Estimation of Minimum Mean Squared Error with Variable Metric from Censored Observations

Von der Fakultät der Mathematik und Physik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

vorgelegt von

StR Matthias Strobel

geboren in Heidelberg

Haupberichter:	Prof. Dr. H. Walk
Mitberichterin:	Prof. Dr. B. Kaltenbacher
Tag der Einreichung:	19. Dezember 2007
Tag der mündlichen Prüfung:	13. Februar 2008

Institut für Stochasik und Anwendungen der Universität Stuttgart

2008

The face in the water looks up

And she shakes her head as if to say

That it's the last time you'll look like today.

Genesis – Ripples

Contents

Acknowledgments

First of all, I would like to thank Prof. Dr. Harro Walk for the encouraging discussions during the writing of this thesis. Also, PD Dr. Jürgen Dippon and Dipl. Math. Stefan Winter from the Institute of Stochastics and Applications at the Universität Stuttgart gave helpful advice on computational aspects and the problem of censoring. Thanks also go to Dr. Peter Fritz and the Elternverein Krebskranker Kinder, Stuttgart and Ludwigsburg, who supported the research of this thesis. Finally, I would like to thank my wife Ulrike and my family for patience and encouragement during the last four years.

Deutsche Zusammenfassung

Die in dieser Arbeit untersuchte mathematische Fragestellung hat eine Anwendung im Bereich der Medizin. Angenommen, ein Patient soll in einem Krankenhaus wegen einer bestimmten Krankheit behandelt werden und der behandelnde Arzt soll nun eine Prognose über den Krankheitsverlauf machen. Dies kann einerseits auf Grund seiner Erfahrung und seines Wissens über den Verlauf der Krankheit geschehen. Andererseits stehen in Krankenhäusern oft Datenbanken über Krankheitsverläufe von bereits behandelten Patienten zu Verfügung. Es ist allerdings unwahrscheinlich, dass sich darunter Patienten befinden, die hinsichtlich Alter, bisherigem Krankheitsverlauf usw. mit dem Patienten genau übereinstimmen. Häufiger wird man nur Fälle finden können, deren Situation ähnlich der des zu behandelnden Patienten ist. Für jede Krankheit aber gibt es Faktoren oder Prdiktoren, die für den Verlauf wichtig und solche, die dies weniger sind. Auf Grund der Gewichtung dieser einzelnen Prädiktoren kann man nun aus den gesammelten Daten einen oder mehrere ähnliche Fälle heraussuchen und mit Hilfe dieser eine Prognose über den weiteren Krankheitsverlauf machen. Um ähnliche Fälle zu finden, wird man eine Hierarchie unter den Prädiktoren haben wollen, die angibt, wo zuerst nach Übereinstimmung zu suchen ist. Weiterhin ist es hilfreich zu wissen, wie viel wichtiger ein Prädiktor im Vergleich zu einem anderen ist.

Diese Arbeit sucht dieses Problem als eine statistische Fragestellung zu beantworten. Angenommen, ein Datensatz mit Informationen über den Krankheitsverlauf von Patienten ist gegeben und zu einem neuen Datenpunkt sind mehrere ähnliche Datenpunkte aus dem Datensatz anzugeben. Die Frage nach Ähnlichkeit ist dann die Frage nach der Metrik, mit der die Abstände gemessen werden. Das Ziel ist es, basierend auf einem Datensatz, eine geeignete Metrik anzugeben bzw. zu schätzen.

Für die Modellierung des Problems wurde ein nichtparametrischer Ansatz gewählt. Damit werden keine oder nur schwache Annahmen über die Verteilung der Prädiktoren und

der Überlebenszeit gemacht. Ein weiterer Vorteil des nichtparametrischen Modells ist die Flexibiltät und somit die Anwendbarkeit in anderen Situationen.
Es seien nun $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ unabhänige identisch verteilte $(d+1)$-dimensionale Zufallsvariablen mit $EY^2 < \infty$. Die d-dimensionalen Zufallsvariablen $X_1, X_2, \ldots$ werden als Prädiktorzufallsvariablen bezeichnet. Die Regressionsfunktion $m : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ wird definiert durch $m(x) := E(Y|X = x)$. Als Hilfsmittel zur Beurteilung der Güte einer Metrik im Prädiktorraum $\mathbb{R}^{\mathrm{d}}$ bieten sich in der Theorie der nichtparametrischen Regression Schätzungen des minimalen mittleren quadratische Fehlers

$$L^* = E\left(Y - m(X)\right)^2$$

auf Grund der Beobachtungen $(X_1, Y_1), (X_2, Y_2), \ldots$ an. Der minimale mittlere quadratische Fehler gibt an, wie schwer ein Regressionsproblem zu lösen ist. Schätzt man die Regressionfunktion m mit einem konsistenten Schätzer, etwa mit der Technik der Kreuzvalidierung oder durch Teilung der Stichprobe zur Ermittlung von Design-Parametern (z.B. Würfelseitenlängen in der kubischen Partitionsschätzung), so erhält man unter Regularitätsvoraussetzungen, insbesondere der Lipschitz-Stetigkeit von m, als maximale Quadratmittel-Konvergenzgeschwindigkeitsordnung $n^{-\frac{2}{d+2}}$ für die Schätzung von m und dieselbe Rate mittelbar durch Teilung der Stichprobe für die Schätzung von L^*. Im letzteren Falle kann man eine bessere Konvergenzgeschwindigkeit durch direkte Schätzung von L^* mit Hilfe eines nicht-konsistenten Schätzers erreichen. Mit Hilfe einer Mischung aus Partitionen- und 1-nächstem-Nachbar-Schätzer $\tilde{L}_n$ wird in dieser Arbeit eine bessere Konvergenzgeschwindigkeit erreicht. Dazu partitioniert man den Prädiktorraum $\mathbb{R}^{\mathrm{d}}$ mit Hilfe von Würfeln der Seitenlänge

$$\tilde{h}_n = n^{-1/(1+d)}.$$

Weiter konstruiert man zu jedem Datenpunkt X_i für $i \in \{1, \ldots, n\}$ einen nächsten Nachbarn im weiteren Sinn $X_{i,n}$ (widest sense nearest neighbour, WSNN), indem man zufällig einen anderen Datenpunkt aus dem Quader aussucht, wenn dieser mehr Punkte als nur X_i enthält. Enthält der Quader nur X_i, wählt man den 1-nächsten Nachbarn als

WSNN. Damit konstruiert man den Schätzer

$$\tilde{L}_n = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - Y_{i,n})^2 ,$$

wobei $Y_{i,n}$ die $X_{i,n}$ entsprechende abhängige Variable ist. Die Konvergenz- geschwindigkeit im unzensierten Fall gibt der folgende Satz an. Es bezeichne

$$\sigma^2(x) := E\left((Y - m(X))^2 \middle| X = x \right)$$

die lokale Varianz.

Satz 1 *Angenommen, es gelte* $|Y_i| \leq B$ *mit* $i \in \{1, \ldots, n\}$ *. Weiterhin sollen Pattsituationen (d.h. zwei Datenpunkte haben den gleichen Abstand zu einem dritten Datenpunkt) bzgl. der euklidischen Norm mit Wahrscheinlichkeit 0 auftreten.* $\mu := P_X$ *habe einen kompakten Träger und sowohl* m *als auch* σ^2 *seien Lipschitz-stetig mit Lipschitzkonstanten* k_1 *und* k_2*. Dann gilt*

$$E\left| \tilde{L}_n - L^* \right| \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

Dabei hängt die Konstante in der obigen Abschätzung von d, B, k_1 *und* k_2 *ab.*

Beim Beweis des Satzes (vgl. Abschnitt 1.3) ist eine Stabilitätseigenschaft von $\tilde{L}_n$ wesentlich. Sie besagt, dass eine kleine Änderung in einem Zufallsvektor nur geringen Einfluß auf die WSNN-Struktur hat.
Man betrachtet nun die Norm im $\mathbb{R}^{\mathrm{d}}$ als variabel, d.h. anstatt der euklidischen Norm benutzt man

$$\| x \|_h := \sqrt{(h_1 x_1)^2 + \cdots + (h_d x_d)^2},$$

wobei $h = (h_1, , \ldots, h_d)$ Element einer Parametermenge $\mathcal{Q}^d \subset \mathbb{R}^{\mathrm{d}}$ mit $h_i > 0$ für alle $i \in \{1, \ldots, d\}$ ist. Für jede Metrik $\| \cdot \|_h$ erhält man eine neue WSNN-Struktur und somit einen neuen Schätzer $\tilde{L}_n^h$. Es bezeichne $X_{i,n}^h$ den WSNN von X_i bzgl. $\| \cdot \|_h$.
Die Idee ist nun, dass man

$$h \longmapsto \tilde{L}_n^h$$

minimiert, um eine optimale Metrik zu erhalten. Zwar ist L^* unabhängig von der Wahl der Metrik, dennoch kann man bei einer günstigen Wahl der Metrik erwarten, dass die geschätzte minimale quadratische Abweichung bei gegebenem Datensatz und Prädiktoren L^* möglichst nahe kommt.
Man definiert für $h \in \mathcal{Q}^d$

$$\begin{aligned} R(h) &:= R_n(h) := E\tilde{L}_n^h && (0.1)\\ \bar{h}_n &:= \arg\min_{h\in\mathcal{Q}^d} R(h) && (0.2)\\ H_n &:= \arg\min_{h\in\mathcal{Q}^d} \tilde{L}_n^h. && (0.3) \end{aligned}$$

Der Schätzer $\tilde{L}_n^{H_n}$ weicht bei großem n stochastisch wenig von $\tilde{L}_n^{\bar{h}}$ ab, wie der folgende Satz zeigt.

Satz 2 *Wenn $|Y_i| \le B$ für alle $i \in \{1, ..., n\}$ und Pattsituationen mit Wahrscheinlichkeit Null auftreten, dann gilt (mit einer von d abhängigen Konstante γ_d)*

$$E\left[R(H_n)\right] - R(\bar{h}_n) \le 2\sqrt{2(1+\log\left(2|\mathcal{Q}^d|\right))}B^2(5+2\gamma_d)n^{-\frac{1}{2}}.$$

Es stellt sich die Frage, wann der auf diesem Weg gewonnene Schätzwert $\tilde{L}_n^{H_n}$ zu optimistisch ist. Dies kann dann angenommen werden, wenn die Lipschitz-Konstante der lokalen Varianz wesentlich größer ist als die Lipschitz-Konstante der Regressionsfunktion.

Lemma *Es sei $X_i \in [0,1]^d$ und $\sigma^2(x)$ bzw. $m(x)$ seien Lipschitz-stetig mit Konstanten k_1 und k_2. Für $h \in \mathcal{Q}^d$ bezeichne*

$$0 < h_{min} := \min\left\{h_1, \ldots, h_d\right\}.$$

Gilt

$$k_1 \le \frac{k_2^2}{\left(\frac{\tilde{h}_n}{h_{min}}+1\right)^d + \frac{(\sqrt{d}+1)^d}{d-1}},$$

dann ist für Konstanten $c_1 \le c_2$

$$\begin{aligned} E|\sigma^2(X_{i,n}^h) - \sigma^2(X_i)| &\le c_1 n^{-\frac{2}{d+1}}\\ E(m(X_{i,n}^h) - m(X_i))^2 &\le c_2 n^{-\frac{2}{d+1}}. \end{aligned}$$

Wenn die Abschätzungen der Erwartungswerte im obigen Lemma scharf sind, folgt dass der

$$\text{bias}\left(\tilde{L}_n^h\right) = E\tilde{L}_n^h - L^*$$

durch eine nichtnegative von n abhängige Konstante nach unten abgeschätzt werden kann.

Häufig ist in medizinischen Anwendungen die Überlebenszeit zensiert. Man kann also anstelle von Y nur

$$Z = \min\{Y, C\}$$

beobachten. Zur mathematischen Analyse stellt man folgende Bedingungen an Y und C bzw. ihre Verteilungen. Es seien $F(t) := P(Y > t)$ und $G(t) := P(C > t)$ die Überlebensfunktionen von Y und C.

(**A1**) C und (X, Y) sind unabhängig

(**A2**) Es existiert ein $L > 0$ mit $P(0 \leq Y \leq L) = 1$ und $P(C > L) > 0$.
Weiter sind G und F stetig.

(**A3**) Für p mit $0 < p \leq \frac{1}{2}$ gilt $\int_0^{T_K} F(t)^{-\frac{p}{1-p}} d\,(1 - G(t)) < \infty$

(A1) ist erfüllt, wenn die Zensierung sowohl vom Gesundheitszustand des Patienten als auch von dessen Person unabhängig ist. Die Zensierung hängt nur von externen Faktoren ab, die nicht mit der Überlebenszeit Y und den Prdiktoren in X zusammenhängen. In vielen Anwendungen ist die Überlebenszeit beschränkt, somit der erste Teil von **(A2)** erfüllt. Im zweiten Teil ist gefordert, dass nicht die gesamte Zensierung in $[0, L]$ stattfindet. Der Kaplan-Meier-Schätzer (s.u.) ist instabil links von seiner oberen Schranke T_K. **(A3)** besagt, dass eine gewisse Anzahl von zensierten und unzensierten Beobachtungen in einer kleinen Umgebung von T_K vorliegen. Die Bedingung **(A3)** wird gebraucht, um die Konvergenzgeschwindigkeit des Kaplan-Meiers-Schtzers G_n gegen die Überlebensfunktion G zu bestimmen.

Das wichtigste Werkzeug, um die Verteilung der Überlebenszeit Y aus den zensierten Beobachtungen zu gewinnen, ist der Kaplan-Meier-Schtzer

$$G_n(t) = \begin{cases} \prod_{i=1,\ldots,n:Z_{(i)}\leq t} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}} & ; \quad t \leq Z_{(n)} \\ 0 & ; \quad \text{sonst.} \end{cases}$$

In der Menge $\left\{(Z_{(1)}, \delta_{(1)}), \ldots, (Z_{(n)}, \delta_{(n)})\right\}$ sind die Beobachtungen in aufsteigender Größe geordnet, dabei werden bei gleichem Wert die zensierten vor den unzensierten Beobachtungen angeordnet. Weiterhin setzt man $\delta_i = 1$, wenn $Y_i < C_i$ und $\delta_i = 0$ wenn $C_i \leq Y_i$ gilt.

Der folgende Satz macht eine Aussage über die Konvergenzgeschwindigkeit von G_n gegen G.

Satz 3 *Es sei* $G(T_K) > 0$, G *stetig und für* $0 < p \leq \frac{1}{2}$

$$\int_0^{T_K} F(t)^{-\frac{p}{1-p}} d(1 - G(t)) < \infty. \tag{0.4}$$

a) (Chen/Lo, 1997) Gilt $0 < p < \frac{1}{2}$, *dann ist* (0.4) *äquivalent zu*

$$\limsup_{n \longrightarrow \infty} n^p \sup_{t \leq T_K} |G_n(t) - G(t)| < \infty \quad \textit{fast sicher.}$$

b) (Gu/Lai, 1990) Gilt $p = \frac{1}{2}$, *dann folgt aus* (0.4) *mit*

$$t_k := \sup\{t : P(Y \geq t) \geq n^{-(1+\alpha)}\}$$

für $\frac{1}{3} < \alpha < \frac{1}{2}$

$$\limsup_{n \to \infty} \left(\frac{n}{2 \log\log n}\right)^{\frac{1}{2}} \cdot \sup_{t < t_k} |G_n(t) - G(t)| < \infty \ \textit{f.s.}$$

Man adaptiert den Schätzer für den unzensierten Fall $\tilde{L}_n$ und setzt

$$\bar{L}_n := \frac{1}{2n} \sum_{i=1}^{n} \left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}_{i,n} + \bar{Y}_{i,n}^2\right)$$

mit

$$\bar{Y}_i := \frac{\delta_i Z_i}{G_n(Z_i)}, \ \bar{Y}_{i,n} := \frac{\delta_{i,n} Z_{i,n}}{G_n(Z_{i,n})},$$

$$\bar{Y}_i^2 := \frac{\delta_i Z_i^2}{G_n(Z_i)} \text{ und } \bar{Y}_{i,n}^2 := \frac{\delta_{i,n} Z_{i.n}^2}{G_n(Z_{i,n})}.$$

Für den Schätzer im zensierten Fall erhält man eine ähnliche Konvergenzgeschwindigkeit wie für unzensierte Beobachtungen.

Satz 4 *Es gelte* $|Y_i| \leq B$ *für alle* $i \in \{1, \ldots, n\}$, m *und* σ^2 *seien Lipschitz-stetig und Pattsituationen treten mit der Wahrscheinlichkeit Null auf. Zusätzlich seien* **(A1)** *und* **(A2)** *erfüllt und* $\mu = P_X$ *habe kompakten Träger.*
a) Gilt $d \in \{1, 2, 3\}$ *und* **(A3)** *mit* $p = \frac{1}{2}$, *dann ist*

$$|\bar{L}_n - L^*| = \mathcal{O}_{\mathbf{P}}\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}.$$

b) Gilt $d \geq 4$ *und* **(A3)** *mit* $p = \frac{2}{d+1}$, *dann ist*

$$|\bar{L}_n - L^*| = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2}{d+1}}\right).$$

Hier ist stochastische Konvergenz bewiesen, im Gegensatz zu L^1-Konvergenz im unzensierten Fall. Das Ziel ist es, eine optimale Metrik $\|\cdot\|_h$ und damit eine WSNN-Struktur $(i,n)^h$ zu schätzen, indem man ein h so wählt, dass $\bar{L}_n^h$ minimal ist für alle $h \in \mathcal{Q}^d$. Es ist dann notwendig, den Stichprobenraum

$$D_{2n} = D_n^1 \cup D_n^2$$

zu teilen, um Rückkoppelungsprozesse zwischen dem Kaplan-Meier-Schätzer und der WSNN-Struktur bezüglich $(i,n)^h$ zu verhindern.
Zur Bestimmung des WSNN wird D_n^1 und zur Berechnung des Kaplan-Meier-Schätzers D_n^2 benutzt. Für das Folgende konditioniert man mit D_n^2. Im Fall einer Metrik $\|\cdot\|_h$ und einem Schätzer $\bar{L}_n^{H_n}$ definiert man analog zu $(0.1) - (0.3)$

$$\begin{aligned} R(h) &:= E(\bar{L}_n^h | D_n^2) \\ \bar{h}_n &:= \arg\min_{h \in \mathcal{Q}^d} E\tilde{L}_n^h \\ H_n &:= \arg\min_{h \in \mathcal{Q}^d} \bar{L}_n^h|_{D_n^2}. \end{aligned}$$

Man erhält dann das folgende Result.

Satz 5 *Es sei* $|Y_i| \leq B$ *für alle* $i \in \{1, \ldots, n\}$ *und Pattsituationen treten mit Wahrscheinlichkeit Null auf. Weiterhin gelten* **(A1)**, **(A2)** *und* **(A3)** *mit* $p = \frac{1}{2}$. *Dann ist*

$$R(H_n) - E\tilde{L}_n^{\bar{h}_n} = \mathcal{O}_{\mathbf{P}}\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}.$$

	Gruppe 1	*Gruppe 2*	*Gruppe 3*
Gewichte	0.125-0.5	1-4	8-32
Prädiktoren	HISTO [0.5,0.125] ER (OSP)[0.125]	AGE (RBK)[4] MENO [4] PR [4,1] ER(RBK) [2] PT [1]	CERB [32] AGE (OSP) [32] GR [32,16] PN [8,32]

Tabelle 1: Gewichte der Prädiktoren des RBK- und des OSP-Datensatzes. Wenn zwei Gewichte angegeben sind, ist das erste vom RBK- und das zweite vom OSP-Datensatz.

In der Anwendung auf Daten wird eine logarithmische Skala

$$\mathcal{Q}_k^d := \left\{2^{-k}, 2^{-(k-1)}, \ldots, 2^{-1}, 2^0, 2^1, \ldots, 2^{k-1}, 2^k\right\}^d$$

mit $d, k \in \mathbb{N}$ benutzt. Es ist sinnvoll, den Parameterraum $\mathcal{Q}_k^d$ nicht zu groß zu wählen, da dann die Konstante $\left|\mathcal{Q}_k^d\right|$ kleiner und die in Satz 2 und Satz 4 angegebene Konvergenzgeschwindigkeit besser ist. Andererseits muss man die Rechenzeit mit $\left|\mathcal{Q}_k^d\right|$ Durchläufen zur einmaligen Berechnung des Schätzers berücksichtigen.

Den Anwendungen lagen zwei Datensätze von Brustkrebspatientinnen zugrunde. Der eine Datensatz wurde im Robert-Bosch-Krankenhaus (RBK) in Stuttgart gesammelt, der andere vom Onkologischen Schwerpunkt Stuttgart (OSP) an verschiedenen Krankenhäusern im Großraum Stuttgart. Vom RBK-Datensatz mit $n = 913$ Patientinnen wurden die Prdiktoren AGE (Alter), MENO (Menopausenstatus), HISTO (histologischer Typ), PN (Anzahl der betroffenen Lymphknoten), GR (Einstufung des Tumors), ER (Östrogen-Rezeptorstatus), PR (Progesteron-Rezeptorstatus) und CERB (Menge des cerb2-Proteins in den Zellmembranen des Tumors) verwendet. Im OSP-Datensatz waren die Informationen von $n = 1221$ Patientinnen gesammelt. Hier wurden die Prädiktoren AGE, HISTO, PN, PT (Größe des Tumors), GR, ER und PR herangezogen. Um die bisherigen Ergebnisse anzuwenden, muss man voraussetzen, dass die Zufallsvariable absolut stetig ist, da dann Pattsituationen mit Wahrscheinlichkeit Null auftreten. Weiterhin müssen die Bedingungen **(A1)**, **(A2)** und **(A3)** erfüllt sein. Dies ist dann der Fall, wenn die Zensierung

unabhängig von den Prädiktoren und der Überlebenszeit und auch nach einem möglichen Tod noch stattfinden kann. Weiterhin müssen genug Patientinnen mit zensierter bzw. unzensierter Überlebenszeit nahe der oberen Schranke beobachtbar sein.

In der Realisierung wurden mit jedem Datensatz drei Studien durchgeführt, denen jeweils eine variierte, durch die Vorgängerstudie bestimmte logarithmische Skala (mit $k = 2$) zugrunde lag. Es zeigte sich dabei, dass das geschätzte minimale L^2 Risiko beim RBK Datensatz geringer als bei den OSP Daten ist (1.936605 im Vergleich zu 2.348019, vgl. Tabellen 3.1 und 3.2 in Kapitel 3). Daraus kann man schließen, dass die Prädiktoren MENO und CERB zusammen wichtiger als PT sind. Die Gewichte der Prädiktoren in den beiden Datensätzen variieren zwar, sie lassen sich aber in drei Gruppen einteilen (vgl. Tabelle 1). In der mittleren Gruppe sind MENO und ER vom RBK-Datensatz und PT vom OSP-Datensatz. ER und PR verlieren im OSP-Datensatz an Gewicht, was an einer möglichen Abhängigkeit zu PT liegen könnte. Der Prädiktor HISTO hat in beiden Datensätzen das niedrigste Gewicht. Die Gruppe der Prädiktoren mit dem größten Gewicht enthält den Prädiktor CERB, der im OSP-Datensatz nicht vorliegt. Beim OSP-Datensatz befindet sich hier hingegen der Prädiktor AGE. Dies liegt vermutlich daran, dass dieser Datensatz den Menopausenstatus enthält, der abhängig vom Alter ist. Das größte Gewicht in beiden Datensätzen haben GR und PN.

