,theta);
hazard_i=j(1,n,0);  s_int=j(1,n,1);  hazard_0=0;
do i=1 to n;                              /* in der i-ten Schleife werden    */
                                          /* hazard_i[i] und s_int[i]  berechnet */
                                          /* (die i-te Beo. wird ausgelassen) */

hazard_0=0;                               /* Eine Schaetzung fuer die Hazazrd- */
                                          /* rate zur Zeit 0 wird benoetigt um */
    dummy=0;                              /* s_int[i] zu berechnen           */
    do j=1 to n while(dummy^=1);
       if A0[j,2]=1 then do;
        if j<i then factor=1/(n-j);
        if j>i then factor=1/(n-j+1);     /* siehe naechster Schritt         */
        if j=i then factor=0;
        argu=(-A0[j,1])/nndist_0;
        if abs(argu)<1
         then hazard_0=hazard_0+factor*0.9375*(1-argu*argu)*(1-argu*argu);
        else dummy=1;
       end; /* if A(j) */
    end; /* do j */
hazard_0=hazard_0/nndist_0;
hazardoo=j(1,n,0);                        /* Fuer die Berechnung von s_int[i] */
                                          /* wird die Hazardratenschaetzung bis */
                                          /* zur i-ten Beo. benoetigt        */
  do k=1 to i;                            /* (X_i ausgenommen).              */
   if nndist[k]=0 then do;                /* Fuer kleine theta und Gruppen von */
    r=min(loc(A0[,1]=A0[k,1]));           /* Bindungen kann nndist=0 auftreten. */
```

```
 maxind=max(loc(AO[,1]=AO[k,1]));      /* Als alternative Berechung von       */
                                       /* hazardoo[k] wird punktweise Be-     */
                                       /* rechnung von Cox  benutzt           */
 dummy=0;
 do nu=maxind to r by -1;
  if AO[nu,2]=0 then dummy=dummy+1;
 end; /* do nu */
 l=maxind-dummy-r+1;
 hazardoo[k]=l/(n-r+1);
end;  /* if i */
else do;                               /* Falls nndist>0 wird die Berechnung  */
                                       /* auf Effizienzgruenden in zwei       */
                                       /* Schritten vollzogen: X_k aufwaerts  */
                                       /* bis der Taeger des biquadrtischen   */
                                       /* Kerns erschoepft ist                */
 dummy=0;                              /* (if abs(argu)>=1) und abwaerts      */
 do j=k to n while(dummy^=1);          /* von X_k in der selben Art           */
    if AO[j,2]=1 then do;
     if j<i then factor=1/(n-j);
     if j>i then factor=1/(n-j+1);
     if j=i then factor=0;
     argu=(AO[k,1]-AO[j,1])/nndist[k];
     if abs(argu)<1
      then hazardoo[k]=hazardoo[k]+factor*0.9375*(1-argu*argu)*(1-argu*argu);
     else dummy=1;
    end; /* if A(j) */
 end; /* do j */
if k>1 then do;
 dummy=0;
 do j=k-1 to 1 by -1 while(dummy^=1);    /* abwaerts */
    if AO[j,2]=1 then do;
     factor=1/(n-j);
     argu=(AO[k,1]-AO[j,1])/nndist[k];
```

```
        if abs(argu)<1
         then hazardoo[k]=hazardoo[k]+factor*0.9375*(1-argu*argu)*(1-argu*argu);
        else dummy=1;
      end; /* if A(j) */
    end; /* do j */
   end; /* if k */
   hazardoo[k]=hazardoo[k]/nndist[k];
  end;  /* else do */
  if k=1 then s_int[i]=exp(-A0[k,1]*
           (hazardoo[1]+hazard_0)/2);     /* polygoniale Approximation des       */
                                          /* Integrals in s_int[i],              */
  else s_int[i]=s_int[i]                  /* sukzessive Berechung                */
    *exp(-(A0[k,1]-A0[k-1,1])
    *(hazardoo[k]+hazardoo[k-1])/2);
 end;  /* do k */
 if hazardoo[i]=0 then hazard_i[i]=10##(-100);      /* Um den Fall log(0) zu     */
 else hazard_i[i]=hazardoo[i];                      /* vermeiden:  hazard_i[i]   */
                                                    /* ist hazardoo[i]           */
end;  /* do i */
finish;
```

