Smoothing Spline Regression Estimates for Randomly Right Censored Data

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

Vorgelegt von

Stefan Winter

aus Stuttgart

Hauptberichter:	Priv.-Doz. Dr. J. Dippon
Mitberichter:	Prof. Dr. I. Steinwart
Tag der mündlichen Prüfung:	05. Februar 2013

Institut für Stochastik und Anwendungen der Universität Stuttgart

2013

Acknowledgments

I am deeply grateful to my supervisor Priv.-Doz. Dr. Jürgen Dippon for his continuous support, patience and encouragement during writing this thesis. Furthermore, I like to thank Prof. Dr. Harro Walk, Prof. Dr. Michael Kohler, Dr. Kinga Máthé, Dr. Matthias Strobel, and Dipl.-Ing. Paola Ferrario for helpful discussions and valuable comments and advice. Also, all of my colleagues at the Institute for Stochastics and Applications for the pleasant working atmosphere. Finally, I thank my family for their caring support, endurance, and thoughtfulness.

Contents

Notations

$\emptyset$	empty set
$\mathbb{N}$	set of all natural numbers
$\mathbb{N}_0$	set of all natural numbers and zero
$\mathbb{R}$	set of all real numbers
$\mathbb{R}_+$	set of all nonnegative real numbers
$e; exp$	Euler's number, exponential function
ln, $\log_2$	natural logarithm (base e), binary logarithm (base 2)
$I_{[\mathcal{S}]}$	indicator function of the set $\mathcal{S}$
$\lvert\mathcal{S}\rvert$	cardinality of the set $\mathcal{S}$
$\lceil x \rceil$	upper integer part of $x \in \mathbb{R}$
$\lfloor x \rfloor$	lower integer part of $x \in \mathbb{R}$
$[a_1, a_2], (a_1, a_2)$	closed and open interval from $a_1 \in \mathbb{R}$ to $a_2 \in \mathbb{R}$
$(a_1, a_2], [a_1, a_2)$	left half-open and right half-open interval from $a_1 \in \mathbb{R}$ to $a_2 \in \mathbb{R}$
$\lVert \cdot \rVert$	Euclidean norm in $\mathbb{R}^d$
$\frac{\partial}{\partial x} f,\ f'_x$	partial derivative of the function $f : \mathbb{R}^d \to \mathbb{R}$ with respect to $x \in \mathbb{R}$
$\frac{\partial^{k_1+\ldots+k_d}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} f$	(weak) partial derivative of order $(k_1, \ldots, k_d)$ of the function $f : \mathbb{R}^d \to \mathbb{R}$ with respect to $x = (x_1, \ldots, x_d)$
$x_0 = \arg\min_{x \in D} f(x)$	x_0 satisfies $x_0 \in D$ and $f(x_0) = \min_{x \in D} f(x)$
a.s.; f.s.	almost surely; fast sicher
i.i.d.	independent and identically distributed

$\mathbf{P}$	probability
$\mathbf{P}[\,X = x\,]$	probability that the random variable X equals x
$\mathbf{E}[\,X\,]$	expected value of the random variable X
$\mathbf{Var}[\,X\,]$	variance of the random variable X
F, $F(t) = \mathbf{P}[\,Y > t]$	survival function of the random variable Y (evaluated at $t \in \mathbb{R}$)
G, $G(t) = \mathbf{P}[\,C > t]$	survival function of the random variable C (evaluated at $t \in \mathbb{R}$)
$\tau_F = \sup\{t \in \mathbb{R} : F(t) > 0\}$	upper endpoint of the distribution of Y
$\tau_K = \sup\{t \in \mathbb{R} : \mathbf{P}[Z > t] > 0\}$	upper endpoint of the distribution of Z
F_n; G_n, $\hat{G}^{(KM)}$	Kaplan-Meier estimates of F and G (Sections 1.2 and 2.4)
$\mathbf{P}[\,Y = y \mid X\,]$	conditional probability that the random variable Y equals y given X
$\mathbf{E}[\,Y \mid X\,]$	conditional expectation of Y given X
$\mathbf{Var}[\,Y \mid X\,]$	conditional variance of Y given X
$m(x) = \mathbf{E}[\,Y \mid X = x\,]$	regression function (Section 1.3)
$m^{[h]}(x) = \mathbf{E}[\,h(X,Y) \mid X = x\,]$	(generalized) regression function of $(X, h(X,Y))$ (Section 1.4)
$\sigma^2(x) = \mathbf{Var}[\,Y \mid X = x\,]$	conditional variance of Y given $X = x$ (Section 1.4)
$\sigma(x) = \lvert\sqrt{\sigma^2(x)}\rvert$	conditional standard deviation of Y given $X = x$
$F(\tau \mid x) = \mathbf{P}[\,Y > \tau \mid X = x\,]$	conditional survival function of Y at fixed point $\tau \in \mathbb{R}$ given $X = x$ (Section 1.4)
μ	distribution of X
$T_{[\cdot,\cdot]}$	truncation operator (Section 2.3)
$\mathcal{N}_r(\epsilon, \mathcal{F}, x_1^n)$	$\mathcal{L}_r$–ϵ–covering number of $\mathcal{F}$ on x_1^n (Appendix C)
$W_k([0,1]^d)$	Sobolev space of degree k on $[0,1]^d$ (Section 2.1)
$\mathcal{O}_{\mathbf{P}}$	rate of stochastic convergence (Section 1.3)
IQR	interquartile range (Section 6.3)
MSSE	multivariate smoothing spline estimate(s), possibly truncated (Chapter 2)
$J_k^2(\cdot)$	roughness penalty of MSSE (Section 2.1)

Zusammenfassung

Das Auftreten von zensierten Daten ist ein typisches Phänomen in vielen verschiedenen Bereichen innerhalb der Medizin, der Biologie, der Soziologie, der Qualitätskontrolle, der Risikotheorie oder auch der Demographie. Zensierte Daten entstehen immer dann, wenn die sogenannte *Überlebenszeit* nicht für alle untersuchten Studienteilnehmer oder Objekte in vollem Umfang beobachtet werden kann. Für die anderen liegt dann möglicherweise nur eine Teilinformation, die sogenannte *Zensierungszeit*, vor. Die Überlebenszeitanalyse, oder allgemeiner die Ereigniszeitanalyse, versucht Aussagen über die Überlebenszeit Y von solchen unvollständigen Daten abzuleiten. Von besonderem praktischen Interesse ist dabei die Untersuchung des Zusammenhangs zwischen Y und einem beobachteten Vektor kovariater Größen $X \in \mathbb{R}^d$.

Die nichtparametrische Regressionsanalyse stellt Techniken zur Verfügung, die hilfreich sind, um dieses Ziel zu erreichen. Außer zur Schätzung der bedingten mittleren Überlebenszeit kann sie zum Beispiel auch zur Schätzung der (bedingten) Überlebensfunktion oder der bedingten Varianz von Y angewendet werden. Im Gegensatz zur nichtparametrischen Regressionsanalyse benötigen die beiden anderen üblicherweise in der Überlebenszeitanalyse verwendeten Verfahren im Allgemeinen stärkere Voraussetzungen an die zugrunde liegende Verteilung der zensierten Daten. Diese Verfahren beruhen zum einen auf der Untersuchung des sogenannten *hazard risk* und zum anderen auf der parametrischen Regressionsschätzung.

Aus diesem Grund haben Techniken aus der nichtparametrischen Regressionsanalyse in den letzten beiden Jahrzehnten zunehmend an Aufmerksamkeit in der Überlebenszeitanalyse erlangt. Für zahlreiche Schätzverfahren wurde die schwache beziehungsweise starke Konsistenz unter unterschiedlichen Annahmen an den Zensierungsmechanismus sowie an

die Art der Abhängigkeit zwischen Überlebenszeit und Zensierungszeit nachgewiesen. Weit weniger ist jedoch über die Konvergenzraten solcher Schätzer bekannt, insbesondere wenn man keine Regularitätsvoraussetzungen an die Verteilung von X stellen möchte. Im Fall unzensierter Daten konnte Stone (1982) zeigen, dass für (p, B)-glatte Regressionsfunktionen $n^{-\frac{2p}{2p+d}}$ die optimale Konvergenzrate bezüglich des $\mathcal{L}_2$-Fehlers ist.

Ziel der vorliegenden Arbeit ist die Konstruktion und die Analyse nichtparametrischer Schätzer der bedingten mittleren Überlebenszeit, der bedingten Überlebensfunktion sowie der bedingten Varianz der Überlebenszeit. Besondere Beachtung gilt hierbei der Analyse der Konvergenzgeschwindigkeit dieser Schätzer in der Gegenwart zensierter Daten, ohne Regularitätsvoraussetzungen an die Verteilung von X zu stellen (außer, dass X beschränkt ist).

In dieser Arbeit wird ausschließlich das Modell der zufälligen Rechtszensierung betrachtet, welches auf eine Vielzahl wichtiger Anwendungen zutrifft. Hierbei wird angenommen, dass sowohl die Überlebenszeit Y als auch die Zensierungszeit C nicht-negative Zufallsvariablen mit unbekannter Überlebensfunktion $F(t) := \mathbf{P}[\, Y > t \,]$ beziehungsweise $G(t) := \mathbf{P}[\, C > t \,]$ $(t \in \mathbb{R})$ sind. Aufgrund dieses Zensierungstyps beobachtet man jedoch nur die Zufallsvariable $Z := \min\{Y, C\}$, das Minimum aus Überlebenszeit und Zensierungszeit, sowie die Zufallsvariable $\delta := I_{[\, Y < C \,]}$, welche die Beobachtung als zensiert $(\delta = 0)$ beziehungsweise unzensiert $(\delta = 1)$ kennzeichnet. Ziel ist es nun, ausgehend von einer Stichprobe der Verteilung von (X, Z, δ), Schätzer der bedingten mittleren Überlebenszeit, der bedingten Überlebensfunktion und der bedingten Varianz der Überlebenszeit zu konstruieren.

Es ist bekannt, dass die sogenannte *Regressionsfunktion* $m(X) = \mathbf{E}[\, Y \,|\, X \,]$ die Funktion des minimalen $\mathcal{L}_2$-Risiko bezüglich Y ist; in der Überlebenszeitanalyse entspricht m der bedingten mittleren Überlebenszeit. Weiterhin kann man zeigen, dass die bedingte Varianz σ^2 sowie die bedingte Überlebensfunktion $F(\, \tau | \cdot \,)$ $(\tau \in \mathbb{R}$ fest gewählt) von Y gegeben X mit den Regressionsfunktionen zu $(X, Y^2 - m(X)^2)$ beziehungsweise $(X, I_{[Y > \tau]})$ übereinstimmen. Dies bedeutet, dass $\sigma^2(X)$ und $F(\, \tau | \, X \,)$ die besten Approximationen von $Y^2 - m(X)^2$ beziehungsweise $I_{[Y > \tau]}$ bezüglich den entsprechenden $\mathcal{L}_2$-Risiken sind.

Im Rahmen dieser Arbeit erfolgt die Schätzung von m, σ^2 und $F(\, \tau | \cdot \,)$ daher mittels eines speziellen "Kleinste-Quadrate"-Ansatzes, den sogenannten multivariaten Smoothing-

Splineschätzern (MSSE).

In der gewöhnlichen Regressionanalyse werden diese Schätzer definiert durch Minimierung der Summe aus betrachtetem empirischem $\mathcal{L}_2$-Risiko und einem geeignet gewählten Strafterm $\lambda_n \cdot J_k^2(\cdot)$, der einer reinen Interpolation der Daten entgegenwirken soll, über einem Sobolevraum W_k vom Grad k ($k \in \mathbb{N}$ mit $2k > d$, $\lambda_n > 0$). Anschließend erfolgt im Fall fast sicher beschränkter abhängiger Variablen noch eine Stutzung der resultierenden Schätzfunktion. Im Gegensatz zu gewöhnlichen Kleinste-Quadrate-Schätzern haben MSSE den Vorteil, dass sie durch Lösen eines linearen Gleichungssystems einfach zu bestimmen und daher schneller zu berechnen sind.

Um MSSE von m, σ^2 und $F(\,\tau|\,\cdot\,)$ für zufällig rechtszensierte Daten zu konstruieren, wird in dieser Arbeit ein verallgemeinerter Ansatz der *censoring unbiased transformation* (siehe dazu z.B. Fan und Gijbels (1994, 1996) und El Ghouch und Van Keilegom (2008)) verwendet. Die Grundidee dieser Vorgehensweise ist, die zensierten Daten mittels einer geeignet gewählten Transformation in im Prinzip unzensierten Daten zu überführen. Das Verfahren hat den Vorteil, dass Schätzer aus der gewöhnlichen Regressionsanalyse direkt auf diese Daten angewendet werden können. Insbesondere lassen sich damit bereits aus der Literatur bekannte Ergebnisse für diese Schätzer in einfacher Art und Weise auf den Fall rechtszensierter Daten übertragen.

Im Rahmen dieser Arbeit wird für jede der drei Funktionen m, σ^2 und $F(\,\tau|\,\cdot\,)$ eine andere Klasse von Transformationen eingeführt. Die Transformation der Daten erfolgt dabei jeweils so, dass der bedingte Erwartungswert, die bedingte Varianz oder die bedingte Überlebensfunktion der transformierten Zufallsvariablen mit m, σ^2 beziehungsweise $F(\,\tau|\,\cdot\,)$ übereinstimmt. Für die Regressionsfunktion wird dafür der in Fan und Gijbels (1994, 1996) beschriebene Ansatz angewendet und auf Klassen von Transformationen für σ^2 und $F(\tau|\,\cdot\,)$ erweitert (vgl. hierzu El Ghouch und Van Keilegom (2008)).

Die oben genannten transformierten Zufallsvariablen sind jedoch alle von der unbekannten Überlebensfunktion G der Zensierungszeiten abhängig und somit in einer statistischen Anwendung nicht berechenbar. Daher führt man nun zunächst noch Schätzer der transformierten Datenpunkte ein, bei denen G durch den bekannten Kaplan-Meier-Schätzer G_n ersetzt wird. Die MSSE von m, σ^2 und $F(\,\tau|\,\cdot\,)$ für zensierte Daten werden nun analog zum unzensierten Fall auf der Grundlage dieser Daten definiert. Ein entscheidender Schritt in

der Analyse der Regressionsschätzer ist daher die Abschätzung der *Transformationsfehler*, das heißt der Beträge der Differenzen zwischen den transformierten Zufallsvariablen und ihren Schätzern.

In Bezug auf den MSSE der bedingten Varianz gilt es noch zu beachten, dass dieser unter Verwendung des MSSE der bedingten mittleren Überlebenszeit definiert wird, da σ^2 von m abhängt. Die Untersuchung des erstgenannten Schätzverfahrens erfolgt daher mit Hilfe vorher bewiesener Ergebnisse für den zweitgenannten Schätzer.

In der vorliegenden Arbeit werden die folgenden Regularitätsvoraussetzungen an die zugrunde liegende Verteilung von (X, Y, C) benötigt, um das Problem der Regressionsanalyse mit zensierten Daten auf die gewöhnliche nichtparametrische Regressionsanalyse zurückzuführen:

(RA1) $X \in [0,1]^d$ f.s.

(RA2) $\exists\, L \in (0,\infty)$, so dass $0 \leq Y \leq L$ f.s. und $\mathbf{P}\,[\,C > L\,] > 0$

(RA3) C und (X, Y) sind unabhängig

(RA4) G ist stetig.

Die Bedingungen **(RA1)** und **(RA4)** sind übliche Annahmen in der Konvergenzanalyse beziehungsweise der Überlebenszeitanalyse und stellen in einer praktischen Anwendung keine ernsthafte Einschränkung dar. Die Voraussetzungen **(RA2)** und **(RA3)** vereinfachen das mathematische Problem der Konvergenzanalyse. Sie sind zum Beispiel in Bezug auf Studien von fester Dauer und bei Unabhängigkeit von Zensierung und betrachteten kovariaten Größen realistisch.

Die Regularitätsbedingungen **(RA1)** – **(RA4)** sind ausreichend für den Nachweis der fast sicheren Konvergenz der maximalen quadratischen Transformationsfehler. Dieses Resultat wird in den Theoremen 3.2 – 3.4 der vorliegenden Arbeit verwendet, um unter geeigneten Annahmen an die Parameter der Schätzer die starke Konsistenz der MSSE von m, σ^2 und $F(\,\tau|\,\cdot\,)$ für alle Verteilungen von (X, Y, C), die **(RA1)** – **(RA4)** erfüllen, zu beweisen. Ähnliche Ergebnisse für andere nichtparametrische Schätzer von m und $F(\,\tau|\,\cdot\,)$ sind bereits unter schwächeren Voraussetzungen an die Verteilung von (Y, C) bekannt,

insbesondere wenn anstatt **(RA3)** nur gefordert wird, dass Y und C bedingt unabhängig sind gegeben X (siehe dazu Beran (1981), Dabrowska (1987, 1989) und Pintér (2001)).

Im Gegensatz dazu ist Theorem 3.3 nach dem Kenntnisstand des Verfassers der vorliegenden Arbeit das erste veröffentlichte Resultat, welches die starke Konsistenz eines nichtparametrischen Schätzers der bedingten Varianz für rechtszensierte Daten ohne Regularitätsvoraussetzungen an die Verteilung von X zeigt. Im Beweis von Theorem 3.3 werden die oben genannten Ergebnisse für die maximalen quadratischen Transformationsfehler sowie den MSSE von m verwendet. Der Nachweis der starken Konsistenz des Schätzers von σ^2 folgt dann aus einer Untersuchung des Abstandes des empirischen $\mathcal{L}_2$-Fehlers und des $\mathcal{L}_2$-Fehlers des MSSE von m.

Ist ein Schätzer konsistent, so konvergiert sein $\mathcal{L}_2$-Fehler mit wachsendem Stichprobenumfang gegen Null. Für statistische Anwendungen ist es aber sehr oft von entscheidender Bedeutung, wie schnell dies geschieht. Da die Konvergenzrate eines Schätzers ohne starke Einschränkungen an die Verteilung von (X, Y) beliebig langsam sein kann, werden in der gewöhnlichen nichtparametrischen Regressionanalyse üblicherweise Glattheitsbedingungen an m gestellt. Unter der Annahme einer p-mal stetig differenzierbaren Regressionsfunktion — insbesondere auch, dass m ein Element eines Sobolevraums vom Grad p ist — wurde im Fall unzensierter Daten für zahlreiche Schätzer gezeigt, dass sie die oben genannte optimale Konvergenzrate $n^{-\frac{2p}{2p+d}}$ erreichen (oder fast erreichen). Falls zensierte Daten auftreten, sind Glattheitsbedingungen in der Regel jedoch nicht ausreichend, um diese Eigenschaft nachzuweisen.

In Korollar 4.1 wird gezeigt, dass die stochastischen Konvergenzraten der in dieser Arbeit betrachteten MSSE von den Raten dieser Schätzer im unzensierten Fall sowie den Raten der mittleren quadratischen Transformationsfehler abhängen. Um ein nichttriviales Resultat für letztere zu erhalten, wird neben **(RA1)** – **(RA4)** eine zusätzliche Bedingung an die Verteilung von (Y, C) benötigt.

Unter den in dieser Arbeit getroffenen Regularitätsannahmen impliziert ein Ergebnis von Chen und Lo (1997) für den Kaplan-Meier-Schätzer G_n, dass $n^{-\gamma}$ mit $\gamma \in (0, 1)$ die Konvergenzrate der mittleren quadratischen Transformationsfehler ist, falls

$$-\int_0^{\tau_F} F(t)^{\frac{-\gamma}{2-\gamma}} dG(t) < \infty. \tag{1}$$

Mit Bedingung (1) wird gefordert, dass ausreichend viele zensierte Beobachtungen in der Nähe der oberen Schranke $\tau_F := \sup\{t \in \mathbb{R} : F(t) > 0\}$ der Verteilung von Y vorhanden sind, so dass G_n ein stabiler Schätzer von G ist.

Für beliebiges $p \in \mathbb{N}$ mit $2p > d$ wird in Theorem 4.2 gezeigt, dass der MSSE von m im Fall geeignet gewählter Parameter die stochastische Konvergenzrate

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}$$

erreicht, und zwar für jede Verteilung von (X, Y, C), die **(RA1)** – **(RA4)**, $m \in W_p([0,1]^d)$ mit $0 < J_p^2(m) < \infty$ und (1) mit $\gamma = \frac{2p}{2p+d}$ erfüllt. Hierbei gilt es zu beachten, dass in Theorem 4.2 bis auf **(RA1)** keine Annahme an die zugrunde liegende Verteilung von X benötigt wird. Insbesondere wird nicht gefordert, dass X eine Dichte bezüglich des Lebesgue-Borel-Maßes besitzt. Darüber hinaus kann man folgern, dass die Konvergenzrate in Theorem 4.2 bis auf den logarithmischen Faktor optimal ist.

Analog zu Theorem 4.2 wird mit Theorem 4.4 für beliebiges $p \in \mathbb{N}$ mit $2p > d$ und für beliebiges, jedoch festes $\tau \in \mathbb{R}$ nachgewiesen, dass der geeignet gewählte MSSE von $F(\tau|\cdot)$ für jede Verteilung von (X, Y, C), die **(RA1)** – **(RA4)**, $F(\tau\,|\cdot) \in W_p([0,1]^d)$ mit $0 < J_p^2(F(\tau\,|\cdot)) < \infty$ und (1) mit $\gamma = \frac{2p}{2p+d}$ erfüllt, die bis auf den logarithmischen Faktor optimale Konvergenzrate

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}$$

erreicht.

Um eine schnelle Konvergenzrate für den Schätzer von σ^2 zu gewährleisten (welcher wie oben beschrieben von dem MSSE von m abhängt), wird in dieser Arbeit sowohl eine Glattheitsbedingung an σ^2 als auch an m gestellt. In Theorem 4.3 wird für beliebige $p_1 \in \mathbb{N}$ und $p_2 \in \mathbb{N}$ mit $2p_1 > d$ sowie $2p_2 > d$ gezeigt, dass die stochastische Konvergenzrate des geeignet gewählten MSSE von σ^2 für jede Verteilung von (X, Y, C), die **(RA1)** – **(RA4)**, $m \in W_{p_1}([0,1]^d)$ mit $0 < J_{p_1}^2(m) < \infty$, $\sigma^2 \in W_{p_2}([0,1]^d)$ mit $0 < J_{p_2}^2(\sigma^2) < \infty$ sowie (1) mit $\gamma = \frac{2p_{max}}{2p_{max}+d}$ erfüllt, gegeben ist durch

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_{min}}{2p_{min}+d}} \tag{2}$$

Hierbei sind p_{min} und p_{max} gegeben durch $p_{min} := \min\{p_1, p_2\}$ und $p_{max} := \max\{p_1, p_2\}$.

Die Rate in (2) ist optimal bis auf den logarithmischen Faktor, falls $p_2 \leq p_1$. Wenn jedoch $p_2 > p_1$ in Theorem 4.3 gilt, dann wird zugelassen, dass m "rauer" ist als σ^2. In diesem Fall wird (2) von der Rate des MSSE der Regressionsfunktion bestimmt und kann sich deutlich von der optimalen Rate $n^{-\frac{2p_2}{2p_2+d}}$ unterscheiden (vgl. Cai, Levine und Wang (2009) für ein ähnliches Ergebnis im Fall unzensierter Daten).

In den Theoremen 4.2 – 4.4 wird angenommen, dass die Parameter der MSSE geeignet gewählt wurden. Diese Wahl hängt jedoch von der unbekannten Glattheit von m, σ^2 beziehungsweise $F(\,\tau|\,\cdot\,)$ ab. Die Schätzer aus den Theoremen 4.2 – 4.4, welche die fast optimalen Konvergenzraten erreichen, sind daher in einer statistischen Anwendung nicht berechenbar. Aus diesem Grund werden die bisher betrachteten MSSE in einem weiteren Schritt so modifiziert, dass die resultierenden Schätzer die Parameter rein datenbasiert wählen und sich automatisch an die Glattheit von m, σ^2 beziehungsweise $F(\,\tau|\,\cdot\,)$ anpassen.

Ein häufig verwendetes und einfach zu handhabendes Verfahren, das es ermöglicht, diese Ziele zu erreichen, ist die *splitting-of-the-sample*-Technik. Für jede mögliche Parameterkombination aus einer geeignet gewählten Menge werden dazu auf einem Teil der Daten, der sogenannten *Lernmenge*, MSSE von m, σ^2 und $F(\,\tau|\,\cdot\,)$ wie oben beschrieben berechnet. Für jede der drei Funktionen wird dann derjenige Schätzer ausgewählt, der auf dem übriggeblieben Teil der Daten, der sogenannten *Testmenge*, die beste Vorhersage liefert. In den Theoremen 5.2 – 5.4 wird gezeigt, dass die durch die splitting-of-the-sample-Technik definierten Schätzer von m, σ^2 und $F(\,\tau|\,\cdot\,)$ die selben Konvergenzraten wie die entsprechenden MSSE aus den Theoremen 4.2 – 4.4 erreichen, ohne dass man zusätzliche Bedingungen an die Verteilung von (X, Y, C) stellen muss. Da die Parameter der Schätzer hierbei rein datenabhängig gewählt werden, kann man daraus schließen, dass diese MSSE sich automatisch an die unbekannte Glattheit von m, σ^2 beziehungsweise $F(\,\tau|\,\cdot\,)$ anpassen.

Zu Theorem 4.2 und Theorem 5.2 analoge Resultate für Kleinste-Quadrate-Schätzer der Regressionsfunktion wurden bereits in Máthé (2006) nachgewiesen, jedoch unter etwas stärkeren Voraussetzungen an die Verteilung von (X, Y, C). Genauer gesagt wurde in Máthé (2006) neben **(RA1)** – **(RA4)** die Regularitätsannahme

$$\limsup_{t \uparrow \tau_F} \frac{G(t) - G(\tau_F)}{F(t)} < \infty \tag{3}$$

anstatt der schwächeren Bedingung (1) verwendet und vorausgesetzt, dass m eine (p, B)-

glatte Funktion ist ($p = k + \psi$ mit $k \in \mathbb{N}_0$ und $0 < \psi \leq 1$, $B \in (0, \infty)$), im Gegensatz zur Forderung $m \in W_p([0,1]^d)$ mit $0 < J_p^2(m) < \infty$ ($p \in \mathbb{N}$ mit $2p > d$) im Rahmen dieser Arbeit.

Soweit dem Verfasser der vorliegenden Arbeit bekannt, sind Theorem 4.3, Theorem 4.4, Theorem 5.3 und Theorem 5.4 dagegen die ersten veröffentlichten Resultate auf dem Gebiet der nichtparametrischen Regressionsschätzung mit zufälligem Design, welche die Konvergenzrate von Schätzern von σ^2 und $F(\tau|\cdot)$ für zensierte Daten untersuchen, ohne Regularitätsannahmen an die Verteilung von X zu stellen.

Die oben vorgestellten Ergebnisse für die MSSE von m, σ^2 und $F(\tau|\cdot)$ sind rein asymptotische Resultate. Für kleine bis mittlere Stichprobenumfänge wird die Güte der Schätzverfahren daher abschließend anhand realer und simulierter rechtszensierter Daten bewertet. Für diese Datensätze wird gezeigt, dass die in dieser Arbeit betrachteten Schätzer gute Approximationseigenschaften zeigen, die vergleichbar zu anderen aus der Literatur bekannten Schätzern für zufällig rechtszensierte Daten sind. Darüberhinaus werden verschiedene Modifikationen der MSSE vorgestellt und untersucht, die darauf abzielen, die Güte der Vorhersage bei begrenztem Stichprobenumfang weiter zu verbessern oder praktischen Problemen in statistischen Anwendungen Rechnung zu tragen.

Kurz zusammengefasst weisen die im Rahmen dieser Arbeit erzielten Resultate unter anderem nach, dass geeignet definierte nichtparametrische Regressionsschätzer von m, σ^2 und $F(\tau|\cdot)$ in der Gegenwart zufällig rechtszensierter Daten die optimale Konvergenzrate bis auf einen logarithmischen Faktor erreichen. Für die Beweise der vorgestellten Theoreme werden dabei weniger starke Voraussetzungen als für aus der Literatur bekannte verwandte Ergebnisse benötigt, sie gelten insbesondere ohne Annahmen an die zugrunde liegende Verteilung von X (außer, dass X beschränkt ist).

Summary

The occurrence of censored data is a typical phenomenon in many different areas of, e.g., medicine, biology, sociology, quality control, risk theory, and demography. Censored data arise whenever the time of interest, the so-called *lifetime* cannot be fully observed for all subjects or objects under study. For those only a partial information, the so-called *censoring time* may be available. In survival analysis or more general event history analysis, one seeks to draw conclusions for the lifetime Y from such incomplete data. The identification of the relationship between Y and a vector of covariates $X \in \mathbb{R}^d$ is of particular interest in this context.

Nonparametric regression analysis provides techniques which can help to achieve this aim. Besides the estimation of the conditional mean lifetime, it may be applied in order to estimate the (conditional) survival function and the conditional variance of Y. In contrast to nonparametric regression analysis the two other commonly used methods in survival analysis, hazard risk approaches and approaches based on parametric regression, usually require stronger regularity conditions on the underlying distribution of the censored data.

Due to this fact, the application of nonparametric regression techniques in survival analysis has attracted more and more attention during the last two decades. Many results have been obtained which show the weak and strong consistency of various estimates concerning different censoring mechanisms and modes of dependency of the lifetime and the censoring time. However, far less is known about the rate of convergence of such estimates in the presence of censored data, especially if one does not impose regularity conditions on the distribution of X. In case that no censoring arises, Stone (1982) proved that for (p, B)-smooth regression functions, the optimal rate of convergence with respect to the $\mathcal{L}_2$ error is given by $n^{-\frac{2p}{2p+d}}$.

The aim of the present work is the construction and the investigation of nonparametric estimates of the conditional mean lifetime, the conditional survival function, and the conditional variance of the lifetime. Special attention is devoted to the analysis of the rate of convergence of these estimates in the presence of censored data without assuming anything on the distribution of the design (beside that X is bounded).

In this thesis, we focus on the *randomly right censorship model*, which is valid for many important applications. This model assumes that the lifetime Y and the censoring time C are both non-negative random variables with unknown survival function $F(t) := \mathbf{P}[Y > t]$ and $G(t) := \mathbf{P}[C > t]$ ($t \in \mathbb{R}$), respectively. Due to the censoring mechanism, one only observes $Z := \min\{Y, C\}$, the minimum of the lifetime and the censoring time, and $\delta := I_{[Y<C]}$, which stores the information whether an observation is censored ($\delta = 0$) or uncensored ($\delta = 1$). Given a sample of the distribution of (X, δ, Z), we now seek to construct estimates of the conditional mean lifetime, the conditional survival function, and the conditional variance of the lifetime in the presence of censored data.

It is well known that the so-called *regression function* $m(x) = \mathbf{E}[Y \,|\, X = x]$ is the function of minimal $\mathcal{L}_2$ risk with respect to Y; in survival analysis, m represents the conditional mean lifetime. In addition, one can show that the conditional variance σ^2 and the conditional survival function $F(\tau|\cdot)$ ($\tau \in \mathbb{R}$ fixed) of Y given X are identical to the regression function of $(X, Y^2 - m(X)^2)$ and $(X, I_{[Y>\tau]})$, respectively. Hence, $\sigma^2(X)$ and $F(\tau|X)$ are the best approximations of $Y^2 - m(X)^2$ or $I_{[Y>\tau]}$ in terms of the corresponding $\mathcal{L}_2$ risks. Therefore, estimates of m, σ^2, and $F(\tau|\cdot)$ may be constructed by applying a (modified) least squares approach. In this thesis, multivariate smoothing spline estimates (MSSE) are considered.

In usual nonparametric regression analysis, these estimates are defined by minimizing the sum of the corresponding empirical $\mathcal{L}_2$ risk and a suitably chosen penalty term $\lambda_n J_k^2(\cdot)$, which is added in order to avoid overfitting, over a Sobolev space W_k of degree k ($k \in \mathbb{N}$ with $2k > d$, $\lambda_n > 0$). Subsequently, the resulting estimates are truncated in case that the dependent variables are bounded almost surely. In contrast to usual least squares estimates, MSSE have the advantage that they can simply be calculated by solving a linear system of equations and are therefore much faster to compute.

In order to derive MSSE of m, σ^2, and $F(\tau|\cdot)$ in the presence of randomly right

censored data, the present work uses a generalized approach of the *censoring unbiased transformation* (vide, e.g., Fan and Gijbels (1994, 1996) and El Ghouch und Van Keilegom (2008)). Here, the idea is that the censored data is first converted in a suitable manner to virtually uncensored data. The advantage of this approach is that estimates from usual regression analysis can simply be applied to these new data. In particular, results for these estimates to be found in literature can be easily transferred to regression in the presence of right censored data.

To be more precise, for each of the three functions m, σ^2, and $F(\,\tau|\,\cdot\,)$, we convert the censored data with a different kind of transformation such that the conditional mean, the conditional variance, and the conditional survival function of the transformed random variables is identical to m, σ^2, and $F(\tau|\,\cdot\,)$, respectively. For the estimation of the regression function, the approach of Fan and Gijbels (1994, 1996) is used and extended to classes of transformations for σ^2 and $F(\tau|\,\cdot\,)$ (see also El Ghouch und Van Keilegom (2008)).

However, the random variables of these transformations all depend on the unknown survival function G of the censoring time and are therefore not calculable in a statistical application. Therefore, in a second step, we construct estimates of the converted data points by replacing G with the well-known Kaplan-Meier estimate G_n. The MSSE of m, σ^2, and $F(\,\tau|\,\cdot\,)$ in the presence censored data are now defined on the basis of these data in analogy to the uncensored case. In our analysis, a basic requirement is therefore that the *transformation errors,* i.e., the absolute differences between the converted random variables and their estimates, are small.