List of Abbreviations

$\mathbb{N}$ set of natural numbers

$\mathbb{R}$ set of real numbers

$\mathbb{R}^{\mathrm{d}}$ set of real d-dimensional vectors

EX expectation of X

EX^2 short for $E(X^2)$

sup supremum

e Euler number

log natural logarithm

$|A|$ cardinality of a set A

I_A indicator function of a set A

a.s almost surely

argmin argument of the minimum of an expression

m regression function

Var variance

bias bias of an estimator

μ distribution P_X of X

$\|\cdot\|$ Euclidean norm

$\mathcal{O}_{\mathbf{P}}(Z)$ convergence in measure against Z

Introduction

The research of this thesis has been motivated by a problem frequently considered in medical science. Assume we have information about a number of patients that is collected during the treatment of an illness. Apart from this data set, we have information of how the illness developed for each patient. In both data sets information may be incomplete due to a variety of reasons. Now, a new patient with the same illness is being treated. Can we predict the development of the illness on the basis of the information about the patient and our collected data? It does not seem very likely that we will have patients that fit exactly our new patient. Therefore we will have to find patients that are similar to the new case. Essentially, this depends on the importance of the different categories in our set of data. If two patients differ largely only with respect to an unimportant category, they may still be considered similar. Small differences in an important category indicate that the circumstances under which treatment is to be received are very different.

Mathematically speaking, if you want to examine similarity in a set of data you will have to find a suitable metric. A metric indicates the importance of a category (also called predictor) in a data set. Different approaches have been discussed in the literature (for an overview see Devroye, Györfi and Lugosi [6] (1996), chapter 26). Most recently Dippon, Fritz and Kohler [9] (2002) analyzed the importance of predictors that are used in the treatment of breast cancer patients. Common to all these studies is the usage of a nonparametric approach. This means that no assumptions about the common distribution of the predictors and the development of the illness (also called survival time) are made. This ensures that the model and the solution have a wide range of possible applications. In the theory of nonparametric regression there is a tool that indicates how difficult it is to solve a given regression problem: It is the minimum mean squared error. A small value of this indicator means that a good solution is possible, a very large value means that we can not expect a satisfying solution. In a nonparametric model it is impossible to

find the exact value of the minimum mean squared error, but we have to approximate it with the help of an estimator. Therefore, different approaches to estimate the minimum mean squared error are discussed in Chapter One. Furthermore a new estimator, that has a better rate of convergence than the estimators used in the studies mentioned before, is introduced (cf. Section 1.3).

The idea in this thesis is the following: to find for a given set of data a metric for which the estimated minimum mean squared error is smallest (cf. Section 1.4). This means that we construct a metric that makes the solution of the regression as easy as possible and thereby generate a notion of similarity on the data set. Furthermore we get information about how difficult the regression problem is.

An issue that complicates the analysis in medical applications is that the follow-up program of the patients treated in a hospital may be incomplete. Frequently, due to a whole range of reasons (medical reasons, personal reasons, etc.), a patient drops out of the program. Thus, after a certain point in time, there is no information any longer about the state of health of the patient any longer. In regression theory this is known as censoring. We have to estimate what happened during the time for which no information is available. A very important tool in handling this situation is the product-limit estimator, which is a well-known tool in estimating the survival time in the censoring model. With the aid of the product-limit estimator the approach described above can be extended to the censoring model and a suitable metric in the case information is lost by censoring can be found (cf. Section 2.4).

A medical application is at the core of Chapter Three. Based on two sets of data of of breast cancer patients, a suitable metric for predictors used in the treatment of the patients is computed. Three groups of predictors can be distinguished. Among the predictors with the greatest weight are CERB (amount of cerb2 proteine in the cell membrane of the tumor), GR (grading of the tumor) and PN (number of affected lymph nodes).

Chapter 1

Estimation of Minimum Mean Squared Error

In this work we are interested in finding a metric in $\mathbb{R}^{\mathrm{d}}$ by estimating the minimum mean squared error. There have been approaches to estimate it via estimation of the regression function (e.g. in the case of a binary response variable in Devroye, Györfi and Lugosi (1996) [6], chapter 26). Estimation based on consistent estimators suffers from the setback that the rate of convergence cannot exceed $n^{-\frac{2}{d+2}}$. In this chapter estimators based on non-consistent estimators that have a better rate of convergence are discussed. It turns out that a suitable estimator should be stable in the sense that an erroneous observation may have only small impact on the value of the estimator. An estimator based on partitioning and First-NN estimation is constructed and proved to have a better rate of convergence than estimators based on a consistent estimator. In general the rate of convergence is $n^{-\frac{2}{d+1}}$ for $d \geq 2$ and $n^{-\frac{2}{3}}$ for $d = 1$.

1.1 Nonparametric Regression

In regression analysis one considers a sequence of d-dimensional observation vectors $X_1, \ldots, X_n$ and corresponding real-valued response variables $Y_1, \ldots, Y_n$. We assume the random variables in both sequences to be independent and identically distributed (i.i.d) copies of X and Y. The d coordinates of $X = (X^{(1)}, \ldots, X^{(d)})$ are called predictors or covariates. The problem in regression analysis is to find a (Borel-)measurable function $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ that is 'close' to the response variable Y. It is a well-known fact that conditional expectation, also called regression function,

$$m(x) = E\,(Y|X = x)$$

is closest to Y in the sense that

$$m = \arg\min_f E\,(Y - f(X))^2$$

(cf. Bauer (1991) [2]). Therefore m has the smallest mean squared error

$$L^* = E\,(Y - m(X))^2$$

among all measurable functions. L^* is a measure of how close we can get to Y using any measurable function. It also indicates how difficult a regression problem is. Large L^* means that no matter how close we get to m, the expected deviation from the response variable will still be large.

As the distribution (X, Y) is generally not known, the regression function m is not calculable. Therefore, in nonparametric regression, we are looking for estimators $m_n\,(x|D_n)$ (where $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a given set of data) that are good approximations of m regardless of the distribution of (X, Y). This idea motivates the notion of universal consistency.

Definition: 1.1 *A sequence of estimators* $\{m_n\}$ *is called weakly universally consistent if for all distributions* (X, Y) *with* $EY^2 < \infty$ *we have*

$$E\left\{\int_{\mathbb{R}} (m_n(x) - m(x))^2\,\mu(dx)\right\} \to 0.$$

Stone proved in 1977 (cf. [20]) that there are universally consistent estimators. He used a nearest neighbour estimator in the proof and thus proved universal consistency for this class of estimators. Since then, similar results concerning strong consistency have been proved for many regression estimates (cf. Devroye and Györfi (1983) [4], Devroye and Krzyzak (1989) [8], Devroye et al. (1994) [5] and Walk (2005) [21]). See Györfi et al. (2002) [14] for further references.

However, to derive rates of convergence one has to impose certain conditions on the distribution of (X, Y) as the convergence can be arbitrarily slow (cf. Györfi et al (2002) [14]). Again Stone proved in 1982 that the optimal rate of convergence in case of a (p, Γ)-smooth regression function is $n^{-\frac{2p}{2p+d}}$. If m is Lipschitz-continuous the optimal rate is $n^{-\frac{2}{2+d}}$ (cf. Stone [19]).

Most of nonparametric deals with estimation of the regression function. In case of estimating a metric, it can be argued that it is more reasonable to estimate L^* which indicates how difficult a regression problem is. Thus we are shifting our focus from m to L^*. The next lemma states results on L^* that are needed in the study of how to estimate the minimum mean squared error. Note that c) even holds for bijective measurable transformations of $\mathbb{R}^{\mathrm{d}}$.

Lemma 1.1 *a) One has* $E\left(Y - f(X)\right)^2 \geq L^*$ *for any measurable function* f. *Therefore* L^* *is a lower bound for the expected deviation of any regression estimate from* Y.

b) One has $EY^2 \geq Em(X)^2$.

c) L^* *is invariant under dilation of* $\mathbb{R}^{\mathrm{d}}$ *with positive weights.*

Proof: a) See Theorem 15.8, Bauer (1991) [2] .

b) This follows from

$$\begin{aligned} L^* &= E\left(Y - m(X)\right)^2 \\ &= E\left(Y^2 - 2m(X)Y + m(x)^2\right) \\ &= E(Y^2 - 2\underbrace{E(m(X)Y|X)}_{=m(X)^2} + m(x)^2) \\ &= EY^2 - Em(X)^2, \end{aligned}$$

as $m(x)$ is measurable.

c) Using weights $h = (h_1, \ldots, h_d)$ (here $h_i > 0$ for $i = 1, \ldots, d$) and thus dilating the i-th coordinate of x by h_i, we define

$$x^h := (h_1 x_1, \ldots, h_d x_d).$$

Let $m(x) = E(Y|X = x)$ be the regression function with the respect to X and Y, and $m^h(x) = E\left(Y|X^h = x\right)$ be the regression function in the dilated model with $X^h = \left(h_1 X^{(1)}, \ldots, h_d X^{(d)}\right)$. We have with $x = (x_1, \ldots, x_n)$

$$\begin{aligned} m^h(x) &= E\left(Y|X^h = x\right) \\ &= E\left(Y|X = (\frac{x_1}{h_1}, \ldots, \frac{x_d}{h_d}\right) \\ &= m\left(\frac{x_1}{h_1}, \ldots, \frac{x_d}{h_d}\right). \end{aligned}$$

As a consequence and because of the Transformation Theorem for Measures (cf. Bauer (1991) [1], chapter 19) we have

$$\begin{aligned} E(m^h(X^h)^2) &= \int_{\mathbb{R}^d} m^h(x)^2 P_{X^h}(dx) \\ &= \int_{\mathbb{R}^d} m\left(\frac{x_1}{h_1}, \ldots, \frac{x_d}{h_d}\right)^2 P_{X^h}(dx) \\ &= \int_{\mathbb{R}^d} m(x_1, \ldots, x_d)^2 P_X(dx) \\ &= E\left(m(X)^2\right). \end{aligned}$$

Using part b, we arrive at

$$L^* = E(Y^2) - m^h(X^h)^2. \square$$

1.2 Existing Approaches

In the selection process of the metric it is crucial to have a good estimator of L^*. This section gives an overview of the different approaches discussed in the literature. Two different types of estimators can be distinguished. The first uses a consistent regression estimate and the second makes use of a nonconsistent one. It can be proved (cf. Györfi

et al. (2002) [14]) that the convergence

$$E\{\int (m_n(x) - m(x))^2 \mu(dx)\} \to 0$$

may be arbitrarily slow. If the regression function is Lipschitz-continuous, a regression estimate can converge at the most with an order of $n^{-\frac{2}{2+d}}$ against m (cf. Stone (1982) [20]). This rate is for example attained by the partitioning estimate with side length $h_n \approx n^{-\frac{1}{d+2}}$ and the k_n nearest neighbour estimate with $k_n \approx n^{\frac{2}{2+d}}$ (in the latter case $d \geq 2$).

The rate of convergence of the regression estimate m_n against the regression function m turns out to be crucial for the convergence of an estimator L_n , based on a consistent estimate, against L^*. Devroye et al. (2003) [7] (cf. Theorem 3.1) proved for the splitting of the sample technique that for any weakly consistent estimator m_n the rate of convergence of the so-called empirical L^2 risk

$$L_n = \frac{1}{n} \sum_{i=n+1}^{2n} (Y_i - m_n(X_i))^2.$$

cannot exceed $n^{-\frac{2}{2+d}}$ ($d \geq 2$) and $n^{-\frac{1}{2}}$ ($d = 1$).

A similar result holds for the cross-validation method. It is proved here for cross-validation based on the k_n nearest neighbour rule

$$m_{k_n}(x) = \frac{1}{k_n} \sum_{k=1}^{k_n} \left(Y_{(k,n)} - Y_k\right)^2 .$$

with tie occurring with probability zero. Here $Y_{(k,n)}$ is the Y_j corresponding to the k_n nearest neighbour X_j of X_k in a set of data D_n . Given D_n , we set

$$D_{in} := D_n \setminus (X_i, Y_i).$$

and define the k_{n-1} nearest neighbour rule $m_{ik_{n-1}}$ that leaves out the i -th observation

$$m_{ik_{n-1}}(X_i) := m_{k_{n-1}}(X_i | D_{in}).$$

The cross-validation estimate is defined as

$$CV_{k_n} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{ik_{n-1}}(X_i))^2.$$

Theorem 1.1 *For the k_n nearest neighbour rule with $k_n \sim n^{\frac{2}{2+d}}$ we have*

$$E\left|CV_{k_n} - L^*\right| \leq max\left\{n^{-\frac{1}{2}}, n^{-\frac{2}{d+2}}\right\},$$

if m is Lipschitz-continuous, $Var(Y|X=x)$ is bounded and ties occur with probability 0.

PROOF: By using Proposition $A.1$ one proves

$$\text{bias}\left(CV_{k_n}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(m_{ik_{n-1}}(X_i) - m(X_i)\right)^2.$$

We then have

$$E\left|CV_{k_n} - L^*\right| \leq E\left|CV_{k_n} - ECV_{k_n}\right| - bias\left(CV_{k_n}\right).$$

For the first term one uses McDiarmid's inequality (Theorem $A.1$) and for the second the fact that $m_{k_{n-1}}$ attains its optimal rate of convergence for the given choice of k_{n-1}.□

For estimators based on a consistent regression estimate, the convergence $m_n \to m$ is limiting the performance of the estimators. To avoid this problem a new type of estimator can be constructed that is not fit to estimate the regression m but is used to construct a direct estimate of L^*. The idea in Devroye et al. (2003) [7] is to estimate L^* by modifying the first nearest neighbour estimate in order to avoid X_i being the first nearest neighbour to more than one X_j.

Define for a set of data D_n a permutation $\pi_n(\cdot, D_n)$ on $\{1, ..., n\}$ with $\pi_n(i, D_n) \neq i$ by splitting the nearest neighbour graph into smaller graphs where nodes have only one predecessor. The estimator

$$\hat{L}_n := \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Y_{\pi_n(i)})^2. \tag{1.1}$$

achieves a rate of convergence of $n^{-\frac{1}{2}}$ for $d = 2, 3, 4$ and $n^{-\frac{2}{d}}$ for $d \geq 5$ (cf. Devroye et al. (2003) [7], Theorem 4.1, Györfi et al. (2002) [14], Lemma 6.4; as to d=2 see Liitiäinen et al. (2007) [16], section 2.2, especially Proposition 2.3 and taking expectation there). In addition the estimator has nonnegative bias.

Lemma 1.2 $\text{bias}\left(\hat{L}_n^h\right) = \frac{1}{2n}\sum_{i=1}^n E(m(X_i) - m(X_{\pi_n^h(i)}))^2$

PROOF: For any $i \in \{1, ..., n\}$ we have

$$\begin{aligned}
& E\left(Y_i - Y_{\pi_n(i)}\right)^2 \\
= \; & E\left((Y_i - m(X_i) + m(X_i)) - \left(Y_{\pi_n^h(i)} - m(X_{\pi_n^h(i)}) - m(X_{\pi_n^h(i)})\right)\right)^2 \\
= \; & E\left(Y_i - m\left(X_i\right)\right)^2 + Em(X_i)^2 - 2E\left(m(X_i)\cdot m(X_{\pi_n^h(i)})\right) \\
& + E\left(Y_{\pi_n^h(i)} - m(X_{\pi_n^h(i)})\right)^2 + E\left(m(X_{\pi_n^h(i)})^2\right) \\
= \; & E\left(Y_i - m(X_i)\right)^2 + E\left(m(X_i) - m(X_{\pi_n^h(i)})\right)^2 + E\left(Y_{\pi_n^h(i)} - m(X_{\pi_n^h(i)})\right)^2,
\end{aligned}$$

where the second equality follows from the fact that for every $k \in \{1, ..., n\}$ we have

$$E\left(Y_k - m\left(X_k\right) - m\left(X_k\right)\right)^2 \;=\; E\left(Y_k - m\left(X_k\right)\right)^2 + E\left(m(X_i)\right)^2$$

and

$$E\left(Y_i \cdot Y_{\pi_n^h(i)}\right) = E\left(m(X_i)\cdot m(X_{\pi_n^h(i)})\right).$$

Since π_n^h is a permutation, it is clear that

$$\sum_{i=1}^n E\left(Y_i - m\left(X_i\right)\right)^2 = \sum_{i=1}^n \left(Y_{\pi_n^h(i)} - m(X_{\pi_n^h(i)})\right)^2.$$

This results in

$$\begin{aligned}
E\hat{L}_n^h &= \frac{1}{2n}\sum_{i=1}^n \left(2L^* + E\left(m(X_i) - m(X_{\pi_n^h(i)})\right)^2\right) \\
&= L^* + \frac{1}{2n}\sum_{i=1}^n E\left(m(X_i) - m(X_{\pi_n^h(i)})\right)^2.\square
\end{aligned}$$

Here, the rate of convergence is better than in the cases of splitting the sample and cross-validation. However, $\hat{L}_n^h$ turns out to be unstable in the sense that a small change in one covariate may change the whole permutation and $\hat{L}_n^h$ considerably. Therefore, small changes, as induced by changing one parameter in this setting, may result in large changes of $\hat{L}_n^h$ which may influence the convergence in a negative way . As a consequence, at this point no exponential inequality for bounding the variance of $\hat{L}_n^h$ is available.

Remark 1.1 Since the rate of convergence of the estimator $\hat{L}_n^h$ is better, a study similar to one undertaken in Dippon, Fritz and Kohler (2002) [9] (cf. Table 3) could be done with it. One considers subsets of the whole set of predictors and calculates $\hat{L}_n^h$ on the basis of this subset. The subset with the lowest estimated L_2 error then contains the most important predictors. However, it is difficult to assign a numerical value this way.

1.3 Balancing First–NN and Partitioning Estimate

In this section a new estimator is introduced that has a better rate of convergence than estimators based on cross-validation or splitting of the sample technique. Furthermore, it is stable in the sense that changes in one observation vector have only local impact. For the following we assume that ties occur with probability zero. This is an assumption on the distribution of X. However, it can be satisfied by including a uniformly distributed random variable that is independent from the variables in the predictor set.
To define the estimator $\tilde{L}_n$ we use a combination of simplified partitioning and 1-NN estimation. Let $\|\cdot\|$ be the Euclidean Norm given by

$$\| x \| = \sqrt{x_1^2, \cdots + x_d^2},$$

with $x = (x_1, \ldots, x_d)$ and $|\cdot|$ be the Euclidean norm on $\mathbb{R}$ (i.e. the absolute value). Assume compact support of $\mu := P_X$ with respect to $\|\cdot\|$, bounded Y and Lipschitz continuity of m and σ^2 with respect to $\|\cdot\|$. Here

$$\sigma^2(x) := E\left((Y - m(X))^2 \middle| X = x\right)$$

is the local variance. Let

$$\{A_{n,k} : k \in \mathbb{N}\} \tag{1.2}$$

be a cubic partition of $\mathbb{R}^{\mathrm{d}}$ with respect to $\|\cdot\|$ with cubes of side length

$$\tilde{h}_n = n^{-1/(1+d)}. \tag{1.3}$$

Denote the $A_{n,j}$ with $x \in A_{n,j}$ by $A_n(x)$. Let $i \in \{1, \ldots, n\}$. If X_i is the only random variable inside $A_n(X_i)$, then $j(i)$ is the 1-NN of X_i among $\{X_1, \ldots, X_n\} \backslash \{X_i\}$. If more

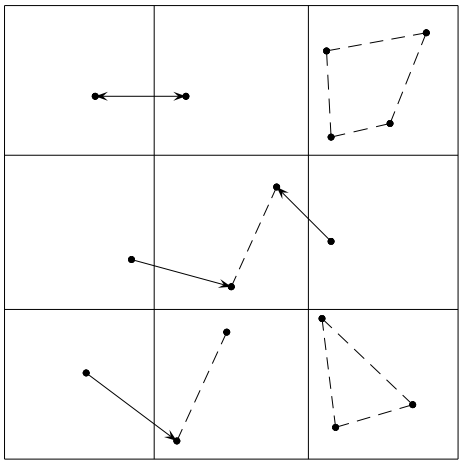

Figure 1.1: Among nodes connected with dashed lines the WSNN is chosen at random. Arrows indicate 1 NN selection.

than one random variable is inside $A_n(X_i)$, $j(i)$ is chosen at random among the other random variables inside $A_n(X_i)$ (cf. Figure 1.1). More precisely, let U be independent of X and uniformly distributed over $[0,1]$. Let $U_i, \dots U_n$ be independent copies of U. Furthermore, let $\{X_i, X_{i_1}, \dots X_{i_r}\}$ be the random variables inside $A_n(X_i)$. $j(i)$ is then chosen as the index of the 1-NN of U_i among $\{U_{i_1}, \dots U_{i_r}\}$. Set

$$X_{i,n} := X_{j(i)} \tag{1.4}$$

$$m_n(X_i) := Y_{i,n} := Y_{j(i)} \tag{1.5}$$

$$\tilde{L}_n := \frac{1}{2n} \sum_{i=1}^{n} \left(m_n(X_i) - Y_i\right)^2. \tag{1.6}$$

$X_{i,n}$ will be called later *wide sense nearest neighbour* (WSNN) of X_i.

Theorem 1.2 *Assume $|Y_i| \le B$ for $i \in \{1, \dots, n\}$. Furthermore, assume ties occur with probability zero, compact support M_μ of μ and Lipschitz-continuity of m and σ^2 with respect to $\|\cdot\|$. Then we have*

$$E\left|\tilde{L}_n - L^*\right| \le \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \ge 2. \end{cases}$$

We prove the theorem with the aid of two lemmas.

Lemma 1.3 *Assume* $|Y_i| \leq B$ *for any* $i \in \{1, \ldots, n\}$, *and ties occur with probability zero. Then one has*

$$\begin{aligned} a) \quad & P\left(\left|\tilde{L}_n - E\tilde{L}_n\right| \geq \epsilon\right) \leq 2\exp\left(\tfrac{-\epsilon^2 n}{2B^4(5+2\gamma_d)^2}\right) \\ b) \quad & E\left|\tilde{L}_n - E\tilde{L}_n\right| \leq \sqrt{2(1+log2)}B^2(5+2\gamma_d)n^{-\frac{1}{2}}. \end{aligned}$$

PROOF OF PART a: For arbitrary $l \in \{1, \ldots, n\}$ we show

$$\sum_{i=1;i\neq l}^{n} I_{\left\{X_l \text{ is WSNN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}\right\}} \leq \gamma_d + 2 \text{ a. s.} \tag{1.7}$$

Let

$$\left\{X_{i_1}, \ldots, X_{i_{r(i)}}\right\} = A_n(X_i) \cap (\{X_i, \ldots, X_n\} \setminus \{X_i\})$$

be all the random variables inside $A_n(X_i)$. The right hand side of equation (1.7) is equal to

$$\begin{aligned} & \sum_{i=1;i\neq l}^{n} \Big(I_{\{\text{none of } X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n \text{ is an element of } A_n(X_i)\}} \\ & \cdot I_{\{X_l \text{ is 1-NN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}\}} \\ & + I_{\{\text{at least one of } X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n \text{ is an element of } A_n(X_i)\}} \\ & \cdot I_{\left\{U_l \text{ is 1-NN of } \tilde{U}_i \text{ in } \left\{U_{r_i},\ldots,U_{i_{r(i)}}\right\}\right\}} \Big) \\ \leq \; & I_{\{X_l \text{ is 1-NN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}\}} \\ & + I_{\left\{U_l \text{ is 1-NN of } \tilde{U}_i \text{ in } \left\{U_{r_i},\ldots,U_{i_{r(i)}}\right\}\right\}} \\ \leq \; & \gamma_d + \gamma_1 \\ = \; & \gamma_d + 2. \end{aligned}$$

Here γ_d is the number of cones centered at 0 with angle $\theta = \frac{\pi}{3}$ that are needed to cover $\mathbb{R}^{\mathrm{d}}$. The last two equations follow from Corollary 6.1 in Györfi et al. (2002) [14] as ties occur with probability zero with respect to X_i and U_i for $i \in \{1, \ldots, n\}$. By the definition of γ_d we have $\gamma_1 = 2$.