```
/*****************************************************************************/
/* Modul zur Berechnung des optimalen Bandbreitenparameters bzgl. des glm.   */
/* Fehlers mit Berueecksichtigung der Daten, siehe Kapitel "Bandbreitenwahl" */
/*       Eingabeparameter: A0    : n*2 Matrix der nach Zeit geordneten Zeiten */
/*                                 und zugehoerigen Zensierungsindikatoren   */
/*                                                                           */
/*       Ausgabeparameter: pn2     : 1*1 Matrix des optimalen                */
/*                                   Banbreitenparameters                    */
/*       Nebenwirkungen  : keine                                             */
/*       validiert       : bei gegeben D1,2,3,L_K,L_tilpsi, M,tilm,m, 30.6. mit T*/
/*                         Konstanten, bzw. x-Werte validiert (7.6.99) mit Daten */
/*                         DATA_0={0.1 1, 1 1, 2 0, 3 1, 4 1, 106 0, 109 1,  */
/*                                 109.3 1, 111 1};                          */
/*                                                                           */
/*****************************************************************************/

start pn_data(A0, pn2);
nobs=nrow(A0);

supK=0.652173913;     /* Konstanten des Biquadratischen Kerns (Traeger [-.5,.5]) */
L_K=2.008174849;
VvonK=0.326086957;


uncens=A0[loc(A0[,2]), 1];                          /* Vektor unzensierter Daten */
nuncens=nrow(uncens);
contrast = (I(nuncens-1)||j(nuncens-1,1,0)) - (j(nuncens-1,1,0)||I(nuncens-1)) ;
delta =  ((I(nuncens-1)||j(nuncens-1,1,0))*uncens)||(abs(contrast * uncens));  ;

xklM=delta[loc(rank(delta[,2])=1)[1],1];        /* fuer Hazardraten-M-Schaetzung */
xgrM=delta[loc(rank(delta[,2])=nuncens-1)[1],1];
                                    /* Wert am linken Rand der Intervalle waehlen */
tildelta =  ((I(nuncens-1)||j(nuncens-1,1,0))*uncens+ 0.5#(abs(contrast * uncens)))
```

```
||(abs(contrast * uncens));


xkltilM=tildelta[loc(rank(delta[,2])=1)[1],1];      /* fuer Dichten-M-Schaetzung */
xgrtilM=tildelta[loc(rank(delta[,2])=nuncens-1)[1],1];
                                    /* Wert am linken Rand der Intervalle waehlen */
                    /* liefert naechste Nachbarnanzahl als Start fuer Konstanten */
run maxtheta(A0, start);
print start;
run hazard(A0, start, 2, xklm, klm);
run hazard(A0, start, 2, xgrm, grm);

run dichte(A0, start, 2, xkltilm, tilklm);
run dichte(A0, start, 2, xgrtilm, tilgrm);

contras1 = ((I(nuncens-2)||j(nuncens-2,1,0))*(abs(contrast * uncens)))/
           ((j(nuncens-2,1,0)||I(nuncens-2))*(abs(contrast * uncens)));

 /* contras1 hat bei Bindungen Definitionsluecken, da durch 0 (Intervalllaengen) */
                                                          /*  geteilt wird!  */
deltaquo =  ((j(nuncens-2,1,0)||I(nuncens-2)||j(nuncens-2,1,0))*uncens)||contras1;

index=loc(rank(abs(log(deltaquo[ ,2])))=nuncens -2)[1]; *print index;
xL_tpsi=deltaquo[index, 1];
run dichte(A0, start, 2, uncens[index, 1] , fdachl);
run dichte(A0, start, 2, uncens[index +2, 1] , fdachr);

run hazard(A0, start, 2, uncens[index, 1] , hdachl);
run hazard(A0, start, 2, uncens[index +2, 1] , hdachr);