Note that since σ^2 depends on m, the MSSE of the conditional variance is defined via the MSSE of the conditional mean. Hence, the investigation of the first estimate is based on results previously obtained for the second estimate.

In this thesis, the following regularity assumptions on the distribution of (X, Y, C) are required to reduce the problem of censored regression to usual nonparametric regression:

(RA1) $X \in [0,1]^d$ a.s.

(RA2) $\exists\, L \in (0,\infty)$ such that $0 \leq Y \leq L$ a.s. and $\mathbf{P}\,[\,C > L\,] > 0$

(RA3) C and (X, Y) are independent

(RA4) G is continuous.

Conditions **(RA1)** and **(RA4)** are common assumptions in the analysis of the rate of convergence or in survival analysis and are no serious constraints in a statistical application. Assumptions **(RA2)** and **(RA3)** simplify the analysis of the rate of convergence. They are realistic if, e.g., the data are collected during a study of fixed duration and the mechanism of censoring is independent of the covariates under study.

The regularity conditions **(RA1)** – **(RA4)** are sufficient in order to show the almost sure convergence of the maximum squared transformation errors. Based on this result, it is proven in Theorems 3.2 – 3.4 of the present work that under suitable conditions on the parameters of the estimates, the MSSE of m, σ^2, and $F(\tau|\cdot)$ are strongly consistent for all distributions of (X, Y, C) satisfying **(RA1)** – **(RA4)**. For other nonparametric estimates of m and $F(\tau|\cdot)$, similar findings were already obtained under weaker conditions on the distribution of (Y, C), especially if one only demands that Y and C are conditionally independent given X (vide Beran (1981), Dabrowska (1987, 1989), and Pintér (2001)).

In contrast, Theorem 3.3 is to the best knowledge of the author of this thesis the first result to be published, which shows the strong consistency of a nonparametric regression estimate of the conditional variance in the presence of censored data without imposing regularity assumptions on the distribution of X. In the proof of Theorem 3.3, we use the aforementioned results on the maximum squared transformation errors and the MSSE of m. The verification of the strong consistency of the estimate of σ^2 then follows from an analysis of the deviation of the empirical $\mathcal{L}_2$ error from the $\mathcal{L}_2$ error of the MSSE of m.

Consistency of an estimate means that its $\mathcal{L}_2$ error converges to zero while the sample size increases. But for statistical applications, it is very often essential to know how fast this happens. Since the rate of convergence may be arbitrarily slow if one does not impose strong restrictions on the distribution of (X, Y), one typically imposes smoothness conditions on m in usual nonparametric regression. Assuming a p times continuously differentiable regression function — or especially that m is a function in a Sobolev space of degree p — many different types of estimates have been shown to achieve (or to nearly achieve) the optimal rate $n^{-\frac{2p}{2p+d}}$ in case that censoring does not arise. However, in censored regression, smoothness conditions are generally not sufficient in order to verify this property.

In Corollary 4.1, it is shown that the rates of convergence of our MSSE depend on

the rate of these estimates in usual nonparametric regression and the rate of the mean squared transformation errors. In order to derive a nontrivial result for the latter one, an additional assumption to **(RA1)** – **(RA4)** on the distribution of (Y, C) is required.

From a result of Chen and Lo (1997) for the Kaplan-Meier estimate G_n we deduce that if **(RA1)** – **(RA4)** hold, the rate of convergence of the mean squared transformation errors is given by $n^{-\gamma}$ with some $\gamma \in (0,1)$ if

$$-\int_0^{\tau_F} F(t)^{\frac{-\gamma}{2-\gamma}} dG(t) < \infty. \tag{1}$$

Condition (1) basically demands that near the upper endpoint $\tau_F := \sup\{t \in \mathbb{R} : F(t) > 0\}$ of the distribution of Y, there are enough censored observations such that G_n is a stable estimate of G.

For arbitrary $p \in \mathbb{N}$ with $2p > d$, we show in Theorem 4.2 that for suitably chosen parameters, the stochastic rate of convergence of the MSSE of m is given by

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}$$

for every distribution of (X, Y, C) which satisfies **(RA1)** – **(RA4)**, $m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$, and (1) with $\gamma = \frac{2p}{2p+d}$. Observe that in Theorem 4.2 no assumption on the underlying distribution besides **(RA1)** is required. Especially, we do not demand that X has a density with respect to the Lebesgue-Borel measure. Moreover, one can conclude that the rate of convergence in Theorem 4.2 is optimal up to the logarithmic factor.

In analogy to Theorem 4.2, we deduce in Theorem 4.4 for arbitrary $p \in \mathbb{N}$ with $2p > d$ and for arbitrary, but fixed $\tau \in \mathbb{R}$, that the suitably chosen MSSE of $F(\tau|\cdot)$ derives the nearly optimal nonparametric rate

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**, $F(\tau\,|\,\cdot\,) \in W_p([0,1]^d)$ with $0 < J_p^2(F(\tau\,|\,\cdot\,)) < \infty$, and (1) with $\gamma = \frac{2p}{2p+d}$.

In order to guarantee a fast rate of convergence for the MSSE of σ^2 (which depends on the MSSE of m as mentioned above), we impose a smoothness condition on m as well as on σ^2. In Theorem 4.3, it is shown for arbitrary $p_1 \in \mathbb{N}$ and $p_2 \in \mathbb{N}$ with $2p_1 > d$ and

$2p_2 > d$, that the rate of stochastic convergence of the suitably defined MSSE of σ^2 is given by

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_{min}}{2p_{min}+d}} \tag{2}$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**, $m \in W_{p_1}([0,1]^d)$ with $0 < J^2_{p_1}(m) < \infty$, $\sigma^2 \in W_{p_2}([0,1]^d)$ with $0 < J^2_{p_2}(\sigma^2) < \infty$, and (1) with $\gamma = \frac{2p_{max}}{2p_{max}+d}$. Here, p_{min} and p_{max} are defined by $p_{min} := \min\{p_1, p_2\}$ and $p_{max} := \max\{p_1, p_2\}$.

In case that $p_2 \leq p_1$, the rate in (2) is optimal up to the logarithmic factor. On the other hand, if $p_2 > p_1$ in Theorem 4.3, then m is "rougher" than σ^2. In this case, (2) is dominated by the rate of the MSSE of the regression function and may be far from the optimal rate $n^{-\frac{2p_2}{2p_2+d}}$ (cf. Cai, Levine, and Wang (2009) for a similar finding in the case of uncensored data).

Note that in Theorems 4.2 – 4.4, it is assumed that the parameters of our MSSE have been suitably chosen. However, these choices depend on the unknown smoothness of m, σ^2, and $F(\,\tau|\,\cdot\,)$, respectively. Hence, the MSSE in Theorems 4.2 – 4.4 which achieve the nearly optimal rates of convergence are not calculable in a statistical application. Therefore, we further modify our MSSE in order to derive estimates which choose the parameters in a totally data-dependent way and are able to adapt automatically to the unknown smoothness of m, σ^2, and $F(\tau|\,\cdot\,)$, respectively.

The splitting of the sample technique is a widely used and simple method, which allows to achieve these aims. For each combination of parameters in a suitably defined set, MSSE of m, σ^2, and $F(\,\tau|\,\cdot\,)$ are computed on the basis of one part of the censored data, the so-called *learning data*. Then, for each of the three functions, that estimate out of all calculated MSSE is chosen which performs best on the remaining part of the data, the so-called *testing data*. In Theorems 5.2 – 5.4, it is shown that the estimates of m, σ^2, and $F(\,\tau|\,\cdot\,)$, which are defined via the splitting of sample technique, achieve the same rate of convergence as the corresponding estimates in Theorem 4.2 – 4.4 without further assumptions on the distribution of (X, Y, C). Since the parameters of these MSSE are chosen in a totally data-dependent way, one can conclude that the former estimates adapt automatically to the unknown smoothness of m, σ^2, and $F(\tau|\,\cdot\,)$.

Analogous results to Theorem 4.2 and Theorem 5.2 for least squares estimates of the

regression function were verified in Máthé (2006), yet using somewhat stronger conditions on the distribution of (X, Y, C). To be more precise, besides **(RA1)** – **(RA4)**, Máthé (2006) applied the regularity assumption

$$\limsup_{t \uparrow \tau_F} \frac{G(t) - G(\tau_F)}{F(t)} < \infty \tag{3}$$

instead of the weaker constraint (1) and supposed that m is (p, B)-smooth ($p = k + \psi$ with $k \in \mathbb{N}_0$ and $0 < \psi \leq 1$, $B \in (0, \infty)$), while we require that $m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$ ($p \in \mathbb{N}$ with $2p > d$).

To the best knowledge of the author of the present work, Theorem 4.3, Theorem 4.4, Theorem 5.3, and Theorem 5.4 are the first results to be published in the context of nonparametric regression with random design, which examine the rate of convergence of estimates of σ^2 and $F(\tau|\cdot)$ in the presence of censored data without imposing regularity assumptions on the distribution of X.

The results for the MSSE of m, σ^2, and $F(\tau|\cdot)$ discussed above are valid for n tending to infinity. For small to moderate sample sizes, the performance of the estimation procedures is finally analyzed by applying them to real and simulated right censored data. For these data sets, it is demonstrated that the estimates show good approximation characteristics which are comparable to other estimates for randomly right censored data discussed in literature. Furthermore, different modifications of the MSSE, which aim at the improvement of the performance of the estimates for limited sample sizes or allow for practical considerations in statistical applications, are introduced and investigated.

In a nutshell, among other things, the results stated in this thesis verify that in the presence of right censored data, suitably defined nonparametric regression estimates of m, σ^2, and $F(\tau|\cdot)$ achieve the optimal rate of convergence up to a logarithmic factor. Compared to related results to be found literature, weaker assumptions on the distribution of (X, Y, C) are required in order to prove the presented theorems. In particular, they are valid without any condition on the distribution of X (beside that X is bounded).

Chapter 1

Introduction

Chapter 1 presents the basic mathematical concepts of this thesis. A general introduction to the area of survival analysis is given in Section 1.1, while Section 1.2 then focuses on the randomly right censored data model. The latter one also contains results to be found in literature, which are useful in the analysis of the asymptotic properties of our estimates. Section 1.3 describes the problem and the basic concepts of nonparametric regression. A simple extension of these concepts in order to estimate a generalized regression function is outlined in Section 1.4. In Section 1.5, published results of regression analysis with randomly right censored data are discussed. Section 1.6 specifies the basic regularity assumptions on the underlying distribution of the data which are required in this thesis.

1.1 Survival analysis

Survival analysis or more general event history analysis is a branch of statistics that investigates (functions of) the elapsed time since a certain event under study occurred and/or until it occurs. This area is also known as reliability theory in engineering and as duration modeling in economics. The term survival analysis refers to its application in biology or medicine. Examples for the event under study are:

- Death of biological organisms
- Failure of mechanical systems
- Occurrence of a certain disease or its relapse in humans

- First use of tobacco by smokers
- Re-entry into employment of non-workers

The presence of censored data is characteristic of survival analysis. These arise whenever the time of interest cannot be fully observed for all subjects or objects under study, i.e., for some we may have complete observations, but for the others only partial information may be available. Censoring can occur due to several facts, e.g., termination of a study, drop outs or loss to follow-up.

Throughout this thesis, the times which are caused by the censoring mechanism are denoted as *censoring times* and the times of interest as *lifetimes* (although they may not be lifetimes but, e.g., other time-to-events, as can be seen in the examples above). Furthermore, the observed censoring times and lifetimes will be termed as *censored* and *uncensored observations*, respectively.

In literature, the type of the censoring mechanism is classified according to two categories: the way the censoring times are operating on the lifetimes and when subjects or objects are entering or leaving the study. In the first category, one generally distinguishes between:

- **Right censoring:** For each subject or object under study, the observed time is the minimum of its lifetime and its censoring time. Typically, the aim of the study is to analyze functions of the time until a certain event occurs (e.g., death of a patient). For all subjects or objects for which the event of interest does not occur before leaving the study, only a censored observation is reported.
- **Left censoring:** For each subject or object under study, the observed time is the maximum of its lifetime and its censoring time. Here, one is usually interested in the time since an event of interest occurred or when it did happen in the past (e.g., age of first use of tobacco by smokers).
- **Interval censoring:** This kind of censoring occurs when the times of interest cannot be measured exactly, but they are only known to lie in a certain time interval (e.g., if the state of health of patients is only observed during physical examinations, and

so the occurrence of a disease or a relapse is only known to have happened between two inspections).

Moreover, in literature many other different censoring types of the first category are described, such as **double censoring** (i.e., when the time of interest can be both, left and right censored), **mid censoring** (i.e., when the lifetimes are available at two extremes, but some of the observations in the middle are censored), or combinations and extensions of these types.

Each of the types of the first category may occur in combination with one of the following types of the second category:

- **Time censoring (also known as Type I censoring):** The censoring times are fixed and equal for all subjects or objects under study. This type of censoring occurs, e.g., when all subjects enter the study at time point $t_1 = 0$ and their observation times are stopped at the same fixed censoring time $t_2 > 0$. In this case, the number of censored and uncensored observations is random.

- **Failure censoring (also known as Type II censoring):** The study is designed in such a way that the times of interest are observed for a predetermined percentage p of the subjects or objects under study, while leaving $1-p$ percent of the observations censored. I.e., one fixes the number of censored and uncensored observations. This type of censoring occurs, e.g., when the study starts for all subjects at the same time point $t_1 = 0$ and terminates after having observed the event of interest for a predetermined percentage p of the subjects. Here, the censoring time is random.

- **Random censoring:** Neither the time points when the subjects or objects enter the study nor when they leave the study are predetermined, only the observation period for the whole study may be fixed. In this case, the number of censored and uncensored observation is random and the censoring times are random variables, which are usually assumed to be independent and identically distributed.

This thesis deals with *randomly right censored data*, i.e., with data which are subject to both, random censoring and right censoring. In this model, at least two different modes of dependency of the lifetimes and the censoring times are analyzed in literature. For each

subject or object under study, the first one requires that the lifetime and the censoring time are independent, while the second one expects conditional independence of lifetime and censoring time given the vector of covariates. Note that neither of these both modes implies the other (vide Dippon (2011) for an example where the second one holds, but not the first).

The next section explains the first scenario in more detail, including results from literature which are relevant for the formulation and the proofs of our results. In order to analyze asymptotic properties of our estimates, we will assume that for each subject or object under study, the random vector containing the covariates and the lifetime is independent of the censoring time. Obviously, this scenario is weakened by the first (and the second) mode mentioned above. Therefore, all results presented in the next section hold under our assumption.

1.2 Analysis of randomly right censored data

Let $Y, Y_1, \ldots, Y_n$ and $C, C_1, \ldots, C_n$ be independent sequences of i.i.d. nonnegative random variables. In our model, $Y_1, \ldots, Y_n$ represent the lifetimes of n subjects or objects under study, while $C_1, \ldots, C_n$ denote the censoring times. Due to the right censoring mechanism, one only observes the sequence $(Z, \delta), (Z_1, \delta_1), \ldots, (Z_n, \delta_n)$, where

$$Z := \min\{Y, C\}, Z_i := \min\{Y_i, C_i\} \quad (i = 1, \ldots, n) \tag{1.1}$$

and

$$\delta = I_{[Y<C]}, \delta_i = I_{[Y_i<C_i]} \quad (i = 1, \ldots, n). \tag{1.2}$$

I.e., Z_i is the observed time of subject or object $i \in \{1, \ldots, n\}$ under study, while δ_i stores the information whether an observation is censored or uncensored. In case of an uncensored observation, we have $Z_i = Y_i$ and $\delta_i = 1$, for a censored observation $Z_i = C_i$ and $\delta_i = 0$. Note that $Z, Z_1, \ldots, Z_n$ and $\delta, \delta_1, \ldots, \delta_n$ are both sequences of nonnegative i.i.d. random variables.

Denote the unknown survival function of the lifetimes, the censoring times, and the observed times as $F(t) := \mathbf{P}[Y > t]$, $G(t) := \mathbf{P}[C > t]$, and $K(t) := \mathbf{P}[Z > t]$ $(t \in \mathbb{R})$,

respectively. Moreover, define

$$\tau_F := \sup\{t \in \mathbb{R} : F(t) > 0\},$$
$$\tau_G := \sup\{t \in \mathbb{R} : G(t) > 0\},$$

and

$$\tau_K := \sup\{t \in \mathbb{R} : K(t) > 0\}.$$

Due to the assumed independence of Y and C, we can conclude that $K(t) = F(t) \cdot G(t)$ $(t \in \mathbb{R})$ which, in turn, implies $\tau_K = \min\{\tau_F, \tau_G\}$.

An important task in survival analysis is the estimation of F and G. In the case of uncensored data, a nonparametric estimate of F is given by the empirical survival function

$$\hat{F}_n(t) := \hat{F}(t, \{Y_1, \ldots, Y_n\}) := \frac{1}{n} \sum_{i=1}^{n} I_{[Y_i > t]}.$$

Due to the Glivenko-Cantelli theorem, $\hat{F}_n$ is a strongly uniform consistent estimate of F, i.e.,

$$\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \to 0 \quad \text{a.s.} \quad (n \to \infty).$$

But for censored data, $\hat{F}_n(t)$ is in general not calculable, since the number of lifetimes larger than t is not known exactly for all $t \in [0, \tau_F]$. A nonparametric estimate of F in this case is given by the well known Kaplan-Meier product-limit estimate (vide, e.g., Kaplan and Meier (1958))

$$F_n(t) := F_n(t, \{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}) := \prod_{\substack{i=1,\ldots,n: \\ Z_{(i)} \le t}} \left[\frac{n-i}{n-i+1}\right]^{\delta_{(i)}} \quad (t \in \mathbb{R}), \tag{1.3}$$

where we set $0^0 := 1$. In (1.3), $(Z_{(i)}, \delta_{(i)})$ $(i = 1, \ldots, n)$ denote the observed pairs (Z_i, δ_i), arranged in such a way that

$$Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)}.$$

In the case of ties, this ordering is defined such that censored observations $(\delta_i = 0)$ occur before uncensored observations $(\delta_i = 1)$, i.e.,

$$\text{if} \quad Z_{(i)} = Z_{(j)}, \delta_{(i)} = 0, \delta_{(j)} = 1 \quad \Rightarrow \quad i < j. \tag{1.4}$$

Note that we have $0 \leq Z_{(1)} \leq Z_{(n)} \leq \tau_K$ and $Z_{(n)} \to \tau_K$ a.s. $(n \to \infty)$ (cf. Peterson (1977)).

Since F is arbitrary, some of the Y_i $(i = 1, \ldots, n)$ may be identical. In this case the ordering of $Z_1, \ldots, Z_n$ into $Z_{(1)}, \ldots, Z_{(n)}$ is not unique. However, one can see from (1.3) that F_n is unique. Furthermore, the definition of the Kaplan-Meier estimate implies that F_n is a monotonically decreasing step function with jumps solely at the uncensored data points. It equals $\hat{F}_n$ in the case that no censored observations occur, i.e., if $Z_i = Y_i$ and $\delta_i = 1$ for all $i = 1, \ldots, n$.

Similar to (1.3), the Kaplan-Meier estimate of G based on the set $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$ is defined by

$$G_n(t) := G_n(t, \{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}) := \prod_{\substack{i=1,\ldots,n: \\ Z_{(i)} \leq t}} \left[\frac{n-i}{n-i+1}\right]^{1-\delta_{(i)}} \quad (t \in \mathbb{R}). \tag{1.5}$$

G_n is a monotonically decreasing step function with jumps solely at the censored data points. Obviously, (1.3) merges into (1.5) by replacing $\delta_i = I_{[Y_i < C_i]}$ with $1 - \delta_i = I_{[C_i \leq Y_i]}$ for all $i = 1, \ldots, n$. This together with a redefinition of Y as censoring time and C as lifetime shows that (1.3) and (1.5) equal the definitions of the Kaplan-Meier estimates of F and G in literature.

It is not evident why F_n and G_n should be consistent estimates. A first interpretation in this respect was the proof of the maximum likelihood property of F_n and G_n (Kaplan and Meier (1958)). Peterson (1977) showed that if F and G have no jumps in common, the Kaplan-Meier estimates are strongly consistent for all $t < \tau_K$, i.e.,

$$F_n(t) \to F(t) \quad \text{a.s.} \quad (n \to \infty)$$

and

$$G_n(t) \to G(t) \quad \text{a.s.} \quad (n \to \infty).$$

Since the definitions of the estimates considered in this thesis are based on the Kaplan-Meier estimate of G (and not of F, vide Chapter 2), the following results are solely formulated for G_n.

Under additional assumptions to the one of Peterson, Stute and Wang (1993) were able to verify the analogon to the Glivenko-Cantelli theorem for G on $[0, \tau_K]$ in presence of randomly right censored data.

Theorem 1.1. *(Stute and Wang (1993)) Assume that F and G do not have jumps in common. Then*

$$\sup_{0\le t\le \tau_K} |G_n(t) - G(t)| \to 0 \quad (n \to \infty) \quad a.s. \tag{1.6}$$

if and only if either $\mathbf{P}\{C = \tau_K\} = 0$ *or* $\mathbf{P}\{C = \tau_K\} > 0$ *but* $\mathbf{P}\{Y \ge \tau_K\} > 0$.

For the proof of Theorem 1.1, see Stute and Wang (1993), Corollary 1.3.

Theorem 1.1 implies that (1.6) holds if G is continuous on $\mathbb{R}$. This result will be used in order to show strong consistency of suitably defined regression estimates (vide Chapter 3).

In Chapter 4, we apply the following result of Chen and Lo (1997) to prove that our estimates nearly achieve the optimal rate of convergence. Theorem 1.2 below states a sufficient and necessary condition for the rate of strong uniform convergence of $G_n - G$ on $[0, \tau_K]$.

Theorem 1.2. *(Chen and Lo (1997)) Assume G is continuous and $G(\tau_K) > 0$. Let $\gamma \in (0,1)$. Then*

$$n^{\frac{\gamma}{2}} \sup_{0\le t\le \tau_K} |G_n(t) - G(t)| \to 0 \quad (n \to \infty) \quad a.s. \tag{1.7}$$

if and only if

$$-\int_0^{\tau_K} F(t)^{\frac{-\gamma}{2-\gamma}} dG(t) < \infty. \tag{1.8}$$

For the proof of Theorem 1.2, see Chen and Lo (1997), Theorem 2.1.

The second remark after Theorem 2.1 in Chen and Lo (1997) suggests that Theorem 1.2 also holds for discontinuous G (cf. Remark 4.3). Note that the assumption $G(\tau_K) > 0$ in Theorem 1.2 together with the independence of Y and C implies $\tau_K = \tau_F < \infty$.

Since F and G are unknown, τ_F, τ_G, and τ_K are in general unknown, too. Therefore, in order to apply (1.7) in the analysis of the rate of convergence of our estimates, we assume that $Y \in [0, L]$ a.s. for some known upper bound $L \in [\tau_K, \infty)$.

In Theorem 1.2, (1.8) represents a condition on the underlying distribution of (Y, C) near its endpoint τ_K, where γ reflects the heaviness of the censoring near τ_K. Setting $\gamma = 0$, the left hand side of (1.8) equals $G(0) - G(\tau_K) = 1 - G(\tau_K) < 1$. In this case, (1.7) would state (if $\gamma = 0$ was allowed in Theorem 1.2) the strong uniform consistency of G_n for continuous G (cf. Theorem 1.1). The larger $\gamma \in (0,1)$ can be chosen such that (1.8) still

holds, the more censored observations are likely to be near the endpoint. As mentioned by Chen and Lo (1997), the Kaplan-Meier estimates (1.3) and (1.5) are rather unstable near τ_K. Since we seek to estimate G, the survival function of the censoring times, more censored observations near the endpoint therefore means that G_n becomes a more reliable estimate of G. Consequently, this is reflected in the rate of convergence in 1.7: The larger γ, the faster is the rate of convergence, $n^{-\frac{\gamma}{2}}$.

The following lemma gives a sufficient assumption under which condition (1.8) is fulfilled:

Lemma 1.1. *(Chen and Lo (1997)) Assume G is continuous, $G(\tau_K) > 0$, and*

$$0 < \liminf_{t \uparrow \tau_K} \frac{\mathbf{P}[t < C \leq \tau_K]^{\tilde{\gamma}}}{\mathbf{P}[t < Y \leq \tau_K]} \leq \limsup_{t \uparrow \tau_K} \frac{\mathbf{P}[t < C \leq \tau_K]^{\tilde{\gamma}}}{\mathbf{P}[t < Y \leq \tau_K]} < \infty \tag{1.9}$$

for some $\tilde{\gamma} > 0$. Then (1.8) holds for $\gamma \in (0,1)$ if and only if $\gamma < \frac{2}{1+\tilde{\gamma}}$. In particular, if $\tilde{\gamma} \leq 1$, (1.8) holds for all $\gamma < 1$.

For the proof of Lemma 1.1, vide Theorem 2.1 and Corollary 2.2 in Chen and Lo (1997).

There are basically two methods to determine the functional interrelationship between covariates and censored response: regression based approaches and hazard risk approaches, which include classical Cox regression as well as extensions to nonparametric models.

In the second approach, one estimates the conditional hazard rate function, which can be used to reconstruct the conditional survival function under suitable assumptions on its structure and under regularity conditions on the underlying distribution. Details can, e.g., be found in the books of Andersen, Borgan, Gill, and Keiding (1993), Fleming and Harrington (1991) or Cox and Oakes (1984), and in the works of Dippon (2011) or Huang and Stone (1998) as well as in the literature cited therein. Concerning estimates which are not based on structural assumptions on the hazard rate see, e.g., Kooperberg, Stone, and Troung (1995a, 1995b) or Döhler and Rüschendorf (2002).

This thesis presents an approach for the analysis of censored data which is based on nonparametric regression estimation. Here, the goal is to estimate the conditional expectation of the lifetime Y given the vector X of covariates, the so-called *regression function* (we extend this approach in Section 1.4 in order to estimate the conditional

expectation of a function of Y and X). In general, this is an easier task than estimating the conditional hazard rate function. Therefore, hazard risk approaches require stronger regularity assumptions on the underlying distribution than regression based approaches.

The next section gives an introduction to nonparametric regression in the presence of uncensored data. Results of regression analysis with censored data are presented in Section 1.5.

1.3 Nonparametric regression

Let (X, Y) be a $\mathbb{R}^d \times \mathbb{R}$–valued random vector with $\mathbf{E}Y^2 < \infty$. No assumptions are made on the distribution functions of the coordinates of X: Some of them may be (absolutely) continuous, others may be step functions or a mixture of these types.

In *regression analysis* one wants to estimate Y (e.g., the lifetime of a patient) after having observed X (e.g., the medical file of the patient), i.e., one wishes to determine a function f such that $f(X)$ is a "good" approximation of Y. Here, we measure the "distance" between $f(X)$ and Y by the $\mathcal{L}_2$ *risk* of f,

$$\mathbf{E}\left[\,|f(X) - Y|^2\,\right], \tag{1.10}$$

which we now want to minimize.

Let μ denote the distribution of X. It is well known that the $\mathcal{L}_2$ risk of every measurable function f is the sum of the $\mathcal{L}_2$ risk of the *regression function*

$$m : \mathbb{R}^d \to \mathbb{R} : x \mapsto \mathbf{E}\left[\,Y \,|\, X = x\,\right]$$

and the $\mathcal{L}_2$ *error* :

$$\mathbf{E}\left[\,|f(X) - Y|^2\,\right] = \mathbf{E}\left[\,|m(X) - Y|^2\,\right] + \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx). \tag{1.11}$$

Indeed, (1.11) follows from

$$\begin{aligned}
\mathbf{E}\left[\,|f(X) - Y|^2\,\right] &= \mathbf{E}\left[\,|f(X) - m(X) + m(X) - Y|^2\,\right] \\
&= \mathbf{E}\left[\,|f(X) - m(X)|^2\,\right] + \mathbf{E}\left[\,|m(X) - Y|^2\,\right] \\
&\quad +2 \cdot \mathbf{E}\left[\,(f(X) - m(X)) \cdot (m(X) - Y)\,\right],
\end{aligned} \tag{1.12}$$

taking into account that the last term on the right hand side (1.12) equals zero:

$$\begin{aligned}
\mathbf{E}\,[\,(f(X)-m(X))\cdot(m(X)-Y)\,] &= \mathbf{E}\,[\,\mathbf{E}\,[\,(f(X)-m(X))\cdot(m(X)-Y)\,|\,X\,]\,] \\
&= \mathbf{E}\,[\,(f(X)-m(X))\cdot\mathbf{E}\,[\,(m(X)-Y)\,|\,X\,]\,] \\
&= \mathbf{E}\,[\,(f(X)-m(X))\cdot(m(X)-\mathbf{E}\,[\,Y\,|\,X\,]\,)\,] \\
&= \mathbf{E}\,[\,(f(X)-m(X))\cdot(m(X)-m(X))\,] \\
&= \mathbf{E}\,[\,(f(X)-m(X))\cdot 0\,] \\
&= 0. \qquad (1.13)
\end{aligned}$$

Since the $\mathcal{L}_2$ error is always non-negative, (1.11) implies that the regression function m is the optimal predictor of Y in view of the minimization of the $\mathcal{L}_2$ risk:

$$\mathbf{E}\left[\,|m(X)-Y|^2\,\right] = \min_{\substack{f:\mathbb{R}^d\to\mathbb{R},\\ f\,\text{measurable}}} \mathbf{E}\left[\,|f(X)-Y|^2\,\right]. \qquad (1.14)$$

In practical applications, the distribution of (X,Y) and hence also m are usually unknown. But it is often possible to observe a sample $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ of this distribution, and one can construct estimates

$$m_n(\cdot) := m_n(\cdot,(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)) : \mathbb{R}^d \to \mathbb{R}$$

of the regression function. It follows from (1.11) that such an estimate m_n is a good approximation of Y in the sense that the $\mathcal{L}_2$ risk of m_n is close to the optimal value $\mathbf{E}\left[\,|m(X)-Y|^2\,\right]$ if and only if the $\mathcal{L}_2$ error $\int_{\mathbb{R}^d}|m_n(x)-m(x)|^2\mu(dx)$ is small. Consequently, the error caused by using an estimate m_n instead of m will be measured by the $\mathcal{L}_2$ error, whose convergence to zero with n tending to infinity is required.

Definition 1.1. (Consistency) *A sequence of measurable regression estimates $(m_n)_{n\in\mathbb{N}}$ is called* **weakly consistent** *for a certain distribution of (X,Y) if*

$$\mathbf{E}\left[\int_{\mathbb{R}^d}|m_n(x)-m(x)|^2\mu(dx)\right]\to 0 \quad (n\to\infty)$$

and it is called **strongly consistent** *for a certain distribution of (X,Y) if*

$$\int_{\mathbb{R}^d}|m_n(x)-m(x)|^2\mu(dx)\to 0 \quad (n\to\infty) \quad \textit{with probability one.}$$

Moreover, the sequence $(m_n)_{n\in\mathbb{N}}$ is called **weakly (strongly) universally consistent** *if it is* **weakly (strongly) consistent** *for all distributions of (X,Y) with $\mathbf{E}Y^2<\infty$.*

At first, Stone (1977) raised the question whether a sequence of regression estimates is universally consistent and answered it positively (i.e., proved weak universal consistency) for a class of local averaging estimates. For multivariate smoothing spline estimates, strong universal consistency was verified by Kohler and Krzyżak (2001). A survey of various weakly and strongly consistent regression estimates can be found in Györfi, Kohler, Krzyżak, and Walk (2002).

By demonstrating convergence of an estimate m_n, one has not drawn a conclusion about how fast the $\mathcal{L}_2$ error of m_n converges to zero when n increases. But this might be important if one thinks of a practical statistical application with a limited sample size. Disappointingly, due to a result of Devroye (1982), without imposing strong restrictions on the distribution of (X, Y), the rate of convergence for the $\mathcal{L}_2$ error may be arbitrarily slow. The question of a nontrivial rate of convergence can therefore solely be answered positively for certain classes of distributions (cf. Devroye and Wagner (1980) and Devroye, Györfi, and Lugosi (1996)).

Stone (1982) proved that for (p, B)-smooth regression functions and d-dimensional covariates, no nonparametric regression estimate converges in suitably defined minimax sense in $\mathcal{L}_2$ faster than $n^{-\frac{2p}{2p+d}}$ (vide Remark 4.1). He showed that the $\mathcal{L}_2$ error of a local polynomial kernel estimate m_n converges to zero in probability with this rate, i.e., that there exists a constant $b > 0$ such that

$$\lim_{n\to\infty} \mathbf{P}\left\{\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) > b \cdot n^{-\frac{2p}{2p+d}}\right\} = 0 \tag{1.15}$$

for all distributions of (X, Y) with $X \in [0, 1]^d$ a.s., X having a density with respect to the Lebesgue-Borel measure which is bounded away from zero and infinity, $\mathbf{E}Y^2 < \infty$, and for (p, B)-smooth m.

A big drawback of this result in many statistical applications is that the assumption of X having a density with respect to the Lebesgue-Borel measure cannot be verified reliably or is even known to be inappropriate. Kohler (2000) demonstrated that suitably defined least squares spline estimates achieve the rate $n^{-\frac{2p}{2p+d}}$ in $\mathcal{L}_2$ for (p, B)-smooth regression functions without any assumptions on the distribution of X (beside X is bounded a.s.). In addition, this result also holds for estimates which do not depend on the smoothness of the regression function m. For multivariate smoothing spline estimates (MSSE), which

are subject of this thesis, similar results were proven by Kohler, Krzyżak, and Schäfer (2002). To be more precise, they showed that if m is a function in the Sobolev space $W_p([0,1]^d)$, totally data-driven MSSE achieve in $\mathcal{L}_2$ the optimal rate of convergence up to some logarithmic factor. Note that this conclusion is valid without any regularity conditions on the distribution of the design.