We use McDiarmid's inequality (Theorem $A.1$) and show that for any $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$

$$\sup_{\bar{D}_n, D_n} \left|\tilde{L}_n(D_n) - \tilde{L}_n(\bar{D}_{n,l})\right| \leq \frac{2B^2(5+2\gamma_d)}{n} \text{ a. s.} \tag{1.8}$$

Here $\bar{D}_{n,l} = (\bar{X}_1, \bar{Y}_1), ..., (\bar{X}_n, \bar{Y}_n)$ differs only with respect to $(\bar{X}_l, \bar{Y}_l)$ from D_n:

$$(\bar{X}_k, \bar{Y}_k) = (X_k, Y_k) \quad \text{for k} \in \{1, ..., n\} \setminus \{l\}$$

Let $\bar{m}_n(X_i)$ be the Y_j for which X_j is WSNN to X_i among

$$X_1, X_2, \ldots, X_{l-1}, \bar{X}_l, X_{l+1}, \ldots, X_n.$$

The absolute value on the right side of (1.8) can be calculated as

$$\begin{aligned}
& \frac{1}{2n} \left| \sum_{i=1}^{n} \left(Y_i - m_n\left(X_i\right)\right)^2 - \sum_{i=1}^{n} \left(\bar{Y}_i - \bar{m}_n\left(\bar{X}_i\right)\right)^2 \right| \\
= \; & \frac{1}{2n} \Bigg(\left| \left(Y_l - m_n\left(X_l\right)\right)^2 - \left(\bar{Y}_l - \bar{m}_n\left(\bar{X}_l\right)\right)^2 \right| \\
& + \left| \sum_{i=1;i\neq l}^{n} \left(Y_i - m_n\left(X_i\right)\right)^2 - \sum_{i=1;i\neq l}^{n} \left(Y_i - \bar{m}_n\left(\bar{X}_i\right)\right)^2 \right| \Bigg) \\
= \; & \frac{1}{2n} \Bigg(\left| \left(Y_l - m_n\left(X_l\right)\right)^2 - \left(\bar{Y}_l - \bar{m}_n\left(\bar{X}_l\right)\right)^2 \right| \\
& + \left| \sum_{i:m_n(X_i)=Y_l} \left(Y_i - Y_l)\right)^2 - \left(Y_i - \bar{m}_n\left(\bar{X}_i\right)\right)^2 \right| \\
& + \left| \sum_{i:\bar{m}_n(X_i)=\bar{Y}_l} \left(Y_i - m_n(X_i))\right)^2 - \left(Y_i - \bar{Y}_l\right)^2 \right| \Bigg) \\
\leq \; & \frac{1}{2n} \left(4B^2 + 4B^2 \sum_{i=1}^{n} I_{\{i:m_n(X_i)=Y_l\}} + 4B^2 \sum_{i=1}^{n} I_{\left\{i:\bar{m}_n(X_i)=\bar{Y}_l\right\}} \right) \\
& \quad (\text{as } |Y_i| \leq B \text{ for all } i \in \{1, \ldots, n\}) \\
\leq \; & \frac{1}{2n} \left(4B^2 + 2 \cdot 4B^2 \left(\gamma_d + 2\right)\right) \quad (\text{by (1.7)}) \\
= \; & \frac{2B^2}{n} \left(5 + 2\gamma_d\right) \text{ a. s.}
\end{aligned}$$

Since the random vectors $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent and identically distributed, we obtain with McDiarmid's inequality (Theorem $A.1$) and (1.8)

$$P\left(\left|\tilde{L}_n - E\tilde{L}_n\right| \geq \epsilon\right) \leq 2 \exp\left(\frac{-\epsilon^2 n}{2B^4(5 + 2\gamma_d)^2}\right).$$

Proof of Part b: For any random variable Z with

$$P\left(Z > \epsilon\right) \leq C \exp\left(\frac{-\epsilon^2 n}{c}\right) \quad \text{for all} \epsilon > 0,$$

provided $C > 1$ and $c > 0$, we have

$$EZ \leq \sqrt{\frac{1 + \log C}{n} \cdot c}$$

(compare Györfi et al (2002) [14], problem 8.2). Putting $Z := |\tilde{L}_n - E\tilde{L}_n|$ we derive

$$E\left|\tilde{L}_n - E\tilde{L}_n\right| \leq \sqrt{\frac{2(1+log2)}{n}} B^2 \left(5 + 2\gamma_d\right). \square$$

Lemma 1.4 *Assume compact support M_μ of μ, and Lipschitz-continuity of m and σ^2 with constants k_1 and k_2 with respect to $\|\cdot\|$. For any $i \in \{1, ..., n\}$ one has*

$$a)\ \left|E\sigma^2(X_{i,n}) - E\sigma^2(X_i)\right| \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

$$b)\ E\left(m\left(X_{i,n}\right) - m\left(X_i\right)\right)^2 \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

Proof: For $\epsilon > 0$ and $l \in \{1, \ldots, n\}$ we show for some constant $c \geq 0$

$$\begin{aligned} & P\left(\| X_{l,n} - X_l \| > \epsilon; X_j \notin A_n(X_l) \text{ for } j \in \{1, \ldots, n\} \setminus \{l\}\right) \\ \leq & \begin{cases} \frac{c}{n \cdot \tilde{h}_n^d} & ; \quad generally \\ \frac{c}{n\epsilon^d} & ; \quad \epsilon > \tilde{h}_n \end{cases} \end{aligned} \tag{1.9}$$

Here $\tilde{h}_n$ is the side length of the cubic partition (cf. (1.3)). For the general case we expand the left-hand side to

$$\begin{aligned} & P\left(\ \| X_j - X_l \| > \epsilon; X_j \notin A_n(X_l) \text{ for } j \in \{1, \ldots, n\} \setminus \{l\}\right) \\ = & \int\limits_{\mathbb{R}^d} P\left(\ \| X_j - x \| > \epsilon; X_j \notin A_n(x) \text{ for } j \in \{1, \ldots, n\} \setminus \{l\}\right) \mu(dx) \\ = & \int\limits_{\mathbb{R}^d} P\left(\ \| X - x \| > \epsilon; X \notin A_n(x)\right)^{n-1} \mu(dx), \end{aligned}$$

because $X_1, \ldots, X_n$ are independent copies of X.

For the cubic partition $\{A_{n,1}, \ldots, A_{n,N_n}\}$ of the bounded support of μ (with cubes of side length $\tilde{h}_n$) we have

$$N_{\tilde{h}_n} = \frac{c_{\tilde{h}_n}}{\tilde{h}_n^d}$$

satisfying $\sup c_{\tilde{h}_n} < \infty$. Observe, that for any $x \in [0,1]$ we have

$$x(1-x)^{n-1} \leq \frac{1}{n}.$$

As a consequence and since M_μ is compact, it follows

$$\begin{aligned}
&\int_{\mathbb{R}^d} P\left(\parallel X - x \parallel > \epsilon; X \notin A_n(x)\right)^{n-1} \mu(dx) \\
\leq\ &\int_{\mathbb{R}^d} \left(1 - \mu\left(A_n\left(x\right)\right)\right)^{n-1} \mu\left(dx\right) \\
=\ &\sum_{j=1}^{N_{\tilde{h}_n}} \int_{A_{n,j}} \left(1 - \mu\left(A_n\left(x\right)\right)\right)^{n-1} \mu\left(dx\right) \\
=\ &\sum_{j=1}^{N_{\tilde{h}_n}} \underbrace{\mu\left(A_{n,j}\right)\left(1 - \mu\left(A_{n,j}\right)\right)^{n-1}}_{\leq \frac{1}{n}} \\
\leq\ &\frac{N_{\tilde{h}_n}}{n} \\
=\ &\frac{c_{\tilde{h}_n}}{n \cdot \tilde{h}_n^d}.
\end{aligned}$$

Now, let $\epsilon > \tilde{h}_n$ and set

$$B^\epsilon(x) := \left\{x' \in \mathbb{R}^d : |x - x'| > \epsilon\right\} \tag{1.10}$$

Let $B_1^\epsilon, \ldots, B_{N(\epsilon)}^\epsilon$ be the cubic partition of the bounded support of μ with sets as defined by (1.10). Furthermore, define $c_\epsilon \geq 0$ by

$$N_\epsilon = \frac{c_\epsilon}{\epsilon^d} \tag{1.11}$$

with some $c_\epsilon \geq 0$. It holds $\sup_{\epsilon>0} c_\epsilon < \infty$ Then by definition of B^ϵ and

$$\begin{aligned}
&\int_{\mathbb{R}^d} P\left(\parallel X - x \parallel > \epsilon; X \notin A_n(x)\right)^{n-1} \mu(dx) \\
\leq\ &\int_{\mathbb{R}^d} P\left(X \notin B^\epsilon(x)\right)^{n-1} \mu(dx) \\
=\ &\sum_{j=1}^{N_\epsilon} \int_{B_j^\epsilon} P\left(X \notin B_j^\epsilon\right)^{n-1} \mu(dx)
\end{aligned}$$

$$\begin{aligned} &= \sum_{j=1}^{N_\epsilon} \underbrace{\mu(B_j^\epsilon)(1-\mu(B_j^\epsilon))^{n-1}}_{\leq \frac{1}{n}} \\ &\leq \frac{N_\epsilon}{n} \\ &= \frac{c_\epsilon}{n\epsilon^d}. \end{aligned}$$

Proof of part a: For $i \in \{1,\dots,n\}$

$$\begin{aligned} & E\sigma^2(X_{i,n}) - E\sigma^2(X_i) \\ = \; & E\left(\left(\sigma^2(X_{i,n}) - \sigma^2(X_i)\right) \cdot I_{\{X_j \notin A_n(X_i) \text{ for } j\in\{1,\dots,n\}\setminus\{i\}\}}\right) \\ & + E\left(\left(\sigma^2(X_{i,n}) - \sigma^2(X_i)\right) \cdot I_{\{X_j \in A_n(X_i) \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\}}\right) \\ =: \; & A + B. \end{aligned}$$

As σ^2 is Lipschitz-continuous with a constant k_2 we have for $d \geq 2$ and $\bar{\epsilon} := diam\{M_\mu\}$

$$\begin{aligned} |A| &\leq k_2 \cdot E\left(\|X_{i,n} - X_i\| \cdot I_{\{X_j \notin A_n(X_i) \text{ for } j\in\{1,\dots,n\}\setminus\{i\}\}}\right) \\ &= k_2 \int_0^\infty P\left(\; \|X_{i,n} - X_i\| \cdot I_{\{X_j \notin A_n(X_i) \text{ for } j\in\{1,\dots,n\}\setminus\{i\}\}} > \epsilon\right) d\epsilon \\ &= k_2 \int_0^\infty \cdot P\left(\|X_{i,n} - X_i\| > \epsilon; X_j \notin A_n(X_i) \text{ for } j \in \{1,\dots,n\}\setminus\{i\}\right) d\epsilon \\ &= k_2 (\underbrace{\int_0^{\tilde{h}_n} P\left(\|X_{i,n} - X_i\| > \epsilon; X_j \notin A_n(X_i) \text{ for } j \in \{1,\dots,n\}\setminus\{i\}\right) d\epsilon}_{\leq \tilde{h}_n \cdot \frac{c_{\tilde{h}_n}}{n\cdot\tilde{h}_n^d} \text{ by part one of (1.9)}} \\ &\quad + \int_{\tilde{h}_n}^\infty \underbrace{P\left(\|X_{i,n} - X_i\| > \epsilon; X_j \notin A_n(X_i) \text{ for } j \in \{1,\dots,n\}\setminus\{i\}\right)}_{\leq \frac{c_\epsilon}{n\cdot\epsilon^d} \text{ by part two of (1.9)}} d\epsilon) \\ &\leq \frac{k_2}{n}\left(c_{\tilde{h}_n}\tilde{h}_n^{1-d} + \bar{c}_\epsilon \int_{\tilde{h}_n}^\infty \frac{1}{\epsilon^d} d\epsilon\right) \\ &\qquad (\text{with } \bar{c}_\epsilon := \sup\{c_\epsilon : \tilde{h}_n \leq \epsilon \leq \bar{\epsilon}\}) \qquad (1.12) \\ &\leq \frac{k_2}{n}\left(c_{\tilde{h}_n}\tilde{h}_n^{1-d} \quad + \quad \bar{c}_\epsilon \left[\frac{1}{1-d}\epsilon^{1-d}\right]_{\tilde{h}_n}^\infty\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{k_2}{n}\tilde{h}_n^{1-d}\left(c_{\tilde{h}_n}+\frac{1}{d-1}\bar{c}_\epsilon\right)\\
&= k_2 n^{-\frac{2}{d+1}}\left(c_{\tilde{h}_n}+\frac{1}{d-1}\bar{c}_\epsilon\right) \quad \text{(by side length of } \tilde{h}_n\text{)}. \qquad (1.13)
\end{aligned}
$$

We prove $B=0$. On the one hand we derive

$$
\begin{aligned}
&E\left(\sigma^2(X_{i,n})\cdot I_{\{X_j\in A_n(X_i) \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\}}\right)\\
&= \sum_{l=1}^{N_n} E\left(\sigma^2(X_{i,n})\cdot I_{\{X_j\in A_{n,l} \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\}}\cdot I_{\{X_i\in A_{n,l}\}}\right)\\
&= \sum_{l=1}^{N_n} E\left(\sigma^2(X_{i,n})\middle| X_i\in A_{n,l}; X_j\in A_{n,l} \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\right)\\
&\quad \cdot P(X_i\in A_{n,l})\cdot P(X_j\in A_{n,l} \text{ for some } j\in\{1,\dots,n\}\setminus\{i\})\\
&= \sum_{l=1}^{N_n}\frac{\int_{A_{n,l}}\sigma^2(x)\,\mu(dx)}{\mu(A_{n,l})}\cdot\mu(A_{n,l})\left(1-(1-\mu(A_{n,l}))^{n-1}\right).
\end{aligned}
$$

On the other hand

$$
\begin{aligned}
&E\left(\sigma^2(X_i)\cdot I_{\{X_j\in A_n(X_i) \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\}}\right)\\
&= \int_{\mathbb{R}^d}\sigma^2(x)\cdot EI_{\{X_j\in A_n(X_i) \text{ for some } j\in\{1,\dots,n\}\setminus\{i\}\}}\mu(dx)\\
&= \int_{\mathbb{R}^d}\sigma^2(x)\left(1-(1-\mu(A_n(X_i)))^{n-1}\right)\mu(dx)\\
&= \sum_{l=1}^{N_n}\int_{A_{n,l}}\sigma^2(x)\,\mu(dx)\left(1-(1-\mu(A_{n,l}))^{n-1}\right).
\end{aligned}
$$

Therefore, (1.13) shows the desired rate of convergence in the case $d\geq 2$.

In the case of $d=1$ enlarge $x\in\mathbb{R}$ with some $a\in\mathbb{R}$ to become

$$
\hat{x} := (x,a). \qquad (1.14)
$$

Let $D_n := \{(X_1,Y_1),\dots,(X_n,Y_n)\}$, then $\hat{X}_1,\dots,\hat{X}_n$ are $\mathbb{R}^2$-valued random variables. As ties occur with probability zero with respect to $X_1,\dots,X_n$, this is also true for $\hat{X}_1,\dots,\hat{X}_n$. Furthermore for any $i\in\{1,\dots,n\}$ the index of the WSNN of X_i among $X_1,\dots,X_{i-1},X_{i+1},X_n$ is the same as of $\hat{X}_i$ among $\hat{X}_1,\dots,\hat{X}_{i-1},\hat{X}_{i+1},\hat{X}_n$. Let for any

$z \in \mathbb{R}^2$ $\hat{m} : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$\hat{m}(z) := E\left(Y|\hat{X} = z\right). \tag{1.15}$$

Then $m(x) = \hat{m}(\hat{x})$ for any $x \in \mathbb{R}$. Consequently, one obtains by what has been shown above for $d \geq 2$

$$\begin{aligned} E\left(m(X_{i,n}) - m(X_i)\right)^2 &= E\left(\hat{m}(\hat{X}_{i,n}) - m(\hat{X}_i)\right)^2 \\ &\leq n^{-\frac{2}{3}} \quad \text{(as d=2)}. \end{aligned} \tag{1.16}$$

This completes the proof of part a.

Proof of Part b: By Lipschitz-continuity of m and the boundedness of the support of μ ,we have for $d \geq 2$ as ties occur with probability zero

$$\begin{aligned} & E\left(m\left(X_{i,n}\right) - m\left(X_i\right)\right)^2 \\ \leq\ & k_1^2 E\big(\underbrace{\|X_{i,n} - X_i\|^2 \cdot I_{\{X_j \notin A_n(X_i) \text{ for } j \in \{1,\ldots,n\}\setminus\{i\}\}}}_{\leq cn^{-\frac{2}{d}} \text{ (with some constant } c)}\big) \\ & \quad + k_1^2 E\big(\underbrace{\|X_{i,n} - X_i\|^2 \cdot I_{\{X_j \in A_n(X_i) \text{ for some } j \in \{1,\ldots,n\}\setminus\{i\}\}}}_{\leq \tilde{h}_n^2 \text{ by side length of } A_n(X_i)}\big) \end{aligned} \tag{1.17}$$

$$\leq\ cn^{-\frac{2}{d}} + cn^{-\frac{2}{d+1}}. \tag{1.18}$$

The bound of the product in the expectation of (1.17) is derived for $d = 2$ by using Proposition 2.3 in Liitiäinen et al. (2007) [16] and in the case of $d \geq 3$ with the aid of Lemma 6.4 in Györfi et al. (2002) [14].

For $d = 1$ we define (cf. (1.14) and (1.15)) for any $z \in \mathbb{R}^2$

$$\hat{\sigma}^2(z) := E\left(\left(Y_1 - \hat{m}(\hat{X}_1)\right)^2 \middle| \hat{X}_1 = z\right).$$

Then by the same reasoning as was used to derive (1.16), one obtains

$$\left|E\sigma^2(X_{i,n}) - E\sigma^2(X_i)\right| = \left|E\hat{\sigma}^2(\hat{X}_{i,n}) - E\sigma^2(\hat{X}_i)\right| \leq n^{-\frac{2}{3}}. \tag{1.19}$$

This completes the proof of part b. □

The rate of convergence $n^{-\frac{2}{3}}$, that was proved in Lemma 1.4, was derived by proving that the rate of convergence for $d = 2$ also holds for $d = 1$. A different technique may improve on this result. However, for practical purposes the low dimensional case is not too important. In many applications the dimension of the observation vectors is rather high.

PROOF OF THEOREM 1.2: First, we derive a decomposition for $E\left|\tilde{L}_n - L^*\right|$. Observe that we have for an $i \in \{1, \ldots, n\}$

$$\begin{aligned} E\left(Y_{i,n} - Y_i\right)^2 &= E\left(Y_{i,n} - m\left(X_{i,n}\right) + m\left(X_{i,n}\right) - Y_i\right)^2 \\ &= E\{\left(Y_{i,n} - m\left(X_{i,n}\right)\right)^2 + \left(m\left(X_{i,n}\right) - Y_i\right)^2\} \\ &= E\{\left(Y_{i,n} - m\left(X_{i,n}\right)\right)^2 + \left(m\left(X_{i,n}\right) - m\left(X_i\right) + m\left(X_i\right) - Y_i\right)^2\} \\ &= E\{\left(Y_{i,n} - m\left(X_{i,n}\right)\right)^2 + \left(m\left(X_{i,n}\right) - m\left(X_i\right)\right)^2 \\ &\quad + \left(m\left(X_i\right) - Y_i\right)^2\}. \end{aligned} \tag{1.20}$$

In this equation array we used from the first to the second line the fact that

$$\begin{aligned} & E\left(\left(Y_{i,n} - m\left(X_{i,n}\right)\right)\left(m\left(X_{i,n}\right) - Y_i\right)\right) \\ = \; & E\left(Y_{i,n} m\left(X_{i,n}\right) - Y_i Y_{i,n} - m\left(X_{i,n}\right)^2 + m\left(X_{i,n}\right) Y_i\right) \\ = \; & E\left(E\left(Y_{i,n} \middle| X_{i,n}\right) m\left(X_{i,n}\right) - Y_i Y_{i,n} - m\left(X_{i,n}\right)^2 + m\left(X_{i,n}\right) Y_i\right) \\ = \; & E\left(Y_i m\left(X_{i,n}\right) - m\left(X_i\right) m\left(X_{i,n}\right)\right. \\ & \quad \text{(by Proposition } A.1) \\ = \; & E\left(m\left(X_{i,n}\right) E(\underbrace{E\left(Y_i - m\left(X_i\right) \middle| X_i\right)}_{=0 \text{ a. s.}} | X_{i,n})\right) \\ = \; & 0. \end{aligned}$$

And from the third to the fourth line

$$\begin{aligned} E(m_{X_{i,n}}) - m(X_i))(m(X_i) - Y_i) &= E(m(X_{i,n})m(X_i) - m(X_i)Y_i) \\ &\quad + \underbrace{E(m(X_i)Y_i - m(X_i)^2)}_{=0} \\ &= E(m(X_{i,n})E(m(X_i) - Y_i|X_{i,n})) \\ &= 0. \end{aligned}$$

Using equation (1.20), it follows that

$$\begin{aligned} E\tilde{L}_n &= \frac{1}{2n}\sum_{i=1}^{n}\big\{E(Y_{i,n}-m(X_{i,n}))^2 \\ &\quad +E(m(X_{i,n})-m(X_i))^2+E(m(X_i)-Y_i)^2\big\} \\ &= \frac{1}{2}L^*+\frac{1}{2n}\sum_{i=1}^{n}E\sigma^2(X_{i,n}) \\ &\quad +\frac{1}{2n}\sum_{i=1}^{n}E(m(X_{i,n})-m(X_i))^2. \end{aligned} \tag{1.21}$$

Then we obtain

$$\begin{aligned} E|\tilde{L}_n-L^*| &\le E|\tilde{L}_n-E\tilde{L}_n|+|E\tilde{L}_n-L^*| \\ &\le E|\tilde{L}_n-E\tilde{L}_n|+\frac{1}{2n}\sum_{i=1}^{n}|E\sigma^2(X_{i,n})-E\sigma^2(X_i)| \\ &\quad +\frac{1}{2n}\sum_{i=1}^{n}E(m(X_{i,n})-m(X_i))^2. \end{aligned} \tag{1.22}$$

Lemma 1.3 part b and Lemma 1.4 are used on the sums on the right hand side. □

1.4 Approximation of an Optimal Metric

Different approaches of how to select a suitable metric based on set of data D_n are discussed in the literature. Generally, the idea is to estimate a $d \times d$ transformation matrix A. The distance of $x, y \in \mathbb{R}^{\mathrm{d}}$ is then $\| A^T(x-y) \|$, where $\| \cdot \|$ is the Euclidean norm. For estimation of A different methods were proposed (cf. Fukanaga and Flick (1984) [12] and Myles and Hand (1990) [18]). Devroye, Györfi and Lugosi proved in [6] (Theorem 26.3) a consistency result for the k_n nearest neighbour rule based on a metric of this type and a binary valued response variable. However no rates of convergence were derived.

In this section rates of convergence of the estimator $\tilde{L}_n$ based on a norm $\| \cdot \|$ to L^* are derived (cf. Theorem 1.3). Minimizing $\tilde{L}_n^h$ over a parameter space $\mathcal{Q}^d$ yields $\tilde{L}_n^{H_n}$. In Theorem 1.4 a rate of convergence of a distance between $\tilde{L}_n^{H_n}$ and $\tilde{L}_n^{\bar{h}_n}$ to zero is proved. Here $\tilde{L}_n^{\bar{h}_n}$ is the best approximation of L^* by any $\tilde{L}_n^h$ with $h \in \mathcal{Q}^d$.