L_tilpsi = abs((fdachl - fdachr) / (uncens[index , 1] - uncens[index + 2, 1]));

L_psi= abs((hdachl - hdachr) / (uncens[index , 1] - uncens[index + 2, 1]));
```

```
uncens=AO[loc(AO[,2]), 1];  /* Vektor unzensierter Daten */
nuncens=nrow(uncens);
cens=AO[loc((AO[,2]-1)), 1];  /* Vektor zensierter Daten */
ncens=nrow(cens);

B=(uncens[nuncens] + uncens[nuncens - 1] + cens[ncens] + cens[ncens - 1])/4;
print B;

              /* stabilisierte sup(Traeger)-Schaetzung */



      /* Berechnung der Verteilungfunktion der Zensierungen (nach Schaefer-Diss) */

AOcens=AO;
AOcens[,2]=abs(AO[,2] - 1);                    /* Zensierungen werden umgedreht */
run s_n(AOcens, kaplan);

GvonB= 1 - kaplan[nrow(loc((B - kaplan[,1]) > 0)), 2];   print GvonB;

/*************************************************/
/* Approximativer Ersatz                         */
/*************************************************/

*tildeD= 9 # (1 - GvonB )##(-0.5);
tildeD=9;

 /* Verteilungfunktion der Obs (B) */

HvonB= nrow(loc((B - AO[,1]) >= 0))/nobs;   print HvonB;
```

```
*D=  9 # (1 - HvonB )##(-0.5);

D=9;

/*****************************************************/

D1 = 2# tildeD # (grm # tilgrm##2)/(tilklm##2) # (supK/tilklm + L_K #
(tilgrm/tilklm));
print D1;
D2 =  (D # tilgrm # sqrt(grm) # VvonK)/(sqrt(tilklm));
print D2;
D3 = (2# tilgrm##3 # grm # L_K # L_tilpsi)/(tilklm##5)
+ (2 # supK # L_tilpsi # tilgrm##2 # grm)/(tilklm##4)
+ (L_psi)/(tilklm);
print D3;

pn2= ((D1 + D2)/(2#D3))##(2/3)
      # (log(nobs)/nobs)##(1/3);


finish;
```

```
/*****************************************************************************/
/*    Rule of Thumb Bandbreite (oder Anzahl naechster Nachbarn)              */
/*       in Adaption der Normal Scale Rule (siehe Wand and Jones (1995))     */
/*       Eingabeparameter: A0    : n*2 Matrix der nach Zeit geordneten Zeiten */
/*                                 und zugehoerigen Zensierungsindikatoren   */
/*       Ausgabeparameter: pn      : 1*1 Matrix des optimalen                */
/*                                    Banbreitenparameters                   */
/*       Nebenwirkungen  : keine                                             */
/*       validiert       : 17.9.99                                           */
/*****************************************************************************/

start smootRoT(A0, pn3);
nobs=nrow(A0);

R_K=5/7;                /* fuer bi-square Kern, siehe Jones und Wand S. 172,176 */
mu2_K=1/7;

varianz=(1/(nobs-1))#j(1,nobs,1)*(A0[,1] - (j(nobs,nobs,1/nobs)*A0[,1]))##2 ;
streuung=sqrt(varianz);
b_RoT=((8 # sqrt(3.1415) # R_K)/(3 # mu2_K##2 # nobs))##(0.2) # streuung;
print b_RoT;
run s_n(A0, kaplan);
kaplan=(j(nobs,1,0)||I(nobs))*kaplan;

beta_hat=- (t(kaplan[,2] - (j(nobs,nobs,1/nobs)*kaplan[,2]))*
          (kaplan[,1] - (j(nobs,nobs,1/nobs)*kaplan[,1])))/
(j(1,nobs,1)*(kaplan[,2] - (j(nobs,nobs,1/nobs)*kaplan[,2]))##2);

*beta_hat= - Regressionskoeffizient der Kaplanmeier Kurve (X_i, S_n(X_i));
                                                    /* Gemaess Krengel Seite 164 */
pn3=beta_hat # b_RoT;
finish;
```