In Theorem 4.2 and 5.2, we extend the results of Kohler, Krzyżak, and Schäfer (2002) to censored data and show that under additional assumptions on the distribution of the censoring times, the $\mathcal{L}_2$ error of our estimates converges to zero in probability with a rate which is optimal up to some logarithmic factor. There, the following notation is used. Let V_n be a nonnegative random variable and b a positive constant. If the $\mathcal{L}_2$ error of an estimate m_n converges to zero in probability with rate V_n, i.e.,

$$\lim_{n\to\infty} \mathbf{P}\left[\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) > b \cdot V_n\right] = 0,$$

then we write

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(V_n\right).$$

1.4 Regression estimates for transformed data

Denote by (X,Y), (X_1,Y_1), (X_2,Y_2), $\ldots$, $(X_n,Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. random vectors with $\mathbf{E}Y^2 < \infty$. (1.14) shows that the best approximation of Y based on the observation of X in terms of the $\mathcal{L}_2$ risk is given by the regression function $m(x) = \mathbf{E}\left[\, Y \mid X = x \,\right]$ $(x \in \mathbb{R}^d)$.

In the following, we discuss how this approach can be extended to estimate $h(X,Y)$ after having observed X, where $h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is some measurable function with $\mathbf{E}\left[\, h(X,Y)^2 \,\right] < \infty$. This will enable us to construct estimates of, e.g., the higher order conditional moments, the conditional variance, and the conditional survival function of Y given $X = x$ (see below). In analogy to Section 1.3, the $\mathcal{L}_2$ risk is used to measure the quality of the approximation to $h(X,Y)$. Obviously, by replacing Y with $h(X,Y)$ in (1.11), (1.12), and (1.13), one can conclude for the regression function of $(X, h(X,Y))$,

$$m^{[h]}(x) := \mathbf{E}\left[\, h(X,Y) \mid X = x \,\right] \quad (x \in \mathbb{R}^d), \tag{1.16}$$

and for any measurable function $f : \mathbb{R}^d \to \mathbb{R}$ that

$$\mathbf{E}\left[\, |f(X) - h(X,Y)|^2 \,\right] = \mathbf{E}\left[\, |m^{[h]}(X) - h(X,Y)|^2 \,\right] + \int_{\mathbb{R}^d} |f(x) - m^{[h]}(x)|^2 \mu(dx).$$

Similar to (1.14), this implies that $m^{[h]}$ minimizes the $\mathcal{L}_2$ risk of f and h, i.e.,

$$\mathbf{E}\left[\,|m^{[h]}(X)-h(X,Y)|^2\,\right]=\min_{\substack{f:\mathbb{R}^d\to\mathbb{R},\\ f\text{measurable}}}\mathbf{E}\left[\,|f(X)-h(X,Y)|^2\,\right].$$

Therefore $m^{[h]}$ is the best approximation of $h(X,Y)$ in terms of the $\mathcal{L}_2$ risk.

Beside trivial choices of h (e.g., $h\equiv 0$), $m^{[h]}$ will be unknown, and the task is to construct estimates $m_n^{[h]}$ of $m^{[h]}$, which depend on h. If $h:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$ is a known function, then an estimate $m_n^{[h]}$ may simply be derived by first transforming the data (X_1,Y_1), (X_2,Y_2), $\dots$, (X_n,Y_n) to $(X_1,h(X_1,Y_1))$, $(X_2,h(X_2,Y_2))$, $\dots$, $(X_n,h(X_n,Y_n))$ and then applying a regression estimate of m to the new data. In general, under the assumption that $\mathbf{E}\left[\,h(X,Y)^2\,\right]<\infty$, all attributes of the estimate m_n transfer directly to the estimate $m_n^{[h]}$.

Examples for such functions are $h:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}:(x,y)\mapsto y^q$ ($q\in\mathbb{N}$), which results in the problem of estimating the qth conditional moment of Y,

$$m^{(q)}(x):=\mathbf{E}\left[\,Y^q\,|\,X=x\,\right]\quad(x\in\mathbb{R}^d)$$

(where we assume that $\mathbf{E}\left[\,Y^{2q}\,\right]<\infty$), or $h:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}:(x,y)\mapsto I_{[Y>\tau]}$ ($\tau\in\mathbb{R}$ fixed), aiming at the estimation of the conditional survival function of Y at point τ,

$$F(\tau\,|\,x):=\mathbf{P}\left[\,Y>\tau\,|\,X=x\,\right]=\mathbf{E}\left[\,I_{[Y>\tau]}\,\middle|\,X=x\,\right]\quad(x\in\mathbb{R}^d).$$

If, in contrast, h is not known, it may still be estimable from the data (X_1,Y_1),$\dots$, (X_n,Y_n). The probably most important example in this case is the estimation of the conditional variance of Y,

$$\sigma^2(x):=\mathbf{Var}\left[\,Y\,|\,X=x\,\right]=\mathbf{E}\left[\,|Y-m(X)|^2\,\middle|\,X=x\,\right]\quad(x\in\mathbb{R}^d).\tag{1.17}$$

Here, we set $h:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}:(x,y)\mapsto|y-m(x)|^2$ and assume that $\mathbf{E}Y^4<\infty$. Since m is unknown in a statistical application, h is unknown, too. But using the fact that

$$\sigma^2(x)=\mathbf{E}\left[\,Y^2\,\middle|\,X=x\,\right]-m(x)^2=\mathbf{E}\left[\,Y^2-m(X)^2\,\middle|\,X=x\,\right]\quad(x\in\mathbb{R}^d),\tag{1.18}$$

one can obviously derive an estimate of σ^2 by applying a regression estimate to the data

$$(X_1,Y_1^2-m_n(X_1)^2),\dots,(X_n,Y_n^2-m_n(X_n)^2),$$

where m_n is an estimate of m.

The conditional variance is sometimes also termed local variance since it locally measures the prediction quality by the regression function (observe that conditional variance is a local measure of the minimal $\mathcal{L}_2$ risk). In this context, it is widely used to construct (local) confidence intervals of confidence bands for m (for the construction of confidence bands in the presence of censored data, see, e.g., Deheuvels and Mason (2004) and the literature cited therein).

A broad survey of regression estimation on the basis of transformed variables (especially within parametric models), including estimates of the conditional variance, can be found in Carroll and Ruppert (1988). Ferraty, Laksaci, Tadj, and Vieu (2010) analyze the uniform rate of convergence of kernel estimates of the generalized regression function $m^{[h]}$ for known h and covariates in a semi-metric space. Concerning the construction of nonparametric estimates of the conditional variance as well as their applications see, e.g., Müller and Stadtmüller (1987), Carroll and Hall (1989), Neumann (1994), Stadtmüller and Tsybakov (1995), Wang, Brown, Cai, and Levine (2008), and Cai, Levine, and Wang (2009). In particular, Wang, Brown, Cai, and Levine (2008) (for $d = 1$) and Cai, Levine, and Wang (2009) show that for nonparametric estimates of σ^2 in fixed design regression, the optimal rate of convergence (pointwise and in $\mathcal{L}_2$) in case of (p_1, B_1)-smooth m and (p_2, B_2)-smooth σ^2 is given by

$$\max\left\{n^{-\frac{4p_1}{d}}, n^{-\frac{2p_2}{2p_2+d}}\right\}.$$

Notably, they thereby correct results of Carroll and Hall (1989) and Stadtmüller and Tsybakov (1995) incorrectly denoting a slower rate as optimal (under the same assumptions as given above).

In Section 2.2, a similar idea as the one described in this section will be presented which aims at the construction of regression estimates in the presence of censored data. First a special kind of transformation is applied to the censored data, leaving the regression function (or more general $m^{[h]}$) unchanged. Since these transformed data points depend on the unknown survival function G of the censoring times, they have to be estimated using the original data. In this step, the Kaplan-Meier estimate G_n (vide (1.5)) plays a crucial role. Now, an estimate for censored regression can be constructed by applying a usual nonparametric regression estimate to the new, virtually uncensored data. Obviously,

the asymptotic behavior of the first estimate is determined by the behavior of the second one and of the estimates of the transformed data, the latter being in turn ruled by the convergence of the Kaplan-Meier estimate G_n. This is demonstrated in the proofs of our main results in Chapters 3 – 5.

1.5 Results of regression analysis with censored data

Throughout the present work, we will assume that $(X, Y, C), (X_1, Y_1, C_1), \ldots, (X_n, Y_n, C_n)$ are i.i.d. $\mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}_+$–valued random vectors, where C is a right censoring variable operating on Y. Set $Z := \min\{Y, C\}$, $\delta := I_{[Y<C]}$, $Z_i := \min\{Y_i, C_i\}$, and $\delta_i = I_{[Y_i<C_i]}$ $(i = 1, \ldots, n)$. The problem of censored regression is now to estimate the regression function $m(x) := \mathbf{E}[Y \mid X = x]$ $(x \in \mathbb{R}^d)$ or more general $m^{[h]}(x)$ (vide Section 1.4) from the data

$$\mathcal{D}_n := \{(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)\}. \tag{1.19}$$

In 1979, Buckley and James introduced an estimate of a linear regression function, whose consistency was investigated by James and Smith (1984). For a slight modification of this estimate, Ritov (1990) and Lai and Ying (1991) established the asymptotic normality. Other estimates for the linear regression model are due to Miller (1976), Leurgans (1987), and Koul, Sousarla, and Van Ryzin (1981).

The first fully nonparametric estimate of the conditional survival function in the presence of right censored data was established by Beran (1981). Under the assumption that Y and C are conditionally independent given X, Beran (1981) and Dabrowska (1987, 1989) showed that this estimate is weakly and strongly consistent.

Without conditions on the structure of the regression function or regularity assumptions on the distribution of the design, Zheng (1987) showed that suitably defined nearest neighbor estimates of m for censored regression are strongly pointwise consistent. He required that (X, Y) and C are independent. In the same setting, strong consistency of suitably defined partitioning estimates with respect to the $\mathcal{L}_2$ error was proven by Carbonez (1992). A survey of corresponding results for further nonparametric estimates of the regression function is given in Pintér (2001). Beyond, in the more general model according to which Y and C are conditionally independent given X, Pintér (2001) showed that

one can use the nonparametric estimate of the conditional survival function introduced by Beran (1981) to construct suitably defined local averaging estimates of m which are strongly consistent with respect to the $\mathcal{L}_2$ error.

Though, far less is known about the rate of convergence regarding regression based approaches for the analysis of censored data, especially without assuming regularity conditions on the distribution of X. Supposing that X has a density with respect to the Lebesgue-Borel measure, Fan and Gijbels (1994) showed that suitably defined local polynomial estimates of m achieve pointwise the optimal rate of convergence. In the presence of right censoring and possible left truncation, Park (2004) proved that for (p, B)-smooth regression functions, suitably defined weighted least squares estimates reach the optimal rate of convergence if X has a bounded density with respect to the Lebesgue-Borel measure. However, these estimates are not calculable, since they depend on p, which is unknown in a statistical application. Supposing that X has a bounded density, Guessoum and Ould-Saïd (2008, 2010) derived a uniform almost sure rate of convergence for kernel estimates of m and established its optimality for $d = 1$. Using the assumption that X has a strictly positive density, Gneyou (2005), El Ghouch and Van Keilegom (2008), and Maillot and Viallon (2009) investigated the rate of convergence of nonparametric estimates of the generalized regression function, and deduced rates of estimates of m and the conditional survival function.

Without regularity assumptions on the distribution of X, Máthé (2006) showed that suitably defined least squares estimates achieve for (p, B)-smooth regression functions the optimal global rate of convergence up to a logarithmic factor. Furthermore, this result also holds for estimates which adapt automatically to the unknown smoothness p of m. In order to derive this rate of convergence, Máthé (2006) assumed that (1.9) holds with $\tilde{\gamma} = 1$.

In this thesis, it is shown that under the somewhat weaker condition (1.8) and if m is a function in a Sobolev space of degree p suitably defined multivariate smoothing spline estimates (MSSE) also achieve the optimal rate of convergence up to some logarithmic factor, without assuming regularity conditions on the distribution of X. Moreover, this rate even holds if the parameters of these estimates are chosen in a totally data-dependent way, i.e., for MSSE which do not depend on the smoothness of m.

Beyond, these results are extended to estimates of the conditional variance $\sigma^2(X)$ and the conditional survival function $F(\tau|X)$ ($\tau \in \mathbb{R}$ fixed) in the presence of censored data. To be more precise, we show that under appropriate conditions on the smoothness of $m(X)$, $\sigma^2(X)$ and $F(\tau|X)$, suitably defined MSSE of the conditional variance and the conditional survival function also achieve their optimal rate of convergence up to some logarithmic factor. Furthermore, this conclusion is still valid for estimates which adapt automatically to the unknown smoothness of $\sigma^2(X)$ and $F(\tau|X)$, respectively. To the knowledge of the author of this thesis, there are so far no results published, which analyze the rate of convergence for nonparametric estimates of $\sigma^2(X)$ and $F(\tau|X)$ in the presence of censored data without imposing regularity assumptions on the distribution of X.

1.6 Regularity assumptions

Consider the situation of the randomly right censored data model in Section 1.2. In the following, we present the regularity assumptions on the underlying distribution of (X, Y, C) which are required in this thesis in order to generalize known bounds on the $\mathcal{L}_2$ error of our estimates from nonparametric regression with random design to censored regression. These conditions can be stated as follows:

(RA1) $X \in [0,1]^d$ a.s.

(RA2) $\exists\, L \in (0,\infty)$ such that $0 \leq Y \leq L$ a.s. and $\mathbf{P}[C > L] > 0$

(RA3) C and (X, Y) are independent

(RA4) G is continuous.

Regularity assumption **(RA1)** demands that with probability one, X takes only values in some bounded set, which we choose without loss of generality equal to $[0,1]^d$. Boundedness of X is a common assumption in the analysis of the rate of convergence of regression estimates and is not a serious constraint in a statistical application. For instance, in survival analysis, the vector X may store the personal and the medical data of a subject under study, such as age, weight, or the expression levels of several hundred genes measured in

a microarray experiment. Note that if no censoring arises, one can show strong consistency of a modified multivariate smoothing spline estimate without assumption **(RA1)** (cf. Remark 3.2).

Regularity assumption **(RA2)** simplifies the analysis of the rate of convergence. It is satisfied in statistical applications where the data are collected during a study of fixed duration or an upper bound $L > 0$ on the lifetimes is known. Once this bound is determined, one can make a more or less rough guess of τ_F and τ_G, since **(RA2)** implies that $\tau_F \leq L < \tau_G$. Note that in general, τ_F and τ_G are unknown in a statistical application. Therefore, we define our estimates with the more general and known upper bound L.

Furthermore we want to stress that in **(RA2)** $\mathbf{P}[C > L] = 1$ is allowed. Therefore the results presented in this thesis are still valid if censoring does not occur. They can be regarded as generalizations of results for multivariate smoothing splines in usual nonparametric regression with random design (vide Kohler and Krzyżak (2001) and Kohler, Krzyżak, and Schäfer (2002)).

Assumption **(RA3)** also simplifies the mathematical problem in the analysis of the rate of convergence. As mentioned at the end of Section 1.1, there are at least two other modes of dependency of Y and C commonly considered in the analysis of randomly right censored data, which both weaken **(RA3)**. The first one demands that Y and C are independent, while the second one supposes that Y and C are conditionally independent given X. In addition to the first mode, **(RA3)** requires that C and X are independent. This condition is realistic whenever the mechanism of censoring is independent of the covariates under study (e.g., the personal and genetic data of a subject). Of course there exist applications where this is not satisfied, but without assumption **(RA3)** the analysis of the rate of convergence seems to be much more difficult.

Regularity assumption **(RA4)** is used to simplify the presentation of our main results and their proofs. As discussed in Section 1.2, we will estimate the unknown survival function G by the Kaplan-Meier estimate G_n. Assume that F (the survival function of the lifetimes) and G have no jumps in common and that **(RA2)** – **(RA3)** hold. If G is continuous in τ_K, this together with Theorem 1.1 implies that G_n is a uniform consistent estimate for G on $[0, \tau_K]$. If, in contrast, G is discontinuous in τ_K, then the assertion of Theorem 1.1 is fulfilled if and only if Y and C both equal the upper bound of the distribu-

tion of Y with a non-zero probability. I.e., unless in this fairly unrealistic situation, G can only be consistently estimated by G_n on the whole interval $[0, \tau_K]$ if G is continuous in τ_K. Vide Remark 3.1 for further details.

Assumptions **(RA1)** – **(RA4)** are sufficient in order to prove that our estimates are strongly consistent. However, from Theorem 1.2 one can conclude that we need an additional assumption on the distribution of (Y, C) in the analysis of the rate of convergence.

Let $\gamma \in (0, 1)$. Observe that the regularity assumptions **(RA2)** and **(RA3)** yield $G(\tau_K) = G(\tau_F) \geq G(L) > 0$. This together with **(RA4)** and Theorem 1.2 implies that the rate of strong uniform convergence of $G_n - G$ on $[0, \tau_K]$ is given by $n^{-\frac{\gamma}{2}}$ if and only if (1.8) holds, i.e.,

$$-\int_0^{\tau_K} F(t)^{\frac{-\gamma}{2-\gamma}} dG(t) < \infty.$$

Note that the left hand side of the last inequality is always non-negative, since G is monotonically decreasing. As mentioned in Section 1.2, (1.8) reflects the heaviness of the censoring near τ_K.

As discussed in Section 1.3, for (p, B)-smooth regression functions and $d-$dimensional covariates, the optimal rate of convergence of an estimate in usual nonparametric regression is given by $n^{-\frac{2p}{2p+d}}$. In order to show that in the presence of censored data, our estimates achieve this rate up to some logarithmic factor, Theorem 1.2 will be applied (vide Section 4.2). Therefore, in the analysis of the rate of convergence, we will assume that (1.8) holds with γ suitably chosen. As mentioned in Section 1.3, Chen and Lo (1997) argue that Theorem 1.2 also holds if **(RA4)** is violated. In this case, one has to handle the discontinuous points of G carefully (vide Remark 4.3).

Chapter 2

Definition of the estimates

This chapter introduces the estimates for censored regression, which are analyzed in this thesis. The multivariate smoothing spline estimates (MSSE) of the regression function $m(X)$ in the presence of uncensored data are defined in Section 2.1. Section 2.2 presents a valuable class of transformations in order to reduce the censored regression problem to a usual nonparametric regression problem. Based on three different transformations in this class, MSSE of $m(X)$, the conditional variance $\sigma^2(X)$, and the conditional survival function $F(\tau|X)$ ($\tau \in \mathbb{R}$ fixed) are defined in the subsequent sections of this chapter. In Section 2.3, our estimates of $m(X)$ in the presence of censored data are introduced. These MSSE depend on parameters, which control the smoothness of the estimates. An estimate of $m(X)$ with a totally data-driven choice of these parameters is presented in Section 2.4. In Section 2.5, the estimates of Sections 2.3 and 2.4 are used to define MSSE of $\sigma^2(X)$. Finally, Section 2.6 presents our estimates of $F(\tau|X)$ in the presence of censored data.

2.1 Multivariate smoothing spline estimates (MSSE) for uncensored data

Let

$$\mathcal{D}_n := \{(X_1, Y_1), \dots, (X_n, Y_n)\} \tag{2.1}$$

be a i.i.d. sample of the $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$. Since the regression function minimizes the $\mathcal{L}_2$ risk (cf. (1.14)), a natural estimate of $m(X)$ can be

obtained by minimizing an estimate of the $\mathcal{L}_2$ risk, the empirical $\mathcal{L}_2$ risk

$$\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2. \tag{2.2}$$

But if one would minimize (2.2) over all (measurable) functions, this would lead to a function which interpolates the data (at least if the $X_1,\dots,X_n$ are all distinct). There are basically two different strategies to avoid this. For least squares estimates one minimizes the empirical $\mathcal{L}_2$ risk over some suitably chosen class of functions which depends on the sample size n. For penalized least squares estimates or smoothing spline estimates one minimizes the sum of the empirical $\mathcal{L}_2$ risk and a penalty term which penalizes the roughness of a function, over basically all functions:

Definition 2.1. (Multivariate smoothing spline estimates (MSSE)) *Let $d,k\in\mathbb{N}$ with $2k>d$, $X\in[0,1]^d$ a.s., $\mathcal{D}_n$ be given by (2.1), and denote by $W_k([0,1]^d)$ the Sobolev space*

$$\left\{f\in\mathcal{L}_2([0,1]^d):\frac{\partial^\kappa}{\partial x_1^{\kappa_1}\dots\partial x_d^{\kappa_d}}f\in\mathcal{L}_2([0,1]^d)\ \forall\kappa_1,\dots,\kappa_d\in\mathbb{N}_0,\sum_{i=1}^{d}\kappa_i=\kappa\le k\right\}. \tag{2.3}$$

The **multivariate smoothing spline estimate (MSSE)** *$\tilde{m}_{n,(k,\lambda_n)}$ is defined by*

$$\tilde{m}_{n,(k,\lambda_n)}(\cdot,\mathcal{D}_n):=\operatorname*{arg\,min}_{f\in W_k([0,1]^d)}\left(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2+\lambda_n J_k^2(f)\right) \tag{2.4}$$

with smoothing parameter $\lambda_n>0$ and penalty term

$$J_k^2(f):=\sum_{\substack{\kappa_1,\dots,\kappa_d\in\mathbb{N}_0:\\ \kappa_1+\dots+\kappa_d=k}}\frac{k!}{\kappa_1!\cdot\ldots\cdot\kappa_d!}\int_{[0,1]^d}\left|\frac{\partial^k}{\partial x_1^{\kappa_1}\dots\partial x_d^{\kappa_d}}f(x)\right|^2dx. \tag{2.5}$$

The condition $2k>d$ implies that the functions in $W_k([0,1]^d)$ are continuous and hence the evaluation of a function at a point is well defined. Observe that the (partial) derivatives in (2.3) do not need to exist in a classical sense, but rather only weak derivatives are required.

Moreover, note that in (2.4), we do not demand that the minimizer is unique. Duchon (1976) and (under some additional assumptions) Wahba (1990) showed that a function of the form

$$\sum_{i=1}^{n}a_{1,i}R(\|x-X_i\|)+\sum_{j=1}^{M}a_{2,j}\Psi_j(x)\quad\left(x\in\mathbb{R}^d,M=\tbinom{d+k-1}{d}\right)$$

achieves the minimum in (2.4), where the so-called *radial basis functions* R are given by

$$R : \mathbb{R}_+ \to \mathbb{R} : t \mapsto \begin{cases} t^{2k-d} \ln t & \text{if} \quad 2k-d \quad \text{is even} \\ t^{2k-d} & \text{if} \quad 2k-d \quad \text{is odd} \end{cases}$$

and $\Psi_1, \ldots, \Psi_M$ are all monomials $x_1^{\kappa_1} \cdot \ldots \cdot x_d^{\kappa_d}$ of total degree $\sum_{i=1}^d \kappa_i \leq k-1$. Furthermore, Duchon and Wahba demonstrated that the coefficients $a_{1,1}, \ldots, a_{1,n}, a_{2,1}, \ldots, a_{2,M} \in \mathbb{R}$ can be computed by solving a linear system of equations.

2.2 Transformation of the censored data

Let (X, Y, C), (X_1, Y_1, C_1), (X_2, Y_2, C_2), $\ldots$, (X_n, Y_n, C_n) be i.i.d. $\mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}_+$–valued random vectors and assume that the regularity conditions **(RA1)** – **(RA3)** hold. Note that **(RA1)** and **(RA2)** imply that there exists a known constant $L \in [\tau_F, \infty)$ such that with probability one $(X, Y) \in [0,1]^d \times [0, L]$.

In order to define estimates of the regression function $m(X) = \mathbf{E}[\, Y \mid X \,]$ in the presence of censored data, several authors, in particular Buckley and James (1979), Koul, Susarla, and Van Ryzin (1981), Leurgans (1987), Zheng (1987), and Fan and Gijbels (1994, 1996), developed and investigated the approach of the so-called *censoring unbiased transformation*. The idea behind this is to convert the censored data in an appropriate way to virtually uncensored data, leaving m unchanged. Now estimates from usual regression analysis can be applied to this new data.

Let $B \in \mathbb{R}_+$, $h : [0,1]^d \times [0, \tau_F] \to [0, B] : (x, y) \mapsto h(x, y)$ be a measurable function, and $m^{[h]}$ be defined by (1.16), i.e.,

$$m^{[h]}(X) = \mathbf{E}[\, h(X, Y) \mid X \,]. \tag{2.6}$$

In the following, we present a generalization of the censoring unbiased transformation, which accounts for the estimation of $m^{[h]}$ (cf. El Ghouch and Van Keilegom (2008)). Similar to this approach, our aim is to convert the censored data such that the regression function to this new data is identical to $m^{[h]}$. This enables us to define not only estimates of m but also of σ^2 and $F(\tau \mid \cdot)$ ($\tau \in \mathbb{R}$ fixed) in censored regression similar to those in usual nonparametric regression (cf. Sections 1.3, 1.4, and 2.1).

For this purpose, we first note that $m^{[h]}$ is equal to the conditional expectation of

$$\frac{\delta h(X,Z)}{G(Z)} \tag{2.7}$$

given X, where $\delta = I_{[Y<C]}$ and $Z = \min\{Y, C\}$ (vide (1.1) and (1.2)). Indeed, **(RA2)** implies $G(Z) \geq G(L) > 0$. Using this together with **(RA3)** and properties of conditional expectation, one gets

$$\begin{aligned}
\mathbf{E}\left[\frac{\delta h(X,Z)}{G(Z)}\,\middle|\,X\right] &= \mathbf{E}\left[I_{[Y<C]}\,\frac{h(X,Y)}{G(Y)}\,\middle|\,X\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[I_{[Y<C]}\,\middle|\,X,Y\right]\,\frac{h(X,Y)}{G(Y)}\,\middle|\,X\right] \\
&= \mathbf{E}\left[\,h(X,Y)\,\middle|\,X\,\right] \\
&= m^{[h]}(X).
\end{aligned} \tag{2.8}$$

Now assume that for all $(x,y) \in [0,1]^d \times [0,\tau_F]$

$$h(x,0) = 0,\ h(x,y) \text{ is continuously differentiable with respect to } y, \text{ and } \frac{\partial}{\partial y}\,h(x,y) \geq 0. \tag{2.9}$$

Set

$$h'_t(x,t) := \frac{\partial}{\partial t}\,h(x,t) \quad (x \in [0,1]^d, t \in [0,\tau_F]).$$

Since (2.9) implies

$$\int_0^Y h'_t(X,t)\,dt = h(X,Y) \quad \text{a.s.}, \tag{2.10}$$

one can conclude similar to (2.8) from **(RA2)** and **(RA3)**

$$\begin{aligned}
\mathbf{E}\left[\int_0^Z \frac{h'_t(X,t)}{G(t)}\,dt\,\middle|\,X\right] &= \mathbf{E}\left[\int_0^Y \frac{I_{[t<C]}\,h'_t(X,t)}{G(t)}\,dt\,\middle|\,X\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\int_0^Y \frac{I_{[t<C]}\,h'_t(X,t)}{G(t)}\,dt\,\middle|\,X,Y\right]\,\middle|\,X\right] \\
&= \mathbf{E}\left[\int_0^Y \frac{h'_t(X,t)}{G(t)}\,\mathbf{E}\left[\,I_{[t<C]}\,\middle|\,X,Y\,\right]\,dt\,\middle|\,X\right] \\
&= \mathbf{E}\left[\int_0^Y h'_t(X,t)\,dt\,\middle|\,X\right] \\
&= \mathbf{E}\left[\,h(X,Y)\,\middle|\,X\,\right] \\
&= m^{[h]}(X).
\end{aligned} \tag{2.11}$$

Therefore, $m^{[h]}$ is in this case also identical to the conditional expectation of

$$\int_0^Z \frac{h'_t(X,t)}{G(t)}\,dt \tag{2.12}$$

given X. Observe that the random variable (2.12) is well defined since **(RA2)** implies $G(t) \geq G(\tau_F) \geq G(L) > 0$ for all $t \in [0, \tau_F]$.

Let $\alpha_{[h]} \in \mathbb{R}$. The transformation of the censored data is now defined as follows. Replace (X, Z, δ) by $(X, Y^{[h]})$, where

$$Y^{[h]} := \begin{cases} (1+\alpha_{[h]}) \displaystyle\int_0^Z \frac{h'_t(X,t)}{G(t)}\,dt - \alpha_{[h]} \frac{\delta h(X,Z)}{G(Z)} & \text{if (2.9) holds} \\ \\ \displaystyle\frac{\delta h(X,Z)}{G(Z)} & \text{otherwise} \end{cases} \tag{2.13}$$

and for all $i = 1, \ldots, n$, the datum point (X_i, Z_i, δ_i) by $(X_i, Y_i^{[h]})$, with

$$Y_i^{[h]} := \begin{cases} (1+\alpha_{[h]}) \displaystyle\int_0^{Z_i} \frac{h'_t(X_i,t)}{G(t)}\,dt - \alpha_{[h]} \frac{\delta_i h(X_i,Z_i)}{G(Z_i)} & \text{if (2.9) holds} \\ \\ \displaystyle\frac{\delta_i h(X_i,Z_i)}{G(Z_i)} & \text{otherwise.} \end{cases} \tag{2.14}$$

From (2.8), (2.11), and (2.13), one can conclude that

$$\mathbf{E}\left[\, Y^{[h]} \,\middle|\, X \,\right] = m^{[h]}(X) = \mathbf{E}\left[\, h(X,Y) \mid X \,\right]. \tag{2.15}$$

Moreover, observe that with probability one, $Y^{[h]}$ is bounded in absolute value by some constant $L^\star_{[h]} > 0$. Indeed, $0 \leq Y \leq L < \infty$ a.s., $1 \geq G(t) \geq G(\tau_F) \geq G(L) > 0$ ($t \in [0, \tau_F]$), and $h(x,y) \in [0, B]$ ($(x,y) \in [0,1]^d \times [0, \tau_F]$) imply in case that (2.9) holds

$$\begin{aligned} |Y^{[h]}| &\leq \left|(1+\alpha_{[h]}) \int_0^Z \frac{h'_t(X,t)}{G(t)}\,dt\right| + \left|\alpha_{[h]} \frac{\delta h(X,Z)}{G(Z)}\right| \\ &\leq (1+|\alpha_{[h]}|) \left|\int_0^Z \frac{h'_t(X,t)}{G(t)}\,dt\right| + |\alpha_{[h]}| \left|\frac{\delta h(X,Z)}{G(Z)}\right| \\ &\leq (1+|\alpha_{[h]}|) \frac{1}{G(L)} \left|\int_0^Z h'_t(X,t)\,dt\right| + |\alpha_{[h]}| \frac{h(X,Z)}{G(L)} \\ &= (1+2|\alpha_{[h]}|) \frac{h(X,Z)}{G(L)} \\ &\leq (1+2|\alpha_{[h]}|) \frac{B}{G(L)} =: L^\star_{[h]} < \infty \quad \text{a.s.}, \end{aligned} \tag{2.16}$$

(where we used (2.10)), and if (2.9) is violated

$$|Y^{[h]}| \leq \frac{h(X,Z)}{G(Z)} \leq \frac{B}{G(L)} \leq L^{\star}_{[h]} < \infty \quad \text{a.s.}, \tag{2.17}$$

Due to (2.15), the estimation of $m^{[h]}$ can be based on the transformed data

$$(X_1, Y_1^{[h]}), \ldots, (X_n, Y_n^{[h]})$$

If our goal is to estimate the regression function $m(X) = \mathbf{E}[\, Y \mid X \,]$, then (2.6) demands that we choose the function h such that $h : [0,1]^d \times [0, \tau_F] : (x,y) \mapsto y$. In this case, (2.9) obviously holds with $h'_y : [0,1]^d \times [0,\tau_F] : (x,y) \mapsto 1$. This yields that $Y^{[h]}$ then equals

$$(1 + \alpha_{[h]}) \int_0^Z \frac{1}{G(t)}\, dt - \alpha_{[h]} \frac{\delta Z}{G(Z)}. \tag{2.18}$$

Note that (2.18) agrees with the transformed random variable in the "new class" of transformations introduced in Fan and Gijbels (1994, 1996), which is therefore a special case of our approach (yet observe that El Ghouch and Van Keilegom (2008) use a more general transformation for the estimation of $m^{[h]}$ than the present work by adding a suitably defined random variable $\tilde{Y}$ with $\mathbf{E}[\, \tilde{Y} \,|\, X \,] = 0$ to (2.13)). The next section discusses the construction of estimates of the regression function in the presence of censored data in more detail.

If (2.9) holds, then (2.13) and (2.14) depend on the transformation parameter $\alpha_{[h]} \in \mathbb{R}$. For the estimation of m, one could, e.g., choose $\alpha_{[h]}$ such that $Y^{[h]} \geq 0$ a.s. (corresponding to $Y \geq 0$ a.s.), which is for example fulfilled for $\alpha_{[h]} = 0$ (vide Leurgans (1987)), $\alpha_{[h]} = -1$ (vide Koul, Susarla, and Van Ryzin (1981)) or the data-dependent choice of Fan and Gijbels (1994, 1996). For general functions h, the parameter $\alpha_{[h]}$ may be defined in an analogous way. Note that the results of this thesis are valid for any (fixed) $\alpha_{[h]} \in \mathbb{R}$.