For the following, metrics on $\mathbb{R}^{\mathrm{d}}$ are considered which are based on norms of the type

$$\| x \|_h := \sqrt{(h_1 x_1)^2 + \cdots + (h_d x_d)^2}, \tag{1.23}$$

with $x \in \mathbb{R}^{\mathrm{d}}$ and $h = (h_1, \ldots, h_d)$ with $h_i > 0$ for $i \in \{1, \ldots, d\}$. h_i is the factor by which the i-th coordinate of $x \in \mathbb{R}^{\mathrm{d}}$ is dilated. In applications h_i is the weight of the i-th predictor. It is crucial for the convergence of $\tilde{L}_n^h$ (cf. (1.4)-(1.6)) that ties occur with probability zero with respect to $\| \cdot \|_h$ and X (for a discussion of breaking distance ties see Devroye, Györfi and Lugosi (1996) [6], chapter 11). But this is true as long as $h_i > 0$ holds for all $i \in \{1, \ldots, d\}$ and ties occur with probability zero with respect to the Euclidean norm. The latter occurs if at least one predictor variable has a density.
Set for any norm $\| \cdot \|_h$ as in (1.23)

$$\tilde{L}_n^h := \frac{1}{2n} \sum_{i=1}^{n} (Y_i - Y_{i,n}^h)^2.$$

Here $Y_{i,n}^h$ corresponds to $X_{i,n}^h$, which is the WSNN to X_i with respect to the metric based on $\| \cdot \|_h$.
A change of metric influences $\tilde{L}_n^h$ in two ways. First the cubic partition (cf. (1.2)) becomes a partition of cuboids with side-length $\frac{\tilde{h}_n}{h_1}, \ldots, \frac{\tilde{h}_n}{h_d}$. Second, the nearest neighbour is selected on the basis of $\| \cdot \|_h$ (cf. Figure 1.2). If h_i is small the i-th side length of the cuboid is large and the WSNN of i is chosen at random among the X_j s inside the cuboid. Second, for the nearest neighbour rule based on $\| \cdot \|_h$ the i-th coordinate is rather unimportant. Consequently, small h_i diminishes the weight of the i-th dimension in the selection process, while large h_i indicates that the dimension is rather important.

Theorem 1.3 *Assume $|Y_i| \leq B$ for $i \in \{1, \ldots, n\}$. Furthermore, assume ties occur with probability zero, compact support M_μ of μ and Lipschitz-continuity of m and σ^2 with respect to $\| \cdot \|$. Then we have*

$$E \left| \tilde{L}_n^h - L^* \right| \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

The constant in the inequality above depends on d, B, k_1, k_2 and $h_{min} = \min\{h_1, \ldots h_d\}$.

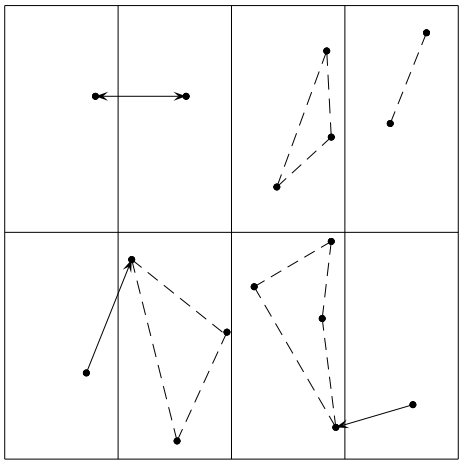

Figure 1.2: Dilated partition with $h = (\frac{2}{3}, \frac{4}{3})$ (cf. Figure 1.1)

The theorem is proved by two lemmas corresponding to Lemma 1.3 and Lemma 1.4.

Lemma 1.5 *Assume* $|Y_i| \leq B$ *for any* $i \in \{1, \ldots, n\}$ *and that ties occur with probability zero with respect to* $\|\cdot\|$. *Then one has*

$$a) \qquad P\left(\left|\tilde{L}_n^h - E\tilde{L}_n^h\right| \geq \epsilon\right) \leq 2\exp\left(\tfrac{-\epsilon^2 n}{2B^4(5+2\gamma_d)^2}\right)$$
$$b) \quad E\left|\tilde{L}_n^h - E\tilde{L}_n^h\right| \leq \sqrt{2(1+log2)}B^2(5+2\gamma_d)n^{-\frac{1}{2}}.$$

Proof of part a: For $l \in \{1, \ldots, n\}$ we have

$$\sum_{i=1;i\neq l}^{n} I_{\{X_l \text{ is WSNN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}\}} \leq \gamma_d + 2 \text{ a. s.}, \tag{1.24}$$

as ties occur with probability zero with respect to $\|\cdot\|_h$ (cf. (1.7)). Using the arguments in the proof of Lemma 1.3, it can be shown that

$$\sup_{\bar{D}_n, D_n} \left|\tilde{L}_n^h(D_n) - \tilde{L}_n^h(\bar{D}_{n,l})\right| \leq \frac{2B^2(5+2\gamma_d)}{n} \text{ a. s.},$$

where $\bar{D}_{n,l}$ differs only with respect to $(\bar{X}_l, \bar{Y}_l)$ from D_n. We then obtain with McDiarmid's inequality (Theorem $A.1$) and (1.8)

$$P\left(\left|\tilde{L}_n^h - E\tilde{L}_n^h\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{2B^4(5+2\gamma_d)^2}\right).$$

This proves part a. Part b follows from part a with arguments from the proof of Lemma 1.3 part a. □

Lemma 1.6 *Assume compact support M_μ of μ and Lipschitz-continuity of m and σ^2 with constants k_1 and k_2 with respect to $\|\cdot\|$. If $i \in \{1, ..., n\}$, then*

$$a) \quad \left|E\sigma^2(X_{i,n}^h) - E\sigma^2(X_i)\right| \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

$$b) \quad E\left(m\left(X_{i,n}^h\right) - m\left(X_i\right)\right)^2 \leq \begin{cases} const \cdot n^{-\frac{2}{3}} & ; \quad d = 1 \\ const \cdot n^{-2/(1+d)} & ; \quad d \geq 2. \end{cases}$$

The constant in the inequality above depends on d, B, k_1, k_2 and $h_{min} := \min\{h_1, \dots h_d\}$.

PROOF: We follow the proof of Lemma 1.4. Let $\left\{A_{n,1}, \dots, A_{n,N_{h,n}}\right\}$ be a partition of M_μ with cuboids of side length $\frac{\tilde{h}_n}{h_1}, \dots, \frac{\tilde{h}_n}{h_d}$. Set

$$h_{min} = \min\left\{h_i : i \in \{1, \dots, d\}\right\} > 0.$$

Choose $c_{h,n} > 0$ in such a way that

$$N_{h,n} = \frac{c_{h,n} h_{min}^d}{\tilde{h}_n^d}. \tag{1.25}$$

Then we have for $\epsilon > 0$ (cf. (1.9))

$$P\left(\| X_{l,n}^h - X_l \| > \epsilon; X_j \notin A_n(X_l) \text{ for } j \in \{1, \dots, n\} \setminus \{l\}\right) \leq \frac{c_{h,n} h_{min}^d}{n \tilde{h}_n^d}.$$

According to the proof of Lemma 6.4 in Györfi et al. (2002) [14], let $B_1^\epsilon, \dots, B_{N(\epsilon)}^\epsilon$ be the cubic partition of the bounded support of μ with diameter ϵ of each B_j^ϵ. Choose c_ϵ for $\epsilon > \frac{\tilde{h}_n}{h_{min}}$ by (cf. (1.11))

$$N_\epsilon = \frac{c_\epsilon}{\epsilon^d}.$$

Then we get

$$P\left(\| X_{l,n}^h - X_l \| > \epsilon; X_j \notin A_n(X_l) \text{ for } j \in \{1, \dots, n\} \setminus \{l\}\right) \leq \frac{c_\epsilon}{n\epsilon^d}.$$

PROOF OF PART a: For $i \in \{1, \dots, n\}$

$$\begin{aligned} & E\sigma^2\left(X_{i,n}^h\right) - E\sigma^2\left(X_i\right) \\ = \; & E\left(\left(\sigma^2\left(X_{i,n}^h\right) - \sigma^2\left(X_i\right)\right) \cdot I_{\{X_j \notin A_n(X_i) \text{ for } j \in \{1,\dots,n\} \setminus \{i\}\}}\right) \\ & + E\left(\left(\sigma^2\left(X_{i,n}^h\right) - \sigma^2\left(X_i\right)\right) \cdot I_{\{X_j \in A_n(X_i) \text{ for some } j \in \{1,\dots,n\} \setminus \{i\}\}}\right) \\ =: \; & A + B. \end{aligned}$$

Again, B=0 and (cf. (1.13)) for $d \geq 2$

$$\begin{aligned} |A| &\leq \frac{k_2}{n}\left(\frac{\tilde{h}_n}{h_{min}}\right)^{1-d}\left(c_{h,n}+\frac{1}{d-1}\bar{c}_\epsilon\right) \\ &= h_{min}^{d-1}k_2 n^{-\frac{2}{d+1}}\left(c_{h,n}+\frac{1}{d-1}\bar{c}_\epsilon\right) \end{aligned}$$

In the case of $d=1$ the same arguments that lead to (1.16) will result in

$$\left|E\sigma^2(X_{i,n}^h)-E\sigma^2(X_i)\right| \leq n^{-\frac{2}{3}}$$

Proof of Part b: By Lipschitz-continuity of m and the boundedness of the support of μ and as ties occur with probability zero, we have for $d \geq 2$,

$$\begin{aligned} E\left(m\left(X_{i,n}^h\right)-m\left(X_i\right)\right)^2 &\leq cn^{-\frac{2}{d}}+c\left(\frac{\tilde{h}_n}{h_{min}}\right)^2 \\ &\leq cn^{-\frac{2}{d}}+h_{min}^{-2}n^{-\frac{2}{d+1}} \end{aligned}$$

In the case of $d=1$ one derives similarly to (1.19)

$$E\left(m\left(X_{i,n}^h\right)-m\left(X_i\right)\right)^2 \leq n^{-\frac{2}{3}}.$$

This concludes the proof. □

Proof of Theorem 1.3: We have (cf. (1.22))

$$\begin{aligned} E|\tilde{L}_n^h-L^*| &\leq E|\tilde{L}_n^h-E\tilde{L}_n^h|+\frac{1}{2n}\sum_{i=1}^n|E\sigma^2(X_{i,n}^h)-E\sigma^2(X_i)| \\ &\quad +\frac{1}{2n}\sum_{i=1}^n E(m(X_{i,n}^h)-m(X_i))^2. \end{aligned}$$

Lemma 1.5 and Lemma 1.6 complete the proof for $d=2$.□

The following lemma gives an idea of how an optimal parameter $h \in \mathcal{Q}^d$ is to be chosen.

Lemma 1.7 *Assume $X_i \in [0,1]^d$, $\sigma^2(x)$ and $m(x)$ are Lipschitz-continuous with constants k_1 and k_2. If we have*

$$k_1 \leq \frac{k_2^2}{\left(\frac{\tilde{h}_n}{h_{min}}+1\right)^d+\frac{(\sqrt{d}+1)^d}{d-1}}, \tag{1.26}$$

then we can deduce for some constants $c_1 \leq c_2$

$$\begin{aligned} E|\sigma^2(X_{i,n}^h) - \sigma^2(X_i)| &\leq c_1 n^{-\frac{2}{d+1}} \\ E(m(X_{i,n}^h) - m(X_i))^2 &\leq c_2 n^{-\frac{2}{d+1}}. \end{aligned}$$

PROOF: $\bar{c}_\epsilon$ and $c_{\tilde{h}_n}$ be the constants as given in (1.12) and (1.25). We evaluate

$$\begin{aligned} N_{\tilde{h}_n} \cdot \left(\frac{\tilde{h}_n}{h_{min}}\right)^d &\leq (1 + \frac{1}{\tilde{h}_n})^d \left(\frac{\tilde{h}_n}{h_{min}}\right)^d \\ &= (\frac{\tilde{h}_n}{h_{min}} + 1)^d \end{aligned}$$

and

$$\begin{aligned} N_\epsilon \cdot \epsilon^d &\leq (1 + \frac{1}{\epsilon})^d \epsilon^d \\ &\leq (\bar{c}_\epsilon + 1)^d \\ &= (\sqrt{d} + 1)^d. \end{aligned}$$

Thus we have

$$\begin{aligned} c_{h,n} &\leq (\frac{\tilde{h}_n}{h_{min}} + 1)^d \\ \bar{c}_\epsilon &\leq (\sqrt{d} + 1)^d. \end{aligned}$$

With (1.18) and (1.13) we have

$$\begin{aligned} E\left|\sigma^2(X_{i,n}^h) - \sigma^2(X_i)\right| &\leq k_1(c_{h,n} + \frac{\bar{c}_\epsilon}{d-1}) n^{-\frac{2}{d+1}} \\ E(m(X_{i,n}^h) - m(X_i))^2 &\leq k_2^2(cn^{-\frac{2}{d}} + n^{-\frac{2}{d+1}}) \\ &\leq k_2^2 \cdot \bar{c} \cdot n^{-\frac{2}{d+1}}, \end{aligned}$$

with some constant $\bar{c} \geq 1$.

With the assumption on the Lipschitz-constants in (1.26) the constant before the leading term in the lower inequality becomes larger than the constant in the upper inequality. □

From (1.21) it follows with $i \in \{1, \ldots, n\}$ for the bias of $\tilde{L}_n^h$

$$E\tilde{L}_n^h - L^* = \frac{1}{2} E(\sigma^2(X_{i,n}^h) - \sigma^2(X_i)) + \frac{1}{2} E(m(X_{i,n}^h) - m(X_i))^2.$$

In case of sharpness of the bounds in Lemma 1.7 and if the Lipschitz-constants of σ^2 and m satisfy the assumption (1.26), the right hand side of the equation above becomes nonnegative. However, the bias is always nonnegative, if σ^2 is constant.
Assuming that both Lipschitz-constants are small, but do not satisfy (1.26), the absolute value of the bias becomes small, and minimizing will lead to quite a good estimator of L^*. Actually, under this assumption most of the $\tilde{L}_n^h$ are quite good estimators.
The reasoning above motivates the choice of a parameter h, because we then have $E\tilde{L}_n^h \geq L^*$ and so it seems reasonable to minimize $\tilde{L}_n^h$ over a given parameter space $\mathcal{Q}^d \subset \mathbb{R}^{\mathrm{d}}$.
We define for any $h \in \mathcal{Q}^d$

$$\begin{aligned} R(h) &:= E\tilde{L}_n^h && (1.27)\\ \bar{h}_n &:= arg\min_{h\in\mathcal{Q}^d} R(h) && (1.28)\\ H_n &:= arg\min_{h\in\mathcal{Q}^d} \tilde{L}_n^h. && (1.29) \end{aligned}$$

We can understand H_n as the best empirical guess of $\bar{h}_n$ which is the under the condition of Lemma 1.7 and the arguments that followed the best value of h that can be achieved by choosing an optimal metric $\|\cdot\|_h$. While $R(H_n)$ is a random variable, since H_n is random, $R(\bar{h}_n)$ is a constant value
As a consequence of the definitions and the inequality above we have for any set of data D_n the result

$$R(\bar{h}) \leq R(H_n(D_n)).$$

Theorem 1.4 *Assume $|Y_i| \leq B$ for all $i \in \{1,...,n\}$ and that ties occur with probability zero. Then we have*

$$E\left[R(H_n)\right] - R(\bar{h}_n) \leq 2\sqrt{2(1+\log\left(2|\mathcal{Q}^d|\right))}B^2(5+2\gamma_d)n^{-\frac{1}{2}}.$$

Proof: With the aid of definitions of H_n and $\bar{h}$ we derive

$$\begin{aligned} R(H_n) - R(\bar{h}_n) &\leq R(H_n) - \tilde{L}_n^{H_n} + \tilde{L}_n^{\bar{h}_n} - R(\bar{h}_n) && (1.30)\\ &\leq \left|R(H_n) - \tilde{L}_n^{H_n}\right| + \left|\tilde{L}_n^{\bar{h}_n} - R(\bar{h}_n)\right| \\ &\leq 2\max_{h\in\mathcal{Q}^d}\left|E\tilde{L}_n^h - \tilde{L}_n^h\right| \end{aligned}$$

$$\leq \; 2 \sum_{h \in \mathcal{Q}^d} \left| E\tilde{L}_n^h - \tilde{L}_n^h \right|.$$

Using Lemma 1.5 part a, we have

$$\begin{aligned} P\left(R(H_n) - R(\bar{h}_n) \geq \epsilon\right) &\leq \sum_{h \in \mathcal{Q}^d} P\left(\left| E\tilde{L}_n^h - \tilde{L}_n^h \right| \geq \frac{\epsilon}{2}\right) \\ &\leq 2\left|\mathcal{Q}^d\right| \exp\left(\frac{-\epsilon^2 n}{8B^4(5+2\gamma_d)^2}\right). \end{aligned}$$

The same argument as used in the proof of part b) of Lemma 1.3 shows that

$$E(R(H_n) - R(\bar{h}_n)) \leq \sqrt{\frac{(1+\log(2|\mathcal{Q}^d)|)}{n} 8B^4(5+2\gamma_d)^2}. \square$$

Using the results from this chapter, one obtains for $d \geq 2$

$$\begin{aligned} & E\left|L^* - R(H_n)\right| \\ \leq \quad & \underbrace{E\left|L^* - \tilde{L}_n^{\bar{h}_n}\right|}_{= \mathcal{O}\left(n^{-\frac{2}{d+1}}\right)} + \underbrace{E\left|\tilde{L}_n^{\bar{h}_n} - R(\bar{h}_n)\right|}_{= \mathcal{O}\left(n^{-\frac{1}{2}}\right)} + \underbrace{E\left|R(\bar{h}_n) - R(H_n)\right|}_{= \mathcal{O}\left(n^{-\frac{1}{2}}\right)}. \\ & \text{by Theorem 1.3} \qquad \text{by Lemma 1.3} \qquad \text{by Theorem 1.4} \end{aligned}$$

For $d \geq 4$ the rate of convergence of a distance between $\tilde{L}_n^{H_n}$ and $\tilde{L}_n^{\bar{h}_n}$ to zero is better than of a distance between $\tilde{L}_n^{\bar{h}_n}$ and L^*. Therefore, we can expect a rather good approximation of an optimal metric $\|\cdot\|_{\bar{h}}$. However, the estimator $\tilde{L}_n^{\bar{h}}$ may still be further away from L^* than $\tilde{L}_n^{\bar{h}_n}$ from $\tilde{L}_n^{H_n}$. In the case of $d = 3$ all terms converge of order $\mathcal{O}\left(n^{-\frac{1}{2}}\right)$ to zero. For $d = 1$ the expected deviation of $\tilde{L}_n^{\bar{h}_n}$ from L^* is of order $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$ and thus converging faster than the other two terms to zero.

Theorem 1.4 also justifies the choice of H_n via minimizing $\tilde{L}_n^h$, since we used that property in (1.30) which results in the rate of convergence $n^{-\frac{1}{2}}$.

Another possibility, which would result in the same rate of convergence, is to maximize $\tilde{L}_n^h$ and $\bar{h}_n$ as maximizing $E\tilde{L}_n^h$ over $\mathcal{Q}^d$. But this would only lead in the unlikely case of $E\tilde{L}_n^{\bar{h}_n} \leq L^*$ to a better result than the described method.

Remark 1.2 In Lemma 1.1 part c, it is shown that L^* is invariant against dilations of $\mathbb{R}^d$ with weights $h_i > 0$. This requirement enables us to use the Transformation Theorem for

Measures. However, assume $h_i = 0$ for some $i \in \{1, \ldots d\}$ and set $m^h(x) := E(Y|X^h = x)$ (where $X^h = (h_1 x_1, \ldots, x_d h_x)$) and assume ties occur with probability zero with respect to $\| \cdot \|_h$ and X. Then set

$$L^{(h)^*} := E(Y - m^h(X))^2 = EY^2 - Em^h(X)^2.$$

Define $\tilde{L}^h_{n,0}$ to be the estimator based only on the predictors with $h_i > 0$. Then $\tilde{L}^h_{n,0}$ is an estimator of $L^{(h)^*}$.

Chapter 2

Censored Observations

In medical applications it is often the survival time of a patient that is censored. As a consequence, the state of health is unknown after a certain point in time. In this chapter, the estimator $\tilde{L}_n^h$ (cf. (1.4)–(1.6)) is generalized with the aid of the product-limit estimator to include the case of censored observations (cf. (2.2)–(2.4)). The rates of convergence of the censored case as compared to the uncensored case differ mainly with respect to a logarithmic factor (cf. Theorems 2.4 and 1.2 and Theorems 2.6 and 1.4). Two results from Gu and Lai (1990) [13] and Chen and Lo (1997) [3] are used to derive rates of convergence. However, convergence in measure is proved here as opposed to L^2 convergence in the previous chapter. Also, the derived rates of convergence apply only after a certain sample size is reached that depends on the distributions of the survival and the censoring time.

2.1 The Censored Model

In the censored model, we have apart from the so-called survival time $Y \geq 0$ a censoring time $C \geq 0$ and are in the position to observe

$$Z = \min\{Y, C\}.$$

In addition. δ contains the information if Z is censored or not. This leaves us with a set of data

$$\{(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)\}$$

with $\delta_i = 1$ if $Y_i < C_i$ and $\delta_i = 0$ if $C_i \leq Y_i$. Fan and Gijbels [10] introduced in 1994 a transformation Y^* for Z with the property

$$E(Y^*|X) = E(Y|X).$$

As the transformed variable Y^* can not be calculated, since it depends the unknown survival function

$$G(t) = P(C > t),$$

they used the product-limit estimator (cf. Kaplan and Meier (1958) [15])

$$G_n(t) = \begin{cases} \prod_{i=1,\ldots,n:Z_{(i)}\leq t} (\frac{n-i}{n-i+1})^{1-\delta_{(i)}} & ; \quad t \leq Z_{(n)} \\ 0 & ; \quad \text{otherwise} \end{cases}$$

(with $\{(Z_{(1)}, \delta_{(1)}), \ldots, (Z_{(n)}, \delta_{(n)})\}$ being the observations ordered on $Z_{(i)}$, where in the case of ties the censored comes before the uncensored observations) to construct an approximation $\bar{Y}$ of Y^*.

2.2 Results on the Product-Limit Estimator

The question if and how G_n converges to G has been the subject of extensive research. Among the many results we use the following in the treatment of the problems that occur when estimating the minimum mean squared error L^* in the censored regression model. Corresponding to the survival function G of C we define

$$F(t) := P(Y > t) \quad \text{and}$$

$$T_K := \sup\{t : F(t) \cdot G(t) > 0\}.$$

T_K is the supremum of all $t > 0$ for which F and G are larger than zero. As a a consequence,

$$Z = \min\{Y, C\} \leq T_K \text{ a. s.}$$

Theorem 2.1 *(Stute/Wang, 1993) Assume F and G have no common jumps. Then*

$$\sup_{t<T_K} |G_n(t) - G(t)| \to 0 \text{ a. s. if } n \to \infty.$$

Moreover, if G is continuous in T_K and F and G have no common jumps, we have

$$\sup_{t\leq T_K} |G_n(t) - G(t)| \to 0 \text{ a. s. if } n \to \infty.$$

The theorem establishes uniform convergence $G_n \longrightarrow G$. However, no rates of convergence are derived.

In order to establish rates of convergence, we have to impose conditions on the distributions of G and F.