```
/*****************************************************************************/
/*****************************************************************************/
/*****************************************************************************/
/***         Erzeugung der Simulationssample                                ***/
/***                                                                        ***/
/***                                                                        ***/
/*****************************************************************************/
/*****************************************************************************/
/*****************************************************************************/

/*****************************************************************************/
/*                     Das Modul "expoweib"                                  */
/*  Erzeugung von nobs expo-weibull verteilten Zufalls Variablen mit          */
/*        censrate Zensierungen                                               */
/*    Notation siehe: The Exponentiated Weibull Family (Mudholkar e.a.)       */
/*                    Technometrics, 37, 1995, 436-445                        */
/*     Eingabeparameter: nobs (Stichprobenumfang)                             */
/*                       censrate (Zensierungsrate)                           */
/*                       alpha (Parameter 1)                                  */
/*                       theta (Parameter 2)                                  */
/*                       sigma (Parameter 3)                                  */
/*     Ausgabeparameter: X (nobs-x-2-Matrix)                                  */
/*                         (Enthaelt nobs ZVs mit Zensierungsindikator)       */
/*            Validiert: 13.10.99                                             */
/*****************************************************************************/


start expoweib(nobs, alpha_T, theta_T, sigma_T, alpha_C, theta_C, sigma_C, X);

X=j(nobs, 2, 55);

do i= 1 to nobs;
 T=ranuni(0);
```

```
 C=ranuni(0);
 *U=ranuni(200); *print u;
 *if c<censrate then X[i,2]=0; *else X[i,2]=1;

 X[i,1]=min(sigma_T#((-log(1 - (T##(1/theta_T))))##(1/alpha_T)),
            sigma_C#((-log(1 - (C##(1/theta_C))))##(1/alpha_C)));

if sigma_T#((-log(1 - (T##(1/theta_T))))##(1/alpha_T))<=
sigma_C#((-log(1 - (C##(1/theta_C))))##(1/alpha_C))
 then X[i,2]=1; else X[i,2]=0;
end;

 print x;
help1=rank(X[ ,1]);
help2=X;
X[help1,]=help2;

a=loc(X[,2]=0);                              /* Berechnung des Zensierungsgrades */
cens=ncol(a)/(nrow(X));

print cens;

b=ncol(loc(X[,1]=0));   /* Loeschen der Bindungsnullen */
if b>1 then do;
        X=(j(nobs-b,b,0)||I(nobs-b))*X;
        print b;
       end;
finish;

run expoweib(100, 5.0, 0.1, 100, 5.0, 0.15, 100, DATA_O);


print data_o;
```

```
/*****************************************************************************/
/*    Feld zum Einstellen des Glaettungsparameters und dessen Art (fix oder NN)  */
/*    sowie der Feinheit der Interpolation                                       */
/*****************************************************************************/
*run maxtheta(DATA_0, nn);                    /* liefert naechste Nachbarnanzahl */
*run pn_ueopt(DATA_0, pn);       /* liefert fixe Bandbreite=Bandbreitenparameter */
*run pn_uevar(DATA_0, pn);       /* liefert fixe Bandbreite=Bandbreitenparameter */
*run pn_data(DATA_0, pn);        /* liefert fixe Bandbreite=Bandbreitenparameter */
run smootRoT(DATA_0, pn);        /* liefert fixe Bandbreite=Bandbreitenparameter */

            /* falls fixe Bandbreite in Anzahl NN umgerechent werden soll gilt: */
                                                  /* NN=Gaussklammer(nobs # bb) */
nobs=nrow(DATA_0);
print pn;
*bb=nn;      /* Anzahl der NN falls nnofbˆ=1  bzw. fixe Bandbreite falls nnofb=1 */

bb= floor(pn#nobs);                          /* Anzahl NN bei fixer Bandbreite pn */
*bb=pn;
print bb;
nnofb=2;                               /* =1, d.h. fix und ˆ=1 d.h. nn-Definition */
fine=100;    /* Anzahl der aequidistanten Interpolationspunkte (beginnend bei 0, */
            /* endend bei der maximalen Ueberlebenszeit (-1)                     */
```