Now assume that the values of $h(x,y)$ are known for all $(x,y) \in [0,1]^d \times [0,\tau_F]$. But even in this case, the random variables $Y_1^{[h]}, \ldots, Y_n^{[h]}$ are not calculable, since the survival function G of the censoring time is unknown in a statistical application. An obvious idea is to replace G in (2.14) by the Kaplan-Meier product-limit estimate G_n (vide (1.5)).

For all $i = 1, \ldots, n$, this results in

$$\hat{Y}_i^{[h]} := \begin{cases} (1 + \alpha_{[h]}) \displaystyle\int_0^{Z_i} \frac{h'_t(X_i, t)}{G_n(t)}\, dt - \alpha_{[h]} \frac{\delta_i h(X_i, Z_i)}{G_n(Z_i)} & \text{if (2.9) holds} \\ \dfrac{\delta_i h(X_i, Z_i)}{G_n(Z_i)} & \text{otherwise,} \end{cases} \tag{2.19}$$

where we set $\frac{0}{0} := 0$. Moreover, let

$$\hat{\mathcal{D}}_n^{[h]} := \left\{ (X_1, \hat{Y}_1^{[h]}), \ldots, (X_n, \hat{Y}_n^{[h]}) \right\}. \tag{2.20}$$

Observe that $\hat{Y}_1^{[h]}, \ldots, \hat{Y}_n^{[h]}$ depend on the sample size n and this is suppressed in our notation. Furthermore, we want to stress that these random variables are in general neither independent nor identically distributed or even fulfill an equality similar to (2.15). The key step in the proof of our main results (vide Sections 3.2 and 4.2) is rather to control the squared differences $|Y_1^{[h]} - \hat{Y}_1^{[h]}|^2, \ldots, |Y_n^{[h]} - \hat{Y}_n^{[h]}|^2$, which we term as squared *transformation errors*.

In the next sections, we will specify three choices of h, each resulting in a different transformed data set (2.20). Based on this virtually uncensored data, we then define estimates of the regression function m, the conditional variance σ^2, and the conditional survival function $F(\tau \,|\, \cdot)$ ($\tau \in \mathbb{R}$ fixed) of Y given X in the presence of censored data.

2.3 Estimating the regression function from censored data

Let the assumptions of Section 2.2 hold. In this section, our estimates of the regression function $m(X) = \mathbf{E}[\, Y \mid X \,]$ in the presence of censored data are defined. Consequently, we set $h = h_1$ in (2.6), where

$$h_1 : [0,1]^d \times [0, \tau_F] \to [0, B_1] : (x, y) \mapsto y$$

and $B_1 := L \geq \tau_F$. In this case, it is obvious that (2.9) holds with $h'_y(x, y) = h'_{1,y}(x, y) = 1$ for all $(x, y) \in [0,1]^d \times [0, \tau_F]$. As mentioned in Section 2.2, (2.13) and (2.14) thus coincide with the transformed random variables in the "new class" of transformations defined in Fan and Gijbels (1994, 1996).

Set $\alpha_1 := \alpha_{[h_1]}$. Next, the following notations for the transformed random variables $Y^{[h_1]}$, $Y_1^{[h_1]}$, ..., $Y_n^{[h_1]}$ are introduced:

$$U^{(1)} := Y^{[h_1]} = (1+\alpha_1)\int_0^Z \frac{1}{G(t)}\,dt - \alpha_1 \frac{\delta Z}{G(Z)} \tag{2.21}$$

and for all $i = 1, \ldots, n$

$$U_i^{(1)} := Y_i^{[h_1]} = (1+\alpha_1)\int_0^{Z_i} \frac{1}{G(t)}\,dt - \alpha_1 \frac{\delta_i Z_i}{G(Z_i)}. \tag{2.22}$$

Since (2.15) and (2.16) imply that

$$\mathbf{E}\left[\, U^{(1)} \,\middle|\, X \,\right] = m(X) \tag{2.23}$$

and

$$|U^{(1)}| \le (1+2|\alpha_1|)\frac{L}{G(L)} =: L_1^\star < \infty \quad \text{a.s.}, \tag{2.24}$$

the estimation of m will now be based on the estimates of the transformed data (cf. Section 2.2).

Therefore, we set

$$\hat{U}_i^{(1)} := \hat{Y}_i^{[h_1]} := (1+\alpha_1)\int_0^{Z_i} \frac{1}{G_n(t)}\,dt - \alpha_1 \frac{\delta_i Z_i}{G_n(Z_i)} \quad (i = 1, \ldots, n) \tag{2.25}$$

($\frac{0}{0} := 0$) and

$$\hat{\mathcal{D}}_n^{(1)} := \hat{\mathcal{D}}_n^{[h_1]} := \left\{ (X_1, \hat{U}_1^{(1)}), \ldots, (X_n, \hat{U}_n^{(1)}) \right\}. \tag{2.26}$$

For the data $\hat{\mathcal{D}}_n^{(1)}$, multivariate smoothing spline estimates of m for censored regression can now be defined in analogy to Definition 2.1.

Let $d, k \in \mathbb{N}$ with $2k > d$ and $\lambda_n > 0$. Our MSSE of m for censored data is given by

$$\tilde{m}_{n,(k,\lambda_n)}(\cdot) := \tilde{m}_{n,(k,\lambda_n)}(\cdot, \hat{\mathcal{D}}_n^{(1)}) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n}\sum_{i=1}^n |f(X_i) - \hat{U}_i^{(1)}|^2 + \lambda_n J_k^2(f) \right) \tag{2.27}$$

with $W_k([0,1]^d)$ and $J_k^2(\cdot)$ defined as in (2.3) and (2.5), respectively.

Since $0 \le Y = h_1(X,Y) \le L < \infty$ a.s. (vide **(RA2)**), it holds with probability one that $0 \le m(X) = \mathbf{E}[\, Y \mid X \,] \le L$. Hence, we now truncate our estimate (2.27) such that it is bounded in the same way:

$$m_{n,(k,\lambda_n)}(\cdot) := T_{[0,L]}\tilde{m}_{n,(k,\lambda_n)}(\cdot). \tag{2.28}$$

Here, for all $a_1, a_2, t \in \mathbb{R}$ with $a_1 \leq a_2$,

$$T_{[a_1,a_2]}t := \begin{cases} a_2 & \text{if} \quad t > a_2 \\ t & \text{if} \quad a_1 \leq t \leq a_2 \\ a_1 & \text{if} \quad t < a_1, \end{cases} \tag{2.29}$$

and for all functions $f : \mathbb{R}^d \to \mathbb{R}$, we define $T_{[a_1,a_2]}f : \mathbb{R}^d \to \mathbb{R}$ by

$$\left(T_{[a_1,a_2]}f\right)(x) := T_{[a_1,a_2]}(f(x)) \quad (x \in \mathbb{R}^d). \tag{2.30}$$

2.4 Adaptation via splitting of the sample

Assume that the conditions of Section 2.2 hold. The estimates (2.27) and (2.28) depend on the smoothing parameter λ_n and on k, which defines the degree of the Sobolev space $W_k([0,1]^d)$. It is evident that a non-data-dependent choice of these parameters can lead to very unsatisfactory results. Therefore we modify the estimate in a second step and choose k and λ_n in a totally data-dependent way via the splitting of the sample technique.

Let $n \geq 2$. We split the sample $(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)$ into two parts, the so-called *learning* or *training data*

$$\mathcal{D}_{n_1}^{(1)} := \{(X_1, Z_1, \delta_1), \ldots, (X_{n_1}, Z_{n_1}, \delta_{n_1})\} \tag{2.31}$$

and the so-called *testing data*

$$\mathcal{D}_{n_t}^{(1)} := \{(X_{n_1+1}, Z_{n_1+1}, \delta_{n_1+1}), \ldots, (X_n, Z_n, \delta_n)\}, \tag{2.32}$$

where $n_t, n_1 \in \mathbb{N}$ with $n_t + n_1 = n$.

Next, the estimates of the transformed data points (2.22) are computed separately for both of these sets. Let $n_0, n_1 \in \mathbb{N}$ with $n_0 \leq n_1$. In analogy to (1.5), the Kaplan-Meier estimate of G based on the sample $\{(Z_{n_0}, \delta_{n_0}), \ldots, (Z_{n_1}, \delta_{n_1})\}$ is defined by

$$\hat{G}^{(KM)}(t, \{(Z_{n_0}, \delta_{n_0}), \ldots, (Z_{n_1}, \delta_{n_1})\}) := \prod_{\substack{i=n_0,\ldots,n_1: \\ Z_{(i)} \leq t}} \left[\frac{n_1 - i}{n_1 - i + 1}\right]^{1-\delta_{(i)}} \quad (t \in \mathbb{R}), \tag{2.33}$$

where $(Z_{(i)}, \delta_{(i)})$ denote the observed pairs (Z_i, δ_i) $(i = n_0, \ldots, n_1)$, arranged in such a way that $Z_{(n_0)} \leq Z_{(n_0+1)} \leq \ldots \leq Z_{(n_1)}$. Observe that for all $t \in \mathbb{R}$, (1.5) implies that $G_n(t) = \hat{G}^{(KM)}(t, \{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\})$.

Let $\alpha_1 \in \mathbb{R}$ and denote the Kaplan-Meier estimates calculated on the learning data and testing data by

$$G_{n_l}(\cdot) := \hat{G}^{(KM)}(\cdot, \{(Z_1, \delta_1), \ldots, (Z_{n_l}, \delta_{n_l})\}) \tag{2.34}$$

and

$$G_{n_t}(\cdot) := \hat{G}^{(KM)}(\cdot, \{(Z_{n_l+1}, \delta_{n_l+1}), \ldots, (Z_n, \delta_n)\}), \tag{2.35}$$

respectively. Similar to (2.25), the estimates of the random variables (2.22) are now defined as

$$\hat{U}^{(1)}_{i,n_l} := (1+\alpha_1) \int_0^{Z_i} \frac{1}{G_{n_l}(t)}\, dt - \alpha_1 \frac{\delta_i Z_i}{G_{n_l}(Z_i)} \quad (i = 1, \ldots, n_l) \tag{2.36}$$

and

$$\hat{U}^{(1)}_{i,n_t} := (1+\alpha_1) \int_0^{Z_i} \frac{1}{G_{n_t}(t)}\, dt - \alpha_1 \frac{\delta_i Z_i}{G_{n_t}(Z_i)} \quad (i = n_l+1, \ldots, n) \tag{2.37}$$

$\left(\frac{0}{0} := 0\right)$. The transformed learning and testing data are therefore given by

$$\hat{\mathcal{D}}^{(1)}_{n_l} := \left\{(X_1, \hat{U}^{(1)}_{1,n_l}), \ldots, (X_{n_l}, \hat{U}^{(1)}_{n_l,n_l})\right\} \tag{2.38}$$

and

$$\hat{\mathcal{D}}^{(1)}_{n_t} := \left\{(X_{n_l+1}, \hat{U}^{(1)}_{n_l+1,n_t}), \ldots, (X_n, \hat{U}^{(1)}_{n,n_t})\right\}. \tag{2.39}$$

Moreover denote the complete transformed data set by

$$\hat{\mathcal{D}}^{(1)}_{n_l,n_t} := \left\{(X_1, \hat{U}^{(1)}_{1,n_l}), \ldots, (X_{n_l}, \hat{U}^{(1)}_{n_l,n_l}), (X_{n_l+1}, \hat{U}^{(1)}_{n_l+1,n_t}), \ldots, (X_n, \hat{U}^{(1)}_{n,n_t})\right\}. \tag{2.40}$$

Note that $\hat{\mathcal{D}}^{(1)}_{n_l}$ and $\hat{\mathcal{D}}^{(1)}_{n_t}$ are independent (in the sense that the two sequences of random variables in (2.38) and (2.39) are independent). In contrast, if one simply splits the data (2.26) into two parts, each of the random variables $\hat{U}_i$ ($i \in \{1, \ldots, n\}$) given by (2.25) depends on the Kaplan-Meier estimate $G_n(\cdot) = \hat{G}^{(KM)}(\cdot, \{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\})$ and therefore on the whole sample $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$. I.e., as mentioned in Section 2.2, $\hat{U}_1, \ldots, \hat{U}_n$ are in general not independent.

Now consider the set of parameters $K_n \times \Lambda_n$ with

$$K_n := \left\{\left\lfloor \frac{d}{2} \right\rfloor + 1, \left\lfloor \frac{d}{2} \right\rfloor + 2, \ldots, \left\lfloor \frac{d}{2} \right\rfloor + \lceil (\ln n)^2 \rceil\right\} \tag{2.41}$$

and

$$\Lambda_n := \left\{\frac{\ln n}{2^n}, \frac{\ln n}{2^{n-1}}, \ldots, \frac{\ln n}{1}\right\}. \tag{2.42}$$

For each pair of parameters $(k, \lambda) \in K_n \times \Lambda_n$, we first use the data (2.38) to define an estimate $m_{n_1,(k,\lambda)}$ via

$$\tilde{m}_{n_1,(k,\lambda)}(\cdot) := \tilde{m}_{n_1,(k,\lambda)}(\cdot, \hat{\mathcal{D}}_{n_1}^{(1)}) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} |f(X_i) - \hat{U}_{i,n_1}^{(1)}|^2 + \lambda J_k^2(f) \right), \quad (2.43)$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are given by (2.3) and (2.5), and

$$m_{n_1,(k,\lambda)}(\cdot, \hat{\mathcal{D}}_{n_1}^{(1)}) := m_{n_1,(k,\lambda)}(\cdot) := T_{[0,L]} \tilde{m}_{n_1,(k,\lambda)}(\cdot). \quad (2.44)$$

Then we choose that estimate out of all calculated estimates (2.44) which performs best on the data (2.39) in terms of the empirical $\mathcal{L}_2$ risk. To be more precise, our modified MSSE is defined by

$$m_n(\cdot) := m_{n_1,\left(k^{(1)},\lambda^{(1)}\right)}(\cdot), \quad (2.45)$$

where

$$\left(k^{(1)}, \lambda^{(1)}\right) := \operatorname*{arg\,min}_{(k,\lambda) \in K_n \times \Lambda_n} \left(\frac{1}{n_t} \sum_{i=n_1+1}^{n} |m_{n_1,(k,\lambda)}(X_i) - \hat{U}_{i,n_t}^{(1)}|^2 \right). \quad (2.46)$$

2.5 Estimating the conditional variance

Let the assumptions of Section 2.2 hold. Below, we consider the estimation of the conditional variance of Y given X, i.e.,

$$\sigma^2(X) := \mathbf{Var}\left[Y \mid X \right], \quad (2.47)$$

in the presence of censored data. Since

$$\sigma^2(X) = \mathbf{E}\left[Y^2 \,\middle|\, X \right] - m(X)^2 = \mathbf{E}\left[Y^2 - m(X)^2 \,\middle|\, X \right], \quad (2.48)$$

$\sigma^2(X)$ is the regression function to $(X, Y^2 - m(X)^2)$ (cf. (1.18)). Due to (2.48), we can split our problem of estimating σ^2 into two parts: The transformation of the censored data in order to estimate the conditional second moment $\mathbf{E}\left[Y^2 \,\middle|\, X \right]$, and the estimation of the squared regression function $m(X)^2$. For the latter, one may simply use the squared MSSE from Section 2.3 or from Section 2.4, i.e., estimate $m(X)^2$ by $m_{n,(k,\lambda_n)}(X)^2$ or $m_n(X)^2$, where $m_{n,(k,\lambda_n)}$ and m_n are given by (2.28) and (2.45), respectively.

In order to derive a solution for the first problem, set $h = h_2$ in Section 2.2, where

$$h_2 : [0,1]^d \times [0, \tau_F] \to [0, B_2] : (x, y) \mapsto y^2$$

and $B_2 := L^2 \geq \tau_F^2$. In this case, (2.9) obviously holds with $h'_y(x,y) = h'_{2,y}(x,y) = 2y$ for all $(x,y) \in [0,1]^d \times [0,\tau_F]$ (cf. Section 2.3). In analogy to Section 2.3, we define $\alpha_2 := \alpha_{[h_2]}$, $U^{(2)} := Y^{[h_2]}$, $U_i^{(2)} := Y_i^{[h_2]}$, and $\hat{U}_i^{(2)} := \hat{Y}_i^{[h_2]}$ $(i = 1, \ldots, n)$, where $Y^{[h]}$, $Y_i^{[h]}$, and $\hat{Y}_i^{[h]}$ are given by (2.13), (2.14), and (2.19), respectively.

To be more precise, let

$$U^{(2)} := Y^{[h_2]} = (1+\alpha_2) \int_0^Z \frac{2\,t}{G(t)}\, dt - \alpha_2 \frac{\delta Z^2}{G(Z)}, \tag{2.49}$$

$$U_i^{(2)} := Y_i^{[h_2]} = (1+\alpha_2) \int_0^{Z_i} \frac{2\,t}{G(t)}\, dt - \alpha_2 \frac{\delta_i Z_i^2}{G(Z_i)} \quad (i = 1, \ldots, n), \tag{2.50}$$

and

$$\hat{U}_i^{(2)} := \hat{Y}_i^{[h_2]} = (1+\alpha_2) \int_0^{Z_i} \frac{2\,t}{G_n(t)}\, dt - \alpha_2 \frac{\delta_i Z_i^2}{G_n(Z_i)} \quad (i = 1, \ldots, n) \tag{2.51}$$

$\left(\frac{0}{0} := 0\right)$. From (2.15), (2.16), (2.49), and (2.48), one can conclude

$$|U^{(2)}| \leq (1 + 2|\alpha_2|) \frac{L^2}{G(L)} =: L_2^\star < \infty \quad \text{a.s.} \tag{2.52}$$

and

$$\mathbf{E}\left[U^{(2)} - m(X)^2 \,\middle|\, X \right] = \mathbf{E}\left[Y^2 \,\middle|\, X \right] - m(X)^2 = \sigma^2(X). \tag{2.53}$$

Hence, $\sigma^2(X)$ is the regression function to $(X, U^{(2)} - m(X)^2)$. Therefore, similar to the estimation of the regression function (cf. Sections 2.3 and 2.4), the estimation of $\sigma^2(X)$ in the presence of censored data may be based on estimates of the random variables $U_i^{(2)} - m(X_i)^2$ $(i = 1, \ldots, n)$. As mentioned above, we will present two different versions of these estimates. The first one is based on the regression estimate $m_{n,(k,\lambda_n)}$, while the second one is defined via a MSSE, which is given similar to m_n.

We start with the description of the first version. Let $k_1 \in \mathbb{N}$ with $2k_1 > d$ and $\lambda_{1,n} > 0$. For all $i = 1, \ldots, n$, define the estimates $\bar{U}_{i,n,(k_1,\lambda_{1,n})}$ of $U_i^{(2)} - m(X_i)^2$ via

$$\bar{U}_{i,n,(k_1,\lambda_{1,n})} := \hat{U}_i^{(2)} - m_{n,(k_1,\lambda_{1,n})}(X_i)^2, \tag{2.54}$$

where $m_{n,(k_1,\lambda_{1,n})}$ and $\hat{U}_i^{(2)}$ are given by (2.28) and (2.51), respectively. So an obvious idea for constructing our first estimate of σ^2 is to simply apply the MSSE of Definition 2.1 to the data

$$\bar{\mathcal{D}}_n^{(2)} := \bar{\mathcal{D}}_{n,(k_1,\lambda_{1,n})}^{(2)} := \left\{ (X_1, \bar{U}_{1,n,(k_1,\lambda_{1,n})}), \ldots, (X_n, \bar{U}_{n,n,(k_1,\lambda_{1,n})}) \right\}. \tag{2.55}$$

Let $k_2 \in \mathbb{N}$ with $2k_2 > d$ and $\lambda_{2,n} > 0$. Define $\tilde{\sigma}^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$ by

$$\tilde{\sigma}^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(\cdot) := \operatorname*{arg\,min}_{f \in W_{k_2}([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \bar{U}_{i,n,(k_1,\lambda_{1,n})}|^2 + \lambda_{2,n} J^2_{k_2}(f) \right), \quad (2.56)$$

where $W_{k_2}([0,1]^d)$ and $J^2_{k_2}(\cdot)$ are given by (2.3) and (2.5).

Now observe that $0 \leq Y \leq L$ a.s. (vide **(RA2)**) implies $0 \leq m(X) \leq L$ a.s. (cf. Section 2.3). From this together with (1.17), one can conclude with probability one that $0 \leq \sigma^2(X) \leq L^2$. Consequently, we truncate our estimate of the conditional variance such that it is bounded in the same way. Set

$$\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(\cdot) := T_{[0,L^2]} \tilde{\sigma}^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(\cdot), \quad (2.57)$$

where $T_{[0,L^2]}$ is given by (2.29) and (2.30). Note that $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$ depends on the four parameters k_1, k_2, $\lambda_{1,n}$, and $\lambda_{2,n}$, where k_1 and $\lambda_{1,n}$ are the parameters of the underlying estimate $m_{n,(k_1,\lambda_{1,n})}$ of the regression function.

Next, we define our second version of a MSSE of σ^2. For the estimate σ^2_n, the parameters are now chosen in a totally data-dependent way by the splitting of the sample technique (cf. Section 2.4). For this purpose, our sample $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ will now be split into three parts. On the first two parts, we define a MSSE of the regression function in analogy to (2.45), which is used to calculate estimates of σ^2 similar to (2.57). We then choose that estimate out of all calculated MSSE which performs best on the third part in terms of the empirical $\mathcal{L}_2$ risk.

Let $n \geq 3$. Set

$$\mathcal{D}^{(2)}_{n_1} := \{(X_1, Z_1, \delta_1), \ldots, (X_{n_1}, Z_{n_1}, \delta_{n_1})\}, \quad (2.58)$$

$$\mathcal{D}^{(2)}_{N_{\mathtt{r}}} := \{(X_{n_1+1}, Z_{n_1+1}, \delta_{n_1+1}), \ldots, (X_{N_1}, Z_{N_1}, \delta_{N_1})\}, \quad (2.59)$$

and

$$\mathcal{D}^{(2)}_{N_{\mathtt{t}}} := \{(X_{N_1+1}, Z_{N_1+1}, \delta_{N_1+1}), \ldots, (X_n, Z_n, \delta_n)\}. \quad (2.60)$$

Here, $n_{\mathtt{l}}, N_{\mathtt{r}}, N_{\mathtt{t}} \in \mathbb{N}$ with $n_{\mathtt{l}} + N_{\mathtt{r}} + N_{\mathtt{t}} = n$ and $N_{\mathtt{l}} := n_{\mathtt{l}} + N_{\mathtt{r}}$.

Similar to Section 2.4, we compute the Kaplan-Meier estimate for each of the sets (2.58), (2.59), and (2.60) separately. I.e., let $\hat{G}^{(KM)}$ and G_{n_1} be given by (2.33) and (2.34) and set

$$G_{N_{\mathtt{r}}}(\cdot) := \hat{G}^{(KM)}(\cdot, \{(Z_{n_1+1}, \delta_{n_1+1}), \ldots, (Z_{N_1}, \delta_{N_1})\}) \quad (2.61)$$

and

$$G_{N_{\mathtt{t}}}(\cdot) := \hat{G}^{(KM)}(\cdot, \{(Z_{N_{\mathtt{l}}+1}, \delta_{N_{\mathtt{l}}+1}), \ldots, (Z_n, \delta_n)\}). \tag{2.62}$$

Now, we use (2.61) and (2.62) in order to define

$$\hat{U}^{(1)}_{i,N_{\mathtt{r}}} := (1+\alpha_1) \int_0^{Z_i} \frac{1}{G_{N_{\mathtt{r}}}(t)}\, dt - \alpha_1 \frac{\delta_i Z_i}{G_{N_{\mathtt{r}}}(Z_i)} \quad (i = n_{\mathtt{l}}+1, \ldots, N_{\mathtt{l}}) \tag{2.63}$$

as well as

$$\hat{U}^{(2)}_{i,N_{\mathtt{r}}} := (1+\alpha_2) \int_0^{Z_i} \frac{2\,t}{G_{N_{\mathtt{r}}}(t)}\, dt - \alpha_2 \frac{\delta_i Z_i^2}{G_{N_{\mathtt{r}}}(Z_i)} \quad (i = n_{\mathtt{l}}+1, \ldots, N_{\mathtt{l}}) \tag{2.64}$$

and

$$\hat{U}^{(2)}_{i,N_{\mathtt{t}}} := (1+\alpha_2) \int_0^{Z_i} \frac{2\,t}{G_{N_{\mathtt{t}}}(t)}\, dt - \alpha_2 \frac{\delta_i Z_i^2}{G_{N_{\mathtt{t}}}(Z_i)} \quad (i = N_{\mathtt{l}}+1, \ldots, n) \tag{2.65}$$

$\left(\frac{0}{0} := 0\right)$.

Let the MSSE $m_{N_{\mathtt{l}}}$ of m be given by (2.45) with n, $n_{\mathtt{t}}$, and $\hat{U}^{(1)}_{n_{\mathtt{l}}+1,n_{\mathtt{t}}}, \ldots, \hat{U}^{(1)}_{n,n_{\mathtt{t}}}$ replaced by $N_{\mathtt{l}}$, $N_{\mathtt{r}}$, and $\hat{U}^{(1)}_{n_{\mathtt{l}}+1,N_{\mathtt{r}}}, \ldots, \hat{U}^{(1)}_{N_{\mathtt{l}},N_{\mathtt{r}}}$, respectively. To be more precise, define $K_n \times \Lambda_n$ by (2.41) and (2.42). For all $(k,\lambda) \in K_n \times \Lambda_n$, let $m_{n_{\mathtt{l}},(k,\lambda)}$ be given via (2.44) and set

$$m_{N_{\mathtt{l}}}(\cdot) := m_{n_{\mathtt{l}},\left(\bar{k}^{(1)},\bar{\lambda}^{(1)}\right)}(\cdot), \tag{2.66}$$

where

$$\left(\bar{k}^{(1)}, \bar{\lambda}^{(1)}\right) := \operatorname*{arg\,min}_{(k,\lambda)\in K_n\times\Lambda_n} \left(\frac{1}{N_{\mathtt{r}}} \sum_{i=n_{\mathtt{l}}+1}^{N_{\mathtt{l}}} |m_{n_{\mathtt{l}},(k,\lambda)}(X_i) - \hat{U}^{(1)}_{i,N_{\mathtt{r}}}|^2 \right). \tag{2.67}$$

Similar to (2.54), the estimates of $U_i^{(2)} - m(X_i)^2$ $(i = n_{\mathtt{l}}+1, \ldots, n)$ are now defined as

$$\bar{U}_{i,N_{\mathtt{r}}} := \hat{U}^{(2)}_{i,N_{\mathtt{r}}} - m_{N_{\mathtt{l}}}(X_i)^2 \quad (i = n_{\mathtt{l}}+1, \ldots, N_{\mathtt{l}}) \tag{2.68}$$

and

$$\bar{U}_{i,N_{\mathtt{t}}} := \hat{U}^{(2)}_{i,N_{\mathtt{t}}} - m_{N_{\mathtt{l}}}(X_i)^2 \quad (i = N_{\mathtt{l}}+1, \ldots, n). \tag{2.69}$$

Set

$$\hat{\mathcal{D}}^{(2)}_{N_{\mathtt{t}}} := \left\{(X_{N_{\mathtt{l}}+1}, \bar{U}_{N_{\mathtt{l}}+1,N_{\mathtt{t}}}), \ldots, (X_n, \bar{U}_{n,N_{\mathtt{t}}})\right\}. \tag{2.70}$$

Now, we are in the position to present our modified MSSE of σ^2. For each pair of parameters $(k,\lambda) \in K_n \times \Lambda_n$, let the estimates $\tilde{\sigma}^2_{N_{\mathtt{l}},(k,\lambda)}$ and $\sigma^2_{N_{\mathtt{l}},(k,\lambda)}$ be given by

$$\tilde{\sigma}^2_{N_{\mathtt{l}},(k,\lambda)}(\cdot) := \operatorname*{arg\,min}_{f\in W_k([0,1]^d)} \left(\frac{1}{N_{\mathtt{r}}} \sum_{i=n_{\mathtt{l}}+1}^{N_{\mathtt{l}}} |f(X_i) - \bar{U}_{i,N_{\mathtt{r}}}|^2 + \lambda J_k^2(f) \right), \tag{2.71}$$

and

$$\sigma^2_{N_1,(k,\lambda)}(\cdot) := T_{[0,L^2]}\tilde{\sigma}^2_{N_1,(k,\lambda)}(\cdot), \tag{2.72}$$

respectively. Here, $W_k([0,1]^d)$ and $J^2_k(\cdot)$ are defined by (2.3) and (2.5). Similar to (2.45), we now choose that estimate out of all calculated estimates (2.72) which performs best on the data (2.70) in terms of the empirical $\mathcal{L}_2$ risk, i.e., set

$$\sigma^2_n(\cdot) := \sigma^2_{N_1,\left(k^{(2)},\lambda^{(2)}\right)}(\cdot), \tag{2.73}$$

where

$$\left(k^{(2)},\lambda^{(2)}\right) := \operatorname*{arg\,min}_{(k,\lambda)\in K_n\times\Lambda_n}\left(\frac{1}{N_{\mathtt{t}}}\sum_{i=N_1+1}^{n}|\sigma^2_{N_1,(k,\lambda)}(X_i)-\bar{U}_{i,N_{\mathtt{t}}}|^2\right). \tag{2.74}$$

2.6 Estimating the conditional survival function

Assume that the conditions of Section 2.2 hold. Let $\tau\in\mathbb{R}$ be arbitrary, but fixed. In the following, we define our estimates of the conditional survival function of Y given X, i.e.,

$$F\left(\tau\mid X\right) := \mathbf{P}\left[\,Y>\tau\mid X\,\right], \tag{2.75}$$

in the presence of censored data.

Observe that

$$F\left(\tau\mid X\right) = \mathbf{E}\left[\,I_{[Y>\tau]}\,\middle|\,X\,\right]. \tag{2.76}$$

This implies that $F(\tau|\,X)$ is the regression function to $(X, I_{[Y>\tau]})$ (cf. Section 1.4). In order to define estimates of the conditional survival function $F(\tau|\,X)$, we set $h=h_3$ in (2.6), where

$$h_3 : [0,1]^d\times[0,\tau_F]\to[0,B_3] : (x,y)\mapsto I_{[y>\tau]}$$

and $B_3 := 1$. Now note that if $\tau<\tau_F$, (2.9) is violated – either since h_3 is discontinuous in $y=\tau$ (for $\tau\in[0,\tau_F)$) or $h_3(x,0)=1$ (for $\tau<0$). And if $\tau\geq\tau_F$, then (2.9) holds, but $h_3(x,y)=0$ and hence $h'_{3,y}(x,y)=0$ for all $(x,y)\in[0,1]^d\times[0,\tau_F]$. Observe that in both cases, we derive on the same form of the transformation (in the second case, we get $Y^{[h_3]}=Y_i^{[h_3]}=\hat{Y}_i^{[h_3]}=0$ a.s. $\forall i=1,\ldots,n$).

Let $Y^{[h]}$, $Y_i^{[h]}$, and $\hat{Y}_i^{[h]}$ $(i=1,\ldots,n)$ be given by (2.13), (2.14), and (2.19). Set

$$U^{(3)} := Y^{[h_3]} = \frac{\delta I_{[Z>\tau]}}{G(Z)} \tag{2.77}$$

and

$$U_i^{(3)} := Y_i^{[h_3]} = \frac{\delta_i I_{[Z_i>\tau]}}{G(Z_i)} \quad (i = 1, \ldots, n). \tag{2.78}$$

Furthermore, define $\hat{U}_i^{(3)}$ $(i = 1, \ldots, n)$ by

$$\hat{U}_i^{(3)} := \hat{Y}_i^{[h_3]} = \frac{\delta_i I_{[Z_i>\tau]}}{G_n(Z_i)} \quad \left(\frac{0}{0} := 0\right). \tag{2.79}$$

Note that $U^{(3)}$, $U_i^{(3)}$, and $\hat{U}_i^{(3)}$ depend on τ and we have suppressed this in our notation.

From (2.15), (2.17), (2.76), and (2.77), one can conclude that

$$\mathbf{E}\left[\, U^{(3)} \,\middle|\, X \,\right] = F\left(\tau \,|\, X\right) \tag{2.80}$$

and

$$|U^{(3)}| \leq \frac{1}{G(L)} =: L_3^\star < \infty \quad \text{a.s.} \tag{2.81}$$

Obviously, (2.80) implies that $F(\tau|\, X)$ is the regression function to $(X, U^{(3)})$ (cf. Section 2.2).

Next, we introduce our MSSE of $F(\tau|\, X)$ (cf. Section 2.3) in the presence of censored data. Let $k, d \in \mathbb{N}$ with $2k > d$ and $\lambda_n > 0$. Define $\tilde{F}_{n,(k,\lambda_n)}$ by

$$\tilde{F}_{n,(k,\lambda_n)}(\tau|\,\cdot) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \hat{U}_i^{(3)}|^2 + \lambda_n J_k^2(f) \right), \tag{2.82}$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are given by (2.3) and (2.5). Clearly, the conditional survival function $F(\tau|\, X)$ takes with probability one only values in $[0,1]$. Therefore, we define our truncated estimate $F_{n,(k,\lambda_n)}(\tau|\,\cdot)$ by

$$F_{n,(k,\lambda_n)}(\tau|\,\cdot) := T_{[0,1]}\tilde{F}_{n,(k,\lambda_n)}(\tau|\,\cdot). \tag{2.83}$$

Let $n \geq 2$. In analogy to the both preceding sections, the splitting of the sample technique is applied below in order to choose the parameters of the MSSE of $F(\tau|\, X)$ in a totally data-dependent way.