Theorem 2.2 *Assume $G(T_K) > 0$, G is continuous and that for $0 < p \leq \frac{1}{2}$ we have*

$$\int_0^{T_K} F(t)^{-\frac{p}{1-p}} d(1 - G(t)) < \infty. \tag{2.1}$$

a) (Chen/Lo, 1997) If $0 < p < \frac{1}{2}$, condition (2.1) is equivalent to

$$\limsup_{n\longrightarrow\infty} n^p \sup_{t\leq T_K} |G_n(t) - G(t)| < \infty \quad \text{a. s.}$$

b) (Gu/Lai, 1990) Assume $p = \frac{1}{2}$ and $\frac{1}{3} < \alpha < \frac{1}{2}$. Then from (2.1) with

$$t_k := \sup\{t : P(Y \geq t) \geq n^{-(1+\alpha)}\}$$

it follows

$$\limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{\frac{1}{2}} \cdot \sup_{t<t_k} |G_n(t) - G(t)| < \infty \text{ a. s.}$$

For the proof of part a compare the proof of Theorem 2.1 in Chen and Lo (1997) [3]. Part b is proved in Gu and Lai (1990) [13] as Corollary 3, which is an extension of Theorem

1. Certain conditions have to be imposed on the censored model in order to analyze it mathematically (for a discussion of different models compare [14] , p. 541 ff.).
We use three conditions in the treatment of the censored case:

(**A1**) C and (X, Y) are independent

(**A2**) There exists an $L > 0$ with $P(0 \leq Y \leq L) = 1$ and $P(C > L) > 0$.
Moreover G and F are continuous.

(**A3**) For p with $0 < p \leq \frac{1}{2}$ the inequality $\int_0^{T_K} F(t)^{-\frac{p}{1-p}} d(1 - G(t)) < \infty$ is fullfilled.

In **(A1)** the censoring time C is independent of the common distribution of the survival time Y and the patient data X. The censoring takes place regardless of the condition of the patient and depends thus only on external factors not related to the data of the patient as represented in (X, Y) .
In most applications the survival time Y is bounded, therefore the first part of **(A2)** is fulfilled. For the second part, it is required that not the whole censoring process takes place in $[0, L]$. In fact, $G(L)$ and $G(T_K)$ influence the rate of convergence as expressed in Theorem 2.3 and in the proof of Theorem 2.6 (cf. the handling of the sums $S_{1,n}, S_{2,n}, S_{3,n}$ and $S_{4,n}$). If F or G is continuous, we have uniform convergence $G_n \longrightarrow G$ (c.f. Theorem 2.1) on the whole interval $[0, T_k]$.
To derive rates of convergence, **(A3)** is needed. The product-limit-estimator is rather unstable near T_K and **(A3)** establishes a condition on F and G to analyze the behaviour of G_n over $[0, T_K]$. It requires a certain number of censored and uncensored observations in a small area prior T_K .

2.3 Definition and Convergence

In order to estimate $L^* = E(Y - E(Y|X))^2$ in the censored model, we need to transform $\tilde{L}_n$, which is based on $\{(X_1, Y_1, \ldots, X_n, Y_n)\}$ to become an estimator based on $(X_1, Z_1, \delta_1, \ldots, X_n, Z_n, \delta_n)$. This new estimator L_n^* is required to have the property

$E(L_n^*) = E(\tilde{L}_n)$, which allows us to extend the results of the previous chapters to the censored case.

We define

$$L_n^* := \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i^{2*} - 2Y_i^* Y_{i,n}^* + Y_{i,n}^{2*} \right) \tag{2.2}$$

with

$$Y_i^* := \frac{\delta_i Z_i}{G(Z_i)} = \begin{cases} \frac{Y_i}{G(Y_i)} & : \quad Y_i < C_i \\ 0 & : \quad \text{otherwise} \end{cases} \tag{2.3}$$

and similarly

$$Y_{i,n}^* := \frac{\delta_{i,n} Z_{i,n}}{G(Z_{i,n})}, \; Y_i^{2*} := \frac{\delta_i Z_i^2}{G(Z_i)} \text{ and } Y_{i,n}^{2*} := \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_{i,n})}. \tag{2.4}$$

Here $\delta_{i,n} = 1$ if $Z_{i,n}$ is uncensored and $\delta_{i,n} = 0$ if $Z_{i,n}$ is censored.

Remark 2.1 Dividing by $G(Z_i)$ (or by $G(Z_{i,n})$) will increase the value of Y_i^* in comparison to Z_i for uncensored Z_i, because $G(t) \leq 1$. Furthermore, $G(t)$ becomes smaller for increasing t. If most of the censoring has taken place, $G(t)$ will be close to 0 and therefore Y_i^* will be larger. Therefore, in the transformation $*$ uncensored information is more augmented, if it is observed after most of the censoring has taken place. All together uncensored Z_i is enlarged in such a way, that the conditional expectation of Y_i^* and Y_i is the same as the following lemma claims.

Remark 2.2 The estimator L_n^* may become negative. This is because the mixed term in

$$Y_i^{2*} - 2Y_i^* Y_{i,n}^* + Y_{i,n}^{2*} = \frac{\delta_i Z_i}{G(Z_i)} - 2\frac{\delta_i Z_i \delta_{i,n}}{G(Z_i)G(Z_{i,n})} + \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_{i,n})}$$

is augmented by $(G(Z_i)G(Z_{i,n}))^{-\frac{1}{2}}$ as compared to a binomial sum.

Lemma 2.1 *With L_n^* defined as above and* **(A1)** *and* **(A2)** *one has*

$$EL_n^* = E\tilde{L}_n.$$

PROOF: We have for any $i \in \{1, \ldots, n\}$

$$\begin{aligned} E(Y_i^*|X_i) &= E\left(\frac{\delta_i Z_i}{G(Z_i)}\middle| X_i\right) \\ &= E\left(\frac{\delta_i Y_i}{G(Y_i)}\middle| X_i\right) \\ &= E\left(E\left(\frac{\delta_i Y_i}{G(Y_i)}\middle| X_i, Y_i\right)\middle| X_i\right) \\ &= E\left(\frac{Y_i}{G(Y_i)} \underbrace{E(\delta_i|X_i, Y_i)}_{=G(Y_i) \text{ by } (\mathbf{A1})}\middle| X_i\right) \\ &= E(Y_i|X_i). \end{aligned}$$

The last equality follows from the fact, that by **(A2)**

$$G(Y_i) = P(C_i > Y_i) > P(C_i > L) > 0.$$

Observe that

$$\begin{aligned} G_{i,n}(t) &:= P(C_{i,n} > t) \\ &= \sum_{j=1:j\neq i}^{n} P(C_j > t; X_j \text{ is WSNN to } X_i) \\ &= \sum_{j=1:j\neq i}^{n} P(C_j > t) P(X_j \text{ is WSNN to } X_i) \\ &\quad \text{(since } C_j \text{ is independent of } X_1, \ldots, X_n \text{ by } (\mathbf{A1})) \\ &= P(C_1 > t) \underbrace{\sum_{j=1:j\neq i}^{n} P(X_j \text{ is WSNN to } X_i)}_{=1}. \end{aligned}$$

Therefore, it follows, because the C_i are identically distributed

$$G_{i,n}(t) = G(t)$$

and

$$\begin{aligned} E(Y_{i,n}^*|X_{i,n}) &= E\left(\frac{\delta_{i,n} Z_{i,n}}{G(Z_{i,n})}\middle| X_{i,n}\right) \\ &= E\left(E\left(\frac{\delta_{i,n} Y_{i,n}}{G(Y_{i,n})}\middle| X_{i,n}, Y_{i,n}\right)\middle| X_{i,n}\right) \end{aligned}$$

$$\begin{aligned} &= E\left(\frac{Y_{i,n}}{G(Y_{i,n})}E\left(\delta_{i,n}|\, X_{i,n}, Y_{i,n}\right)\middle|\, X_{i,n}\right) \\ &= E\left(\frac{Y_{i,n}}{G_{i,n}(Y_{i,n})}E(\delta_{i,n}|X_{i,n}, Y_{i,n})\middle|\, X_{i,n}\right) \\ &= E(Y_{i,n}|X_{i,n}) \text{ (again by } (\mathbf{A1})). \end{aligned}$$

Using that $P(C_i^2 > Y_i^2) = G(Y_i)$, because of $C_i, Y_i \geq 0$, and a similar argument to the one used above, we obtain

$$E\left(Y_i^{2*}|X_i\right) = E\left(Y_i^2|X_i\right) \quad \text{and } E\left(Y_{i,n}^{2*}|X_{i,n}\right) = E\left(Y_{i,n}^2|X_{i,n}\right).$$

So the expected value of L_n^* can be calculated as

$$\begin{aligned} EL_n^* &= \frac{1}{2n}\sum_{i=1}^n E(Y_i^{2*} - 2Y_i^*Y_{i,n}^* + Y_{i,n}^{2*}) \\ &= \frac{1}{2n}\sum_{i=1}^n E\{E(Y_i^{2*}|X_i) - 2E(Y_i^*Y_{i,n}^*) + E(Y_{i,n}^{2*})\} \\ &= \frac{1}{2n}EY_i^2 - 2E(E(Y_i|X_i)\cdot E(Y_{i,n}|X_{i,n})) + E(Y_{i,n}^2) \\ &= E(\tilde{L}_n). \end{aligned}$$

In the second equation from the bottom we used the fact that we have in the uncensored case $E(Y_i \cdot Y_{i,n}) = E(E(Y_i|X_i)\cdot E(Y_{i,n}|X_{i,n}))$ (cf. Proposition A.1). Setting $\tilde{Y}_i := \frac{\delta_i Z_i}{G(Z_i)}$ we have $\tilde{Y}_{i,n} = Y_{i,n}^*$ and we can extend this result to the censored case. □

Corollary 2.1 *Assuming* **(A1)** *and* **(A2)** *one has*

$$EL_n^* = \frac{1}{2}L^* + \frac{1}{2n}\sum_{i=1}^n E\sigma^2(X_{i,n}) + \frac{1}{2n}\sum_{i=1}^n E(m(X_{i,n}) - m(X_i))^2$$

Proof: The corollary is a consequence of Lemma 2.1 and equation (1.21). □

The following theorems establish the rates of convergence in the censored model. Since only almost everywhere convergence of the product-limit estimator is available, the final result states the convergence in terms of convergence in measure.

Theorem 2.3 *a) Assume* **(A1)**, **(A2)** *and ties with probability zero. Then one has*

$$P\left(|EL_n^* - L_n^*| \geq \epsilon\right) \leq 2exp\left(\frac{-2\epsilon^2 n G(T_K)^4}{(5+2\gamma_d)^2 T_K^4}\right).$$

b) Under the same assumptions as in a) it holds that

$$E|L_n^* - EL_n^*| \leq \sqrt{\frac{2(1+log2)}{2}} \frac{T_K^2(5+2\gamma_d)}{G(T_K)^2} n^{-\frac{1}{2}}.$$

Proof First, observe that we have by **(A2)**

$$1 \geq G(Z) \geq G(T_K) \geq G(L) > 0.$$

To bound L_n^* we derive on the one hand

$$\begin{aligned} Y_i^{2*} - 2Y_i^* Y_{i,n}^* + Y_{i,n}^{2*} &= \frac{\delta_i Z_i^2}{G(Z_i)} - 2\frac{\delta_i \delta_{i,n} Z_i Z_{i,n}}{G(Z_i)G(Z_{i,n})} + \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_{i,n})} \\ &\leq \frac{\delta_i Z_i^2}{G(Z_i)G(Z_{i,n})} - 2\frac{\delta_i \delta_{i,n} Z_i Z_{i,n}}{G(Z_i)G(Z_{i,n})} + \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_i)G(Z_{i,n})} \\ &= \frac{(\delta_i Z_i - \delta_{i,n} Z_{i,n})^2}{G(Z_i)G(Z_{i,n})} \\ &\leq \frac{T_K^2}{G(T_K)^2} \text{ a. s.} \end{aligned}$$

On the other hand, (cf. Remark 2.2)

$$\begin{aligned} -\left(Y_i^{2*} - 2Y_i^* Y_{i,n}^* + Y_{i,n}^{2*}\right) &= -\frac{\delta_i Z_i^2}{G(Z_i)} + 2\frac{\delta_i \delta_{i,n} Z_i Z_{i,n}}{G(Z_i)G(Z_{i,n})} - \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_{i,n})} \\ &\geq -\frac{\delta_i Z_i^2}{G(Z_i)} - \frac{\delta_{i,n} Z_{i,n}^2}{G(Z_{i,n})} \\ &\geq -\frac{T_K^2}{G(T_K)} \text{ a. s.} \end{aligned}$$

This results in

$$\left|Y_i^{2*} - 2Y_i^* Y_{i,n}^* + Y_{i,n}^{2*}\right| \leq \frac{T_K^2}{G(T_K)^2} \text{ a. s.} \tag{2.5}$$

McDiarmid's inequality (Theorem $A.1$) is used to show that for any $D_n = \{(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n\})$

$$\sup_{D_{n,l}, D_n} |L_n^*(D_n) - L_n^*(D_{n,l})| \leq \frac{T_K^2(5+2\gamma_d)}{nG(T_K)^2} \text{ a. s.},$$

where $D_{n,l} = \{(X_1', Z_1', \delta_1'), ..., (X_n', Z_n', \delta_n')\}$. $D_{n,l}$ differs only with respect to (X_l', Z_l', δ_l') from D_n, in other words

$$(X_k', Z_k', \delta_k') = (X_k, Z_k, \delta_k) \quad \text{for k } \in \{1, ..., n\} \setminus \{l\}.$$

Let $X_{\widehat{i,n}}$ be the X_j for which X_j is WSNN to X_i among

$$\{X_1, X_2, \ldots, X_n\} \setminus \{X_i, X_l\} \cup \{X_l'\}.$$

The absolute value in the the supremum above can be calculated as

$$\begin{aligned}
& \frac{1}{2n}\left|\sum_{i=1}^{n}\left(Y_i^{2*} - 2Y_i^*Y_{i,n}^* + Y_{i,n}^{2*}\right) - \sum_{i=1}^{n}\left(Y_i'^{2*} - 2Y_i'^* \cdot Y_{\widehat{i,n}}'^* + Y_{\widehat{i,n}}'^{2*}\right)\right| \\
\leq\ & \frac{1}{2n}\Bigg\{\left|Y_l^{2*} - 2Y_l^* \cdot Y_{l,n}^* + Y_{l,n}^{2*}\right| + \left|Y_l'^{2*} - 2Y_l'^* \cdot Y_{\widehat{l,n}}^* + Y_{\widehat{l,n}}^{2*}\right| \\
& \quad + \left|\sum_{i:i,n=l}\left(Y_i^{2*} - 2Y_i^* \cdot Y_l^* + Y_l^{2*}\right) - \left(Y_i^{2*} - 2Y_i^* \cdot Y_{\widehat{i,n}}'^* + Y_{\widehat{i,n}}'^{\,2*}\right)\right| \\
& \quad + \left|\sum_{i:\widehat{i,n}=l}\left(Y_i^{2*} - 2Y_i^*Y_{i,n}^* + Y_{i,n}^{2*}\right) - \left(Y_i^{2*} - 2Y_i^* \cdot Y_l'^* + Y_l'^{2*}\right)\right|\Bigg\} \\
\leq\ & \frac{1}{2n}\left(\frac{T_K^2}{G(T_K)^2} + \frac{T_K^2}{G(T_K)^2} + 2\frac{T_K^2}{G(T_K)^2}(\gamma_d + 2) + 2\frac{T_K^2}{G(T_K)^2}(\gamma_d + 2)\right) \\
=\ & \frac{T_K^2(5 + 2\gamma_d)}{G(T_K)^2 n} \text{ a. s.,}
\end{aligned} \tag{2.6}$$

by (1.7), (2.5), as the random vectors $(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)$ are independent and identically distributed and ties occur with probability zero. Using McDiarmid's inequality (Theorem A.1) we arrive at

$$P(|EL_n^* - L_n^*| \geq \epsilon) \leq 2\exp\left(\frac{-2\epsilon^2 n G(T_K)^4}{(5 + 2\gamma_d)^2 T_K^4}\right)$$

and prove part a. The same argument as in Lemma 1.4 proves part b. □

As the survival function $G(t)$ is not calculable, we estimate $G(t)$ by the product-limit estimator $G_n(t)$ and define (cf. (2.2)-(2.4))

$$\bar{L}_n := \frac{1}{2n}\sum_{i=1}^{n}\left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}_{i,n} + \bar{Y}_{i,n}^2\right) \tag{2.7}$$

with

$$\bar{Y}_i := \frac{\delta_i Z_i}{G_n(Z_i)},\ \bar{Y}_{i,n} := \frac{\delta_{i,n} Z_{i,n}}{G_n(Z_{i,n})}, \tag{2.8}$$

$$\bar{Y}_i^2 := \frac{\delta_i Z_i^2}{G_n(Z_i)} \text{ and } \bar{Y}_{i,n}^2 := \frac{\delta_{i,n} Z_{i.n}^2}{G_n(Z_{i,n})} \tag{2.9}$$

Remark 2.3 By estimating G through G_n we are using more of the information of the censored observations. In the definition (cf. (2.3) and (2.8)), we have $Y_i^* = 0$ and $\bar{Y}_i = 0$ if Z_i is censored. (also $Y_{i,n}^* = 0$ and $\bar{Y}_{i,n} = 0$, if $Z_{i,n}$ is censored). The transformation introduced by Fan and Gijbels [10] is capable of using also censored information to estimate Y_i^*.

Lemma 2.2 *a) Assume* **(A1)**, **(A2)** *and* **(A3)** *with* $p = \frac{1}{2}$. *Furthermore assume that ties occur with probability zero. Then*

$$\limsup_{n\longrightarrow\infty} \left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} \left|\bar{L}_n - L_n^*\right| < \infty \ a.\ s.$$

b) Assume **(A1)**, **(A2)** *and* **(A3)** *with* $p = \frac{2}{d+1} < \frac{1}{2}$. *In addition, assume ties occur with probability zero. Then*

$$\limsup_{n\longrightarrow\infty} n^{\frac{2}{d+1}} \left|\bar{L}_n - L_n^*\right| < \infty \ a.\ s.$$

Proof of Part a: Let $\frac{1}{3} < \alpha < \frac{1}{2}$. Observe that with $i \in \{1, \ldots, n\}$ and (cf. Theorem 2.2)

$$t_k := \sup\left\{t : P(Y \geq t) \geq n^{-(1+\alpha)}\right\}$$

we have

$$\begin{aligned}
& \limsup_{n\to\infty} \left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} \frac{1}{2n} \sum_{i=1}^{n} |G(Y_i) - G_n(Y_i)| \\
\leq\ & \limsup_{n\to\infty} \left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} 1_{\{Y_i < t_k\}} \sup_{0\leq t<t_k} |G(t) - G_n(t)| \right. \\
& \left. + \frac{1}{2n} \sum_{i=1}^{n} 1_{\{Y_i \geq t_k\}} \underbrace{\sup_{t_k \leq t \leq T_K} |G(t) - G_n(t)|}_{\leq 1} \right\} \\
\leq\ & \limsup_{n\longrightarrow\infty} \left\{ \left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} \left\{ \frac{1}{2} \sup_{0\leq t<t_k} |G(t) - G_n(t)| \right.\right.
\end{aligned}$$

$$+\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2n}\sum_{i=1}^{n}1_{\{Y_i\geq t_k\}}\Bigg\}\text{ a. s.}$$

The first term on the right hand side is ¡ ∞ by Theorem 2.2 part b. The mean in the second sum on the right hand side converges almost surely against $P(Y_i \geq t_k)$ by the strong law of large numbers. By the definition of t_k, by Theorem 2.2 part b with any $\frac{1}{3} < \alpha < \frac{1}{2}$ and the continuity of F by **(A2)** we have

$$P(Y_i \geq t_k) \leq n^{-\frac{1}{2}}.$$

Therefore we have

$$\limsup_{n\to\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_i)-G_n(Y_i)|<\infty.\text{ a. s.} \tag{2.10}$$

Moreover, using the same arguments as above we arrive at

$$\begin{aligned}
&\limsup_{n\to\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_{i,n})-G_n(Y_{i,n})|\\
\leq\ &\limsup_{n\longrightarrow\infty}\Bigg\{\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2}\sup_{0\leq t<t_k}|G(t)-G_n(t)|\\
&+\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2n}\sum_{i=1}^{n}1_{\{Y_{i,n}\geq t_k\}}\Bigg\}\text{ a. s.}
\end{aligned}$$

The first term is again of order $\mathcal{O}\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}$ by Theorem 2.2 part b. Now, for the second term we can derive

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}1_{\{Y_{i,n}\geq t_k\}}\\
=\ &\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}1_{\{Y_j\geq t_k\}}\cdot 1_{\{X_j\text{ is WSNN to }X_i\}}\\
=\ &\frac{1}{2n}\sum_{j=1}^{n}1_{\{Y_j\geq t_k\}}\underbrace{\sum_{i=1}^{n}1_{\{X_j\text{ is WSNN to }X_i\}}}_{\leq\gamma_d+2\text{ by (1.7)}}\\
\leq\ &(\gamma_d+2)\frac{1}{n}\sum_{j=1}^{n}1_{\{Y_j\geq t_k\}}\text{ a. s.}
\end{aligned}$$

Therefore we have by the strong law of large numbers

$$\begin{aligned}\limsup_{n\longrightarrow\infty}\frac{1}{2n}\sum_{j=1}^{n}1_{\{Y_{i,n}\geq t_k\}} &\leq (\gamma_d+2)\,P(Y_1\geq t_k)\\ &\leq (\gamma_d+2)\,n^{-\frac{1}{2}}\text{ a. s.},\end{aligned}$$

where the last equation follows from Theorem 2.2 part b with any $\frac{1}{3}<\alpha<\frac{1}{2}$. Similarly to (2.10) one can calculate

$$\limsup_{n\to\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_{i,n})-G_n(Y_{i,n})|<\infty\text{ a. s.}\tag{2.11}$$

Furthermore, using (2.3) we derive the following decomposition

$$\begin{aligned}\left|\bar{L}_n-L_n^*\right| \leq\ & \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left|\bar{Y}_i^2-Y_i^{2*}\right|}_{=:T_{1,n}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|Y_i^*\cdot Y_{i,n}^*-\bar{Y}_i\cdot\bar{Y}_{i,n}\right|}_{=:T_{2,n}}\\ &+\underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left|Y_{i,n}^{2*}-\bar{Y}_{i,n}^2\right|}_{=:T_{3,n}}.\end{aligned}\tag{2.12}$$

Then one obtains

$$\begin{aligned}\limsup_{n\to\infty}T_{1,n} &= \limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}\left|\frac{\delta_iY_i^2}{G_n(Y_i)}-\frac{\delta_iY_i^2}{G(Y_i)}\right|\\ &\leq T_K^2\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}\frac{|G(Y_i)-G_n(Y_i)|}{G_n(T_K)G(T_K)}\\ &\leq T_K^2\limsup_{n\to\infty}\frac{1}{G_n(T_K)G(T_K)}\limsup_{n\longrightarrow\infty}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_i)-G_n(Y_i)|\\ &\leq \frac{T_K^2}{G(T_K)^2}\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_i)-G_n(Y_i)|\text{ a. s.}\end{aligned}$$

And therefore by (2.10)

$$\limsup_{n\longrightarrow\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}T_{1,n}<\infty.\text{ a. s.}$$