```
/*****************************************************************************/
/*     Programmaufruf zur Hazardratenschaetzung                              */
/*****************************************************************************/

last=nrow(DATA_0);                              /* Anzahl der Beoachtungen       */
writehel=j(fine+1, 2, 55);                  /* Matrix, die Stelle und Schaetzungen  */
                                            /* behaelt                       */
do i=0 to fine;                             /* Schleife ueber Schaetzpunkte  */
 interpol=(i/fine)*DATA_0[last,1];          /* Stelle der Schaetzung         */
 run hazard(DATA_0, bb, nnofb, interpol, hazard_x);   /* SCHAETZUNG           */
 writehel[i+1,1]=interpol;                  /* Beschreiben der Speichermatrix */
 writehel[i+1,2]=hazard_x;                  /* mit 1. Koordinate=Stelle,     */
end;                                        /* 2.=Schaetzung                 */
create xhx from writehel;                   /* Matrix in Datensatz schreiben */
append from writehel;
close xhx;

quit;
```

```
libname c 'u:\promo\sasprog\simdata';
data c.xhazardx
(RENAME=(COL1=x COL2=h_x)); set xhx; run;
proc print
data=c.xhazardx;          /* Ausgabe der Schaetzung in
Output-Fenster */ run;



/*****************************************************************************/
/*      Darstellung der geschaetzten Hazardwerte als Funktionsgraphen        */
/*                       Ausgabe in Graphik-Fenster                          */
/*****************************************************************************/
symbol1 color=black line=1 v = none i = spline w = 2 ;
                               /* 2-D Graphik fuer Zeit gegen Hazardrate */
symbol2 color=black line=41 v = none i = spline w = 1 ;

goptions device = win vsize=15cm hsize=15cm ctext=bl     ftext=swissl;
*goptions gaccess='sasgastd>u:\Promo\Pictures\haztyp3.eps' device=psepsf
                                                    htext=2 ftext=swiss;
                                          /* option zur Erstellung von .eps-file */
                                          /* fuer Einbinden der Graphik in LaTeX */

axis1 /* order=(0 to 1 by 0.2)  */  label=(f=swiss justify=right h=2 "Zeit");
axis2 /* order=(0 to 1 by 0.2) */   label=(f=swiss h=2 "Hazardrate");