Let $\mathcal{D}_{n_1}^{(1)}$, $\mathcal{D}_{n_t}^{(1)}$, G_{n_1}, and G_{n_t} be given by (2.31), (2.32), (2.34), and (2.35). Similar to (2.36) and (2.37), define the estimates of the transformed random variables (2.78) separately for $\mathcal{D}_{n_1}^{(1)}$ and $\mathcal{D}_{n_t}^{(1)}$, i.e.,

$$\hat{U}_{i,n_1}^{(3)} := \frac{\delta_i I_{[Z_i>\tau]}}{G_{n_1}(Z_i)} \quad (i = 1, \ldots, n_1) \tag{2.84}$$

and

$$\hat{U}^{(3)}_{i,n_{\mathfrak{t}}} := \frac{\delta_i I_{[Z_i>\tau]}}{G_{n_{\mathfrak{t}}}(Z_i)} \quad (i = n_1+1,\dots,n). \tag{2.85}$$

$\left(\frac{0}{0} := 0\right)$. Moreover, set

$$\hat{\mathcal{D}}^{(3)}_{n_1} := \left\{(X_1, \hat{U}^{(3)}_{1,n_1}), \dots, (X_{n_1}, \hat{U}^{(3)}_{n_1,n_1})\right\} \tag{2.86}$$

and

$$\hat{\mathcal{D}}^{(3)}_{n_{\mathfrak{t}}} := \left\{(X_{n_1+1}, \hat{U}^{(3)}_{n_1+1,n_{\mathfrak{t}}}), \dots, (X_n, \hat{U}^{(3)}_{n,n_{\mathfrak{t}}})\right\}. \tag{2.87}$$

Denote the whole transformed data set by

$$\hat{\mathcal{D}}^{(3)}_{n_{\mathfrak{t}}} := \left\{(X_1, \hat{U}^{(3)}_{1,n_1}), \dots, (X_{n_1}, \hat{U}^{(3)}_{n_1,n_1}), (X_{n_1+1}, \hat{U}^{(3)}_{n_1+1,n_{\mathfrak{t}}}), \dots, (X_n, \hat{U}^{(3)}_{n,n_{\mathfrak{t}}})\right\}. \tag{2.88}$$

Let $K_n \times \Lambda_n$ be given by (2.41) and (2.42). For each pair of parameters $(k,\lambda) \in K_n \times \Lambda_n$, we define a MSSE $F_{n_1,(k,\lambda)}(\tau\,|\,\cdot)$ by

$$\tilde{F}_{n_1,(k,\lambda)}(\tau\,|\,\cdot) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} |f(X_i) - \hat{U}^{(3)}_{i,n_1}|^2 + \lambda J_k^2(f) \right), \tag{2.89}$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are given by (2.3) and (2.5), and

$$F_{n_1,(k,\lambda)}(\tau\,|\,\cdot) := T_{[0,1]} \tilde{F}_{n_1,(k,\lambda)}(\tau\,|\,\cdot) \tag{2.90}$$

(cf. 2.83). Similar to (2.44), we now choose that estimate out of all calculated estimates (2.90) which performs best on the data (2.87) in terms of the empirical $\mathcal{L}_2$ risk. I.e., our modified MSSE is given by

$$F_n(\tau\,|\,\cdot) := F_{n_1,\left(k^{(3)},\lambda^{(3)}\right)}(\tau\,|\,\cdot), \tag{2.91}$$

where

$$\left(k^{(3)},\lambda^{(3)}\right) := \operatorname*{arg\,min}_{(k,\lambda)\in K_n\times\Lambda_n} \left(\frac{1}{n_{\mathfrak{t}}} \sum_{i=n_1+1}^{n} |F_{n_1,(k,\lambda)}(\tau\,|\,X_i) - \hat{U}^{(3)}_{i,n_{\mathfrak{t}}}|^2 \right). \tag{2.92}$$

Chapter 3

Consistency

Chapter 3 presents conditions on the parameters of our MSSE (2.28), (2.57), and (2.83) which ensure that these estimates are strongly consistent for all distributions of (X, Y, C) satisfying **(RA1)** – **(RA4)**. Section 3.1 introduces an estimate, whose definition covers the definitions of (2.28), (2.57), and (2.83). As shown in Theorem 3.1, a key step in order to transfer the consistency property of our MSSE from usual nonparametric regression to censored regression is the analysis of the squared transformation errors (cf. Section 2.2). This is done in Section 3.2. The strong consistency of the suitably defined MSSE is proven in Sections 3.3 – 3.5. Finally, Section 3.6 contains the proof of Theorem 3.1.

3.1 A general result

Here, we shall put the setting of Sections 2.2, 2.3, 2.5, and 2.6 in a more general context. Let therefore $(X, Y^\star), (X_1, Y_1^\star), \ldots, (X_n, Y_n^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s. be i.i.d. random vectors, where $\beta^\star \in (0, \infty)$. Assume that there exists a constant $\beta \in (0, \beta^\star]$ such that

$$m^\star(X) := \mathbf{E}\,[\, Y^\star \mid X \,] \in [0, \beta] \quad \text{a.s.} \tag{3.1}$$

Define $U^{(1)}$, $U^{(2)}$, and $U^{(3)}$ by (2.21) (2.49), and (2.77), respectively. From (2.23), (2.53), and (2.80) we know that $m^\star(X)$ equals $m(X)$, $\sigma^2(X)$ or $F(\tau|\, X)$ ($\tau \in \mathbb{R}$ fixed), if one replaces $Y^\star$ in (3.1) by $U^{(1)}$, $U^{(2)} - m(X)^2$, and $U^{(3)}$, respectively.

Let $k \in \mathbb{N}$ with $2k > d$ and $\lambda_n > 0$. In this section, a MSSE $m^\star_{n,(k,\lambda_n)}$ of $m^\star$ is introduced whose definition covers the definitions of our estimates (2.28), (2.57), and (2.83).

Since $U_i^{(1)}$, $U_i^{(2)} - m(X_i)^2$, and $U_i^{(3)}$ $(i = 1, \ldots, n)$ are not calculable, we assume that the random variables $Y_i^\star$ are unknown, too. As in Sections 2.3, 2.5, and 2.6, the definition of $m^\star_{n,(k,\lambda_n)}$ is rather based on some observable real-valued random variables $\bar{Y}_1^{(n)}, \ldots, \bar{Y}_n^{(n)}$, where each $\bar{Y}_i^{(n)}$ may depend on the whole sample and a finite number of further, suitably chosen parameters (cf. (2.54) and (2.79)). Note that we neither demand the random variables $\bar{Y}_1^{(n)}, \ldots, \bar{Y}_n^{(n)}$ to be independent nor to be identically distributed. In fact, the only assumptions needed in this section are that the sequence

$$(X_1, \bar{Y}_1^{(n)}), \ldots, (X_n, \bar{Y}_n^{(n)}) \text{ is independent of } (X, Y^\star) \tag{3.2}$$

and that the mean squared difference

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i^\star - \bar{Y}_i^{(n)}|^2 \tag{3.3}$$

between $Y_1^\star, \ldots, Y_n^\star$ and $\bar{Y}_1^{(n)}, \ldots, \bar{Y}_n^{(n)}$ is "small". In the analysis of the MSSE (2.28) and (2.83), $|Y_1^\star - \bar{Y}_1^{(n)}|^2, \ldots, |Y_n^\star - \bar{Y}_n^{(n)}|^2$ represent the squared transformation errors (cf. Section 2.2). For the estimate (2.57), they depend on (2.50), (2.51), and the MSSE (2.28).

Similar to (1.11), one can show that the minimal $\mathcal{L}_2$ risk of a regression estimate based on the data

$$\bar{\mathcal{D}}_n := \left\{ (X_1, \bar{Y}_1^{(n)}), \ldots, (X_n, \bar{Y}_n^{(n)}) \right\} \tag{3.4}$$

is the same as in case of i.i.d. data. Here, we used that assumption (3.2) yields

$$\mathbf{E}\left[Y^\star \mid X, \bar{\mathcal{D}}_n \right] = \mathbf{E}\left[Y^\star \mid X \right] = m^\star(X). \tag{3.5}$$

Next, the MSSE of $m^\star$ in this general setting will be defined. In analogy to (2.27), (2.56), and (2.82), the estimate $\tilde{m}^\star_{n,(k,\lambda_n)}$ is given by

$$\tilde{m}^\star_{n,(k,\lambda_n)}(\cdot) := \tilde{m}^\star_{n,(k,\lambda_n)}(\cdot, \bar{\mathcal{D}}_n) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \bar{Y}_i^{(n)}|^2 + \lambda_n J_k^2(f) \right), \tag{3.6}$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are defined as in (2.3) and (2.5). Since $m^\star(X) \in [0, \beta]$ a.s., the MSSE (3.6) will be truncated in the same way (cf. (2.28), (2.57), and (2.83)), i.e., set

$$m^\star_{n,(k,\lambda_n)}(\cdot) := T_{[0,\beta]} \tilde{m}^\star_{n,(k,\lambda_n)}(\cdot). \tag{3.7}$$

The following theorem states the conditions on k, the smoothing parameter λ_n, and (3.3) for which $m^\star_{n,(k,\lambda_n)}$ is strongly consistent for all distributions of $(X, Y^\star)$ satisfying

$(X, Y^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s. and $m^\star(X) \in [0, \beta]$ a.s. $(0 < \beta \leq \beta^\star)$. From this result, we will then derive in Sections 3.3 – 3.5 the strong consistency of our MSSE (2.28), (2.57), and (2.83) in the presence of censored data.

Theorem 3.1. (Consistency) *Let $k, d \in \mathbb{N}$ with $2k > d$ and $0 < \beta \leq \beta^\star < \infty$. For $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n \to \infty) \qquad \text{and} \qquad \frac{n \lambda_n^{\frac{d}{2k}}}{\ln n} \to \infty \quad (n \to \infty). \tag{3.8}$$

Let the estimate $m^\star_{n,(k,\lambda_n)}$ be defined by (3.6) and (3.7). If assumption (3.2) holds and

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i^\star - \bar{Y}_i^{(n)}|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.} \tag{3.9}$$

then

$$\int_{\mathbb{R}^d} \left| m^\star_{n,(k,\lambda_n)}(x) - m^\star(x) \right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad \text{a.s.} \tag{3.10}$$

for every distribution of $(X, Y^\star)$ with $(X, Y^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s. and $m^\star(X) \in [0, \beta]$ a.s.

The proof of Theorem 3.1 will be given in Section 3.6.

3.2 Maximum squared transformation errors

Condition (3.9) in Theorem 3.1 indicates that an important task in the proofs of the strong consistency of the MSSE (2.28), (2.57), and (2.83) will be the control of the mean squared transformation errors. But in the analysis of the rate of convergence of our estimates, the stochastic convergence of the maximum squared transformation errors is demanded. Therefore, we now prove a slightly more general result than required by (3.9), which implies the almost sure convergence of the mean squared transformation errors.

Lemma 3.1. *Let $\alpha_1, \alpha_2 \in \mathbb{R}$ and let $\tau \in \mathbb{R}$ be arbitrary, but fixed. For $j \in \{1,2,3\}$ define $U_i^{(j)}$ and $\hat{U}_i^{(j)}$ $(i = 1, \ldots, n)$ by (2.22), (2.25), (2.50), (2.51), (2.78) and (2.79), respectively. Then, for every $j \in \{1,2,3\}$,*

$$\max_{i=1,\ldots,n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

for all distributions of (Y, C) with Y and C independent, which satisfy **(RA2)** *and* **(RA4)**.

PROOF OF LEMMA 3.1. First note that from **(RA2)** and due to the independence of Y and C, we have

$$\tau_F = \tau_K \leq L \quad \text{and} \quad Z_i \in [0, \tau_F] \text{ a.s. } (i = 1, \ldots, n). \tag{3.11}$$

If $j \in \{1, 2\}$, then one can conclude from (2.22), (2.25), (2.50), and (2.51) for all $i = 1, \ldots, n$

$$U_i^{(j)} = (1 + \alpha_j) \int_0^{Z_i} \frac{j \cdot t^{j-1}}{G(t)} \, dt - \alpha_j \frac{\delta_i Z_i^j}{G(Z_i)} \tag{3.12}$$

and

$$\hat{U}_i^{(j)} = (1 + \alpha_j) \int_0^{Z_i} \frac{j \cdot t^{j-1}}{G_n(t)} \, dt - \alpha_j \frac{\delta_i Z_i^j}{G_n(Z_i)}, \tag{3.13}$$

where $\frac{0}{0} := 0$. For $j \in \{1, 2\}$, (3.11) – (3.13) and the relation $(a_1 + a_2)^2 \leq 2a_1^2 + 2a_2^2$ $(a_1, a_2 \in \mathbb{R})$ imply

$$\begin{aligned}
&\max_{i=1,\ldots,n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \\
&= \max_{i=1,\ldots,n} \left| (1 + \alpha_j) \int_0^{Z_i} \left(\frac{j \cdot t^{j-1}}{G(t)} - \frac{j \cdot t^{j-1}}{G_n(t)} \right) dt - \alpha_j \left(\frac{\delta_i Z_i^j}{G(Z_i)} - \frac{\delta_i Z_i^j}{G_n(Z_i)} \right) \right|^2 \\
&\leq 2 \max_{i=1,\ldots,n} \left[\left| (1 + \alpha_j) \int_0^{Z_i} \left(\frac{j \cdot t^{j-1}}{G(t)} - \frac{j \cdot t^{j-1}}{G_n(t)} \right) dt \right|^2 + \left| \alpha_j \left(\frac{\delta_i Z_i^j}{G(Z_i)} - \frac{\delta_i Z_i^j}{G_n(Z_i)} \right) \right|^2 \right] \\
&\leq 2 (1 + |\alpha_j|)^2 \max_{i=1,\ldots,n} \left[\left| \int_0^{Z_i} j \cdot t^{j-1} \left(\frac{1}{G(t)} - \frac{1}{G_n(t)} \right) dt \right|^2 + \left| Z_i^j \left(\frac{1}{G(Z_i)} - \frac{1}{G_n(Z_i)} \right) \right|^2 \right] \\
&\leq 2 (1 + |\alpha_j|)^2 \max_{i=1,\ldots,n} \left[\left(\left| \int_0^{Z_i} j \cdot t^{j-1} \, dt \right|^2 + \left| Z_i^j \right|^2 \right) \left(\sup_{0 \leq t \leq Z_i} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2 \right] \\
&= 2 (1 + |\alpha_j|)^2 \max_{i=1,\ldots,n} \left[2 Z_i^{2j} \left(\sup_{0 \leq t \leq Z_i} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2 \right] \\
&\leq 4 (1 + |\alpha_j|)^2 \tau_F^{2j} \left(\sup_{0 \leq t \leq \tau_F} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2 \\
&\leq 4 (1 + |\alpha_j|)^2 L^{2j} \left(\sup_{0 \leq t \leq \tau_F} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2 \quad \text{a.s.}
\end{aligned}$$

If, in contrast, $j = 3$, then (2.78), (2.79), and (3.11) yield with probability one

$$\max_{i=1,\ldots,n} |U_i^{(3)} - \hat{U}_i^{(3)}|^2 = \max_{i=1,\ldots,n} \left| \frac{\delta_i I_{[Z_i > \tau]}}{G(Z_i)} - \frac{\delta_i I_{[Z_i > \tau]}}{G_n(Z_i)} \right|^2 \leq \left(\sup_{0 \leq t \leq \tau_F} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2.$$

Hence we have shown that for every $j \in \{1,2,3\}$

$$\max_{i=1,\ldots,n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \leq B^{(j)} \cdot \left(\sup_{0\leq t\leq \tau_F} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| \right)^2 \quad \text{a.s.} \tag{3.14}$$

holds, where $B^{(1)} := 4\,(1+|\alpha_1|)^2\,L^2$, $B^{(2)} := 4\,(1+|\alpha_2|)^2\,L^4$, and $B^{(3)} := 1$.

The survival function G of the censoring times is monotonically decreasing. Therefore, one can conclude from **(RA2)** and (3.11):

$$1 \geq G(t) \geq G(\tau_K) = G(\tau_F) \geq G(L) > 0 \quad \forall t \in [0, \tau_F]. \tag{3.15}$$

For fixed $n \in \mathbb{N}$, the Kaplan-Meier estimate G_n is also monotonically decreasing (vide (1.5)), i.e.,

$$1 \geq G_n(t) \geq G_n(\tau_K) = G_n(\tau_F) \quad \text{a.s.} \quad \forall t \in [0, \tau_F].$$

The last two inequalities imply

$$\begin{aligned} \sup_{0\leq t\leq \tau_F} \left| \frac{1}{G(t)} - \frac{1}{G_n(t)} \right| &= \sup_{0\leq t\leq \tau_F} \frac{|\,G_n(t) - G(t)\,|}{G(t)\cdot G_n(t)} \\ &\leq \sup_{0\leq t\leq \tau_F} \frac{1}{G(t)\cdot G_n(t)} \cdot \sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \\ &\leq \frac{1}{G(\tau_F)\cdot G_n(\tau_F)} \cdot \sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \quad \text{a.s.} \end{aligned} \tag{3.16}$$

For every $j \in \{1,2,3\}$, (3.14) and (3.16) yield with probability one

$$\max_{i=1,\ldots,n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \leq \frac{B^{(j)}}{G^2(\tau_F)\cdot G_n^2(\tau_F)} \cdot \left(\sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \right)^2. \tag{3.17}$$

Next, we will apply Theorem 1.1 to the right hand side of (3.14). From **(RA4)**, one can conclude that F and G have no jumps in common and that $\mathbf{P}\,[\,C = \tau_K\,] = 0$. First we note that this together with Theorem 1.1, $G(\tau_F) > 0$ (cf. (3.15)), **(RA4)** and

$$\begin{aligned} \mathbf{P}\left[\limsup_{n\to\infty} \frac{1}{G_n(\tau_F)} > \frac{2}{G(\tau_F)} \right] &= \mathbf{P}\left[\liminf_{n\to\infty} G_n(\tau_F) < \frac{G(\tau_F)}{2} \right] \\ &= \mathbf{P}\left[\limsup_{n\to\infty} \left(G(\tau_F) - G_n(\tau_F) \right) > \frac{G(\tau_F)}{2} \right] \\ &\leq \mathbf{P}\left[\limsup_{n\to\infty} \sup_{0\leq t\leq \tau_F} |G(t) - G_n(t)| > \frac{G(\tau_F)}{2} \right] \end{aligned}$$

implies

$$\limsup_{n\to\infty} \frac{1}{G_n^2(\tau_F)} \leq \frac{4}{G^2(\tau_F)} \quad \text{a.s.} \tag{3.18}$$

Finally, we have from (3.17) and (3.18) for every $j \in \{1,2,3\}$:

$$\begin{aligned}
&\limsup_{n\to\infty} \max_{i=1,\dots,n} \left| U_i^{(j)} - \hat{U}_i^{(j)} \right|^2 \\
&\leq \limsup_{n\to\infty} \left[\frac{B^{(j)}}{G^2(\tau_F) \cdot G_n^2(\tau_F)} \cdot \left(\sup_{0\leq t\leq \tau_F} |\, G_n(t) - G(t)\,| \right)^2 \right] \\
&\leq \frac{B^{(j)}}{G^2(\tau_F)} \cdot \left[\limsup_{n\to\infty} \frac{1}{G_n^2(\tau_F)} \right] \cdot \left[\limsup_{n\to\infty} \left(\sup_{0\leq t\leq \tau_F} |\, G_n(t) - G(t)\,| \right)^2 \right] \\
&\leq \frac{4B^{(j)}}{G^4(\tau_F)} \cdot \left[\limsup_{n\to\infty} \sup_{0\leq t\leq \tau_F} |\, G_n(t) - G(t)\,| \right]^2 \quad \text{a.s.}
\end{aligned}$$

This together with Theorem 1.1, **(RA4)**, $G(\tau_F) > 0$, and $B^{(j)} \in \mathbb{R}_+$ $(j \in \{1,2,3\})$ implies the assertion of Lemma 3.1.

□

3.3 Consistent MSSE of the regression function

Now we are in the position to formulate and prove the strong consistency of our MSSE of the regression function in the presence of censored data.

Theorem 3.2. (Consistency) *Let $k, d \in \mathbb{N}$ with $2k > d$ and let $\alpha_1 \in \mathbb{R}$. For $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n \to \infty) \qquad \text{and} \qquad \frac{n\lambda_n^{\frac{d}{2k}}}{\ln n} \to \infty \quad (n \to \infty). \tag{3.19}$$

Let $\hat{U}_1^{(1)}, \dots, \hat{U}_n^{(1)}$ be given by (2.25) and define the estimate $m_{n,(k,\lambda_n)}$ via (2.27) and (2.28). Then

$$\int_{\mathbb{R}^d} \left| m_{n,(k,\lambda_n)}(x) - m(x) \right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**.

Remark 3.1. It follows from assumption **(RA2)**, Theorem 1.1, and the proof of Lemma 3.1 that assumption **(RA4)** may be dropped in Theorem 3.2 if we assume that F and G do not have common jumps and that either $\mathbf{P}[\,C = \tau_F\,] = 0$ or $\mathbf{P}[\,C = \tau_F\,] > 0$ but $\mathbf{P}[\,Y = \tau_F\,] > 0$. The condition $\mathbf{P}[\,C = \tau_F\,] = 0$ demands that G is continuous in τ_F, while $\mathbf{P}[\,Y = \tau_F\,] > 0$ requires that the lifetime Y equals the upper endpoint τ_F with a

non-zero probability. Note that the latter assumption is fairly unrealistic in a statistical application.

Remark 3.2. Since in **(RA2)** $\mathbf{P}[\,C > L\,] = 1$ is allowed, Theorem 3.2 is still valid if no censoring occurs. In this case, assumption **(RA1)** may be abandoned in Theorem 3.2 if we slightly modify the estimate (vide Kohler and Krzyżak (2001), Remark 3).

Remark 3.3. The strong universal consistency of a weighted MSSE of the regression function in the presence of right censored data has already been demonstrated by Pintér (2001). Moreover, Pintér (2001) showed that suitably defined local averaging estimates of m are strongly consistent with respect to the $\mathcal{L}_2$ error in case that **(RA3)** is violated but Y and C are conditionally independent given X.

Remark 3.4. From $0 \leq m(x) \leq L$ (cf. **(RA2)**) and $0 \leq m_{n,(k,\lambda_n)}(x) \leq L$ (vide (2.28)), one can conclude

$$\int_{\mathbb{R}^d} \left| m_{n,(k,\lambda_n)}(x) - m(x) \right|^2 \mu(dx) \leq L^2 < \infty \quad \text{a.s.}$$

Hence the strong consistency of $m_{n,(k,\lambda_n)}$ and Lebesgue's dominated convergence theorem imply the weak consistency of $m_{n,(k,\lambda_n)}$.

Proof of Theorem 3.2. Let $U^{(1)}$, $U_i^{(1)}$, $\hat{U}_i^{(1)}$, and $\hat{\mathcal{D}}_n^{(1)}$ be defined by (2.21), (2.22), (2.25), and (2.26), respectively ($i = 1, \ldots, n$). First notice that (2.23) and (2.24) state

$$m(X) = \mathbf{E}\left[\, U^{(1)} \,\middle|\, X \,\right]$$

and

$$|U^{(1)}| \leq (1 + 2|\alpha_1|)\frac{L}{G(L)} = L_1^\star < \infty \quad \text{a.s.}$$

This together with **(RA2)** yields with probability one that $0 \leq m(X) \leq L \leq L_1^\star$. Moreover, since (X, Y, C), (X_1, Y_1, C_1), $\ldots$, (X_n, Y_n, C_n) are i.i.d. random vectors, we deduce from (1.5), (2.21), (2.25), and (2.26) that $(X, U^{(1)})$ and $\hat{\mathcal{D}}_n^{(1)}$ are independent.

Clearly, if one sets $\beta^\star = L_1^\star$, $\beta = L$, $Y^\star = U^{(1)}$, $Y_i^\star = U_i^{(1)}$, and $\bar{Y}_i^{(n)} = \hat{U}_i^{(1)}$ ($i = 1, \ldots, n$) in Section 3.1, then $\bar{\mathcal{D}}_n$ equals $\hat{\mathcal{D}}_n^{(1)}$, $m^\star$ equals m, $m^\star_{n,(k,\lambda_n)}$ equals $m_{n,(k,\lambda_n)}$, and

$$\frac{1}{n}\sum_{i=1}^n \left| Y_i^\star - \bar{Y}_i^{(n)} \right|^2 = \frac{1}{n}\sum_{i=1}^n |\, U_i^{(1)} - \hat{U}_i^{(1)} \,|^2 \leq \max_{i=1,\ldots,n} |\, U_i^{(1)} - \hat{U}_i^{(1)} \,|^2.$$

This together with Theorem 3.1 and Lemma 3.1 implies the assertion of Theorem 3.2.

□

3.4 Consistent MSSE of the conditional variance

In this section, we show that the suitably defined MSSE of the conditional variance $\sigma^2(X)$ is strongly consistent for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)**. Since the definition of this estimate depends on the MSSE of the regression function, we require that the assumptions of Theorem 3.2 are fulfilled.

Theorem 3.3. (Consistency) *Let $k_1, k_2, d \in \mathbb{N}$ with $2k_1, 2k_2 > d$ and let $\alpha_1, \alpha_2 \in \mathbb{R}$. For $n \in \mathbb{N}$ and $j \in \{1, 2\}$ choose $\lambda_{j,n} > 0$ such that*

$$\lambda_{j,n} \to 0 \quad (n \to \infty) \qquad \text{and} \qquad \frac{n \lambda_{j,n}^{\frac{d}{2k_j}}}{\ln n} \to \infty \quad (n \to \infty). \tag{3.20}$$

Let $\bar{U}_{1,n,(k_1,\lambda_{1,n})}, \ldots, \bar{U}_{n,n,(k_1,\lambda_{1,n})}$ be given by (2.54) and define the estimate $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$ via (2.56) and (2.57). Then

$$\int_{\mathbb{R}^d} \left| \sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(x) - \sigma^2(x) \right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**.

Remark 3.5. As stated in Remark 3.2, we can conclude that Theorem 3.3 is still valid in case that no censoring arises. In addition, Remark 3.1 and Remark 3.4 (with m replaced by σ^2, $m_{n,(k,\lambda_n)}$ replaced by $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$, and L replaced by L^2) also hold for Theorem 3.3.

Proof of Theorem 3.3. First observe that **(RA2)** implies that with probability one $m(X) \in [0, L]$ and $\sigma^2(X) \in [0, L^2]$.

Define $U^{(2)}$, $U_i^{(2)}$, and $\bar{U}_{i,n,(k_1,\lambda_{1,n})}$ $(i = 1, \ldots, n)$ by (2.49), (2.50), and (2.54). Moreover, let $\bar{\mathcal{D}}_n^{(2)}$ be given by (2.55) . From (2.52) and (2.53), one gets

$$\sigma^2(X) = \mathbf{E}\left[U^{(2)} - m(X)^2 \,\middle|\, X \right]$$

and

$$|U^{(2)} - m(X)^2| \le |U^{(2)}| + m(X)^2 \le L_2^\star + L^2 =: \bar{L}_2 < \infty \quad \text{a.s.},$$

where

$$L_2^\star = \frac{(1 + 2\,|\alpha_2|)\, L^2}{G(L)} \in (0, \infty).$$

Now note that (X,Y,C), (X_1,Y_1,C_1), $\ldots$, (X_n,Y_n,C_n) are i.i.d. random vectors and this together with (2.49), (2.54), and (2.55) yields that $(X, U^{(2)} - m(X)^2)$ and $\bar{\mathcal{D}}_n^{(2)}$ are independent.

If we set $\beta^\star = \bar{L}_2$, $\beta = L^2$, $Y^\star = U^{(2)} - m(X)^2$, $Y_i^\star = U_i^{(2)} - m(X_i)^2$, $\bar{Y}_i^{(n)} = \bar{U}_{i,n,(k_1,\lambda_{1,n})}$ $(i = 1,\ldots,n)$, $k = k_2$, and $\lambda_n = \lambda_{2,n}$ in Section 3.1, then one can conclude similar to the proof of Theorem 3.2 that $\bar{\mathcal{D}}_n$ equals $\bar{\mathcal{D}}_n^{(2)}$, $m^\star$ equals σ^2, $m^\star_{n,(k_2,\lambda_{2,n})}$ equals $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$, and that

$$\frac{1}{n}\sum_{i=1}^{n}\left| Y_i^\star - \bar{Y}_i^{(n)} \right|^2 = \frac{1}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2. \tag{3.21}$$

Therefore, Theorem 3.1 implies that in order to prove Theorem 3.3, it suffices to show

$$\frac{1}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 \to 0 \quad (n\to\infty) \quad \text{a.s.}$$

Let $m_{n,(k_1,\lambda_{1,n})}$ and $\hat{U}_i^{(2)}$ $(i = 1,\ldots,n)$ be given by (2.28) and (2.51), respectively. Since $m_{n,(k_1,\lambda_{1,n})}(x) \in [0,L]$ and $m(x) \in [0,L]$ $(x \in [0,1]^d)$, we have from (2.54) with probability one

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - m(X_i)^2 - \left(\hat{U}_i^{(2)} - m_{n,(k_1,\lambda_{1,n})}(X_i)^2\right) \right|^2 \\
&\le \frac{2}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 + \frac{2}{n}\sum_{i=1}^{n}\left| m_{n,(k_1,\lambda_{1,n})}(X_i)^2 - m(X_i)^2 \right|^2 \\
&= \frac{2}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\left(\left| m_{n,(k_1,\lambda_{1,n})}(X_i) + m(X_i) \right|^2 \cdot \left| m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i) \right|^2 \right) \\
&\le \frac{2}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 + \frac{2}{n}(2L)^2\sum_{i=1}^{n}\left| m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i) \right|^2.
\end{aligned} \tag{3.22}$$

Here, we used the facts that $(a+b)^2 \le 2a^2 + 2b^2$ and $(a^2 - b^2) = (a+b)\cdot(a-b)$ $\forall a,b \in \mathbb{R}$. From Lemma 3.1 and **(RA2)** – **(RA4)**, one can conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 \le \max_{i=1,\ldots,n}\left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 \to 0 \quad (n\to\infty) \quad \text{a.s.} \tag{3.23}$$

This together with (3.22) implies that it suffices to show

$$\frac{1}{n}\sum_{i=1}^{n} \left| m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i) \right|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.} \tag{3.24}$$

From Theorem 3.2, we know that under the assumptions of Theorem 3.3, the estimate $m_{n,(k_1,\lambda_{1,n})}$ of the regression function m is strongly consistent with respect to the $\mathcal{L}_2$ error, i.e.,

$$\int_{\mathbb{R}^d} \left| m_{n,(k_1,\lambda_{1,n})}(x) - m(x) \right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad \text{a.s.},$$

for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)**.

So it remains to prove that with probability one

$$\frac{1}{n}\sum_{i=1}^{n} \left| m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i) \right|^2 - \int_{\mathbb{R}^d} \left| m_{n,(k_1,\lambda_{1,n})}(x) - m(x) \right|^2 \mu(dx) \to 0 \quad (n \to \infty). \tag{3.25}$$

Let $U_i^{(1)}$ and $\hat{U}_i^{(1)}$ $(i = 1, \ldots, n)$ be defined by (2.22) and (2.25), respectively. First observe that the relation $(a+b)^2 \le 2a^2 + 2b^2$ $(a, b \in \mathbb{R})$, (2.24), and (2.27) yield

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \left| \tilde{m}_{n,(k_1,\lambda_{1,n})}(X_i) - \hat{U}_i^{(1)} \right|^2 + \lambda_{1,n} J_{k_1}^2(\tilde{m}_{n,(k_1,\lambda_{1,n})}) \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} \left| 0 - \hat{U}_i^{(1)} \right|^2 + \lambda_{1,n} J_{k_1}^2(0) = \frac{1}{n}\sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} - U_i^{(1)} \right|^2 \\
&\quad \le \frac{2}{n}\sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 + \frac{2}{n}\sum_{i=1}^{n} \left| U_i^{(1)} \right|^2 \\
&\quad \le \frac{2}{n}\sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 + 2\,(L_1^\star)^2\,,
\end{aligned} \tag{3.26}$$

where $L_1^\star = (1 + 2|\alpha_1|)\frac{L}{G(L)} \in (0, \infty)$ (cf. (2.24)). Similar to (3.23), one can conclude from Lemma 3.1 and **(RA2)** – **(RA4)** that

$$\frac{1}{n}\sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

This together with (2.27) and (3.26) implies that with probability one, we have for n sufficiently large

$$\tilde{m}_{n,(k_1,\lambda_{1,n})} \in \tilde{\mathcal{F}}_{3(L_1^\star)^2/\lambda_{1,n}} := \left\{ \tilde{f} : \tilde{f} \in W_{k_1}([0,1]^d), J_{k_1}^2(\tilde{f}) \le \frac{3(L_1^\star)^2}{\lambda_{1,n}} \right\}. \tag{3.27}$$

From (2.28) and (3.27), it follows that in order to prove (3.25), it suffices to show

$$\sup_{g\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbf{E}g(X)\right|\to 0\quad(n\to\infty)\quad\text{a.s.},\tag{3.28}$$

where

$$\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}:=\left\{g:g(x)=|T_{[0,L]}\tilde{f}(x)-m(x)|^2,\tilde{f}\in\tilde{\mathcal{F}}_{3(L_1^\star)^2/\lambda_{1,n}},x\in[0,1]^d\right\}.$$

For this purpose, Lemma C.1 will now be applied. Observe that for any $g\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}$, $0\le m(x)\le L$ implies $0\le g(x)\le L^2$ $(x\in[0,1]^d)$. This together with Lemma C.1 yields for all $\epsilon>0$

$$\begin{aligned}&\mathbf{P}\left[\sup_{g\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbf{E}g(X)\right|>\epsilon\right]\\&\quad\le\quad 8\exp\left(-\frac{n\epsilon^2}{128\,L^4}\right)\cdot\mathbf{E}\mathcal{N}_1\left(\frac{\epsilon}{8},\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}},X_1^n\right).\end{aligned}\tag{3.29}$$

Here, $X_1^n:=(X_1,\ldots,X_n)$ and $\mathcal{N}_1(\cdot,\cdot,\cdot)$ denotes the $\mathcal{L}_1$–covering number (vide Definiton C.1).