With (2.3), (2.4) and Theorem 2.1 we have

$$\begin{aligned}
\limsup_{n\to\infty} T_{2,n} &= \limsup_{n\to\infty}\left|\frac{1}{2n}\sum_{i=1}^{n}\frac{\delta_i\delta_{i,n}Y_iY_{i,n}}{G(Y_i)G(Y_{i,n})}-\frac{\delta_i\delta_{i,n}Y_iY_{i,n}}{G_n(Y_i)G_n(Y_{i,n})}\right| \\
&\le \frac{T_K^2}{G(T_K)^4}\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}|G_n(Y_i)G_n(Y_{i,n})-G(Y_i)G(Y_{i,n})| \\
&\le \frac{T_K^2}{G(T_K)^4}\limsup_{n\to\infty}\Big\{|G_n(Y_i)G_n(Y_{i,n})-G_n(Y_i)G(Y_{i,n})| \\
&\quad +|G_n(Y_i)G(Y_{i,n})-G(Y_i)G(Y_{i,n})|\Big\} \\
&\le \frac{T_K^2}{G(T_K)^4}\left\{\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}|G_n(Y_{i,n})-G(Y_{i,n})|\right. \\
&\quad \left.+\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}|G_n(Y_i)-G(Y_i)|\right\}\ \text{a. s.} \\
&\quad (\text{as } G(t)\le 1).
\end{aligned}$$

This together with (2.10) and (2.11) implies

$$\limsup_{n\longrightarrow\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} T_{2,n} < \infty \text{ a. s.}$$

Now, for the third sum again by Theorem 2.1

$$\begin{aligned}
\limsup_{n\to\infty} T_{3,n} &= \limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}\left|\frac{\delta_{i,n}Y_{i,n}^2}{G_n(Y_{i,n})}-\frac{\delta_iY_{i,n}^2}{G(Y_{i,n})}\right| \\
&\le T_K^2\limsup_{n\to\infty}\left|\frac{G(Y_{i,n})-G_n(Y_{i,n})}{G_n(Z_{i,n})G(Z_{i,n})}\right| \\
&\le \frac{T_K^2}{G(T_K)^2}\limsup_{n\to\infty}|G(Y_{i,n})-G_n(Y_{i,n})|\ \text{a. s.}
\end{aligned}$$

Hence from (2.11) we can conclude

$$\limsup_{n\longrightarrow\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} T_{3,n} < \infty \text{ a. s.}$$

Combining all the results from above, we arrive at

$$\limsup_{n\to\infty}\left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}}|\bar{L}_n - L_n^*| < \infty \text{ a. s.}$$

Proof of part b: Using Theorem 2.2 part a with $p=\frac{2}{d+1}<\frac{1}{2}$, equation (2.10) turns into

$$\limsup_{n\to\infty} n^{\frac{2}{d+1}}\frac{1}{2n}\sum_{i=1}^{n}|G(Y_i)-G_n(Y_i)| < \infty \text{ a. s.}$$

and (2.11) becomes

$$\limsup_{n\to\infty} n^{\frac{2}{d+1}} \frac{1}{2n} \sum_{i=1}^{n} |G(Y_{i,n}) - G_n(Y_{i,n})| < \infty \text{ a. s.}$$

Using these two results on the sums $T_{1,n}, T_{2,n}$ and $T_{3,n}$ in (2.12), part b of the Lemma is proved. □

Theorem 2.4 *Assume* $|Y_i| \le B$ *for all* $i \in \{1, \dots, n\}$, *Lipschitz-continuity of* m *and* σ^2, *ties occur with probability zero. Furthermore, assume* **(A1)** *and* **(A2)** *hold and* $\mu = P_X$ *has compact support.*

a) Let $d \in \{1, 2, 3\}$ *and assume* **(A3)** *with* $p = \frac{1}{2}$. *Then as a consequence*

$$\left|\bar{L}_n - L^*\right| = \mathcal{O}_{\mathbf{P}}\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}.$$

b) For $d \ge 4$, *and if* **(A3)** *with* $p = \frac{2}{d+1}$ *holds, one has*

$$\left|\bar{L}_n - L^*\right| = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2}{d+1}}\right).$$

Proof Using Lemma 2.1 and the triangle equation we have

$$\left|\bar{L}_n - L^*\right| \le \underbrace{\left|\bar{L}_n - L_n^*\right|}_{=:S_{1,n}} + \underbrace{\left|L_n^* - EL_n^*\right|}_{=:S_{2,n}} + \underbrace{\left|E\tilde{L}_n - \tilde{L}_n\right|}_{=:S_{3,n}} + \underbrace{\left|\tilde{L}_n - L^*\right|}_{=:S_{4,n}}.$$

We prove almost everywhere convergence for $S_{1,n}$ and L^2-convergence for $S_{2,n}, S_{3,n}$ and $S_{4,n}$.

Using for $d \in \{1, 2, 3\}$ Lemma 2.2 part a, we have for $S_{1,n}$

$$\limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{\frac{1}{2}} \left|\bar{L}_n - L_n^*\right| < \infty \text{ a. s.}$$

and in the case of $d \ge 4$ with the aid of part b with $p = \frac{2}{d+1}$ in **(A3)**

$$\limsup_{n\to\infty} n^{\frac{2}{d+1}} \left|\bar{L}_n - L_n^*\right| < \infty \text{ a. s.}$$

For $S_{2,n}$ we have by Theorem 2.3 for any $d \ge 1$

$$E\left|L_n^* - EL_n^*\right| \le const \cdot n^{-\frac{1}{2}}.$$

For $S_{3,n}$, we proved in Lemma 1.3 part b for any $d \geq 1$

$$E|E\tilde{L}_n - \tilde{L}_n| \leq const \cdot n^{-\frac{1}{2}}.$$

As m and σ^2 are Lipschitz-continuous, ties occur with probability zero and Y_i is bounded for all $i \in \{1, \ldots, n\}$, we have for $S_{4,n}$ by Theorem 1.2 if $d = 1$

$$E|\tilde{L}_n - L^*| \leq const \cdot n^{-\frac{2}{3}}.$$

For $d \geq 2$ it was shown there that

$$E|\tilde{L}_n - L^*| \leq const \cdot n^{-\frac{2}{d+1}}.$$

This implies that for $d \in \{1, 2, 3\}$ the rate of convergence for $n \to \infty$ of $|\bar{L}_n - L^*|$ is dominated by S_1, so we have a rate of $\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}$. This proves part a of the theorem.
For $d \geq 4$, the term S_4 is leading and the rate of convergence is $n^{-\frac{2}{d+1}}$. This concludes the proof. □

2.4 Approximation of an Optimal Metric

Assume $\|\cdot\|_h$ is a metric on $\mathbb{R}^{\mathrm{d}}$ as defined in (1.23), where $h = (h_1, \ldots, h_d)$ is taken from some parameter space $\mathcal{Q}^d$ with $h_j > 0$ for all $j \in \{1, \ldots, d\}$.
Define the estimators based on D_n and for $h \in \mathcal{Q}^d$ as

$$L_n^{*h} := \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i^{2*} - 2Y_i^* \cdot Y_{(i,n)^h}^* + Y_{(i,n)^h}^{2*} \right) \quad \text{and} \tag{2.13}$$

$$\bar{L}_n^h := \frac{1}{2n} \sum_{i=1}^{n} \left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}_{(i,n)^h} + \bar{Y}_{(i,n)^h}^2 \right),$$

where $X_{(i,n)^h}$ is the WSNN based on $\|\cdot\|_h$.

Lemma 2.3 *a) Assume* **(A1)**, **(A2)** *and* **(A3)** *with* $p = \frac{1}{2}$. *In addition, assume that ties occur with probability zero. Then as a consequence*

$$\limsup_{n \longrightarrow \infty} \left(\frac{\log\log n}{n} \right)^{-\frac{1}{2}} \left| \bar{L}_n^h - L_n^* \right| < \infty \ \textit{a. s.}$$

b) Assume **(A1)**, **(A2)** *and* **(A3)** *with* $p = \frac{2}{d+1} < \frac{1}{2}$ *and ties occur with probability zero. Then one has*

$$\limsup_{n \longrightarrow \infty} n^{\frac{2}{d+1}} \left| \bar{L}_n^h - L_n^* \right| < \infty \ a.\ s.$$

PROOF: The proof of Lemma 2.2 can be modified to include the WSNN-rule with respect to $\| \cdot \|_h$ with weights $h_i > 0$ for all $i \in \{1, \ldots, d\}$. The only properties of $\| \cdot \|_h$ used are that ties occur with probability zero with respect to $\| \cdot \|_h$. Furthermore holds $\left| Z_{(i,n)^h} \right| \leq T_K$.□

Theorem 2.5 *a) Assume* **(A1)**, **(A2)** *and ties with probability zero. Then one has*

$$P\left(\left| EL_n^{*h} - L_n^{*h} \right| \geq \epsilon \right) \leq 2exp\left(\frac{-2\epsilon^2 n G(T_K)^4}{(5 + 2\gamma_d)^2 T_K^4} \right).$$

b) Under the same assumptions as in a) it holds

$$E|L_n^{*h} - EL_n^{*h}| \leq \sqrt{\frac{2(1 + log2)}{2}} \frac{T_K^2 (5 + 2\gamma_d)}{G(T_K)^2} n^{-\frac{1}{2}}.$$

PROOF: In the proof of Theorem 2.3 we used only the boundedness of $Z_{(i,n)^h}$ and the fact that ties occur with probability zero. Therefore it extends to the WSNN with respect to $\| \cdot \|_h$.□

Corollary 2.2 *Let* L_n^{*h} *be defined as in equation 2.13) and assume that* **(A1)** *and* **(A2)** *hold. Then*

$$EL_n^{*h} = \frac{1}{2} L^* + \frac{1}{2} E(\sigma^2(X_{(i,n)^h}) + \frac{1}{2} E(m(X_{(i,n)^h}) - m(X_i))^2.$$

PROOF: The equation is a consequence of Lemma 2.1 and Corollary 2.1. □

From Corollary 2.2 we can deduce that under the same conditions as mentioned in Lemma 1.7 and the discussion following it, that $EL_n^{*h} \geq L^*$.

As we want to estimate an empirical optimal metric $\| \cdot \|_h$ and thus $(i,n)^h$ by choosing h in such a way that $\bar{L}_n^h$ is minimal for all $h \in \mathcal{Q}^d$, it is necessary to split the sample

$$D_{2n} = D_n^1 \cup D_n^2,$$

since we have to avoid interplay between the Product-Limit estimator G_n and the calculation of the WSNN $(i,n)^h$ based on $\|\cdot\|_h$. We calculate the WSNN $(i,n)^h$ of $i \in \{1,\ldots,n\}$ by using D_n^1 and the product-limit estimator G_n by using D_n^2.
We condition from now on for the rest of this section on D_n^2. Also $\bar{L}_n^h|_{D_n^2}$ is the estimator $\bar{L}_n^h$ where the WSNN-structure is calculated by using D_n^1 (considered variable) and the Product-Limit $G|_{D_n^2}$ is calculated by using D_n^2 (considered fixed).
Similarly to (1.27)–(1.29) we express the idea of approximation of an optimal metric in the censored model as follows. For $h \in \mathcal{Q}^d$, we set

$$\begin{aligned} R(h) &:= E(\bar{L}_n^h | D_n^2) \\ \bar{h}_n &:= \arg\min_{h\in\mathcal{Q}^d} EL_n^{*h} = \arg\min_{h\in\mathcal{Q}^d} E\tilde{L}_n^h \quad \text{(by Lemma 2.1)} \\ H_n &:= \arg\min_{h\in\mathcal{Q}^d} \bar{L}_n^h|_{D_n^2}. \end{aligned}$$

The following theorem establishes the rate of convergence of the conditional expectation $R(H_n)$ against the expectation of $\tilde{L}_n^{\bar{h}_n}$. The value $G(T_K)$ influences the speed of convergence as expressed in the upper limits of $S_{1,n},\ldots,S_{4,n}$ in the proof.

Theorem 2.6 *Assume* $|Y_i| \le B$ *for all* $i \in \{1,\ldots,n\}$ *and ties occur with probability zero. Moreover, assume* **(A1)**, **(A2)** *and* **(A3)** *with* $p = \frac{1}{2}$. *Then one has*

$$R(H_n) - E\tilde{L}_n^{\bar{h}_n} \quad = \quad \mathcal{O}_{\mathbf{P}}\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}}.$$

Proof: Set

$$\begin{aligned} R^*(h) &:= EL_n^{*h} \\ H_n^* &:= \arg\min_{h\in\mathcal{Q}^d} EL_n^{*h}. \end{aligned}$$

Then it follows

$$\begin{aligned} \left|R(H_n) - E\tilde{L}_n^{\bar{h}_n}\right| &= \left|R(H_n) - EL_n^{*\bar{h}_n}\right| \\ &\le |R(H_n) - R^*(H_n^*)| + \underbrace{R^*(H_n^*) - EL_n^{*\bar{h}_n}}_{\ge 0 \text{ by the definition of } \bar{h}_n} \\ &=: T_{1,n} + T_{2,n}. \end{aligned}$$

For the two sums we have

$$\begin{aligned} T_{1,n} &\leq \left|R(H_n) - \bar{L}_n^{H_n}|_{D_n^2}\right| + \left|\bar{L}_n^{H_n}|_{D_n^2} - L_n^{*H_n^*}\right| + \left|L_n^{*H_n^*} - R^*(H_n^*)\right| \quad \text{and} \\ T_{2,n} &\leq \left|R^*(H_n^*) - L_n^{*H_n^*}\right| + \left|L_n^{*\bar{h}_n} - EL_n^{*\bar{h}_n}\right| \\ &\quad (\text{as } L_n^{*\bar{h}_n} - L_n^{*H_n^*} \geq 0 \text{ by the definition of } H_n^*). \end{aligned}$$

So therefore

$$\begin{aligned} & T_{1,n} + T_{2,n} \\ \leq\ & \left|R(H_n) - \bar{L}_n^{H_n}|_{D_n^2}\right| + \left|\bar{L}_n^{H_n}|_{D_n^2} - L_n^{*H_n^*}\right| + 2\left|L_n^{*H_n^*} - R^*(H_n^*)\right| \\ & + \left|L_n^{*\bar{h}_n} - EL_n^{*\bar{h}_n}\right| \\ =:\ & S_{1,n} + S_{2,n} + 2S_{3,n} + S_{4,n}. \end{aligned}$$

We derive an exponential inequality for $S_{1,n}$. Fixing D_n^2 means that the product-limit estimator $G_n|_{D_n^2}$ based on D_n^2 remains unchanged for any D_n^1.
Using McDiarmid's inequality (Theorem A.1) we show that for any D_n^1

$$\sup_{D_{n,l}^1, D_n^1} \left|\bar{L}_n^h|_{D_n^2} - \bar{L}_n^h|_{D_n^2}\right| \leq \frac{T_K^2(5 + 2\gamma_d)}{n(G_n(T_K) - \beta)^2} \text{ a. s.}, \tag{2.14}$$

where $D_{n,l}^1$ for $l \in \{1, \ldots, n\}$ differs only with respect to (X_l', Z_l', δ_l') from D_n^1. Let $X_{(\widehat{i,n})^h}$ be the X_j for which X_j is WSNN to X_i based on the metric $\|\cdot\|_h$ among

$$\{X_1, X_2, \ldots, X_n\} \cup \{X_l'\} \setminus \{X_i, X_l\}.$$

Then similarly to the computations resulting in (2.6)

$$\begin{aligned} & \sup_{D_{n,l}^1, D_n^1} \left|\bar{L}_n^h|_{D_n^2} - \bar{L}_n^h|_{D_n^2}\right| \\ \leq\ & \frac{1}{2n}\left|\sum_{i=1}^n \left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}_{(i,n)^h} + \bar{Y}_{(i,n)^h}^2\right) - \sum_{i=1}^n \left(\bar{Y}_i'^2 - 2\bar{Y}_i' \cdot \bar{Y}_{(\widehat{i,n})^h} + \bar{Y}_{(\widehat{i,n})^h}'^2\right)\right| \\ \leq\ & \frac{1}{2n}\left|\bar{Y}_l^2 - 2\bar{Y}_l \cdot \bar{Y}_{(l,n)^h} + \bar{Y}_{(l,n)^h}^2\right| + \left|\bar{Y}_l'^2 - 2\bar{Y}_l' \cdot \bar{Y}_{(\widehat{l,n})^h} + \bar{Y}_{(\widehat{l,n})^h}^2\right| \\ & + \left|\sum_{i:(i,n)^h = l} \left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}_l + \bar{Y}_l^2\right) - \left(\bar{Y}_i^2 - 2\bar{Y}_i \cdot \bar{Y}'_{(\widehat{i,n})^h} + \bar{Y}'^2_{(\widehat{i,n})^h}\right)\right| \end{aligned}$$

$$
\begin{aligned}
&+\left|\sum_{i:\widehat{(i,n)}^h=l}\left(\bar{Y}_i^2-2\bar{Y}_i\cdot\bar{Y}_{(i,n)^h}+\bar{Y}_{(i,n)^h}^2\right)-\left(\bar{Y}_i^2-2\bar{Y}_i\cdot\bar{Y}_l+\bar{Y}_l'^2\right)\right|\\
\leq\ & \frac{1}{2n}\left(\frac{T_K^2}{(G_n|_{D_n^2}(T_K))^2}\quad+\frac{T_K^2}{(G_n|_{D_n^2}(T_K))^2}\right.\\
&\left.+2\frac{T_K^2}{(G_n|_{D_n^2}(T_K))^2}(\gamma_d+2)\quad+\quad 2\frac{T_K^2}{(G_n|_{D_n^2}(T_K))^2}(\gamma_d+2)\right)\\
=\ & \frac{T_K^2(5+2\gamma_d)}{(G_n|_{D_n^2}(T_K))^2 n}\quad\text{a. s. by (1.7)},
\end{aligned}
$$

as ties occur with probability zero. By **(A2)** and Theorem 2.1, we have convergence $G_n(T_K)\to G(T_K)$. Therefore, for any $0<\beta<G(T_K)$ there exists an $N_{G(T_K)}$, so that

$$|G(T_K)-G_n(T_K)|<\beta \text{ a. s.}$$

for all $n>N_{G(T_K)}$. Then we have

$$\sup_{D_{n,l}^1,D_n^1}\left|\bar{L}_n^h|_{D_n^2}-\bar{L}_n^h|_{D_n^2}\right|\leq\frac{T_K^2(5+2\gamma_d)}{n(G(T_K)-\beta)^2}\text{ a. s.}$$

for $n\geq N_{G(T_K)}$.

Using McDiarmid's inequality (Theorem $A.1$) we arrive at

$$P\left(\left|E\left(\bar{L}_n^h\middle|D_n^2\right)-\bar{L}_n^h|_{D_n^2}\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{2\epsilon^2 n(G(T_K)-\beta)^4}{(5+2\gamma_d)^2T_K^4}\right). \tag{2.15}$$

Apart from this, we obtain

$$
\begin{aligned}
\left|R(H_n)-\bar{L}_n^{H_n}|_{D_n^2}\right| &\leq \max_{h\in\mathcal{Q}^d}\left|E\left(\bar{L}_n^h|D_n^2\right)-\bar{L}_n^h|_{D_n^2}\right|\\
&\leq \sum_{h\in\mathcal{Q}^d}\left|E\left(\bar{L}_n^h|D_n^2\right)-\bar{L}_n^h|_{D_n^2}\right|.
\end{aligned}
$$

Using (2.15) we get

$$
\begin{aligned}
P\left(\left|R(H_n)-\bar{L}_n^{H_n}|_{D_n^2}\right|\geq\epsilon\right) &\leq \sum_{h\in\mathcal{Q}^d}\left|P\left(\left|E(\bar{L}_n^h|D_n^2)-\bar{L}_n^h|_{D_n^2}\right|\geq\epsilon\right)\right.\\
&\leq 2\left|\mathcal{Q}^d\right|\exp\left(-\frac{2\epsilon^2 n(G(T_K)-\beta)^4}{(5+2\gamma_d)^2T_K^4}\right).
\end{aligned}
$$

The same argument as in Lemma 1.3 proves

$$E\left|R(H_n)-\bar{L}_n^{H_n}|_{D_n^2}\right|\leq\sqrt{\frac{1+\log(2|\mathcal{Q}^d|)}{2}}\frac{T_K^2(5+2\gamma_d)}{(G(T_K)-\beta)^2}n^{-\frac{1}{2}}.$$

For $S_{2,n}$ observe that we have for any $h \in \mathcal{Q}^d$ by Lemma 2.3

$$\limsup_{n\longrightarrow\infty} \left(\frac{\log\log n}{n}\right)^{-\frac{1}{2}} \left|\bar{L}_n^h - L_n^{*h}\right| < \infty \text{ a. s.}$$

Therefore we have the same result for $\bar{L}_n^{H_n}$ and $L_n^{*H_n^*}$ which are the minima over $\mathcal{Q}^d$ of the respective estimators $\bar{L}_n^h$ and L_n^{*h}.

For $S_{3,n}$ we use the exponential inequality of Theorem 2.5 and use the same argument as in the proof of Theorem 1.4 to obtain

$$E\left|L_n^{*H_n^*} - R^*(H_n^*)\right| \leq \sqrt{\frac{2(1+\log\left(2\left|\mathcal{Q}^d\right|\right))}{2}} \frac{T_K^2(5+2\gamma_d)}{G(T_K)^2} n^{-\frac{1}{2}}. \tag{2.16}$$

Finally, for $S_{4,n}$ we use Theorem 2.5 for $L_n^{*\bar{h}_n}$ and have

$$E\left|L_n^{*\bar{h}_n} - EL_n^{*\bar{h}_n}\right| \leq \sqrt{\frac{2(1+\log 2)}{2}} \frac{T_K^2(5+2\gamma_d)}{G(T_K)^2} n^{-\frac{1}{2}}. \tag{2.17}$$

Thus, $S_{2,n}$ is the leading term and $R(H_n) - E\tilde{L}_n^{\bar{h}_n}$ converges with a rate of convergence $\left(\frac{\log\log(n)}{n}\right)^{\frac{1}{2}}$.□

Chapter 3

Application to Data

In this chapter the estimator developed in Chapters One and Two is applied to two sets of data of breast cancer patients. These sets of data have been collected in the Stuttgart area since the 1980s. The data entries are naturally censored as patients drop out of the follow-up programme for a variety of reasons. Under some mathematical assumptions the data can be understood as satisfying the conditions of the previous chapters. The splitting of the data influences the product-limit estimator considerably. Therefore a set of data is split repeatedly to diminish this effect. A logarithmic scale is used that is adapted after a certain number of runs are implemented. The weight of a predictor changes if the set of predictors is altered. Six studies of the data reveal that the predictors can be divided into three groups of increasing importance (cf. Table 3.3). Among the most important predictors are amount of cerb2 proteine in the cell membrane of the tumor, grading of the tumor and number of affected lymph nodes.

3.1 Logarithmic Scale and Normalization

In view of Theorem 1.4 and equations (2.16) and (2.17) it is reasonable to choose the size of the parameter space $\mathcal{Q}^d$ as small as possible. For in this case, the constant $\left|\mathcal{Q}^d\right|$ is small and thus Theorem 1.4 and Theorem 2.6 guarantee a better speed of convergence. Another argument for a moderate size of $\mathcal{Q}^d$ is computational time. To minimize $\bar{L}_n^h$ over $\mathcal{Q}^d$ it is necessary to perform $\left|\mathcal{Q}^d\right|$ computations of $\bar{L}_n^h$. On the other hand, too small a parameter space may not deliver satisfying results. For the analysis of empirical data a logarithmic scale

$$\mathcal{Q}_k^d := \left\{2^{-k}, 2^{-(k-1)}, \ldots, 2^{-1}, 2^0, 2^1, \ldots, 2^{k-1}, 2^k\right\}^d \tag{3.1}$$

with $k, d \in \mathbb{N}$ was used.