proc gplot data=c.xhazardx;
plot h_x * x  / vaxis=axis2  haxis=axis1 /* overlay frame */;
run;
quit;
```

Literaturverzeichnis

[1] Aalen, O.O.: *Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models*, Annals of Statistics, **6**, 1978, 534-545.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N.: *Statistical Models Based on Counting Processes*, Springer-Verlag New York, 1993.

[3] Bennett, G.: *Probability inequalities for the sum of independent random variables*, Journal of the American Statistical Association , **57**, 1962, 33-45.

[4] Bernstein, S.: *Sur une modification de l'inéqualité de Tchebichef*, Annals Science Institute Sav. Ukraine, Sect. Math.1., 1924,

[5] Bickel, P.J., Rosenblatt, M.: *On some Global Measures of the Deviations of Density Function Estimates*, Annals of Statistics, **1**, 1973, 1071-1095.

[6] Borgan, Ø.: *Three Contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aalen-Johansen Estmators*, Statistical Research Report, Department of Mathematics, University of Oslo , **5/3**, 1997.

[7] Breiman, L., Meisel, W., Purcell, E.: *Variable kernel estimates of multivariate densities*, Technometrics, **19**, 1977, 135-144.

[8] Chuang-Stein, C., DeMasi, R.: *Surrogate Endpoints in AIDS Drug Development: Current Status*, Drug Information Journal, **32**, 1998, 439-448.

[9] Chow, Y.S., Geman, S., Wu, L.D.: *Consistent Cross-Validated density estimation*, Annals of Statistics, **11**, 1983, 25-38.

[10] DEHEUVELS, P., HOMINAL, P.: *Estimation Automatique de la Densité*, Revue de Statistique Appliquée, **18**, 1980, 25-55.

[11] DESCARTE, R.: *Abhandlung über die Methode des richtigen Vernunftgebrauchs*, Reclam, 1961.

[12] DETTE, H., GEFELLER, O.: *The Impact of Different Definitions of Nearest Neighbour Distances for Censored Data on the Nearest Neighbour Kernel Estimators of the Hazard Rate*, Nonparametric Statistics, **4**, 1995, 271-282.

[13] DEVROYE, L.: *Exponetial Inequalities in Nonparametric Estimation*, in Functional Estimation, ed. G. Roussas, NATO ASI Series Dordrecht, Kluwer Academic Publishers, 1991, 31-44.

[14] DEVROYE, L., GYÖRFI, L.: *Nonparametric Density Estimation, The L_1 View*, John Wiley & Sons, New York, 1985.

[15] DIEHL, S., STUTE, W.: *Kernel Density and Hazard Function Estimation in the Presence of Censoring*, Journal of Multivariate Analysis, **25**, 1988, 299-310.

[16] EPANECHNIKOV, V.A.: *Nonparametric Estimate of a Multivariate Probability Density*, Theory of Probability and Its Applications **14**, 1969, 153-158.

[17] ESTÉVEZ-PÉREZ, G., QUINTELA-DEL-RÍO, A.: *Nonparametric Estimation of the Hazard Function Under Dependence Conditions*, Communications in Statistics Theory and Methods **28(10)**, 1999, 2297-2331.

[18] FIX, E., HODGES, J.L.JR.: *Discriminatory Analysis, Nonparametric Discrimination: Consistency Property*, Report No. 4, USAF School of Aviation Medicine, Texas, 1951.

[19] FLEMING, T.R., HARRINGTON, D.P.: *Counting Processes and Survival Analysis*, John Wiley & Sons, New York, 1991.

[20] FRYER, M.J.: *Some Errors Associated with the Non-parametric Estimation of Density Functions*, J. Inst. Math. Applics. **18**, 1976, 371-380.

[21] Gasser, T., Müller, H.-G., Mammitzsch, V.: *Kernels for Nonparametric Curve Estimation*, Journal of the Royal Statistical Society. Series B, **47**, 1985, 238-352.

[22] Gefeller, O.: *Kernschätzung für die Hazardfunktion bei zensierten Daten*, Diplomarbeit, Fachbreich Statistik, Universität Dortmund, 1986.

[23] Gefeller, O., Dette, H.: *Nearest Neighbour Kernel Estimation of the Hazard Function from Censored Data*, Journal of Statistical Computation and Simulation, **43**, 1992, 93-101.

[24] Gefeller, O., Hjort, N.L. : *A New Look at the Visual Performance of Nonparametric Hazard Rate Estimators*, In: I. Balderjahn, R. Mathar, M. Schrader (eds.): Classification, Data Analysis, and Data Highways, Springer, 1998, 139-146.