In order to bound the $\mathcal{L}_1$–covering number on the right hand side of (3.29) from above, first note that for all $a,b\in\mathbb{R}$ and all $B>0$, one gets from (2.29)

$$\left|T_{[0,B]}a-T_{[0,B]}b\right|\le\left|T_{[-B,B]}a-T_{[-B,B]}b\right|.\tag{3.30}$$

Now, let $g_1,g_2\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}$ be two arbitrary functions with $g_j(x)=|T_{[0,L]}\tilde{f}_j(x)-m(x)|^2$ $(x\in[0,1]^d)$, where $\tilde{f}_j\in\tilde{\mathcal{F}}_{3(L_1^\star)^2/\lambda_{3,n}}$ $(j\in\{1,2\})$. Set $f_1:=T_{[0,L]}\tilde{f}_1$ and $f_2:=T_{[0,L]}\tilde{f}_2$. From $0\le m(X)\le L$ a.s. (cf. **(RA2)**), one can conclude

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|g_1(X_i)-g_2(X_i)|&=\frac{1}{n}\sum_{i=1}^{n}\left|(f_1(X_i)-m(X_i))^2-(f_2(X_i)-m(X_i))^2\right|\\&=\frac{1}{n}\sum_{i=1}^{n}\left|f_1^2(X_i)-2f_1(X_i)\,m(X_i)-f_2^2(X_i)+2f_2(X_i)\,m(X_i)\right|\\&=\frac{1}{n}\sum_{i=1}^{n}|(f_1(X_i)+f_2(X_i)-2m(X_i))\cdot(f_1(X_i)-f_2(X_i))|\\&\le\frac{1}{n}\sum_{i=1}^{n}(|f_1(X_i)+f_2(X_i)-2m(X_i)|\cdot|f_1(X_i)-f_2(X_i)|)\\&\le 2L\cdot\frac{1}{n}\sum_{i=1}^{n}|f_1(X_i)-f_2(X_i)|\quad\text{a.s.}\end{aligned}\tag{3.31}$$

This together with (3.30) implies with probability one

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|g_1(X_i)-g_2(X_i)| &\leq 2L\cdot\frac{1}{n}\sum_{i=1}^{n}\left|T_{[0,L]}\tilde{f}_1(X_i)-T_{[0,L]}\tilde{f}_2(X_i)\right| \\ &\leq 2L\cdot\frac{1}{n}\sum_{i=1}^{n}\left|T_{[-L,L]}\tilde{f}_1(X_i)-T_{[-L,L]}\tilde{f}_2(X_i)\right|. \qquad (3.32)\end{aligned}$$

Define

$$\mathcal{F}_{3(L_1^\star)^2/\lambda_{1,n}} := \left\{T_{[-L,L]}\tilde{f} : \tilde{f}\in W_{k_1}([0,1]^d), J_{k_1}^2(\tilde{f})\leq\frac{3(L_1^\star)^2}{\lambda_{1,n}}\right\}.$$

For all $\epsilon>0$, (3.32) yields

$$\mathcal{N}_1\left(\frac{\epsilon}{8},\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}},X_1^n\right)\leq\mathcal{N}_1\left(\frac{\epsilon}{16L},\mathcal{F}_{3(L_1^\star)^2/\lambda_{1,n}},X_1^n\right)\quad\text{a.s.}$$

The last inequality, (3.20), (3.29), and Lemma C.2 imply that for all $0<\epsilon<16L^2$ and n sufficiently large, it holds

$$\begin{aligned}&\mathbf{P}\left[\sup_{g\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbf{E}g(X)\right|>\epsilon\right] \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{128\,L^4}\right)\cdot\mathbf{E}\mathcal{N}_1\left(\frac{\epsilon}{16L},\mathcal{F}_{3(L_1^\star)^2/\lambda_{1,n}},X_1^n\right) \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{128\,L^4}+\left[B_1\left(\sqrt{\frac{3(L_1^\star)^2}{\lambda_{1,n}}}\cdot\frac{16L}{\epsilon}\right)^{\frac{d}{k_1}}+B_2\right]\ln\left(B_3\frac{16L^2n}{\epsilon}\right)\right) \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{128\,L^4}+2B_1\left(\frac{16\sqrt{3}\,L\cdot L_1^\star}{\epsilon}\right)^{\frac{d}{k_1}}\lambda_{1,n}^{-\frac{d}{2k_1}}\ln n+2B_2\ln n\right) \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{256L^4}\right), \qquad (3.33)\end{aligned}$$

where $B_1,B_2,B_3>0$ are constants which only depend on k_1 and d. From this, we deduce that

$$\sum_{n=1}^{\infty}\mathbf{P}\left[\sup_{g\in\mathcal{G}_{3(L_1^\star)^2/\lambda_{1,n}}}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbf{E}g(X)\right|>\epsilon\right]<\infty \qquad (3.34)$$

for all $\epsilon>0$. Finally, (3.28) follows from (3.34) and an application of the Borel-Cantelli lemma (Lemma D.2).

□

3.5 Consistent MSSE of the conditional survival function

The following theorem shows that our estimates of the conditional survival function of the lifetimes, evaluated at fixed point $\tau \in \mathbb{R}$, are strongly consistent for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)**.

Theorem 3.4. (Consistency) *Let $k, d \in \mathbb{N}$ with $2k > d$ and let $\tau \in \mathbb{R}$ be arbitrary, but fixed. For $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n \to \infty) \qquad \text{and} \qquad \frac{n\lambda_n^{\frac{d}{2k}}}{\ln n} \to \infty \quad (n \to \infty). \tag{3.35}$$

Let $\hat{U}_1^{(3)}, \ldots, \hat{U}_n^{(3)}$ be given by (2.79) and define the estimate $F_{n,(k,\lambda_n)}(\tau\,|\,\cdot)$ via (2.82) and (2.83). Then

$$\int_{\mathbb{R}^d} \left|F_{n,(k,\lambda_n)}(\tau\,|\,x) - F(\tau\,|\,x)\right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**.

Remark 3.6. As stated in Remark 3.2, we can conclude that Theorem 3.4 is still valid if no censoring arises. In this case, one may drop condition **(RA2)** and allow that Y is unbounded if we assume that $\mathbf{E}Y^2 < \infty$. However, for censored data, we require that $G(\tau_F) > 0$ (cf. Theorem 1.1), i.e., that C exceeds with non-zero probability the upper endpoint τ_F of the distribution of Y.

Moreover, note that Remark 3.1 and Remark 3.4 (with m replaced by $F(\tau\,|\,\cdot)$, $m_{n,(k,\lambda_n)}$ replaced by $F_{n,(k,\lambda_n)}(\tau\,|\,\cdot)$, and L replaced by 1) hold for Theorem 3.4, too.

Proof of Theorem 3.4. Fix $\tau \in \mathbb{R}$. Obviously, it holds that

$$F\,(\,\tau\,|\,X\,) = \mathbf{P}\,[\,Y > \tau\,|\,X\,] \in [0,1] \quad \text{a.s.}$$

Let $U^{(3)}$ and $U_i^{(3)}$ $(i = 1, \ldots, n)$ be defined by (2.77) and (2.78). From (2.80) and (2.81), we have

$$F\,(\,\tau\,|\,X\,) = \mathbf{E}\left[\,U^{(3)}\,\middle|\,X\,\right]$$

and

$$|U^{(3)}| \le L_3^\star < \infty \quad \text{a.s.},$$

where $L_3^\star = \frac{1}{G(L)} \in [1, \infty)$ (cf. (3.15)).

Furthermore, since (X, Y, C), (X_1, Y_1, C_1), $\dots$, (X_n, Y_n, C_n) are i.i.d. random vectors, we deduce that the sequence $(X_1, \hat{U}_1^{(3)})$, $\dots$, $(X_1, \hat{U}_n^{(3)})$ is independent of $(X, U^{(3)})$.

Similar to the proof of Theorem 3.2, one can conclude that if we set $\beta^\star = L_3^\star$, $\beta = 1$, $Y^\star = U^{(3)}$, $Y_i^\star = U_i^{(3)}$, and $\bar{Y}_i^{(n)} = \hat{U}_i^{(3)}$ $(i = 1, \dots, n)$ in Section 3.1, then $m^\star(\cdot)$ equals $F(\tau\,|\,\cdot)$ and $m^\star_{n,(k,\lambda_n)}(\cdot)$ equals $F_{n,(k,\lambda_n)}(\tau\,|\,\cdot)$. Furthermore, it holds that

$$\frac{1}{n}\sum_{i=1}^{n} |\, Y_i^\star - \bar{Y}_i^{(n)} \,|^2 = \frac{1}{n}\sum_{i=1}^{n} |\, U_i^{(3)} - \hat{U}_i^{(3)} \,|^2 \le \max_{i=1,\dots,n} |\, U_i^{(3)} - \hat{U}_i^{(3)} \,|^2.$$

This together with Theorem 3.1 and Lemma 3.1 implies the assertion of Theorem 3.4.

□

3.6 Proof of Theorem 3.1

In the proof of Theorem 3.1, we will apply the following lemma which investigates the difference between the $\mathcal{L}_2$ risk and the empirical $\mathcal{L}_2$ risk of a MSSE, which is truncated at $[-\beta^\star, \beta^\star]$. Here, it is shown that if the conditions (3.8) and (3.9) hold, this difference converges almost surely to zero with n tending to infinity.

Lemma 3.2. *Let $k, d \in \mathbb{N}$ with $2k > d$ and $\beta^\star \in (0, \infty)$. Define $\bar{\mathcal{D}}_n$ by (3.4). For $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n \to \infty) \qquad \text{and} \qquad \frac{n\lambda_n^{\frac{d}{2k}}}{\ln n} \to \infty \quad (n \to \infty). \tag{3.36}$$

Let the estimate $\tilde{m}^\star_{n,(k,\lambda_n)}$ be given by (3.6) and set

$$\hat{m}^\star_{n,(k,\lambda_n)}(\cdot) := T_{[-\beta^\star,\beta^\star]}\tilde{m}^\star_{n,(k,\lambda_n)}(\cdot). \tag{3.37}$$

If

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i^\star - \bar{Y}_i^{(n)}|^2 \to 0 \quad (n \to \infty) \quad a.s. \tag{3.38}$$

then

$$\mathbf{E}\left[\, |\hat{m}^\star_{n,(k,\lambda_n)}(X) - Y^\star|^2 \,\middle|\, \bar{\mathcal{D}}_n \,\right] - \frac{1}{n}\sum_{i=1}^{n} |\hat{m}^\star_{n,(k,\lambda_n)}(X_i) - Y_i^\star|^2 \to 0 \quad (n \to \infty) \quad a.s.$$

for every distribution of $(X, Y^\star)$ with $(X, Y^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s.

PROOF OF LEMMA 3.2. We mimic the arguments which we used in order to show (3.25) in the proof of Theorem 3.3.

In analogy to (3.26), one can conclude from Definition (3.6) of the estimate $\tilde{m}^{\star}_{n,(k,\lambda_n)}$, $|Y_i^{\star}| \leq \beta^{\star}$ a.s. $(i = 1, \ldots, n)$, and (3.38)

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}|^2 + \lambda_n J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)}) \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n} |0 - \bar{Y}_i^{(n)}|^2 + \lambda_n J_k^2(0) = \frac{1}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)} - Y_i^{\star}|^2 \\
&\quad \leq \frac{2}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 + \frac{2}{n}\sum_{i=1}^{n} |Y_i^{\star}|^2 \\
&\quad \leq \frac{2}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 + 2(\beta^{\star})^2 \to 2(\beta^{\star})^2 \quad (n \to \infty) \quad \text{a.s.,}
\end{aligned} \tag{3.39}$$

where the second inequality in (3.39) follows from $(a+b)^2 \leq 2a^2 + 2b^2$ $(a, b \in \mathbb{R})$.

This implies that with probability one, for sufficiently large n,

$$\hat{m}^{\star}_{n,(k,\lambda_n)} \in \mathcal{F}_{3(\beta^{\star})^2/\lambda_n} := \left\{ T_{[-\beta^{\star},\beta^{\star}]} f : f \in W_k([0,1]^d), J_k^2(f) \leq \frac{3(\beta^{\star})^2}{\lambda_n} \right\}. \tag{3.40}$$

Similar to (3.28), we deduce from (3.40) that it suffices to show

$$\sup_{g \in \mathcal{G}_{3(\beta^{\star})^2/\lambda_n}} \left| \mathbf{E}\left[g(X, Y^{\star})\right] - \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i^{\star}) \right| \to 0 \quad (n \to \infty) \quad \text{a.s.,} \tag{3.41}$$

where

$$\mathcal{G}_{3(\beta^{\star})^2/\lambda_n} := \left\{ g : g(x,y) = |f(x) - y|^2, f \in \mathcal{F}_{3(\beta^{\star})^2/\lambda_n}, x \in [0,1]^d, y \in [-\beta^{\star}, \beta^{\star}] \right\}.$$

As in the proof of (3.28) (vide (3.29)), we may use Lemma C.1 to bound the probability that the right hand side of (3.41) exceeds some arbitrary $\epsilon > 0$:

$$\begin{aligned}
&\mathbf{P}\left[\sup_{g \in \mathcal{G}_{3(\beta^{\star})^2/\lambda_n}} \left| \mathbf{E}\left[g(X, Y^{\star})\right] - \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i^{\star}) \right| > \epsilon \right] \\
&\quad \leq 8 \exp\left(-\frac{n\epsilon^2}{128\left(4(\beta^{\star})^2\right)^2} \right) \cdot \mathbf{E}\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}_{3(\beta^{\star})^2/\lambda_n}, (X, Y^{\star})_1^n \right).
\end{aligned} \tag{3.42}$$

Here, $(X, Y^{\star})_1^n := ((X_1, Y_1^{\star}), \ldots, (X_n, Y_n^{\star}))$.

Now, let $g_1, g_2 \in \mathcal{G}_{3(\beta^{\star})^2/\lambda_n}$ be two arbitrary functions with $g_j(x,y) = |f_j(x) - y|^2$ $((x,y) \in [0,1]^d \times [-\beta^{\star}, \beta^{\star}])$, where $f_j \in \mathcal{F}_{3(\beta^{\star})^2/\lambda_n}$ $(j \in \{1,2\})$. Then one can conclude in

analogy to (3.31)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|g_1(X_i,Y_i^\star)-g_2(X_i,Y_i^\star)| &= \frac{1}{n}\sum_{i=1}^{n}|(f_1(X_i)+f_2(X_i)-2Y_i^\star)\cdot(f_1(X_i)-f_2(X_i))| \\ &\leq 4\beta^\star\cdot\frac{1}{n}\sum_{i=1}^{n}|f_1(X_i)-f_2(X_i)| \quad \text{a.s.}\end{aligned}$$

This implies that with probability one, the $\mathcal{L}_1$–covering number on the right hand side of (3.42) is for all $\epsilon>0$ bounded from above by

$$\mathcal{N}_1\left(\frac{\epsilon}{32\beta^\star},\mathcal{F}_{3(\beta^\star)^2/\lambda_n},X_1^n\right), \tag{3.43}$$

where $X_1^n := (X_1,\ldots,X_n)$.

Similar to (3.33), Lemma C.2, (3.36), (3.42), and (3.43) yield for all $0<\epsilon<32\,(\beta^\star)^2$ and n sufficiently large

$$\begin{aligned}&\mathbf{P}\left[\sup_{g\in\mathcal{G}_{3(\beta^\star)^2/\lambda_n}}\left|\mathbf{E}\left[g(X,Y^\star)\right]-\frac{1}{n}\sum_{i=1}^{n}g(X_i,Y_i^\star)\right|>\epsilon\right] \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{2048(\beta^\star)^4}+\left[B_1\left(\sqrt{\frac{3(\beta^\star)^2}{\lambda_n}}\cdot\frac{32\beta^\star}{\epsilon}\right)^{\frac{d}{k}}+B_2\right]\ln\left(B_3\frac{32\,(\beta^\star)^2\,n}{\epsilon}\right)\right) \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{2048(\beta^\star)^4}+2B_1\left(\frac{32\sqrt{3}(\beta^\star)^2}{\epsilon}\right)^{\frac{d}{k}}\lambda_n^{-\frac{d}{2k}}\ln n+2B_2\ln n\right) \\ &\leq 8\exp\left(-\frac{n\epsilon^2}{4096(\beta^\star)^4}\right).\end{aligned}$$

Here, $B_1,B_2,B_3>0$ are constants which only depend on k and d. From this, the assertion of Lemma 3.2 follows by an application of Lemma D.2.

□

Now we are in the position to prove Theorem 3.1.

Proof of Theorem 3.1. Let $\epsilon>0$ be arbitrarily chosen. From Lemma D.1 in Appendix D, one can conclude that there exists a function $g_\epsilon\in W_k([0,1]^d)$ such that (cf. Kohler and Krzyżak (2001) , Pintér (2001), and Corollary A.1 in Györfi, Kohler, Krzyżak, and Walk (2002))

$$\int_{\mathbb{R}^d}|g_\epsilon(x)-m^\star(x)|^2\mu(dx)\leq\epsilon \quad\text{and}\quad J_k^2(g_\epsilon)<\infty. \tag{3.44}$$

Define $\tilde{m}^{\star}_{n,(k,\lambda_n)}$, $m^{\star}_{n,(k,\lambda_n)}$, and $\hat{m}^{\star}_{n,(k,\lambda_n)}$ by (3.6), (3.7), and (3.37), respectively. Note that $0 \leq m^{\star}(x) \leq \beta \leq \beta^{\star}$ implies $|m^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x)|^2 \leq |\hat{m}^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x)|^2$ ($x \in [0,1]^d$). Therefore, it suffices to show

$$\int_{\mathbb{R}^d} \left|\hat{m}^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x)\right|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad \text{a.s.} \tag{3.45}$$

Let $\bar{\mathcal{D}}_n$ be given by (3.4). In order to prove (3.45), we first decompose the $\mathcal{L}_2$ error of $\hat{m}^{\star}_{n,(k,\lambda_n)}$ in the following way

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left|\hat{m}^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x)\right|^2 \mu(dx) \\
&= \mathbf{E}\left[\, |\hat{m}^{\star}_{n,(k,\lambda_n)}(X) - Y^{\star}|^2 \,\middle|\, \bar{\mathcal{D}}_n \right] - \mathbf{E}\left[\, |m^{\star}(X) - Y^{\star}|^2 \,\right] \\
&= \mathbf{E}\left[\, |\hat{m}^{\star}_{n,(k,\lambda_n)}(X) - Y^{\star}|^2 \,\middle|\, \bar{\mathcal{D}}_n \right] - \frac{1}{n}\sum_{i=1}^{n} |\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y^{\star}_i|^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} |\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y^{\star}_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y^{\star}_i|^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y^{\star}_i|^2 - (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}^{(n)}_i|^2 \\
&\quad + (1+\epsilon)\left[\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}^{(n)}_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |g_{\epsilon}(X_i) - \bar{Y}^{(n)}_i|^2\right] \\
&\quad + (1+\epsilon)\left[\frac{1}{n}\sum_{i=1}^{n} |g_{\epsilon}(X_i) - \bar{Y}^{(n)}_i|^2 - (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} |g_{\epsilon}(X_i) - Y^{\star}_i|^2\right] \\
&\quad + (1+\epsilon)^2\left[\frac{1}{n}\sum_{i=1}^{n} |g_{\epsilon}(X_i) - Y^{\star}_i|^2 - \mathbf{E}\left[\, |g_{\epsilon}(X) - Y^{\star}|^2 \,\right]\right] \\
&\quad + (1+\epsilon)^2\,\mathbf{E}\left[\, |g_{\epsilon}(X) - Y^{\star}|^2 \,\right] - (1+\epsilon)^2\,\mathbf{E}\left[\, |m^{\star}(X) - Y^{\star}|^2 \,\right] \\
&\quad + ((1+\epsilon)^2 - 1)\,\mathbf{E}\left[\, |m^{\star}(X) - Y^{\star}|^2 \,\right] \\
&=: \sum_{j=1}^{8} H_{j,n}.
\end{aligned} \tag{3.46}$$

Here, we used that assumption (3.2) yields (3.5) and this, in turn, the first equality in (3.46) in analogy to (1.11) – (1.13) .

Below, it is shown how each of the eight terms on the right hand side of (3.46) is bounded from above when the sample size n increases. By an application of Lemma 3.2,

we get

$$H_{1,n} = \mathbf{E}\left[\, |\hat{m}^{\star}_{n,(k,\lambda_n)}(X) - Y^{\star}|^2 \middle| \bar{\mathcal{D}}_n \,\right] - \frac{1}{n}\sum_{i=1}^{n} |\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y_i^{\star}|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

Definition (3.37) of the truncated estimate $\hat{m}^{\star}_{n,(k,\lambda_n)}$ and $|Y^{\star}| \le \beta^{\star}$ a.s. imply

$$H_{2,n} = \frac{1}{n}\sum_{i=1}^{n} |\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y_i^{\star}|^2 - \frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y_i^{\star}|^2 \le 0$$

In order to bound the third term

$$H_{3,n} = \frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - Y_i^{\star}|^2 - (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}|^2$$

from above, observe that for all $a, b \in \mathbb{R}$, we have

$$(a+b)^2 \le a^2(1+\epsilon) + b^2\left(1 + \frac{1}{\epsilon}\right). \tag{3.47}$$

From (3.9) and (3.47), one can conclude

$$H_{3,n} \le \left(1 + \frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

From Definition (3.6) of $\tilde{m}^{\star}_{n,(k,\lambda_n)}$, (3.44), and $\lambda_n \to 0$ $(n \to \infty)$ it follows that

$$\begin{aligned} H_{4,n} &= (1+\epsilon)\left[\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}|^2 - \frac{1}{n}\sum_{i=1}^{n} |g_\epsilon(X_i) - \bar{Y}_i^{(n)}|^2\right] \\ &\le (1+\epsilon)\left[\frac{1}{n}\sum_{i=1}^{n} \left|g_\epsilon(X_i) - \bar{Y}_i^{(n)}\right|^2 + \lambda_n J_k^2(g_\epsilon) - \frac{1}{n}\sum_{i=1}^{n} \left|g_\epsilon(X_i) - \bar{Y}_i^{(n)}\right|^2\right] \\ &= (1+\epsilon)\,\lambda_n J_k^2(g_\epsilon) \to 0 \quad (n \to \infty). \end{aligned}$$

Using again (3.9) and (3.47), we have

$$\begin{aligned} H_{5,n} &= (1+\epsilon)\left[\frac{1}{n}\sum_{i=1}^{n} |g_\epsilon(X_i) - \bar{Y}_i^{(n)}|^2 - (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} |g_\epsilon(X_i) - Y_i^{\star}|^2\right] \\ &\le (1+\epsilon)\left(1 + \frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 \to 0 \quad (n \to \infty) \quad \text{a.s.}, \end{aligned}$$

and – since $(X, Y^{\star}), (X_1, Y_1^{\star}), \ldots, (X_n, Y_n^{\star}) \in [0,1]^d \times [-\beta^{\star}, \beta^{\star}]$ a.s. are i.i.d. random vectors – by the strong law of large numbers

$$H_{6,n} = (1+\epsilon)^2\left[\frac{1}{n}\sum_{i=1}^{n} |g_\epsilon(X_i) - Y_i^{\star}|^2 - \mathbf{E}\left[\, |g_\epsilon(X) - Y^{\star}|^2 \,\right]\right] \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

Another application of (3.44) yields

$$\begin{aligned} H_{7,n} &= (1+\epsilon)^2\,\mathbf{E}\left[\,|g_\epsilon(X)-Y^\star|^2\,\right] - (1+\epsilon)^2\,\mathbf{E}\left[\,|m^\star(X)-Y^\star|^2\,\right] \\ &= (1+\epsilon)^2\,\mathbf{E}\left[\,|g_\epsilon(X)-m^\star(X)+m^\star(X)-Y^\star|^2\,\right] - (1+\epsilon)^2\,\mathbf{E}\left[\,|m^\star(X)-Y^\star|^2\,\right] \\ &= (1+\epsilon)^2\,\mathbf{E}\left[\,|g_\epsilon(X)-m^\star(X)|^2\,\right] \\ &\quad +2\,(1+\epsilon)^2\,\mathbf{E}\left[\,(g_\epsilon(X)-m^\star(X))\cdot(m^\star(X)-Y^\star)\,\right] \\ &= (1+\epsilon)^2\,\mathbf{E}\left[\,|g_\epsilon(X)-m^\star(X)|^2\,\right] \\ &\le \epsilon\,(1+\epsilon)^2, \end{aligned} \tag{3.48}$$

where the last equality in (3.48) follows from a conversion similar to (1.13):

$$\begin{aligned} \mathbf{E}\left[\,(g_\epsilon(X)-m^\star(X))\cdot(m^\star(X)-Y^\star)\,\right] &= \mathbf{E}\left[\,(g_\epsilon(X)-m^\star(X))\cdot\mathbf{E}\left[\,(m^\star(X)-Y^\star)\mid X\,\right]\right] \\ &= \mathbf{E}\left[\,(g_\epsilon(X)-m^\star(X))\cdot(m^\star(X)-\mathbf{E}\left[\,Y^\star\mid X\,\right])\right] \\ &= 0. \end{aligned}$$

Finally, we get with $|Y^\star|\le\beta^\star$ a.s. and $|m^\star(X)|\le\beta\le\beta^\star$ a.s. for the last of the eight terms

$$H_{8,n} = ((1+\epsilon)^2-1)\,\mathbf{E}\left[\,|m^\star(X)-Y^\star|^2\,\right] \le ((1+\epsilon)^2-1)\,(2\beta^\star)^2.$$

Combining all the results from above, one can conclude

$$\limsup_{n\to\infty}\int_{\mathbb{R}^d}\left|\hat{m}^\star_{n,(k,\lambda_n)}(x)-m^\star(x)\right|^2\mu(dx) \le \epsilon\,(1+\epsilon)^2 + 4\,((1+\epsilon)^2-1)\,(\beta^\star)^2 \quad \text{a.s.}$$

With $\epsilon\to 0$, (3.45) follows.

□

Chapter 4

Rate of convergence

In this chapter, the rates of convergence of our MSSE for censored regression, (2.28), (2.57), and (2.83), are analyzed. Similar to Chapter 3, the first section present results for the estimate (3.7). In order to derive rates of convergence for (2.28), (2.57), and (2.83), we will then use the fact that their definitions are covered by the definition of (3.7). From Theorem 4.1, Corollary 4.1, and Lemma 4.1, one can conclude that the main task in this step is to control the squared transformation errors. A rate of convergence of these errors is given in Section 4.2. Sections 4.3 – 4.5 then present our results for the MSSE (2.28), (2.57), and (2.83) while Section 4.6 contains the proofs of Theorem 4.1 and Lemma 4.1.

4.1 General results

Below, we are dealing with the MSSE $m^{\star}_{n,(k,\lambda_n)}$ of Section 3.1 (cf. (3.7)). As mentioned there, this estimate is a generalization of our MSSE (2.28), (2.57), and (2.83).

In the following, we will recall the definition of $m^{\star}_{n,(k,\lambda_n)}$. Let $\beta^{\star} \in (0,\infty)$ and let $(X, Y^{\star})$, $(X_1, Y_1^{\star})$, $\ldots$, $(X_n, Y_n^{\star}) \in [0,1]^d \times [-\beta^{\star}, \beta^{\star}]$ a.s. be the i.i.d. random vectors of Section 3.1. Furthermore, let $\beta \in (0, \beta^{\star}]$ such that

$$m^{\star}(X) = \mathbf{E}\left[\, Y^{\star} \mid X \,\right] \in [0, \beta] \quad \text{a.s.}$$

Since we assumed that $Y_1^{\star}, \ldots, Y_n^{\star}$ are unknown, the estimate $m^{\star}_{n,(k,\lambda_n)}$ in (3.7) is based on some observable real-valued random variables $\bar{Y}_1^{(n)}, \ldots, \bar{Y}_n^{(n)}$, where each $\bar{Y}_i^{(n)}$ may depend on the whole sample and a finite number of further, suitably chosen parameters.

As mentioned in Section 3.1, we do not demand that $\bar{Y}_1^{(n)}, \ldots, \bar{Y}_n^{(n)}$ are independent or identically distributed. In the following, it is only required that the squared differences $|Y_1^\star - \bar{Y}_1^{(n)}|^2$, $\ldots$, $|Y_n^\star - \bar{Y}_n^{(n)}|^2$ are "small" (cf. (3.9)). In particular, note that in this section we do not assume that condition (3.2) is fulfilled (however, condition (3.2) ensures that $m^\star$ is the function which minimizes the $\mathcal{L}_2$ risk with respect to $(X, Y^\star)$, cf. Section 1.3 and (3.5)).

Let $k \in \mathbb{N}$ with $2k > d$ and $\lambda_n > 0$. According to (3.7) the MSSE $m^\star_{n,(k,\lambda_n)}$ is defined via

$$m^\star_{n,(k,\lambda_n)}(\cdot) = T_{[0,\beta]} \tilde{m}^\star_{n,(k,\lambda_n)}(\cdot),$$

where $\tilde{m}^\star_{n,(k,\lambda_n)}$ is given by (3.6) as

$$\tilde{m}^\star_{n,(k,\lambda_n)}(\cdot) = \tilde{m}^\star_{n,(k,\lambda_n)}(\cdot, \bar{\mathcal{D}}_n) = \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^n |f(X_i) - \bar{Y}_i^{(n)}|^2 + \lambda_n J_k^2(f) \right).$$

Here, $W_k([0,1]^d)$, $J_k^2(\cdot)$, and $\bar{\mathcal{D}}_n$ are defined by (2.3), (2.5), and (3.4), respectively.

The next theorem investigates the rate of stochastic convergence of $m^\star_{n,(k,\lambda_n)}$ for smooth regression functions $m^\star \in W_p([0,1]^d)$ with $0 < J_p^2(m^\star) < \infty$ ($p \in \mathbb{N}$ with $2p > d$).

Theorem 4.1. (Rate of convergence) *Let $d, n \in \mathbb{N}$ and $1 \leq \beta \leq \beta^\star < \infty$. Let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Assume that we have chosen the parameters k and λ_n of the estimate $m^\star_{n,(k,\lambda_n)}$, which is given by (3.6) and (3.7), such that $k = p$ and $\lambda_n > 0$ with*

$$\left(\frac{n}{\ln n}\right)^{\frac{2p}{2p+d}} \lambda_n \to \infty \quad (n \to \infty). \tag{4.1}$$

If there exists a constant $b_2 > 0$ such that

$$\mathbf{P}\left[\max_{i=1,\ldots,n} |Y_i^\star - \bar{Y}_i^{(n)}|^2 > b_2 \right] \to 0 \quad (n \to \infty), \tag{4.2}$$

then it holds that

$$\int_{\mathbb{R}^d} \left| m^\star_{n,(k,\lambda_n)}(x) - m^\star(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n} \sum_{i=1}^n |Y_i^\star - \bar{Y}_i^{(n)}|^2 + \lambda_n + \left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right)$$

for every distribution of $(X, Y^\star)$ satisfying $(X, Y^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s., $m^\star(X) \in [0, \beta]$ a.s., and $m^\star \in W_p([0,1]^d)$ with $0 < J_p^2(m^\star) < \infty$.

The proof of Theorem 4.1 will be given in Section 4.6.

In case that λ_n is chosen such that λ_n is some positive constant times

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}},$$

Theorem 4.1 implies the following result on the rate of convergence of the MSSE $m^{\star}_{n,(k,\lambda_n)}$.

Corollary 4.1. *Let $d, n \in \mathbb{N}$ and $1 \leq \beta \leq \beta^{\star} < \infty$. Let $b_1 > 0$ be an arbitrary constant and let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Choose the parameters k and λ_n of the estimate $m^{\star}_{n,(k,\lambda_n)}$, which is given by (3.6) and (3.7), such that $k = p$ and*

$$\lambda_n = b_1 \cdot \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}. \tag{4.3}$$

If there exists a constant $b_2 > 0$ such that (4.2) holds, then we have

$$\int_{\mathbb{R}^d} \left|m^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x)\right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right)$$

for every distribution of $(X, Y^{\star})$ satisfying $(X, Y^{\star}) \in [0,1]^d \times [-\beta^{\star}, \beta^{\star}]$ a.s., $m^{\star}(X) \in [0, \beta]$ a.s., and $m^{\star} \in W_p([0,1]^d)$ with $0 < J_p^2(m^{\star}) < \infty$.

Proof of Corollary 4.1. Let λ_n be chosen according to (4.3) with some arbitrary constant $b_1 > 0$. For $p, d \in \mathbb{N}$ this yields

$$\left(\frac{n}{\ln n}\right)^{\frac{2p}{2p+d}} \lambda_n = b_1 \, (\ln n)^{\frac{2p}{2p+d}} \to \infty \quad (n \to \infty). \tag{4.4}$$

The assertion of Corollary 4.1 follows from Theorem 4.1 , (4.3), and (4.4) .

□

Since almost sure convergence implies stochastic convergence, one can conclude from Lemma 3.1 that (4.2) holds if we replace

$$\max_{i=1,\ldots,n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 \tag{4.5}$$

by the maximum squared transformation error

$$\max_{i=1,\ldots,n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \quad (j \in \{1,2,3\}).$$

Here, $U_i^{(1)}$, $U_i^{(2)}$, and $U_i^{(3)}$ $(i = 1, \ldots, n)$ are defined by (2.22), (2.50), and (2.78), respectively. Furthermore, $\hat{U}_i^{(1)}$, $\hat{U}_i^{(2)}$, and $\hat{U}_i^{(3)}$ $(i = 1, \ldots, n)$ are given by (2.25), (2.51), and (2.79).