Since we are using a discrete approach, it is reasonable to normalize the predictors. Instead of using the original predictors, for now $\hat{X}_i$, we work with

$$X_i := \frac{\hat{X}_i}{\max\left(\hat{X}_i\right)}$$

for all $i \in \{1, \ldots, n\}$ with $\hat{X}_i \not\equiv 0$, because we do not want one predictor to have more weight than another predictor from the start. If for the normalized predictors the conditions of Lemma 1.7 are fulfilled, one can assume that $E\bar{L}_n^h \geq L^*$. Combined with the mathematical requirements of the previous two chapters, this causes $\bar{L}_n^{H_n}$ with

$$H_n := \arg\min_{h \in \mathcal{Q}^d} \bar{L}_n^h$$

to be a good estimate of the minimum mean squared error.

To calculate $\bar{L}_n^h$, the range of the d-dimensional observation vectors is clustered with a cubic partition $A_{n,j}$ of side-length $\tilde{h}_n = n^{-1/(1+d)}$ (cf. (1.2) and (1.3)). For any metric $\|\cdot\|_h$ a cube with respect to the Euclidean metric becomes a cuboid of side length

$$\left(c_1^h, \ldots, c_d^h\right) := \left(\frac{\tilde{h}_n}{h_1}, \ldots, \frac{\tilde{h}_n}{h_d}\right)$$

if $h_i > 0$ for all $i \in \{1, \ldots, d\}$ (cf. Section 1.4).

The clustering

$$\left\{A_{n,j}^h : i \in \mathbb{N}\right\} \tag{3.2}$$

with cuboids of side length $c_1^h, \ldots, c_d^h$ is feasible because $c_i^h < 1$ if $h_i \geq 1$ for any $i \in \{1, \ldots, d\}$, in case a logarithmic scale is used. This follows from $\tilde{h}_n < 1$ for $n \geq 1$.

3.2 Randomization

To compute the WSNN to one X_i contained in a cuboid with more than one point, it is necessary to randomize among points in the cuboid. This is done by generating independent and on [0,1] identically uniformly distributed random numbers $U_1, \ldots, U_n$. For randomizing in a cell containing $X_i, X_{i_1}, \ldots, X_{i_r}$ (with $r \geq 1$) the WSNN of X_i, chosen at random, has the same index as the NN of U_i among $U_{i_1}, \ldots, U_{i_r}$ (cf. Section 1.3).

A problem that concerns the product-limit estimator is that it is '*inaccurate at the tail* ' (Fan and Gijbels (1996) [11], p.169). This means that the values of the product-limit estimator close to its upper limit T_K are unstable, thus differing considerably for two sets of data even if they are drawn from independent and identically distributed random variables. In the computation of the estimator $\bar{L}_n^h$ (cf. (2.3)) one has to divide by the product-limit estimator. Thus, small differences in the product-limit estimator cause large changes of $\bar{L}_n^h$ especially for the small values of the product-limit estimator close to its upper limit T_K. As we split the sample to calculate the product-limit estimator with the second part and the WSNN with the first part, the results may differ depending on how the sample is split. In order to avoid this, it is necessary to shuffle the data before $\bar{L}_n^h$ is calculated and to perform $s \geq 1$ runs. Each produces results $\bar{L}_{n,1}^h, \ldots, \bar{L}_{n,s}^h$ for each $h \in \mathcal{Q}^d$. This procedure is repeated and

$$H_{n,s} := \arg\min_{h \in \mathcal{Q}^d} \bar{L}_n^{h,s} \tag{3.3}$$

$$:= \arg\min_{h \in \mathcal{Q}^d} \frac{1}{s} \sum_{i=1}^{s} \bar{L}_{n,i}^h. \tag{3.4}$$

is chosen as the optimal vector for weights.

3.3 A Remark about Dependent Random Variables

Assume two random variables $X^{(1)}$ and $X^{(2)}$ are dependent. Furthermore, assume all the information of $X^{(2)}$ can also be derived from $X^{(1)}$. Then (in the uncensored case) $\tilde{L}_n^h$ will be minimal for any $h \in \mathcal{Q}^d$ for which h_1 is the optimal weight of $X^{(1)}$ and h_2 can take almost any value as long as it does not influence h_1. The reason for this is that for a given h_1 all points (x_1, x_2) will be inside a rectangle of side length h_1 and h_2 (by dependency of $X^{(1)}$ and $X^{(2)}$, cf. Figure 3.1). Therefore, if the rectangle is nonempty, any value $h_2 \geq h^*$ will lead to a similar value $\bar{L}_n^h$ with $h = (h_1, h_2)$. The only difference comes possibly from randomization inside clusters. Thus minimization of $\tilde{L}_n^h$ with respect to $h \in \mathcal{Q}^d$ may lead to large weight of $X^{(1)}$ and smaller or undecided weight of $X^{(2)}$ even if $X^{(2)}$ contains important information about Y.

For an example let $X^{(1)}$ have be normal distribution with mean zero and variance one and let

$$X^{(2)} = \begin{cases} 1 & : \quad X^{(1)} < 0 \\ 0 & : \quad otherwise. \end{cases}$$

Set $Y = X^{(1)}$. The minimal value of $\tilde{L}_n^h$ using a logarithmic scale $\mathcal{Q}_7^2$ is attained for $h_1 = 128$ and any $h_1 \in \{2^{-7}, 2^{-6}, \ldots, 2^7\}$. The multiple weights of $X^{(2)}$ are due to the fact that no randomization took place as no cuboids with more than one data point existed for $h_1 = 128$. This example also shows that the weight of a predictor variable may diminish if a relevant dependent predictor is added. $X^{(2)}$ itself contains a lot of information about Y, but becomes insignificant when $X^{(1)}$ is present. Also, if an important predictor variable is removed, the weight of dependent variables will increase.

3.4 Application to Breast Cancer Data

In general, breast cancer patients are diagnosed on the basis of a set of predictors. In addition, the survival time and the censoring time, indicating if and when a patient drops out of the follow-up programme are collected. Since the data is censored, we would want them to satisfy conditions **(A1)** to **(A3)** (cf. Section 2.2). Condition **(A1)** is fulfilled, if

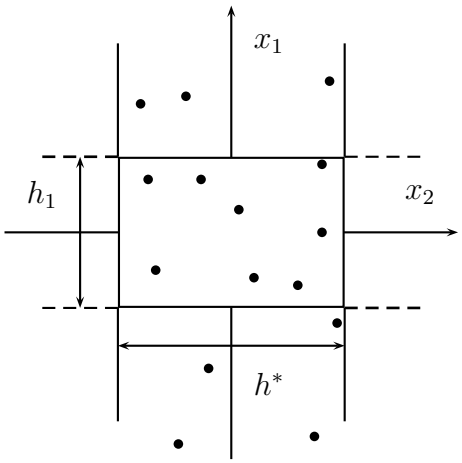

Figure 3.1: For $h_1 \geq h^*$ the WSNN-structure remains unchanged

the censoring is not dependent on the set of predictors or the survival time. We have to assume that the reason for leaving the follow-up programme is not related to the patient's illness and the collected personal data. The survival time is naturally bounded and also the reason for leaving the follow-up programme should still be relevant after the patient's death. In this case **(A2)** is valid. **(A3)** states that the patients that are longest in the follow-up programme are required to have censored as well as uncensored survival time. For the sets of data considered here it can be observed that the uncensored observations are thinning out as the survival time increases. However, theoretically speaking, it seems to be reasonable that even close to the upper bound of the minimum of survival and censoring time there are repeatedly patients with censored and uncensored survival time. Under the conditions described above, the rates of convergence derived in Chapters One and Two apply.

The set of predictors contains AGE which is measured in fractions of years. Furthermore, MENO (menopause status), ER (estrogen receptor status), PR (progesterone receptor status) and PM (occurrence of metastases) are binary (0 for negative and 1 for positive). HISTO (histological type) takes three values (ductal, lobular and other). PN (number of affected lymph nodes) is grouped into four classes {0,1,2,3}. PT is the tumor size (grouped into four classes {1,2,3,4}) and GR (grading of the tumor) takes values in {1,2,3} indicating a well- or poorly differentiated tumor. CERB indicates the amount of cerb2

proteine in the cell membrane of the tumor and takes values in {0,1,2,3}. The survival time is measured in fractions of years.

Another important condition for the rates of convergence derived in Chapters One and Two are that ties occur with probability zero. This holds if at least one predictor variable has a density. For the following discussion it is assumed that AGE (measured in fraction of years) has a density.

As the variable HISTO is nominal in three categories, we define the corresponding metric for patient data $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d)$ and $h \in \mathcal{Q}_k^d$ as

$$\begin{aligned} d(x,y)_h &= \| x - y \|_h \\ &= \sqrt{h_1^2 (p_{x,y}^{(1)})^2 + \ldots h_d^2 (p_{x,y}^{(d)})^2}. \end{aligned}$$

Here we set, if i indicates HISTO,

$$p_{x,y}^{(i)} := \begin{cases} 1 & : \text{ if } x_i \neq y_i \\ 0 & : \text{ otherwise.} \end{cases}$$

With respect to HISTO we only measure if the categories fit. In all other cases (AGE, MENO, PN, PT, GR, ER, PR and CERB), set

$$p_{x,y}^{(i)} = x_i - y_i.$$

Two sets of data have been analyzed. The first set of data (called RBK data) was supplied by the Robert-Bosch-Krankenhaus in Stuttgart, Germany. It contains $n = 913$ patients. It includes the predictors AGE, MENO, HISTO, PM, PN, PT, GR, ER, PR and CERB. The variable PT was not considered, as not enough values in the categories 3 and 4 were found. The second set of data (called OSP data) was collected from different hospitals in the city of Stuttgart by the Onkologischer Schwerpunkt Stuttgart. Here, $n = 1221$ cases with respect to AGE, MENO, HISTO, PM, PN, PT, GR, ER and PR were considered. The predictor CERB is not contained in this database. The predictor MENO was not used since some entrees were unreliable. In both data sets only patients with all values present (no entries with 'non-applicable') were considered.

Furthermore, it is well known that the occurrence of metastases has a large impact on

the treatment of breast cancer. In many cases, when metastases are found, the treatment is very different from the treatment of cases where no metastases are found. Thus only patients without metastases are considered for the studies.

From the RBK set of data AGE, MENO, HISTO, PN, GR, ER, PR and CERB were used. Regarding the OSP set of data, AGE, HISTO, PN, PT, GR, ER and PR were considered. The RBK set of data may expected to be more homogenous as it comes form one hospital. Due to computational considerations a logarithmic scale with $k = 2$ (cf. 3.1) was chosen.

Table 3.1: Weights for patient data from the Robert-Bosch-Krankenhaus.

study	runs	AGE	MENO	HISTO	PN	GR	ER	PR	CERB	L^2 error
1	63	2	4	1	4	2	4	0.25	4	1.971428
2	53	1	4	0.25	4	8	1	1	16	1.917742
3	65	4	4	0.5	8	32	2	4	32	1.936605

With each set of data three studies were performed. In each study a certain number of runs (from 53 to 95) were implemented (cf. Tables 3.1 and 3.2 and Figures 3.2–3.6). In one run for each vector of weights $h \in \mathcal{Q}_2^d$ the set of data was split, the product-limit estimator and the WSNN-structure computed, and thus the estimated minimum mean squared error calculated. One study consisted of s runs on the basis of which $\bar{L}_n^{H_{n,s},s}$ and its argument $H_{n,s}$ were calculated (cf. (3.3) and (3.4)). The logarithmic scale depended on the outcome of the previous study. If $H_{n,s}^{i-1} = (H_1, \ldots, H_d)$ was the argument of the

Table 3.2: Weights for data from the Onkologischer Schwerpunkt Stuttgart.

study	runs	AGE	HISTO	PN	PT	GR	ER	PR	L^2 error
1	95	4	0.5	4	4	4	0.5	1	2.382029
2	70	8	1	8	4	8	0.125	4	2.415973
3	73	32	0.25	32	1	16	0.125	1	2.348019

minimum in the $(i-1)$-th run, the logarithmic scale of the i-th run were multiples of the argument

$$\mathcal{Q}_{2,i}^{d} := \left\{ \hat{h} \in \mathbb{R}^{\mathrm{d}} : \exists h \in \mathcal{Q}_2^d \text{ with } \hat{h}_j = h_j \cdot H_j \text{ for all } j = 1, \ldots, d \right\},$$

where $\hat{h} = \left(\hat{h}_1, \ldots, \hat{h}_d\right)$ and $h = (h_1, \ldots, h_d)$. One run with the OSP data took about one day and with the RBK data about four days on a Pentium 3 GHz Single Core computer.

An examination of the six studies (cf. Figures 3.2–3.6) shows that minima are unique the latest at $i = 21$ (cf. Figure 3.3). In Study Three of RBK and Studies Two and Three of OSP data minima are already unique for a small number of runs. The speed of convergence becomes faster as the process of selecting a minimum is repeated using a different logarithmic scale. However, for both sets of data there is an increase of estimated L^2 error (for the RBK data from Study Two to Three and for the OSP data from Study One to Study Two (cf. Tables 3.1 and 3.2)). In Study Two of the OSP data the minimum is unique rather late and therefore the convergence may set in late (cf Figure 3.3). In Study Two of the OSP data minima are unique from the start and the convergence appears to be smooth (cf. Figure 3.6). However, the estimated L^2 error is over that of Study One. This emphasizes the fact that the product-limit estimator depends very much on how the data is split.

Concerning the estimated minimum mean squared error, the RBK data has smaller risk than the OSP data (1.936605 as compared to 2.348019, cf. Tables 3.1 and 3.2). From this it can be deduced that the predictors MENO and CERB together are more important than PT.

Comparing the results of the two sets of data, the weights of the predictors differ, which is not surprising as the set of predictors are different. However, predictors can be grouped into three categories. In group 2 predictors with weights from 1 to 4 are collected (cf. Table 3.3). Here AGE and MENO (from the RBK set of data) as well as PR and PT are found. The predictor ER from the RBK set of data belongs to this category. It loses weight in the OSP data set (from 2 to 0.125) and dropping into group 1 as a consequence. Also PR loses weight in the OSP set of data (from 4 to 1). This may be due to the fact

Table 3.3: Grouping of predictors of the RBK- and OSP-data according to weight. First weight is from the RBK, second from the OSP set of data if two weights are given.

	group 1	group 2	group 3
weights	0.125-0.5	1-4	8-32
predictors	HISTO [0.5,0.125] ER (OSP)[0.125]	AGE (RBK)[4] MENO [4] PR [4,1] ER(RBK) [2] PT [1]	CERB [32] AGE (OSP) [32] GR [32,16] PN [8.32]

that the OSP set of data includes PT, tumor size, which may depend on the estrogen and progesterone receptor status. The predictor HISTO is in both sets of the lowest category. Of the nine predictors present in the two sets of data, four are in the highest category (group 3, weights from 8 to 32). Among them are CERB, GR and PN. The predictor AGE is in this category if the OSP set of data is used. This may be due to the fact the MENO is not part of this set of data, as menopause pause is dependent on the age of the patient. Also PN has in the RBK set of data less weight than in the OSP set of data (8 as compared to 32). This may indicate that PN is dependent on CERB.

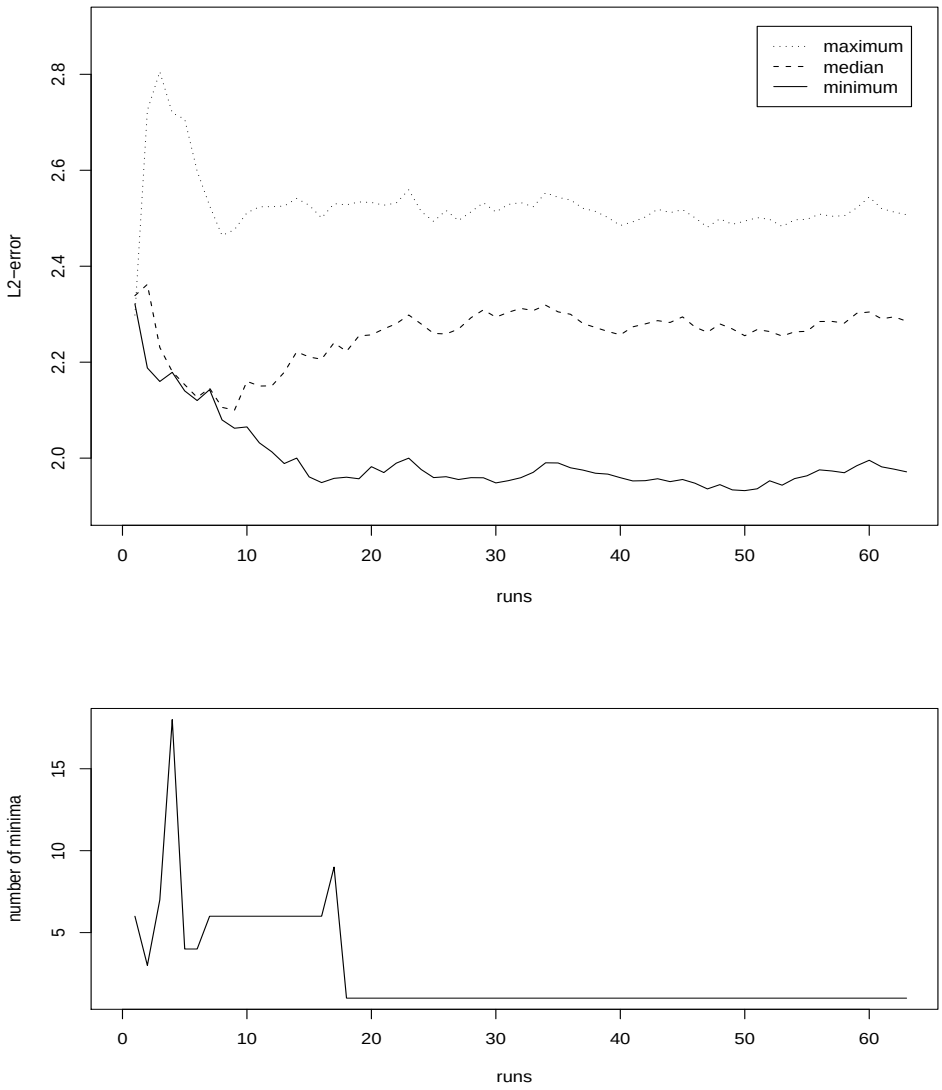

Figure 3.2: For s runs the values $\bar{L}_n^{h,i}$ for $i = 1, \ldots, s$ of the arguments of maximum, median and minimum of $\bar{L}_n^{h,s}$ are plotted in the first graph. The second graph shows the number of minima of $\bar{L}_n^{h,i}$ for each run (see also Figures 3.3 to 3.6). Here $s = 63$ in Study One of RBK data.

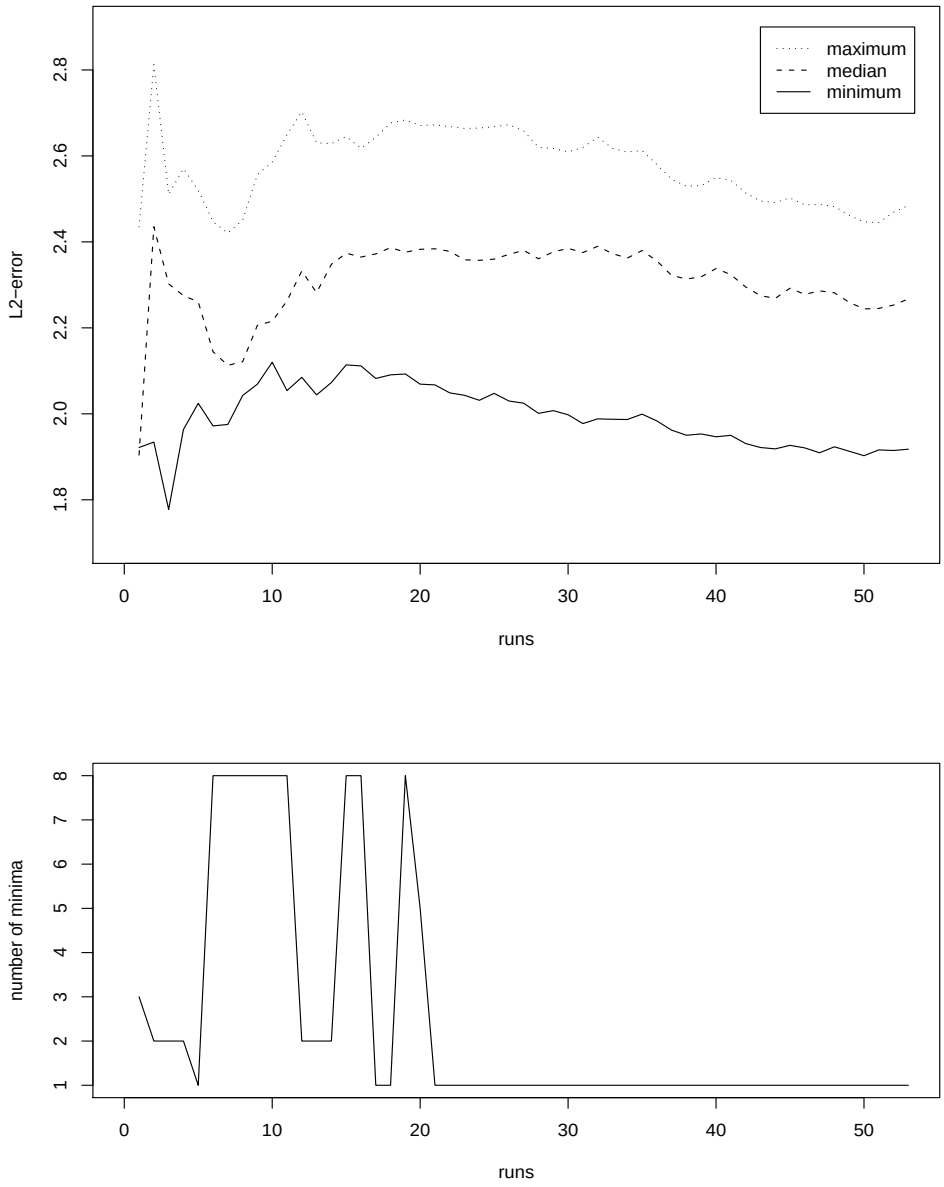

Figure 3.3: Study Two of RBK data with $s = 53$ runs (cf. Figure 3.2). Decreasing estimated L^2 error and unique minima starting at $i = 21$.