[25] Gefeller, O., Michels, P.: *Nichtparametrische Analyse von Verweildauern*, Österreichische Zeitschrift für Statistik und Informatik, **22**, 1992, 37-59.

[26] Gefeller, O., Michels, P.: *A Review on Smoothing Methods for the Estimation of the Hazard Rate Based on Kernel Functions*, COMPSTAT'92, Proceedings of the 10th Symposium on Computational Statistics, Physica, 1992, 459-464.

[27] Gefeller, O., Pflüger, R., Bregenzer, T.: *The Implementation of a Data-Driven Selection Procedure for the Smoothing Parameter in Nonparametric Hazard Rate Estimation Using SAS/IML Software*, SEUGI'95, Proceedings of the 13th SAS European Users Group International Conference, SAS Institute Inc. Carry 1996, 1996, 1288-1300.

[28] Glad, I.K., Hjort, N.L., Ushtchakov, N.G.: *Density Estimation using the Sinc Kernel*, unpublished manuscript, 1998.

[29] Hall, P., Hu, T.C., Marron, J.S.: *On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator*, Annals of Statistics, **15**, 1987, 163-181.

[30] Hall P., Marron, J.S.: *Improved Variable Window Kernel Estimates of Probability Densities*, Annals of Statistics, **23**, 1995, 1-10.

[31] Hall, P., Sheather, S.J., Jones, M.C., Marron, J.S.: *On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation*, Biometrika, **78**, 1991, 263-269.

[32] Hess, K.R., Serachitopol, D.M., Brown, B.W.: *Hazard Function Estimators: A Simulation Study*, Statistics in Medicine, **18**, 1999, 3075-3088.

[33] Hjort, N.L.: *Semiparametric Estimation of the Hazard Rates*, Invited paper presented at the Advanced Study Workshop on Survival Analysis and Related Topics, Columbus, Ohio, June 1991.

[34] Hjort, N.L., Glad, I.K.: *Nonparamtric Density Estimaiton with a Parametric Start*, Annals of Statistics, **23**, 1995, 882-904.

[35] Hjorth, U.: *A Reliability Distribution with Increasing, Decreasing, Constant and Bathtub-Shaped Failure Rates*, Technometrics, **22**, 1980, 99-107.

[36] Hoeffding, W.: *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association , **58**, 1962, 13ff.

[37] Izenman, A.J.: *Recent Developments in Nonparametric Density Estimation*, Journal of the American Statistical Association , **86**, 1991, 205-224.

[38] Jones, M.C., Marron, J.S., Scheather, S.J.: *A Brief Survey on Bandwidth Selection for Density Estimation*, Journal of the American Statistical Association , **91**, 1996, 401-407.

[39] Kaplan, E.L., Meier, P.: *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association , **53**, 1958, 457-481.

[40] Kalbfleisch, J.D., Prentice R.L.: *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York ,1980.

[41] Korostelev, A., Nussbaum, M.: *The Asymptotic Minimax Constant for Sup-norm Loss in Nonparametric Density Estimation*, Bernoulli, London, **5**, 1999, 1099-1113.

[42] Lee, E.T.: *Statistical Methods for Survival Data Analysis*, John Wiley & Sons, New York , 1992.

[43] Loftsgaarden, D.O., Quesenberry, C.P.: *A Nonparametric Estimate of a Multivariate Density Function*, Annals of Mathematical Statistics, **36**, 1965, 320-326.

[44] Marron, J.S.: *A Comparison of Cross-Validation Techniques in Density Estimation*, Annals of Statistics, **15**, 1987, 152-163.

[45] Marron, J.S., Tsybakov, A.B.: *Visual Error for Qualitative Smoothing*, Journal of the American Statistical Association , **90**, 1995, 499-507.

[46] Marron, J.S., Wand, M.P.: *Exact Mean Integrated Squared Error*, Annals of Statistics, **20**, 1992, 712-736.

[47] Minkowski, H.: *Diophantische Approximationen*, Teubner, 1907.

[48] Mudholkar, G.S., Srivastava, D.K., Freimer, M.: *The Exponetial Weibull Family: A Reanalysis of the Bus-Motor-Failure Data*, Technometrics, **37**, 1995, 436-445.

[49] Müller, H.-G., Wang, J.-L.: *Hazard Rate Estimation Under Random Censoring with Varying Kernels and Bandwidths*, Biometrics, **50**, 1994, 61-76.

[50] Nadaraya, E.A.: *On Nonparametric Estimates of Density Funktions and Regression Curves*, Theor. Probability Appl. , **10**, 1965, 186-190.