Assume that $m^\star \in W_p([0,1]^d)$ and $0 < J_p^2(m^\star) < \infty$ for some arbitrary $p \in \mathbb{N}$ with $2p > d$. Corollary 4.1 indicates that the rates of convergence of suitably defined MSSE of $m(X)$ and $F(\tau|X)$ ($\tau \in \mathbb{R}$ fixed) depend on the rate of the mean squared transformation error

$$\frac{1}{n}\sum_{i=1}^{n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \quad (j \in \{1,3\}) \tag{4.6}$$

and the rate of these estimates in usual nonparametric regression, which is given by

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}. \tag{4.7}$$

For the estimation with (2.28) and (2.83), $m^\star(X)$ equals $m(X)$ and $F(\tau|X)$, respectively. While (4.7) is determined by the smoothness of $m^\star$, which is measured by p, the rate of (4.6) depends on the distribution of (Y, C). I.e., the latter one controls the asymptotic behavior of our estimates in the presence of censored data (cf. Theorem 3.1). In the next section, we show that under suitable conditions on the survival functions F and G, the rate of convergence of the mean squared transformation errors is given by $n^{-\gamma}$ for some $\gamma \in (0,1)$. This together with Corollary 4.1 implies that if we can choose γ such that $\gamma \geq \frac{2p}{2p+d}$, the rates of convergence of suitably defined MSSE of $m(X)$ and $F(\tau|X)$ ($\tau \in \mathbb{R}$ fixed) correspond to the rates of these estimates in usual nonparametric regression with random design and without imposing regularity conditions on the distribution of X to be found in literature (cf. Kohler, Krzyżak, and Schäfer (2002)). For the estimates (2.28) and (2.83), these two results are formulated and proven in Section 4.3 and Section 4.5, respectively.

Let $k_1 \in \mathbb{N}$ with $2k_1 > d$ and let $\lambda_{1,n} > 0$. Recall from (2.54) that the MSSE (2.57) of the conditional variance depends on the estimates $\bar{U}_{i,n,(k_1,\lambda_{1,n})} = \hat{U}_i^{(2)} - m_{n,(k_1,\lambda_{1,n})}(X_i)^2$ of the random variables $U_i^{(2)} - m(X_i)^2$ $(i = 1, \ldots, n)$. Here, $m_{n,(k_1,\lambda_{1,n})}$ is the MSSE of the regression function defined by (2.28). I.e., in order to apply Corollary 4.1 in the analysis of the rate of convergence of (2.57), we have to investigate the asymptotic behavior of

$$\frac{1}{n}\sum_{i=1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})}|^2.$$

Observe that $m_{n,(k_1,\lambda_{1,n})}(X) \in [0,L]$ a.s., $m(X) \in [0,L]$ a.s., $(a+b)^2 \leq 2a^2+2b^2$, and $a^2-b^2=(a+b)\cdot(a-b)$ $(a,b\in\mathbb{R})$ yield with probability one

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|U_i^{(2)}-m(X_i)^2-\bar{U}_{i,n,(k_1,\lambda_{1,n})}|^2\\
&= \frac{1}{n}\sum_{i=1}^{n}|U_i^{(2)}-\hat{U}_i^{(2)}+m_{n,(k_1,\lambda_{1,n})}(X_i)^2-m(X_i)^2|^2\\
&\leq \frac{2}{n}\sum_{i=1}^{n}|U_i^{(2)}-\hat{U}_i^{(2)}|^2+\frac{2}{n}\sum_{i=1}^{n}|m_{n,(k_1,\lambda_{1,n})}(X_i)^2-m(X_i)^2|^2\\
&= \frac{2}{n}\sum_{i=1}^{n}|U_i^{(2)}-\hat{U}_i^{(2)}|^2\\
&\quad+\frac{2}{n}\sum_{i=1}^{n}\left(|m_{n,(k_1,\lambda_{1,n})}(X_i)+m(X_i)|^2\cdot|m_{n,(k_1,\lambda_{1,n})}(X_i)-m(X_i)|^2\right)\\
&\leq \frac{2}{n}\sum_{i=1}^{n}|U_i^{(2)}-\hat{U}_i^{(2)}|^2+(2L)^2\cdot\frac{2}{n}\sum_{i=1}^{n}|m_{n,(k_1,\lambda_{1,n})}(X_i)-m(X_i)|^2. \qquad (4.8)
\end{aligned}$$

Hence, beside the rate of the mean squared transformation error

$$\frac{1}{n}\sum_{i=1}^{n}|U_i^{(2)}-\hat{U}_i^{(2)}|^2$$

and the rate implied by (4.7), the rate of convergence of the estimate (2.57) is determined by the asymptotic behavior of the empirical $\mathcal{L}_2$ error of $m_{n,(k_1,\lambda_{1,n})}$,

$$\frac{1}{n}\sum_{i=1}^{n}|m_{n,(k_1,\lambda_{1,n})}(X_i)-m(X_i)|^2.$$

Now assume that $m\in W_p([0,1]^d)$ and $0<J_p^2(m)<\infty$ for some arbitrary $p\in\mathbb{N}$ with $2p>d$. As mentioned above, we can conclude from the result in the following section that under suitable assumptions on F and G, $n^{-\frac{2p}{2p+d}}$ is the rate of convergence of the mean squared transformation error

$$\frac{1}{n}\sum_{i=1}^{n}|U_i^{(1)}-\hat{U}_i^{(1)}|^2.$$

In order to show that the MSSE (2.57) achieves the optimal rate of convergence up to some logarithmic factor, it therefore remains to prove

$$\frac{1}{n}\sum_{i=1}^{n}|m_{n,(k_1,\lambda_{1,n})}(X_i)-m(X_i)|^2=\mathcal{O}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^{n}|U_i^{(1)}-\hat{U}_i^{(1)}|^2+\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right).$$

Since $\lambda_n J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)})\geq 0$ and the definition of $m^{\star}_{n,(k,\lambda_n)}$ covers the definition of $m_{n,(k,\lambda_n)}$, this is implied by the next lemma if λ_n is chosen according to Corollary 4.1.

Lemma 4.1. *Let the conditions of Theorem 4.1 hold. If there exists a constant $b_2 > 0$ such that (4.2) is fulfilled, then one gets*

$$\frac{1}{n}\sum_{i=1}^{n} |m^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i)|^2 + \lambda_n J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)}) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i^{\star} - \bar{Y}_i^{(n)}|^2 + \lambda_n + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right).$$

for every distribution of $(X, Y^{\star})$ satisfying $(X, Y^{\star}) \in [0,1]^d \times [-\beta^{\star}, \beta^{\star}]$ a.s., $m^{\star}(X) \in [0, \beta]$ a.s., and $m^{\star} \in W_p([0,1]^d)$ with $0 < J_p^2(m^{\star}) < \infty$.

The proof of Lemma 4.1 is given in Section 4.6.

4.2 Rate of the maximum squared transformation errors

In Section 3.2, it is shown that under our regularity assumptions on the distribution of (Y, C), the maximum squared transformation error

$$\max_{i=1,\ldots,n} \left|U_i^{(j)} - \hat{U}_i^{(j)}\right|^2 \quad (j \in \{1,2,3\}) \tag{4.9}$$

converges almost surely to zero with n tending to infinity. Here, we derive a result on the rate of convergence of (4.9) under the additional assumption (1.8). Since

$$\frac{1}{n}\sum_{i=1}^{n} |U_i^{(j)} - \hat{U}_i^{(j)}|^2 \leq \max_{i=1,\ldots,n} \left|U_i^{(j)} - \hat{U}_i^{(j)}\right|^2 \quad (j \in \{1,2,3\}),$$

this implies the same rate of convergence for the mean squared transformation error.

Lemma 4.2. *Let $\alpha_1, \alpha_2 \in \mathbb{R}$ and let $\tau \in \mathbb{R}$ be arbitrary, but fixed. Define $U_i^{(1)}$, $U_i^{(2)}$, and $U_i^{(3)}$ $(i = 1, \ldots, n)$ by (2.22), (2.50), and (2.78), respectively. Furthermore, let $\hat{U}_i^{(1)}$, $\hat{U}_i^{(2)}$, and $\hat{U}_i^{(3)}$ $(i = 1, \ldots, n)$ be given by (2.25), (2.51), and (2.79). Let $\gamma \in (0,1)$ and assume that*

$$-\int_0^{\tau_F} F(t)^{\frac{-\gamma}{2-\gamma}} dG(t) < \infty. \tag{4.10}$$

For every $j \in \{1,2,3\}$, there exists a constant $b_3 \in (0, \infty)$ such that

$$\limsup_{n\to\infty} n^{\gamma} \max_{i=1,\ldots,n} \left|U_i^{(j)} - \hat{U}_i^{(j)}\right|^2 \leq b_3 \quad a.s.$$

for all distributions of (Y, C) with Y and C independent, which satisfy **(RA2)** *and* **(RA4)**.

PROOF OF LEMMA 4.2. First we observe that since Y and C are independent, **(RA2)** yields $\tau_K = \tau_F \leq L < \infty$ and $G(\tau_K) = G(\tau_F) \geq G(L) > 0$ (cf. Section 1.2).

Set $B^{(1)} := 4\,(1+|\alpha_1|)^2\,L^2$, $B^{(2)} := 4\,(1+|\alpha_2|)^2\,L^4$, and $B^{(3)} := 1$. From (3.17) and (3.18), one can conclude for all $j \in \{1,2,3\}$ and all $\gamma \in (0,1)$

$$\begin{aligned}
&\limsup_{n\to\infty}\, n^{\gamma} \max_{i=1,\ldots,n} \left|U_i^{(j)} - \hat{U}_i^{(j)}\right|^2 \\
&\leq \limsup_{n\to\infty} \left[n^{\gamma} \cdot \frac{B^{(j)}}{G^2(\tau_F)\cdot G_n^2(\tau_F)} \cdot \left(\sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \right)^2 \right] \\
&\leq \frac{B^{(j)}}{G^2(\tau_F)} \cdot \left[\limsup_{n\to\infty} \frac{1}{G_n^2(\tau_F)} \right] \cdot \left[\limsup_{n\to\infty}\, n^{\gamma} \left(\sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \right)^2 \right] \\
&\leq \frac{4B^{(j)}}{G^4(\tau_F)} \cdot \left[\limsup_{n\to\infty}\, n^{\frac{\gamma}{2}} \sup_{0\leq t\leq \tau_F} |\,G_n(t) - G(t)\,| \right]^2 \quad \text{a.s.}
\end{aligned}$$

This together with Theorem 1.2, $G(\tau_K) > 0$, **(RA4)**, (4.10), and the independence of Y and C implies the assertion of Lemma 4.2.

□

4.3 Rate of the MSSE of the regression function

In this section, our result for the rate of convergence of the MSSE (2.28) is presented. Corollary 4.1 and Lemma 4.2 indicate that for this purpose, we need additional regularity assumptions to **(RA1)** – **(RA4)** on the distribution of (X,Y,C).

Theorem 4.2. (Rate of convergence) *Let $d, n \in \mathbb{N}$, $\alpha_1 \in \mathbb{R}$, and $L \geq 1$. Let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Assume that we have chosen the parameters k and λ_n of the estimate $m_{n,(k,\lambda_n)}$, which is defined by (2.27) and (2.28), such that $k = p$ and λ_n fulfills (4.3) with an arbitrary constant $b_1 > 0$. Then*

$$\int_{\mathbb{R}^d} \left| m_{n,(k,\lambda_n)}(x) - m(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right) \tag{4.11}$$

for every distribution of (X,Y,C) which satisfies **(RA1)** – **(RA4)**, *$m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$, and*

$$-\int_0^{\tau_F} F(t)^{\frac{-p}{p+d}}\, dG(t) < \infty. \tag{4.12}$$

Note that since G is monotonically decreasing, the left hand side of (4.12) is always non-negative.

Remark 4.1. Stone (1982) proved that the optimal rate of convergence (in appropriate minimax sense) in $\mathcal{L}_2$ for nonparametric estimates of (p, B)-smooth regression functions is given by $n^{-\frac{2p}{2p+d}}$. For $p \in \mathbb{N}$ and $B \in (0, \infty)$, a function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, B)-smooth if for all $p_1, \ldots, p_d \in \mathbb{N}_0$ with $p_1 + \ldots + p_d = p - 1$ the partial derivative $\frac{\partial^{(p-1)}}{\partial x_1^{p_1} \ldots \partial x_d^{p_d}} f$ exists and satisfies

$$\left| \frac{\partial^{(p-1)}}{\partial x_1^{p_1} \ldots \partial x_d^{p_d}} f(x) - \frac{\partial^{(p-1)}}{\partial x_1^{p_1} \ldots \partial x_d^{p_d}} f(x_0) \right| \leq B \cdot \|x - x_0\| \quad (x, x_0 \in \mathbb{R}^d). \tag{4.13}$$

In some respects, (4.13) is a much stronger condition than the assumption $J_p^2(f) < \infty$, (cf. Theorem 4.2), because in the latter one, the weak derivatives of order p may vary in such a way that the squared average (2.5) is bounded by some non-negative constant. In contrast, if (4.13) holds, the partial derivatives of total order p are bounded by B, i.e., the function satisfies the smoothness condition on the whole domain.

Since in our setting it is allowed that censoring does not arise, i.e., $\mathbf{P}[C > L] = 1$ (vide **(RA2)**), we deduce from Stone (1982) that the rate of convergence in Theorem 4.2 is optimal up to the logarithmic factor $(\ln n)^{\frac{4p}{2p+d}}$. Note that the rate in Theorem 4.2 corresponds to published rates of MSSE in usual nonparametric regression with random design and without imposing regularity conditions on the distribution of X (cf. Kohler, Krzyżak, and Schäfer (2002)). However, for censored regression, the additional assumptions on the distribution of C are needed.

If $J_p^2(m) = 0$, then m is a multivariate polynomial of degree $p - 1$. In this case, one can deduce similar to the proofs of Theorem 4.1, Theorem 4.2, and Theorem 1 in Kohler, Krzyżak, and Schäfer (2002) that for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)** and $m \in W_p([0, 1]^d)$, the MSSE (2.28) with $k = p$ achieves the optimal parametric rate n^{-1} up to the factor n^{ω} if $n^{\omega} (\ln n)^{-2} \lambda_n \to \infty$ $(n \to \infty)$ and (4.10) holds with $\gamma = 1 - \omega$ $(0 < \omega < 1)$.

Remark 4.2. We want to stress that in Theorem 4.2 no assumption on the underlying distribution of X besides **(RA1)** is required. Especially, we do not demand that X has a density with respect to the Lebesgue-Borel measure.

Remark 4.3. It follows from the proofs of Lemma 3.1, Lemma 4.2, and Theorem 4.2, and the Remark after Theorem 1.1 in Chen and Lo (1997), that our result also holds for discontinuous G if the conditions of Theorem 1.1 are fulfilled (cf. Remark 3.1). In this case, replace G and F in (4.12) by continuous survival functions $\tilde{G}$ and $\tilde{F}$, where $\tilde{G}$ smooths the probability mass of G at its discontinuity points to small intervals and $\tilde{F}$ assigns probability zero to these intervals. For further details, see Chen and Lo (1997).

Remark 4.4. Lemma 1.1 and **(RA2)** imply that assumption (4.12) in Theorem 4.2 holds if there exists some $\tilde{\gamma} \in \left(0, 1 + \frac{d}{p}\right)$ such that

$$0 < \liminf_{t \uparrow \tau_F} \frac{(G(t) - G(\tau_F))^{\tilde{\gamma}}}{F(t)} \leq \limsup_{t \uparrow \tau_F} \frac{(G(t) - G(\tau_F))^{\tilde{\gamma}}}{F(t)} < \infty.$$

Proof of Theorem 4.2. Define $U^{(1)}$, $U_i^{(1)}$, and $\hat{U}_i^{(1)}$ $(i = 1, \ldots, n)$ by (2.21), (2.22), and (2.25). In the first part of this proof, it is shown that we can apply Corollary 4.1 in order to verify (4.11). Note that **(RA2)**, (2.23), and (2.24) imply $m(X) \in [0, L]$ a.s.,

$$m(X) = \mathbf{E}\left[U^{(1)} \,\middle|\, X \right],$$

and

$$|U^{(1)}| \leq (1 + 2|\alpha_1|)\frac{L}{G(L)} = L_1^\star < \infty \quad \text{a.s.},$$

respectively. Since $G(L) \leq 1$, we have $L_1^\star \geq L$.

Hence, one can conclude similar to the proof of Theorem 3.2, that $m^\star$ equals m, $m^\star_{n,(k,\lambda_n)}$ equals $m_{n,(k,\lambda_n)}$, and

$$\left| Y_i^\star - \bar{Y}_i^{(n)} \right| = \left| U_i^{(1)} - \hat{U}_i^{(1)} \right| \quad (i = 1, \ldots, n), \tag{4.14}$$

if we set $\beta^\star = L_1^\star$, $\beta = L$, $Y^\star = U^{(1)}$, $Y_i^\star = U_i^{(1)}$, and $\bar{Y}_i^{(n)} = \hat{U}_i^{(1)}$ $(i = 1, \ldots, n)$ in Section 4.1. Now, Lemma 3.1, **(RA2)** – **(RA4)**, and (4.14) yield for any $\epsilon > 0$

$$\mathbf{P}\left[\max_{i=1,\ldots,n} \left| Y_i^\star - \bar{Y}_i^{(n)} \right|^2 > \epsilon \right] = \mathbf{P}\left[\max_{i=1,\ldots,n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 > \epsilon \right] \to 0 \quad (n \to \infty) \tag{4.15}$$

and therefore (4.2). This implies that the conditions of Corollary 4.1 hold and one gets

$$\int_{\mathbb{R}^d} \left| m_{n,(k,\lambda_n)}(x) - m(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n} \sum_{i=1}^n \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 + \left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right). \tag{4.16}$$

Set $\gamma := \frac{2p}{2p+d}$. Since $p, d \in \mathbb{N}$, we have $\gamma \in (0,1)$. Moreover, it holds that

$$\frac{\gamma}{2-\gamma} = \frac{2p}{2 \cdot (2p+d) - 2p} = \frac{p}{p+d}\,. \tag{4.17}$$

Thus, Lemma 4.2, **(RA2)** – **(RA4)**, and (4.12) imply that there exists a constant $b_3 \in (0,\infty)$ such that

$$\limsup_{n\to\infty} n^{\frac{2p}{2p+d}} \frac{1}{n} \sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 \leq \limsup_{n\to\infty} n^{\frac{2p}{2p+d}} \max_{i=1,\dots,n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 \leq b_3 \quad \text{a.s.}$$

From this, one can conclude

$$\frac{1}{n} \sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p}{2p+d}} \right). \tag{4.18}$$

Now, (4.16) and (4.18) yield the assertion of Theorem 4.2.

□

In the next section, we examine the rate of the MSSE (2.57) of the conditional variance, which is defined via the estimate (2.28). As mentioned in Section 4.1, a crucial step in this analysis is the control of the empirical $\mathcal{L}_2$ error of (2.28).

Corollary 4.2. *Define the MSSE $m_{n,(k,\lambda_n)}$ by (2.27) and (2.28). If the conditions of Theorem 4.2 hold, then we have*

$$\frac{1}{n} \sum_{i=1}^{n} \left| m_{n,(k,\lambda_n)}(X_i) - m(X_i) \right|^2 = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right)$$

for every distribution of (X,Y,C) which satisfies **(RA1)** – **(RA4)**, *$m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$, and (4.12).*

Proof of Corollary 4.2. Let λ_n be chosen according to (4.3) with an arbitrary constant $b_1 > 0$. Define $U_i^{(1)}$ and $\hat{U}_i^{(1)}$ $(i = 1, \dots, n)$ by (2.22) and (2.25), respectively. Similar to the proof of Theorem 4.2, one can conclude from Lemma 4.1, **(RA1)**, **(RA2)**, (2.23), (2.24), (4.4) and (4.15), that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \left| m_{n,(k,\lambda_n)}(X_i) - m(X_i) \right|^2 + \lambda_n J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \\ &\qquad = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n} \sum_{i=1}^{n} \left| U_i^{(1)} - \hat{U}_i^{(1)} \right|^2 + \lambda_n + \left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right). \end{aligned} \tag{4.19}$$

The assertion of Corollary 4.2 follows from $\lambda_n J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \geq 0$, (4.3), (4.18), and (4.19).

□

4.4 Rate of the MSSE of the conditional variance

This section contains our result for the rate of convergence of the MSSE (2.57). Since the definition of (2.57) depends on the estimate (2.28) of the regression function, we therefore require that the assumptions of Theorem 4.2 on m and the parameters of (2.28) hold.

Theorem 4.3. (Rate of convergence) *Let $d, n \in \mathbb{N}$, $\alpha_1, \alpha_2 \in \mathbb{R}$, and $L \geq 1$. Furthermore, let $b_1, b_2 > 0$ be two arbitrary constants and let $p_1, p_2 \in \mathbb{N}$ with $2p_1 > d$, $2p_2 > d$ be arbitrary. Set $p_{min} := \min\{p_1, p_2\}$ and $p_{max} := \max\{p_1, p_2\}$. Choose the parameters k_1, k_2, $\lambda_{1,n}$, and $\lambda_{2,n}$ of the estimate $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$, defined by (2.56) and (2.57), such that $k_1 = p_1$, $k_2 = p_2$,*

$$\lambda_{1,n} = b_1 \cdot \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_1}{2p_1+d}} \tag{4.20}$$

and

$$\lambda_{2,n} = b_2 \cdot \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}. \tag{4.21}$$

Then one gets

$$\int_{\mathbb{R}^d} \left|\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(x) - \sigma^2(x)\right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_{min}}{2p_{min}+d}}\right) \tag{4.22}$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**, *$m \in W_{p_1}([0,1]^d)$ with $0 < J^2_{p_1}(m) < \infty$, $\sigma^2 \in W_{p_2}([0,1]^d)$ with $0 < J^2_{p_2}(\sigma^2) < \infty$, and*

$$-\int_0^{\tau_F} F(t)^{\frac{-p_{max}}{p_{max}+d}} dG(t) < \infty. \tag{4.23}$$

Remark 4.5. From Wang, Brown, Cai, and Levine (2008) (for $d = 1$), and Cai, Levine, and Wang (2009), we deduce that that for (p_1, B_1)-smooth m and (p_2, B_2)-smooth σ^2, the fastest achievable $\mathcal{L}_2$ rate of convergence of a nonparametric estimate of σ^2 is given by

$$\max\left\{n^{-\frac{4p_1}{d}}, n^{-\frac{2p_2}{2p_2+d}}\right\}. \tag{4.24}$$

If $2p_1 > d$ and $2p_2 > d$, then (4.24) equals $n^{-\frac{2p_2}{2p_2+d}}$. In case that $p_{min} = p_2 \leq p_1$, one can therefore conclude in analogy to Remark 4.1 that the rate of convergence in Theorem 4.3 is optimal up to a logarithmic factor. If even $p_2 = p_1$, then this rate corresponds to the rate given in Theorem 4.2, and condition (4.23) is identical to (4.12). However,

in order to show that (4.22) holds, additional assumptions on the smoothness of σ^2, i.e., $\sigma^2 \in W_{p_2}([0,1]^d)$ with $0 < J_{p_2}^2(\sigma^2) < \infty$, are needed.

Moreover, observe that Theorem 4.2 and Theorem 4.3 imply that for $p_2 \leq p_1$, the MSSE (2.57) of the conditional variance achieves the same rate of convergence as in the case when the regression function m is completely known (cf. Cai, Levine, and Wang (2009))

If, in contrast, $p_2 > p_1$ in Theorem 4.3 then we allow m to be "rougher" than σ^2. As a consequence, the rate of convergence in (4.22) is then dominated by the rate of the MSSE of the regression function. I.e., the rate in Theorem 4.3 equals the rate given in Theorem 4.2 but may be far from being (nearly) optimal for the nonparametric estimation of σ^2 in case that $p_2 \gg p_1$.

If $J_{p_1}^2(m) = 0$ and $J_{p_2}^2(\sigma^2) = 0$, then one can conclude similar to Remark 4.1 that for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)**, $m \in W_{p_1}([0,1]^d)$, and $\sigma^2 \in W_{p_2}([0,1]^d)$, the MSSE (2.57) with $k_1 = p_1$ and $k_2 = p_2$ achieves the rate $n^{\omega-1}$ if $n^{\omega} (\ln n)^{-2} \min\{\lambda_{1,n}, \lambda_{2,n}\} \to \infty$ $(n \to \infty)$ and (4.10) holds with $\gamma = 1 - \omega$ $(0 < \omega < 1)$.

Besides, Remarks 4.2, 4.3, and 4.4 (with p_{max} instead of p) also hold for Theorem 4.3.

Proof of Theorem 4.3. In analogy to the proof of Theorem 4.2, we first show that the assumptions of Corollary 4.1 hold.

Let $U^{(2)}$, $U_i^{(2)}$, and $\bar{U}_{i,n,(k_1,\lambda_{1,n})}$ $(i = 1, \ldots, n)$ be given by (2.49), (2.50), and (2.54). Our second regularity assumption **(RA2)** yields that with probability one, $m(X) \in [0, L]$ and $\sigma^2(X) \in [0, L^2]$. Furthermore, (2.52) and (2.53) imply

$$\sigma^2(X) = \mathbf{E}\left[U^{(2)} - m(X)^2 \,\middle|\, X \right]$$

and

$$|U^{(2)} - m(X)^2| \leq |U^{(2)}| + m(X)^2 \leq L_2^\star + L^2 =: \bar{L}_2 < \infty \quad \text{a.s.}, \tag{4.25}$$

respectively. Since

$$L_2^\star = \frac{(1 + 2\,|\alpha_2|)\, L^2}{G(L)} > 0,$$

we have $\bar{L}_2 \geq L^2$.

If we set $\beta^\star = \bar{L}_2$, $\beta = L^2$, $Y^\star = U^{(2)} - m(X)^2$, $Y_i^\star = U_i^{(2)} - m(X_i)^2$, $\bar{Y}_i^{(n)} = \bar{U}_{i,n,(k_1,\lambda_{1,n})}$ $(i = 1, \ldots, n)$, $k = k_2$, and $\lambda_n = \lambda_{2,n}$ in Section 4.1, then one can conclude in analogy

to the proofs of Theorem 3.3 and Theorem 4.2 that $m^\star$ equals σ^2 and $m^\star_{n,(k,\lambda_n)}$ equals $\sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}$. Moreover,

$$\left| Y_i^\star - \bar{Y}_i^{(n)} \right|^2 = \left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 \tag{4.26}$$

holds for all $i = 1, \ldots, n$.

Define $m_{n,(k_1,\lambda_{1,n})}$ and $\hat{U}_i^{(2)}$ $(i = 1, \ldots, n)$ by (2.28) and (2.51). Let $\epsilon > 0$ be arbitrary and set $B_1 := \epsilon + 2L^4$. Lemma 3.1, **(RA2)** – **(RA4)**, (2.54), (4.26), $m(X) \in [0, L]$ a.s., and $m_{n,(k_1,\lambda_{1,n})}(X) \in [0, L]$ a.s. yield

$$\begin{aligned}
&\mathbf{P}\left[\max_{i=1,\ldots,n} \left| Y_i^\star - \bar{Y}_i^{(n)} \right|^2 > B_1 \right] \\
&= \mathbf{P}\left[\max_{i=1,\ldots,n} \left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 > B_1 \right] \\
&= \mathbf{P}\left[\max_{i=1,\ldots,n} \left| U_i^{(2)} - m(X_i)^2 - \left(\hat{U}_i^{(2)} - m_{n,(k_1,\lambda_{1,n})}(X_i)^2 \right) \right|^2 > B_1 \right] \\
&\le \mathbf{P}\left[2 \cdot \max_{i=1,\ldots,n} \left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 + 2 \cdot \max_{i=1,\ldots,n} \left| m_{n,(k_1,\lambda_{1,n})}(X_i)^2 - m(X_i)^2 \right|^2 > \epsilon + 2L^4 \right] \\
&\le \mathbf{P}\left[2 \cdot \max_{i=1,\ldots,n} \left| U_i^{(2)} - \hat{U}_i^{(2)} \right|^2 > \epsilon \right] \to 0 \quad (n \to \infty),
\end{aligned} \tag{4.27}$$

where we used that $(a+b)^2 \le 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$. This implies (4.2).

Therefore, we deduce from Corollary 4.1, $k_2 = p_2$, (4.21), $\sigma^2 \in W_{p_2}([0,1]^d)$, and $0 < J^2_{p_2}(\sigma^2) < \infty$ that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left| \sigma^2_{n,(k_1,k_2,\lambda_{1,n},\lambda_{2,n})}(x) - \sigma^2(x) \right|^2 \mu(dx) \\
&= \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n} \sum_{i=1}^n \left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 + \left(\frac{(\ln n)^2}{n} \right)^{\frac{2p_2}{2p_2+d}} \right).
\end{aligned}$$

Since for all $a_1, a_2 \in \mathbb{N}_0$ with $a_1 \le a_2$

$$\left(\frac{(\ln n)^2}{n} \right)^{\frac{2a_2}{2a_2+d}} \le \left(\frac{(\ln n)^2}{n} \right)^{\frac{2a_1}{2a_1+d}}, \tag{4.28}$$

this yields that in order to prove (4.22), it suffices to show

$$\frac{1}{n} \sum_{i=1}^n \left| U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})} \right|^2 = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n} \right)^{\frac{2p_1}{2p_1+d}} \right). \tag{4.29}$$

An inequality similar to (4.8) implies with probability one

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\, U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})}\,\right|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\left|\, U_i^{(2)} - \hat{U}_i^{(2)}\,\right|^2 + \frac{2}{n}\sum_{i=1}^{n}\left|\, m_{n,(k_1,\lambda_{1,n})}(X_i)^2 - m(X_i)^2\,\right|^2 \\
&= \frac{2}{n}\sum_{i=1}^{n}\left|\, U_i^{(2)} - \hat{U}_i^{(2)}\,\right|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\left(\left|\, m_{n,(k_1,\lambda_{1,n})}(X_i) + m(X_i)\,\right|^2 \cdot \left|\, m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i)\,\right|^2\right) \\
&\leq 2 \cdot \max_{i=1,\ldots,n}\left|\, U_i^{(2)} - \hat{U}_i^{(2)}\,\right|^2 + (2L)^2 \cdot \frac{2}{n}\sum_{i=1}^{n}\left|\, m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i)\,\right|^2. \quad (4.30)
\end{aligned}$$

Set $\gamma := \frac{2p_{max}}{2p_{max}+d}$. Similar to the proof of Theorem 4.2, we deduce that $\gamma \in (0,1)$ and

$$\frac{\gamma}{2-\gamma} = \frac{p_{max}}{p_{max}+d}\,. \quad (4.31)$$

Hence, Lemma 4.2 and (4.23) yield that there exists a constant $b_3 \in (0,\infty)$ such that

$$\limsup_{n\to\infty} n^{\frac{2p_{max}}{2p_{max}+d}}\,\frac{1}{n}\sum_{i=1}^{n}\left|\, U_i^{(2)} - \hat{U}_i^{(2)}\,\right|^2 \leq \limsup_{n\to\infty} n^{\frac{2p_{max}}{2p_{max}+d}} \max_{i=1,\ldots,n}\left|\, U_i^{(2)} - \hat{U}_i^{(2)}\,\right|^2 \leq b_3 \quad \text{a.s.}$$

This together with (4.30) implies

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\, U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,n,(k_1,\lambda_{1,n})}\,\right|^2 \\
&\quad = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p_{max}}{2p_{max}+d}} + \frac{1}{n}\sum_{i=1}^{n}\left|\, m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i)\,\right|^2\right). \quad (4.32)
\end{aligned}$$

From (4.28), (4.29), and (4.32), one can conclude that in order to prove (4.22), it remains to show

$$\frac{1}{n}\sum_{i=1}^{n}\left|\, m_{n,(k_1,\lambda_{1,n})}(X_i) - m(X_i)\,\right|^2 = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_1}{2p_1+d}}\right). \quad (4.33)$$

In the following, Corollary 4.2 will be applied. First note that $F(t) \in [0,1]$ $(t \in \mathbb{R})$. Thus, we have

$$F(t)^{\frac{p_1}{p_1+d}} \geq F(t)^{\frac{p_{max}}{p_{max}+d}} \quad \forall t \in \mathbb{R}. \quad (4.34)$$

Since F and G are monotonically decreasing on $\mathbb{R}$, one gets from (4.23) and (4.34)

$$-\int_0^{\tau_F} F(t)^{\frac{-p_1}{p_1+d}}\,dG(t) \leq -\int_0^{\tau_F} F(t)^{\frac{-p_{max}}{p_{max}+d}}\,dG(t) < \infty. \quad (4.35)$$

Therefore, we have shown that condition (4.12) is fulfilled. Now, one can conclude from Corollary 4.2 (where we set $p = p_1$, $k = k_1$, and $\lambda_n = \lambda_{1,n}$) that (4.33) holds. This implies the assertion of Theorem 4.3.

□

4.5 Rate of the MSSE of the conditional survival function

Let $\tau \in \mathbb{R}$ be arbitrary, but fixed. In the following theorem, we investigate the rate of stochastic convergence of our estimate of the conditional survival function for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)**, $F(\tau\,|\cdot\,) \in W_p([0,1]^d)$ with $0 < J_p^2(F(\tau\,|\cdot\,)) < \infty$, and (4.12).