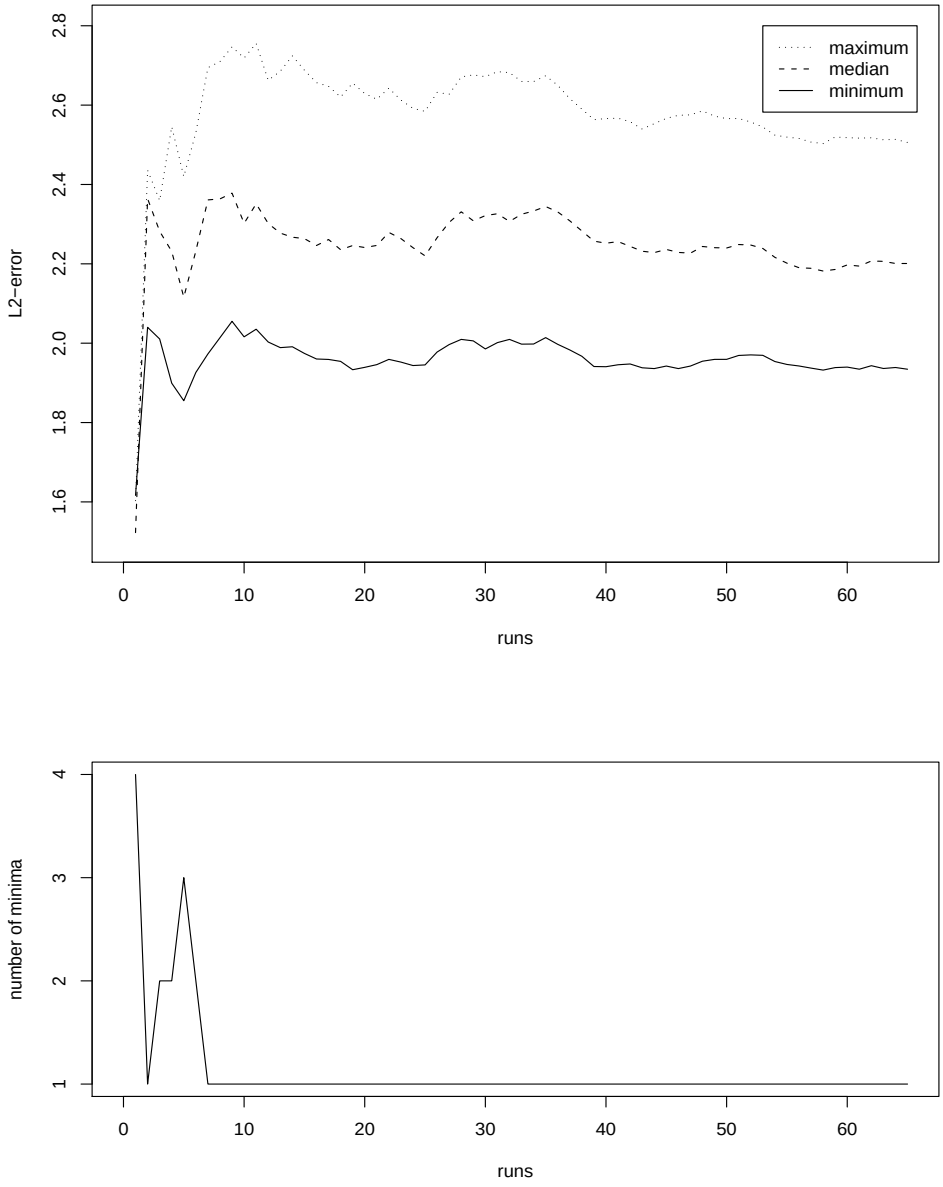

Figure 3.4: Study Three of RBK data with $s = 65$ runs (cf. Figure 3.2). Estimated L^2 error smaller than in Study One but larger than in Study Two.

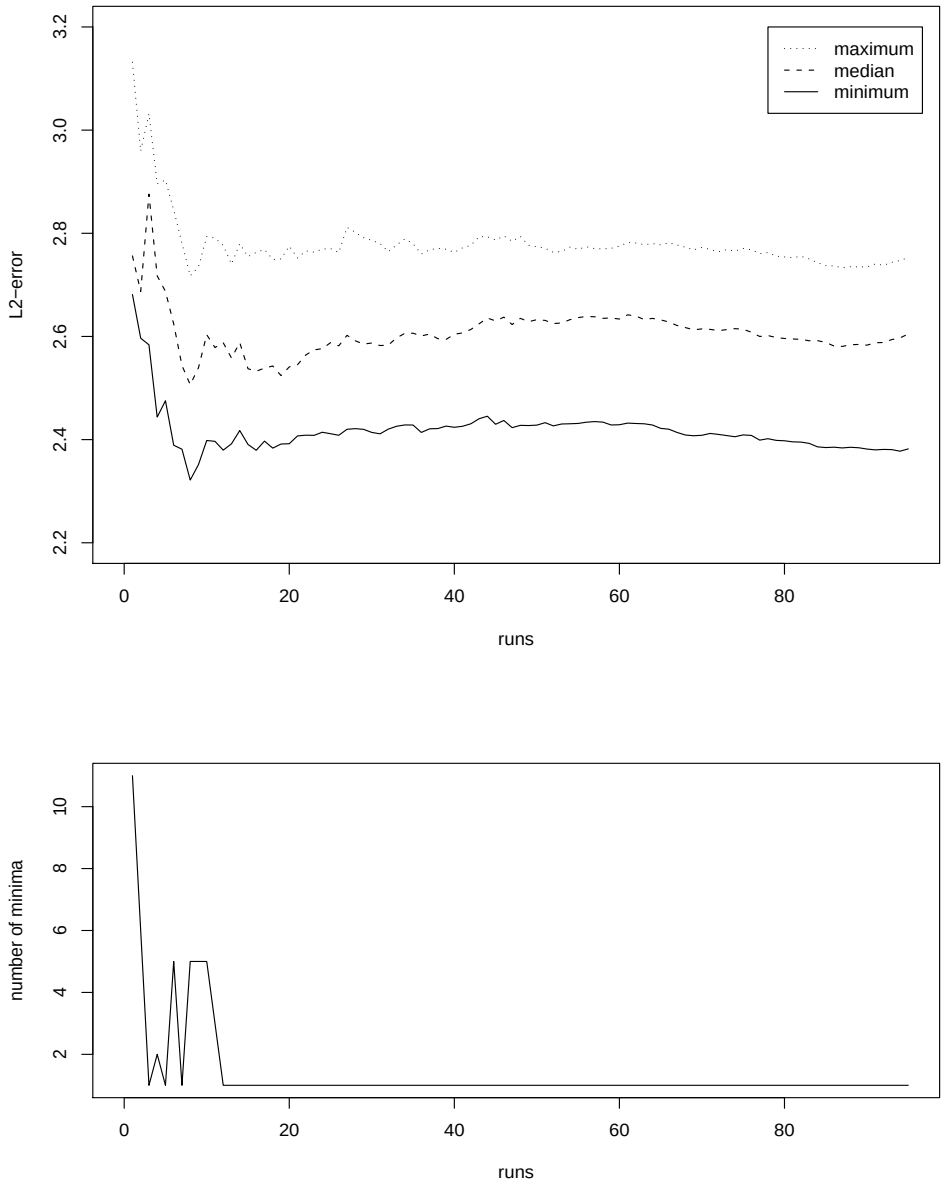

Figure 3.5: Study One of OSP data with $s = 95$ runs (cf. Figure 3.2).

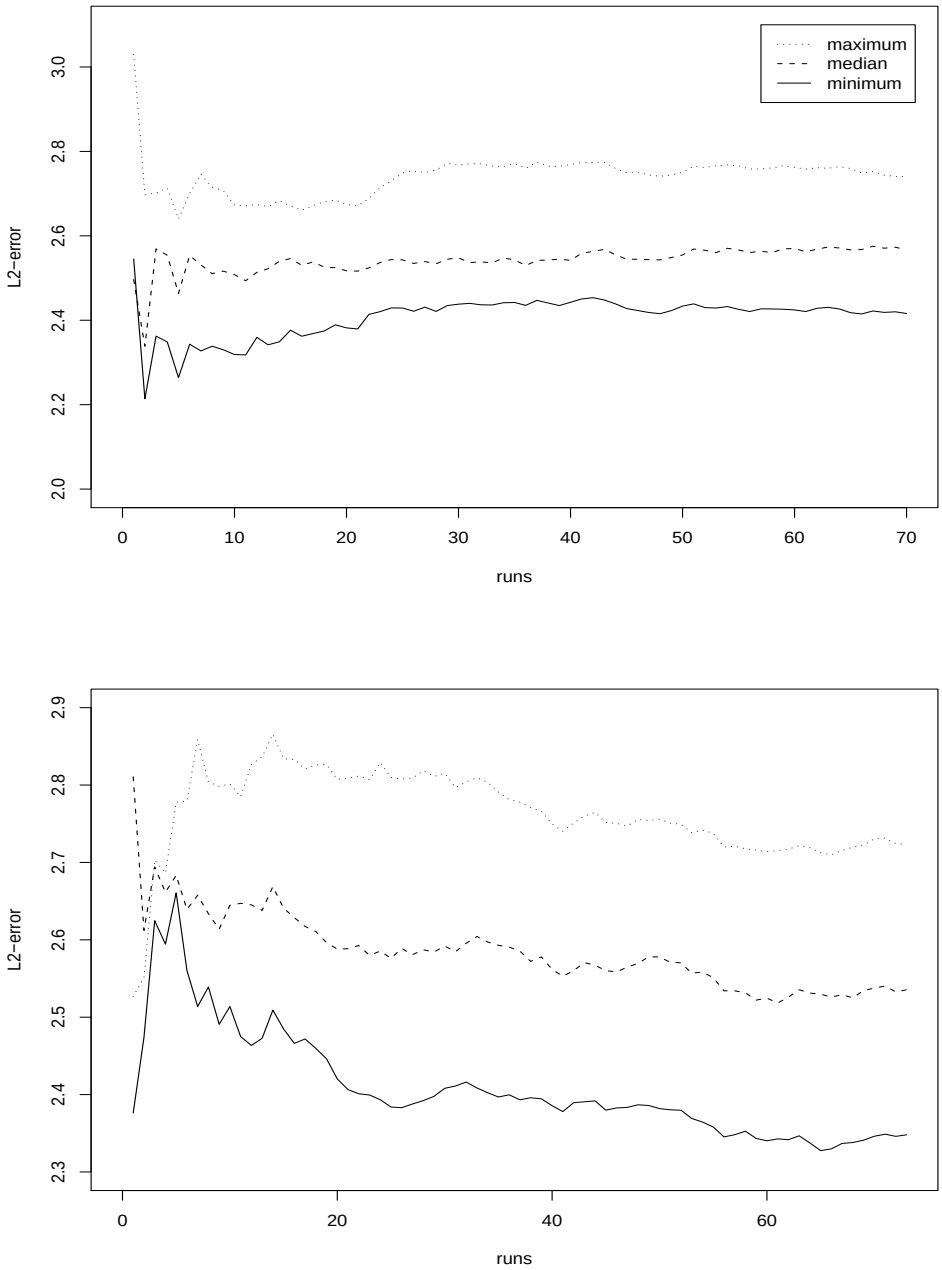

Figure 3.6: Study Two and Three of OSP data with $s = 70$ and $s = 73$ runs (cf. Figure 3.2). Minima are unique throughout.

Appendix A

Tools

Proposition A.1 *With* $N = \{1, \ldots, n\}$ *, one has for any function*

$$f : (N, \mathbb{R}^{\mathrm{d}n}) \longrightarrow N$$

with the property $f(i, x_1, \ldots, x_n) \neq i$ *and* $m(X_{f(i)}) = E\left(Y_{f(i)} \middle| X_{f(i)}\right)$

$$E\left(Y_i Y_{f(i)}\right) = E\left(m(X_i) m\left(X_{f(i)}\right)\right).$$

Proof: With

$$A_{ij} := \{(x_1, x_2, \ldots, x_n) : f(i, x_1, \ldots, x_n) = j\}$$

we have

$$E\left(Y_i Y_{f(i)}\right) \;=\; E\left(\sum_{j=1:j\neq i}^{n} E\left(Y_i Y_j \middle| A_{ij}\right) \cdot P(A_{ij})\right).$$

Therefore, it suffices to show that for all $i \neq j$

$$E\left(Y_i Y_j | A_{ij}\right) = E\left(m(X_i) m(X_j) \middle| A_{ij}\right) \text{ a.s.}$$

This follows from

$$\begin{aligned} E\left(Y_i Y_j \middle| A_{ij}\right) &= E(E(Y_i Y_j | X_i, X_j) | A_{ij}) \\ &= \int\limits_{A_{ij}} E\left(Y_i Y_j | X_i = x_i, X_j = x_j\right) \, dP_{(X_i, X_j)}(x_i, x_j) \\ &= \int\limits_{A_{ij}} E\left(Y_i Y_j | X_i = x_i, X_j = x_j\right) \, dP_{X_i}(x_i) \, dP_{X_j}(x_j), \end{aligned}$$

where we used the independence of X_i and X_j. For any measurable set $B_i \times B_j$ we arrive at

$$\begin{aligned}
&\int\limits_{B_i \times B_j} E(Y_i Y_j | X_i = x_i, X_j = x_j)\, dP_{X_i}(x_i)\, dP_{X_j}(x_j) \\
= \;& E\left(Y_i Y_j \cdot 1_{\{X_i \in B_i, X_j \in B_j\}}\right) \\
= \;& E\left(Y_i \cdot 1_{\{X_i \in B_i\}} Y_j \cdot 1_{\{X_j \in B_j\}}\right) \\
= \;& E\left(Y_i \cdot 1_{\{X_i \in B_i\}}\right) E\left(Y_j \cdot 1_{\{X_j \in B_j\}}\right) \\
& \quad \text{(by independence of } Y_i \text{ and } Y_j\text{)} \\
= \;& \int\limits_{B_i} m(x_i)\, dP_{X_i}(dx_i) \cdot \int\limits_{B_j} m(x_j)\, dP_{X_j}(dx_j) \\
= \;& \int\limits_{B_i \times B_j} m(x_i) m(x_j)\, dP_{X_i}(x_i)\, dP_{X_j}(x_j) \\
& \quad \text{(by independence of } m(X_i) \text{ and } m(X_j)\text{).}\square
\end{aligned}$$

The next inequality is useful when handling a sum of independent random variables. In the context of this work it is used to guarantee a rate of convergence of $n^{-\frac{1}{2}}$ if the supremum below can be bounded by $const \cdot n^{-1}$.

Theorem A.1 (McDiarmid, 1989) *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A and assume that*

$$f : A^n \longrightarrow \mathbb{R}^{\mathrm{d}}$$

satisfies for all $1 \leq i \leq n$

$$\sup_{x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i.$$

Then, for all $\epsilon > 0$

$$P(f(X_1, \ldots, X_n) - Ef(X_1, \ldots, X_n)) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2},$$

and

$$P(E(f(X_1, \ldots, X_n)) - f(X_1, \ldots, X_n)) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2}.$$

Proof: See McDiarmid (1989) [17]. $\square$

Appendix B

R Coding

The program creates a fixed number of runs. It uses a logarithmic scale with $k = 2$ starting at a vector `hstart`. The predictor HISTO is expected to be the third column of the data frame and is split into three additional columns which are dilated by the same factor. The main loop is designed for a data frame of dimension eight. The matrix `results` is saved after each run.

```
# ---initialization ---
library(class); # for nearest neighbour
library(survival) # for Product-Limit estimator

print( "All objects erased.");
rm(list=objects());
w=readline("name of data file (includes dataframe 'data.num'):  ")
load(w);

print("size of matrix:")
d=length(data.num[1,])-2;
numberofpatients=length(data.num[,1]);
n=trunc(numberofpatients/2); # WSNN on 1st half

print(paste("d= ",d));
print(paste("n= ",n));
```

```
#--- number of runs ---

runs=as.numeric(readline("number of runs "));
runsnames=as.numeric(readline("number of 1st run "));

for (runnumber in 1:runs) {

print(paste(runnumber,".  run"));
print(date())

#--- shuffling ---

print("Shuffling 'data.num'.");
print("Results are saved after each run.")
for (j in 1:2){
n2=length(data.num[1,]);
n1=length(data.num[,1]);
dx=as.vector(sample(n1,replace=FALSE));
c=data.frame(data.num)
for (i in 1:n1) {c[i,]=data.num[dx[i],]}
data.num[1:n1,]=c[1:n1,]
}
realrunnumber=runsnames+runnumber-1;

#--- side length of partition ---

cubuslength=n^(-1/(1+d));

#--- preparing data matrix ---

# 1 creating data.numonly with no factors
```

```
# column 3 (Histo) as 3 separate predictors

Histonew=matrix(0,numberofpatients,3)
for (i in 1:numberofpatients)

{if (data.num[i,3]=="ductal") {Histonew[i,1]=1};
if (data.num[i,3]=="lobular") {Histonew[i,2]=1};
if (data.num[i,3]=="other") {Histonew[i,3]=1}
}

dnum=d+2;
data.numonly=cbind(data.num[,1:2],Histonew[,1:3],data.num[,4:d]);
names(data.numonly)[3]="ductal";
names(data.numonly)[4]="lobular";
names(data.numonly)[5]="other";
data.numonly[,10]=as.numeric(data.numonly[,10]);

# 2 normalize data.num

print("Create normalized data.numonly.")

for (t in 1:dnum) {data.numonly[,t]=data.numonly[,t]/max(data.numonly[,t])}

# --- creating event-vector ---

event=vector(mode="logical",n)

print("Category 'unknown' is considered as related event.");

for (k in 1:numberofpatients)
{event[k]=FALSE;
```

```
if (data.num[k,(d+2)]=="mamma ca") {event[k]=TRUE};
if (data.num[k,(d+2)]=="unknown") {event[k]=TRUE}
}

# --- creating function Kaplan-Meier
# on 2nd half of data ---

lengthkm=numberofpatients-n;

eventkm=vector(mode="logical",(numberofpatients-n))

for (j in 1:(numberofpatients-n)) {eventkm[j]=event[n+j];
if (is.na(eventkm[j])==TRUE) {eventkm[j]=FALSE} }

time=data.num[(n+1):numberofpatients,(d+1)];
kmdata=Surv(time,eventkm);
kmfit=survfit(kmdata);

KaplanMeier=function(x)
{if (x <= kmfit$time[1]) {value=1}
else {value=kmfit$surv[sum(kmfit$time<x)]}
value }

kmvector=vector(mode="numeric",length=n);
for (i in 1:n) {kmvector[i]=KaplanMeier(data.num[i,(d+1)])};

# --- nearest neighbour function ---

NN=function(i) {

value=0;
```

```
class=c(1:(n-1));
data=data.numonlyh[,1:dnum];
test=data[i,];

if (i==1)
{train=data[2:n,];
class=c(2:n) }

if (i==n)
{train=data[1:(n-1),];
class=c(1:(n-1)) }

if ((i > 1) & ( i < n))
{ train=rbind(data[1:(i-1),],data[(i+1):n,]);
class=c(c(1:(i-1)),c((i+1):n)) }

value=as.numeric(as.vector(knn1(train,test,class)));
value }

# --- main program ---

z=0; # for results-matrix

h=vector(mode="numeric",length=d);

# --- main loop ---

# variables for loop

print(paste("dimension of results ",d));
print("Logarthimic scale with k=2 starting from")
```

```
hstart=c(1,2,0.25,4, 8 ,1 ,1,16)
print(paste(hstart,"."))
results=matrix(0,5^d,d+1);
hnum=vector(mode="numeric",length=dnum);
print("Creating random numbers for randomization in each cell.");
u=runif(n,0,1);

# --- creating h, dimensin d=8.---

h=NULL;
for (i in 1:d) {h=c(h,1)}

for (z1 in -2:2) {h[1]=(2^z1)*hstart[1];
for (z2 in -2:2) {h[2]=(2^z2)*hstart[2];
for (z3 in -2:2) {h[3]=(2^z3)*hstart[3];
for (z4 in -2:2) {h[4]=(2^z4)*hstart[4];
for (z5 in -2:2) {h[5]=(2^z5)*hstart[5];
for (z6 in -2:2) {h[6]=(2^z6)*hstart[6];
for (z7 in -2:2) {h[7]=(2^z7)*hstart[7];
for (z8 in -2:2) {h[8]=(2^z8)*hstart[8];

# --- creating extended hnum ---

for (i in 1:3) {hnum[i]=h[i]}
hnum[4]=h[3];
hnum[5]=h[3];
for (i in 4:d) {hnum[i+2]=h[i]}

# --- creating address for each data point and datanumonlyh ---

adress=matrix(0,n,dnum);
```

```
for (k in 1:dnum) { adress[,k]=trunc(data.numonly[1:n,k]/cubuslength*hnum[k])};


data.numonlyh=matrix(0,n,dnum);
for (i in 1:dnum) {data.numonlyh[,i]=data.numonly[1:n,i]*hnum[i]}

# --- WSNN-procedure with randomization ----

WSNN=vector(mode="numeric",length=n);
dmatrix=as.matrix(dist(adress));
diag(dmatrix)=1;

for (i in 1:n)

{ if (min(dmatrix[,i])==0)
{
cell=as.vector(which(dmatrix[,i]==0));
WSNN[i]=as.numeric(as.vector(knn1(u[cell],u[i],cell)));
}

else {WSNN[i]=NN(i)} }

# ----approximation of L2-error ---

y=vector(mode="numeric",length=n);
y2=vector(mode="numeric",length=n);
yWSNN=vector(mode="numeric",length=n);
yWSNN2=vector(mode="numeric",length=n);

y[1:n]=(data.num[1:n,(d+1)]/kmvector[1:n])*event[1:n];
y2[1:n]=((data.num[1:n,(d+1)])^2 /kmvector[1:n])*event[1:n];
```

```
yWSNN[1:n]=(data.num[WSNN[1:n],d+1]/kmvector[WSNN[1:n]])*event[WSNN[1:n]];
yWSNN2[1:n]=((data.num[WSNN[1:n],d+1])^2 /kmvector[WSNN[1:n]])*event[WSNN[1:n]];
L2=y2-2*y*yWSNN+yWSNN2;
L2sum=sum(L2);
L2error=L2sum/(2*n);
z=z+1;
results[z,]=c(h,L2error);

}}}}}}}
}

# --- save results ---
now=date();
save(results, data.num, now, file=paste("rnd.", realrunnumber));
}
```

Bibliography

[1] Bauer, H. (1991). *Maß- und Integrationstheorie*. De Gruyter, New York.

[2] Bauer, H. (1991). *Wahrscheinlichkeitstheorie*. De Gruyter, New York.

[3] Chen, K., Lo, S. (1997). On the rate of uniform convergence of the product-limit estimator: strong and weak laws. *Ann. Statist.*, 25, No 3, 1050–1087.

[4] Devroye, L., Györfi, L. (1983). Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function. *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, Konecny F., Mogyorodi, J. and Wertz, W., Eds., 66–76. Akadémiai Kiadó, Budapest.

[5] Devroye, L., Györfi, L., Krzyzak, A., Lugosi, G. (1994). On the strong universal consistency of nearest neighbour regression function estimates. *Ann. Statist.* 22, 1371–1385.

[6] Devroye, L., Györfi, L., Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

[7] Devroye, L., Györfi, L., Schäfer, D., Walk, H. (2003). The estimation problem of minimum mean squared error. *Statistics & Decisions* 21, 15–28.

[8] Devroye, L. Krzyzak, A. (1989). An equivalence theorem for L_1 convergence fo the kernel regression estimate. *Statist. Planning Inference*, 23, 71–82.

[9] Dippon, J., Fritz, P., Kohler, M. (2002). A statistical approach to case based reasoning with application to breast cancer data. *Comput. Statist. Data Anal.* 40 , 579–602.

[10] Fan, J., Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.* 89, 560–570.

[11] Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[12] Fukunaga, K., Flick, P., T. (1984). An optimal global nearest neighbour metric. *IEEE Trans. Pattern Analysis and Machine Intelligence* 3, 314–318.

[13] Gu, M., Lai, T. (1990). Functional laws of the iterated logarithm for the product-limit estimator of a distribution function under random censorship or truncation. *Ann. Statist.* 18, 160–189.

[14] Gyrfi, L., Kohler, M., Krzyzak, A., Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

[15] Kaplan, E., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53, 457–481.

[16] Liitiäinen, E., Corona, F., Lendasse, A. (2007) Nearest neighbor distributions and noise variance estimation. ESANN 2007, *European Symposium on Artificial Neural Networks*, Bruges (Belgium).

[17] McDiarmid, C. (1989). On the method of bounded differences. In *Survey in Combinatorics 1989*, 148–188. Cambridge University Press, Cambridge, UK.

[18] Myles, J. P., Hand, D. J. (1990). The multi-class metric problem in nearest neighbor discrimination rules. *Pattern Recognition* 23 (11), 1291–1297.

[19] Stone, C. J.(1977). Consistent nonparametric regression. *Ann. Statist.* 5, 595–645.

[20] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* 10, 1040–1053.

[21] Walk, H. (2005) Strong universal consistency of smooth kernel regression estimates, *Ann. Inst. Statist. Math.* 57, 665–685.