[51] Nelson, W.: *Theory and Applications of Hazard Plotting for Censored Failure Data*, Technometrics, **14**, 1972, 945-966.

[52] Nussbaum, M.: *Asymptotic Equivalence of Statistical Experiments*, Lecture Notes, Workshop: Nonparametric Functional Estimation, Neural Nets and Risk Asymptotics, Oberwolfach, 2000.

[53] ØKSENDAL, B.: *Stochastic differential equations: an introduction with applications*, Springer, Berlin, 4th ed., 1995.

[54] PARZEN, E.: *On the Estimation of a Probobility Density Function and the Mode*, Annals of Mathematical Statistics, **33**, 1962, 1065-1076.

[55] PARK, B.U., KIM, W.C., MARRON, J.S.: *Asymptotically Best Bandwidth Selectors in Kernel Density Estimation*, Core Discussion Paper No 9154, Université Catholique de Louvain, 1991.

[56] PARK, B.U., MARRON, J.S.: *Comparison of Data-Driven Bandwidth Selectors*, Journal of the American Statistical Association , **85**, 1990, 66-72.

[57] PATIL, P.N.: *On the Least Square Cross-Validation Bandwidth in Hazard Rate Estimation*, Annals of Statistics, **21**, 1993, 1792-1810.

[58] PRAKASA RAO, B.L.S.: *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.

[59] RALESCU, S.S.: *The Law of the Iterated Logarithm for the Multivariate Nearest Neighbor Density Estimators*, Journal of Multivariate Analysis, **53**, 1995, 159-179.

[60] ROSENBLATT, M.: *Remarks on some nonparametric estimates of a density function*, Annals of Mathematical Statistics, **27**, 1956, 832-837.

[61] SAS INSTITUTE INC.: SAS/IML Software: Usage and References, Version 6, First Edition. Cary, NC: SAS Institute Inc. 1990.

[62] SCHÄFER, H.: *Die Konvergenz des variablen Kernschätzers und die Geschwindigkeit der lokalen Konvergenz empirischer Maße bei zufällig zensierten Daten*, Dissertation, Düsseldorf, 1986a.

[63] SCHÄFER, H.: *Local Convergence of Empirical Measures in the Random Censorship Situation with Application to Density and Rate Estimators*, Annals of Statistics, **14**, 1986b, 1240-1245.

[64] Scott, D.W., Terrell, G.R.: *Biased and Unbiased Cross-Validation in Density Estmation*, Journal of the American Statistical Association , **82**, 1987, 1131-1146.

[65] Silverman, B.W.: *Density Estimation*, Chapman & Hall, London, 1986.

[66] Silverman, B.W.: *Weak and Strong Uniform Consistency of the Kernel Estimate of the Density and its Derivates*, Annals of Statistics, **6**, 1978, 177-184.

[67] Simonoff, J.S.: *Smoothing Methods in Statistics*, Springer-Verlag New York, 1996.

[68] Singpurwalla, N.D., Wong, W.H.: *Kernel Estimators of the Failure-rate Function and Density Estimation: an Anology*, Journal of the American Statistical Association , **78**, 1983, 478-481.

[69] Sirjaev, A.N.: *Wahrscheinlichkeit*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1988.

[70] Siu, L.L., Banerjee, D., Khurana, R.J., Pan, X., Pflüger, R., Tannock, I.F., Moore, M.J.: *The Prognostic Role of p53, Metallothionein, P-glycoprotein, and MIB-1 in Muscle-invasive Urothelial Transitional Cell Carcinoma*, Clinical Cancer Research, **4**, 1998, 559-565.

[71] Stute, W.: *The Oscillation Behaviour of Empirical Processes*, Annals of Probability, **10**, 1982a, 86-107.

[72] Stute, W.: *The Law of the Logarithm for Kernel Density Estimators*, Annals of Probability, **10**, 1982b, 414-422.

[73] Tanner, M.A., Wong, W.H.: *Data-Based Nonparametric Estimation of the Hazard Function with Application to Model Diagnostics and Exploratory Analysis*, Journal of the American Statistical Association , **79**, 1984, 174-182.

[74] Wand, M.P., Jones, M.C.: *Kernel Smoothing*, Chapman & Hall, London, 1995.

[75] Yandell, B.S.: *Nonparametric Inference for Rates with Censored Survival Data*, Annals of Statistics, **11**, 1983, 1119-1135.

[76] Zhang, B.: *A Law of the Iterated Logarithm for Kernel Estimators of Hazard Functions Under Random Censorship*, Scandinavian Journal of Statistics, **23**, 1996, 37-47.