Theorem 4.4. (Rate of convergence) *Let $\tau \in \mathbb{R}$ be arbitrary, but fixed and $d, n \in \mathbb{N}$. Let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Assume that we have chosen the parameters k and λ_n of the estimate $F_{n,(k,\lambda_n)}(\tau\,|\cdot)$, which is defined by (2.82) and (2.83), such that $k = p$ and*

$$\lambda_n = b_1 \cdot \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}},$$

where $b_1 > 0$ is an arbitrary constant. Then

$$\int_{\mathbb{R}^d} \left|F_{n,(k,\lambda_n)}(\tau\,|\,x) - F(\tau\,|\,x)\right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right)$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**, *$F(\tau\,|\cdot\,) \in W_p([0,1]^d)$ with $0 < J_p^2(F(\tau\,|\cdot\,)) < \infty$, and (4.12).*

Remark 4.6. Due to (2.76), $F(\tau|X)$ is the regression function to $(X, I_{[Y>\tau]})$. Therefore, we deduce from Remark 4.1 that the rate of convergence in Theorem 4.4 is optimal up to the logarithmic factor $(\ln n)^{\frac{4p}{2p+d}}$.

If $n^\omega (\ln n)^{-2} \lambda_n \to \infty$ $(n \to \infty)$ and (4.10) holds with $\gamma = 1 - \omega$ $(0 < \omega < 1)$, then the estimate (2.83) with $k = p$ derives the rate $n^{\omega-1}$ for all distributions of (X, Y, C) which satisfy **(RA1)** – **(RA4)** and $F(\tau\,|\cdot\,) \in W_p([0,1]^d)$ with $J_p^2(F(\tau\,|\cdot\,)) = 0$ (cf. Remark 4.1).

Moreover, Remarks 4.2, 4.3, and 4.4 hold for Theorem 4.4, too.

PROOF OF THEOREM 4.4. Below, we mimic the proof of Theorem 4.2. First, it is shown that Corollary 4.1 may be applied.

Fix $\tau \in \mathbb{R}$. Let $U^{(3)}$, $U_i^{(3)}$, and $\hat{U}_i^{(3)}$ $(i = 1, \ldots, n)$ be defined by (2.77) – (2.79). Obviously, we have with probability one that $F(\,\tau|\,X) = \mathbf{P}\,[\,Y > \tau\,|\,X\,] \in [0,1]$. Now observe that (2.80) and (2.81) yield

$$F\,(\,\tau\,|\,X\,) = \mathbf{E}\left[\,U^{(3)}\,\middle|\,X\,\right]$$

and

$$|U^{(3)}| \leq \frac{1}{G(L)} = L_3^\star < \infty \quad \text{a.s.},$$

respectively. Since $G(L) \leq 1$, it holds that $L_3^\star \geq 1$.

From this, one can conclude in analogy to the proof of Theorem 3.4, that if we set $\beta^\star = L_3^\star$, $\beta = 1$, $Y^\star = U^{(3)}$, $Y_i^\star = U_i^{(3)}$, and $\bar{Y}_i^{(n)} = \hat{U}_i^{(3)}$ $(i = 1, \ldots, n)$ in Section 4.1, then $m^\star(\cdot)$ equals $F\,(\,\tau\,|\,\cdot\,)$ and $m^\star_{n,(k,\lambda_n)}(\cdot)$ equals $F_{n,(k,\lambda_n)}(\tau\,|\,\cdot)$. Furthermore, one gets

$$\left|\,Y_i^\star - \bar{Y}_i^{(n)}\,\right|^2 = \left|\,U_i^{(3)} - \hat{U}_i^{(3)}\,\right|^2 \quad (i = 1, \ldots, n).$$

Note that Lemma 3.1 yields for any $\epsilon > 0$ (cf. (4.15))

$$\mathbf{P}\left[\max_{i=1,\ldots,n}\left|U_i^{(3)} - \hat{U}_i^{(3)}\right|^2 > \epsilon\right] \to 0 \quad (n \to \infty),$$

which, in turn, implies (4.2). Therefore, we have shown that the assumptions of Corollary 4.1 are fulfilled and it holds that

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|F_{n,(k,\lambda_n)}(\tau\,|\,x) - F(\tau\,|\,x)\right|^2 \mu(dx)\\ &\quad = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^n \left|\,U_i^{(3)} - \hat{U}_i^{(3)}\,\right|^2 + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right). \end{aligned} \tag{4.36}$$

Set $\gamma := \frac{2p}{2p+d}$. Similar to the proof of Theorem 4.2, one can deduce from $p, d \in \mathbb{N}$, Lemma 4.2, (4.12), and (4.17) that there exists a constant $b_3 \in (0, \infty)$ such that

$$\limsup_{n\to\infty}\, n^{\frac{2p}{2p+d}}\,\frac{1}{n}\sum_{i=1}^n \left|\,U_i^{(3)} - \hat{U}_i^{(3)}\,\right|^2 \leq \limsup_{n\to\infty}\, n^{\frac{2p}{2p+d}} \max_{i=1,\ldots,n}\left|\,U_i^{(3)} - \hat{U}_i^{(3)}\,\right|^2 \leq b_3 \quad \text{a.s.}$$

which yields

$$\frac{1}{n}\sum_{i=1}^n \left|\,U_i^{(3)} - \hat{U}_i^{(3)}\,\right|^2 = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p}{2p+d}}\right).$$

This together with (4.36) implies the assertion of Theorem 4.4.

□

4.6 Proofs of Theorem 4.1 and Lemma 4.1

In this section, it is shown that the assertions of Theorem 4.1 and Lemma 4.1 hold. We start with the proof of Lemma 4.1. Subsequently, this result will be applied in order to verify Theorem 4.1. In the proof of Lemma 4.1, we apply a result from fixed design regression, Lemma A.1, which is formulated and proven in Appendix A.

Proof of Lemma 4.1. Define t_n by

$$t_n := \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}} \tag{4.37}$$

and set

$$\begin{aligned} S^\star_{n,(k,\lambda_n)} &:= \frac{1}{n}\sum_{i=1}^n \left|m^\star_{n,(k,\lambda_n)}(X_i) - m^\star(X_i)\right|^2 + \lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) \\ &\quad - \frac{64}{n}\sum_{i=1}^n \left|Y_i^\star - \bar{Y}_i^{(n)}\right|^2 - 2\lambda_n J_p^2(m^\star). \end{aligned}$$

Assume that there exists a constant $b_2 > 0$ such that (4.2) holds and let $l := \beta^\star + \sqrt{b_2}$. Then one can conclude

$$\begin{aligned} &\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^n \left|m^\star_{n,(k,\lambda_n)}(X_i) - m^\star(X_i)\right|^2 + \lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) \right. \\ &\qquad \left. > \frac{64}{n}\sum_{i=1}^n \left|Y_i^\star - \bar{Y}_i^{(n)}\right|^2 + 2\lambda_n J_p^2(m^\star) + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right] \\ &= \mathbf{P}\left[S^\star_{n,(k,\lambda_n)} > t_n\right] \\ &\le \mathbf{P}\left[S^\star_{n,(k,\lambda_n)} > t_n, \max_{i=1,\dots,n}\left|\bar{Y}_i^{(n)}\right| \le l\right] + \mathbf{P}\left[\max_{i=1,\dots,n}\left|\bar{Y}_i^{(n)}\right| > l\right] \\ &=: q_{1,n} + q_{2,n}. \end{aligned} \tag{4.38}$$

For the second term on the right hand side of (4.38), $|Y_i^\star| \le \beta^\star$ a.s. $(i = 1, \dots, n)$ and (4.2) yield

$$\begin{aligned} q_{2,n} &\le \mathbf{P}\left[\max_{i=1,\dots,n}\left|Y_i^\star - \bar{Y}_i^{(n)}\right| + \max_{i=1,\dots,n}|Y_i^\star| > l\right] \le \mathbf{P}\left[\max_{i=1,\dots,n}\left|Y_i^\star - \bar{Y}_i^{(n)}\right| + \beta^\star > l\right] \\ &= \mathbf{P}\left[\max_{i=1,\dots,n}\left|Y_i^\star - \bar{Y}_i^{(n)}\right|^2 > b_2\right] \to 0 \quad (n \to \infty). \end{aligned} \tag{4.39}$$

Set

$$\hat{m}^{\star}_{n,(k,\lambda_n)}(\cdot) := T_{[-l,l]}\tilde{m}^{\star}_{n,(k,\lambda_n)}(\cdot). \tag{4.40}$$

From (3.7), (4.40), and $0 \le m^{\star}(X) \le \beta < l$ a.s., we have

$$\frac{1}{n}\sum_{i=1}^{n}\left|m^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i)\right|^2 \le \frac{1}{n}\sum_{i=1}^{n}\left|\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i)\right|^2 \quad \text{a.s.} \tag{4.41}$$

Similarly, one can conclude from (3.6) and (4.40) that inside of $q_{1,n}$

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2 \le \frac{1}{n}\sum_{i=1}^{n}\left|\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2. \tag{4.42}$$

For the first term on the right hand side of (4.38), (4.41), and (4.42) imply

$$\begin{aligned} q_{1,n} &\le \mathbf{P}\left[S^{\star}_{n,(k,\lambda_n)} > t_n,\ \frac{1}{n}\sum_{i=1}^{n}\left|\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2 \le \frac{1}{n}\sum_{i=1}^{n}\left|\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2\right] \\ &\le \mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i)\right|^2 + \lambda_n J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)})\right. \\ &\qquad > \frac{64}{n}\sum_{i=1}^{n}\left|Y_i^{\star} - \bar{Y}_i^{(n)}\right|^2 + 2\lambda_n J_p^2(m^{\star}) + t_n\,, \\ &\qquad \left.\frac{1}{n}\sum_{i=1}^{n}\left|\hat{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2 \le \frac{1}{n}\sum_{i=1}^{n}\left|\tilde{m}^{\star}_{n,(k,\lambda_n)}(X_i) - \bar{Y}_i^{(n)}\right|^2\right]. \end{aligned} \tag{4.43}$$

Next, Lemma A.1 will be applied (note that we have chosen k such that $k = p$). For this purpose, it is first shown that the conditions (A.4) – (A.6) hold with the special choice of t_n in (4.37).

Obviously, we may deduce from (4.37) that

$$t_n \to 0 \quad (n \to \infty) \tag{4.44}$$

and

$$\frac{nt_n}{\ln n} = n^{\frac{d}{2p+d}}(\ln n)^{\frac{2p-d}{2p+d}} \to \infty \quad (n \to \infty). \tag{4.45}$$

Now, (4.44) and (4.45) imply (A.4) and (A.5). Moreover, (4.1) and (4.45) yield

$$\frac{nt_n}{\ln n}\lambda_n^{\frac{d}{2p}} = \left[\left(\frac{n}{\ln n}\right)^{\frac{2p}{2p+d}}\lambda_n\right]^{\frac{d}{2p}}(\ln n)^{\frac{2p}{2p+d}} \to \infty \quad (n \to \infty). \tag{4.46}$$

and therefore (A.6).

Hence, Lemma A.1 and (4.43) imply for all sufficiently large n that

$$q_{1,n} \leq b_5 \exp(-b_6 n t_n), \tag{4.47}$$

where $b_5, b_6 > 0$ are two constants only depending on $\beta^\star$. This together with (4.45) yields

$$q_{1,n} \leq b_5 \exp(-b_6 \ln n) = b_5\, n^{-b_6} \to 0 \quad (n \to \infty). \tag{4.48}$$

The assertion of Lemma 4.1 follows from (4.38), (4.39), and (4.48).

□

Now we are in the position to prove Theorem 4.1. In order to show that the assertion of Theorem 4.1 holds, we apply Lemma B.2 and Lemma C.4, which are presented in Appendix B and C, respectively.

Proof of Theorem 4.1. From Lemma 4.1, one can conclude that it remains to show

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left| m^\star_{n,(k,\lambda_n)}(x) - m^\star(x) \right|^2 \mu(dx) \\
&= \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n} \sum_{i=1}^n \left| m^\star_{n,(k,\lambda_n)}(X_i) - m^\star(X_i) \right|^2 + \lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) + \lambda_n \right).
\end{aligned} \tag{4.49}$$

An application of the peeling-technique (cf. Section 5.3 in Van de Geer (2000)) yields for all $t > 0$

$$\begin{aligned}
&\mathbf{P}\left[\int_{\mathbb{R}^d} \left| m^\star_{n,(k,\lambda_n)}(x) - m^\star(x) \right|^2 \mu(dx) \right. \\
&\qquad\qquad \left. > \frac{2}{n} \sum_{i=1}^n \left| m^\star_{n,(k,\lambda_n)}(X_i) - m^\star(X_i) \right|^2 + 2\,\lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) + t \right] \\
&= \mathbf{P}\left[Q_{n,(k,\lambda_n)} > t \right] \\
&\leq \sum_{j=0}^{\infty} \mathbf{P}\left[Q_{n,(k,\lambda_n)} > t,\ 2^j t \leq 2\lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) + t < 2^{j+1} t \right],
\end{aligned} \tag{4.50}$$

where

$$\begin{aligned}
Q_{n,(k,\lambda_n)} \;:=\; & \int_{\mathbb{R}^d} \left| m^\star_{n,(k,\lambda_n)}(x) - m^\star(x) \right|^2 \mu(dx) \\
& - 2 \left[\frac{1}{n} \sum_{i=1}^n \left| m^\star_{n,(k,\lambda_n)}(X_i) - m^\star(X_i) \right|^2 + \lambda_n J_k^2(\tilde{m}^\star_{n,(k,\lambda_n)}) \right].
\end{aligned}$$

Let $\bar{\mathcal{D}}_n$ be defined by (3.4). From (4.50), one gets for all $t > 0$

$$
\begin{aligned}
&\mathbf{P}\left[\, Q_{n,(k,\lambda_n)} > t\right] \\
&\quad\le \sum_{j=0}^{\infty} \mathbf{P}\left[2\,\mathbf{E}\left[\left| m^{\star}_{n,(k,\lambda_n)}(X) - m^{\star}(X) \right|^2 \middle| \bar{\mathcal{D}}_n \right] - \frac{2}{n}\sum_{i=1}^{n} \left| m^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i) \right|^2 \right. \\
&\qquad\qquad \left. > \mathbf{E}\left[\left| m^{\star}_{n,(k,\lambda_n)}(X) - m^{\star}(X) \right|^2 \middle| \bar{\mathcal{D}}_n \right] + 2^j t,\ J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)}) < \frac{2^j t}{\lambda_n} \right] \\
&\quad\le \sum_{j=0}^{\infty} \mathbf{P}\left[\exists\, g \in \mathcal{G}_{2^j t/\lambda_n} :\ 2\,\mathbf{E}\, g(X) - \frac{2}{n}\sum_{i=1}^{n} g(X_i) > \mathbf{E}\, g(X) + 2^j t \right] \\
&\quad= \sum_{j=0}^{\infty} \mathbf{P}\left[\exists\, g \in \mathcal{G}_{2^j t/\lambda_n} :\ \frac{\mathbf{E}\, g(X) - \frac{1}{n}\sum_{i=1}^{n} g(X_i)}{\mathbf{E}\, g(X) + 2^j t} > \frac{1}{2} \right] \\
&\quad\le \sum_{j=0}^{\infty} \mathbf{P}\left[\sup_{g \in \mathcal{G}_{2^j t/\lambda_n}} \frac{\left| \mathbf{E}\, g(X) - \frac{1}{n}\sum_{i=1}^{n} g(X_i) \right|}{\mathbf{E}\, g(X) + 2^j t} > \frac{1}{2} \right].
\end{aligned}
\tag{4.51}
$$

Here, for every $j = 0, 1, \dots$

$$
\mathcal{G}_{2^j t/\lambda_n} := \left\{ g : g(x) = |T_{[0,\beta]}\tilde{f}(x) - m^{\star}(x)|^2, \tilde{f} \in \tilde{\mathcal{F}}_{2^j t/\lambda_n}, x \in [0,1]^d \right\}
$$

with

$$
\tilde{\mathcal{F}}_{2^j t/\lambda_n} := \left\{ \tilde{f} : \tilde{f} \in W_k([0,1]^d), J_k^2(\tilde{f}) \le \frac{2^j t}{\lambda_n} \right\}.
$$

Fix $j \in \{0, 1, \dots\}$. In the following, Lemma C.4 will be applied in order to bound the probabilities on the right hand side of (4.51). First, we check that the conditions (C.1) – (C.3) are fulfilled for all sufficiently large n. For this purpose, set $V = X$, $V_i = X_i$ $(i = 1, \dots, n)$, $\mathcal{G} = \mathcal{G}_{2^j t/\lambda_n}$, $\Theta = [0,1]^d$, $\epsilon = \frac{1}{2}$, $\nu = 2^j t$, and $K_1 = K_2 = \beta^2$ in Lemma C.4. Since $\beta \ge 1$ and $|g(X)| \le \beta^2$ a.s. for all $g \in \mathcal{G}_{2^j t/\lambda_n}$, one can conclude that (C.1) holds.

Now, note that (4.1) implies

$$
n\,\lambda_n \ge \left(\frac{n}{\ln n} \right)^{\frac{2p}{2p+d}} \ \lambda_n \to \infty \quad (n \to \infty). \tag{4.52}
$$

For all $t \ge \lambda_n$ and all sufficiently large n, we have from (4.52)

$$
\sqrt{2^j\, n\, t}\cdot \frac{1}{2\sqrt{2}} \ge \sqrt{\frac{n\,\lambda_n}{8}} \ge 576\,\beta^2,
$$

which, in turn, yields (C.2).

Next, it is shown that (C.3) is fulfilled. Let $g_1, g_2 \in \mathcal{G}_{2^j t/\lambda_n}$ be two arbitrary functions with $g_1(x) = |T_{[0,\beta]}\tilde{f}_1(x) - m^\star(x)|^2$ and $g_2(x) = |T_{[0,\beta]}\tilde{f}_2(x) - m^\star(x)|^2$ $(x \in [0,1]^d)$, where $\tilde{f}_1, \tilde{f}_2 \in \tilde{\mathcal{F}}_{2^j t/\lambda_n}$. Set $f_1 := T_{[0,\beta]}\tilde{f}_1$ and $f_2 := T_{[0,\beta]}\tilde{f}_2$. Similar to (3.31) and (3.32) , one can conclude from $m^\star(x) \in [0,\beta]$ $(x \in [0,1]^d)$ and (3.30) for all $x_1, \dots, x_n \in [0,1]^d$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n |g_1(x_i) - g_2(x_i)|^2 &= \frac{1}{n}\sum_{i=1}^n \left|(f_1(x_i) - m^\star(x_i))^2 - (f_2(x_i) - m^\star(x_i))^2\right|^2 \\
&= \frac{1}{n}\sum_{i=1}^n |(f_1(x_i) + f_2(x_i) - 2\,m^\star(x_i)) \cdot (f_1(x_i) - f_2(x_i))|^2 \\
&\le (2\beta)^2 \cdot \frac{1}{n}\sum_{i=1}^n |f_1(x_i) - f_2(x_i)|^2 \\
&\le (2\beta)^2 \cdot \frac{1}{n}\sum_{i=1}^n \left|T_{[-\beta,\beta]}\tilde{f}_1(x_i) - T_{[-\beta,\beta]}\tilde{f}_2(x_i)\right|^2 .
\end{aligned}$$

This implies that for all $s > 0$ and all $x_1, \dots, x_n \in [0,1]^d$

$$\mathcal{N}_2\left(s, \mathcal{G}_{2^j t/\lambda_n}, x_1^n\right) \le \mathcal{N}_2\left(\frac{s}{2\beta}, \mathcal{F}_{2^j t/\lambda_n}, x_1^n\right) \tag{4.53}$$

holds, where $x_1^n = (x_1, \dots, x_n)$ and

$$\mathcal{F}_{2^j t/\lambda_n} := \left\{T_{[-\beta,\beta]}\tilde{f} : \tilde{f} \in W_k([0,1]^d), J_k^2(\tilde{f}) \le \frac{2^j t}{\lambda_n}\right\}.$$

Now observe that

$$\frac{\ln n}{n} \le \left(\frac{\ln n}{n}\right)^{\frac{2p}{2p+d}} \qquad \forall p, d, n \in \mathbb{N}. \tag{4.54}$$

From (4.52), it follows for all $t \ge \lambda_n$, all $\xi \ge \frac{2^j t}{4}$, and all sufficiently large n that $n\,\xi \ge 4\beta^4$. This together with (4.1), (4.53), (4.54), and Lemma B.2 yields

$$\begin{aligned}
\int_0^{\sqrt{\xi}} \sqrt{\ln \mathcal{N}_2\left(s, \mathcal{G}_{2^j t/\lambda_n}, x_1^n\right)}\, ds &\le 2\beta \int_0^{\frac{\sqrt{\xi}}{2\beta}} \sqrt{\ln \mathcal{N}_2\left(s, \mathcal{F}_{2^j t/\lambda_n}, x_1^n\right)}\, ds \\
&\le 2\beta \left(b_8 \left(\frac{4\beta^2}{\xi} \cdot \frac{2^j\, t}{\lambda_n}\right)^{\frac{d}{4p}} \frac{\sqrt{\xi}}{2\beta}\sqrt{\ln n} + b_9 \frac{\sqrt{\xi}}{2\beta}\sqrt{\ln n}\right) \\
&\le b_8\,(4\beta)^{\frac{d}{2p}}\,\lambda_n^{-\frac{d}{4p}}\sqrt{\xi}\sqrt{\ln n} + b_9\sqrt{\xi}\sqrt{\ln n} \\
&\le 2\sqrt{n}\,\xi \left(b_8\,(4\beta)^{\frac{d}{2p}} \sqrt{\frac{\ln n}{n}\lambda_n^{-\frac{2p+d}{2p}}} + b_9\sqrt{\frac{\ln n}{n\lambda_n}}\right) \\
&\le 2\left(b_8\,(4\beta)^{\frac{d}{2p}} + b_9\right)\sqrt{n}\,\xi \sqrt{\left(\frac{\ln n}{n}\right)^{\frac{2p}{2p+d}} \lambda_n^{-1}}
\end{aligned} \tag{4.55}$$

for all $t \geq \lambda_n$, all $\xi \geq \frac{2^j t}{4}$, all $x_1, \ldots, x_n \in [0,1]^d$, and all sufficiently large n. Here, $b_8, b_9 > 0$ are two constants which only depend on p and d. Finally, (4.1), (4.55), and the definiton of covering numbers (Definition C.1) imply that (C.3) is fulfilled for all sufficiently large n.

Therefore, one can conclude from (4.51), Lemma C.4, $2^j \geq j+1$ $(j = 0, 1, 2, \ldots)$, and $\beta \geq 1$ that for all $t \geq \lambda_n$ and all sufficiently large n, it holds

$$\begin{aligned} \mathbf{P}\left[\, Q_{n,(k,\lambda_n)} > t \,\right] &\leq 50 \sum_{j=0}^{\infty} \exp\left(- \frac{2^j \, n \, t}{8 \cdot 128 \cdot 2304 \beta^4} \right) \leq 50 \sum_{j=0}^{\infty} \exp\left(-(j+1) \, B_1 \, n \, t \right) \\ &= 50 \, \frac{\exp\left(-B_1 \, n \, t\right)}{1 - \exp\left(-B_1 \, n \, t\right)}, \end{aligned} \tag{4.56}$$

where

$$B_1 := \frac{1}{8 \cdot 128 \cdot 2304 \beta^4}.$$

From (4.50), (4.52), and (4.56), one gets

$$\begin{aligned} &\mathbf{P}\left[\int_{\mathbb{R}^d} \left| m^{\star}_{n,(k,\lambda_n)}(x) - m^{\star}(x) \right|^2 \mu(dx) \right. \\ &\qquad \left. > \frac{2}{n} \sum_{i=1}^{n} \left| m^{\star}_{n,(k,\lambda_n)}(X_i) - m^{\star}(X_i) \right|^2 + 2\lambda_n J_k^2(\tilde{m}^{\star}_{n,(k,\lambda_n)}) + \lambda_n \right] \\ &= \mathbf{P}\left[\, Q_{n,(k,\lambda_n)} > \lambda_n \,\right] \\ &\leq 50 \, \frac{\exp\left(-B_1 \, n \, \lambda_n\right)}{1 - \exp\left(-B_1 \, n \, \lambda_n\right)} \to 0 \quad (n \to \infty). \end{aligned} \tag{4.57}$$

This yields (4.49) and therefore the assertion of Theorem 4.1.

□

Chapter 5

Adaptation

Chapter 5 investigates the rates of convergence of the MSSE (2.45), (2.73), and (2.91), which are defined via the splitting of the sample technique. Similar to the both preceding chapters, an estimate $m_N^\star$ which covers the definitions of (2.45), (2.73), and (2.91) is introduced in Section 5.1. In Theorem 5.1, Lemma 5.1, and Corollary 5.1, the rate of convergence of $m_N^\star$ is analyzed. Based on these results, we derive in Sections 5.2 – 5.4 nearly optimal rates of the MSSE (2.45), (2.73), and (2.91). Finally, Section 5.5 contains the proofs of Theorem 5.1 and Lemma 5.1.

5.1 General results

In this sequel, a MSSE $m_N^\star$, which is a generalization of our regression estimates (2.45), (2.73), and (2.91), is defined via the splitting of the sample technique. Compared to Sections 3.1 and 4.1, a slightly different notation is introduced below in order to simplify the presentations of the proofs in Sections 5.2 – 5.4.

For this purpose, we now denote the sample size by N, where $N \in \mathbb{N}$ with $N \geq 2$. In addition, let $(X, Y^\star)$, $(X_1^\star, Y_1^\star)$, $\ldots$, $(X_N^\star, Y_N^\star) \in [0,1]^d \times \mathbb{R}$ a.s. be i.i.d. random vectors with $\mathbf{E}\,(Y^\star)^2 < \infty$. Throughout this chapter, it is assumed that there exists a constant $\beta \in \mathbb{R}_+$ such that

$$m^\star(X) = \mathbf{E}\,[\, Y^\star \mid X \,] \in [0, \beta] \quad \text{a.s.}$$

In analogy to Section 3.1 and Section 4.1, we suppose that $Y_1^\star$, $\ldots$, $Y_N^\star$ are unknown. Therefore, the estimate $m_N^\star$ of $m^\star$ is now rather based on some observable real-valued ran-

dom variables $\bar{Y}_1^{(N)}, \dots, \bar{Y}_N^{(N)}$, which need neither to be independent nor to be identically distributed and where each $\bar{Y}_i^{(n)}$ may depend on the whole sample and finite number of further, suitably chosen parameters.

Let $N_\ell, N_\mathrm{T} \in \mathbb{N}$ with $N_\ell + N_\mathrm{T} = N$. According to Section 2.4, we split the data

$$\bar{\mathcal{D}}_N := \left\{(X_1^\star, \bar{Y}_1^{(N)}), \dots, (X_N^\star, \bar{Y}_N^{(N)})\right\} \tag{5.1}$$

in two parts: The learning data

$$\bar{\mathcal{D}}_{N_\ell} := \left\{(X_1^\star, \bar{Y}_1^{(N)}), \dots, (X_{N_\ell}^\star, \bar{Y}_{N_\ell}^{(N)})\right\} \tag{5.2}$$

and the testing data

$$\bar{\mathcal{D}}_{N_\mathrm{T}} := \left\{(X_{N_\ell+1}^\star, \bar{Y}_{N_\ell+1}^{(N)}), \dots, (X_N^\star, \bar{Y}_N^{(N)})\right\}. \tag{5.3}$$

Furthermore, let $K_N^\star \times \Lambda_N^\star$ be a finite, non-empty set of parameters with

$$K_N^\star \subset \left\{\left\lfloor \frac{d}{2} \right\rfloor + 1, \left\lfloor \frac{d}{2} \right\rfloor + 2, \dots\right\} \tag{5.4}$$

and

$$\Lambda_N^\star \subset (0, \infty). \tag{5.5}$$

Similar to Section 2.4, we use the learning data in order to define for each pair of parameters $(k, \lambda) \in K_N^\star \times \Lambda_N^\star$ an estimate $m_{N_\ell,(k,\lambda)}^\star$ via

$$\tilde{m}_{N_\ell,(k,\lambda)}^\star(\cdot) := \tilde{m}_{N_\ell,(k,\lambda)}^\star(\cdot, \bar{\mathcal{D}}_{N_\ell}) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |f(X_i^\star) - \bar{Y}_i^{(N)}|^2 + \lambda J_k^2(f) \right), \tag{5.6}$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are given by (2.3) and (2.5), and

$$m_{N_\ell,(k,\lambda)}^\star(\cdot) := T_{[0,\beta]} \tilde{m}_{N_\ell,(k,\lambda)}^\star(\cdot). \tag{5.7}$$

Then we choose that estimate out of all calculated MSSE (5.7) which performs best on the testing data in terms of the empirical $\mathcal{L}_2$ risk, i.e., our modified estimate is defined by

$$m_N^\star(\cdot) := m_{N_\ell,(k^\star,\lambda^\star)}^\star(\cdot) \tag{5.8}$$

with

$$(k^\star, \lambda^\star) := \operatorname*{arg\,min}_{(k,\lambda) \in K_N^\star \times \Lambda_N^\star} \left(\frac{1}{N_\mathrm{T}} \sum_{i=N_\ell+1}^{N} |m_{N_\ell,(k,\lambda)}^\star(X_i^\star) - \bar{Y}_i^{(N)}|^2 \right). \tag{5.9}$$

The parameters $k^\star$ and $\lambda^\star$ of $m_N^\star$ are chosen in a data-dependent way and are therefore random variables. The next theorem demonstrates in which way the rate of convergence of $m_N^\star = m_{N_\ell,(k^\star,\lambda^\star)}^\star$ depends on the rate of the empirical $\mathcal{L}_2$ error of $m_{N_\ell,(k,\lambda)}^\star$ on the testing data for arbitrary, but deterministic $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$.

Theorem 5.1. *Let d, N, N_ℓ, $N_{\mathrm{T}} \in \mathbb{N}$ with $N = N_\ell + N_{\mathrm{T}}$ and $N_\ell \leq \lceil \frac{N}{2} \rceil$. In addition, let $1 \leq \beta \leq \beta^\star < \infty$ and let the finite, non-empty set of parameters $K_N^\star \times \Lambda_N^\star$ be given by (5.4) and (5.5), where we assume that*

$$\frac{|K_N^\star \times \Lambda_N^\star|}{N^2} \to 0 \quad (N \to \infty). \tag{5.10}$$

For each $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$, define the estimate $m_{N_\ell,(k,\lambda)}^\star$ by (5.6) and (5.7). Furthermore, let the MSSE $m_N^\star$ be given by (5.8) and (5.9).

If

$$\bar{\mathcal{D}}_{N_\ell} \text{ and } (X,Y^\star),(X_{N_\ell+1}^\star,Y_{N_\ell+1}^\star),\dots,(X_N^\star,Y_N^\star) \text{ are independent} \tag{5.11}$$

and moreover

$$\bar{\mathcal{D}}_N \text{ and } (X,Y^\star) \text{ are independent,} \tag{5.12}$$

then we have

$$\int_{\mathbb{R}^d} \left|m_N^\star(x) - m^\star(x)\right|^2 \mu(dx)$$

$$= \mathcal{O}_{\mathbf{P}}\left(\min_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_{N_\ell,(k,\lambda)}^\star(X_i^\star) - m^\star(X_i^\star)|^2 + \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \frac{(\ln N)^2}{N} \right)$$

for every distribution of $(X,Y^\star)$ with $(X,Y^\star) \in [0,1]^d \times [-\beta^\star,\beta^\star]$ a.s. and $m^\star(X) \in [0,\beta]$ a.s.

The proof of Theorem 5.1 is given in Section 5.5.

As in Chapter 3, we deduce that condition (5.12) ensures

$$\mathbf{E}\left[Y^\star \middle| X, \bar{\mathcal{D}}_N\right] = \mathbf{E}\left[Y^\star \middle| X\right] = m^\star(X)$$

(cf. (3.2) and (3.5)). This implies that the minimal $\mathcal{L}_2$ risk of any regression estimate based on the data $\bar{\mathcal{D}}_N$ is the same as in case of i.i.d. data (vide Sections 1.3 and 3.1). Besides, (5.11) is required in the proof of Theorem 5.1 to show that the rate of the empirical $\mathcal{L}_2$

error of $m_N^\star$ on the testing data is identical to the rate of convergence given in Theorem 5.1. This assertion will be reformulated as a fixed design regression problem and verified by an application of Hoeffding's inequality (Lemma D.3), demanding that the sequences of random variables $Y_{N_\ell+1}^\star, \ldots, Y_N^\star$ and $\bar{Y}_1^{(N)}, \ldots, \bar{Y}_{N_\ell}^{(N)}$ are independent. In order to prove Theorem 5.1, it then suffices to show that the rate of the difference between the $\mathcal{L}_2$ error and two times the empirical $\mathcal{L}_2$ error of our MSSE is given by $\frac{(\ln N)^2}{N}$. Since $(k^\star, \lambda^\star) \in K_N^\star \times \Lambda_N^\star$, this is implied by the following lemma:

Lemma 5.1. *Let d, N, N_ℓ, $N_{\mathrm{T}} \in \mathbb{N}$ with $N = N_\ell + N_{\mathrm{T}}$ and $N_\ell \leq \lceil \frac{N}{2} \rceil$. Assume that $X \in [0,1]^d$ a.s., $\mathbf{E}\,(Y^\star)^2 < \infty$, and $m^\star(X) \in [0,\beta]$ a.s. for some $\beta \in (0,\infty)$. Let the finite, non-empty set of parameters $K_N^\star \times \Lambda_N^\star$ be given by (5.4) and (5.5). For each $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$, define the estimate $m_{N_\ell,(k,\lambda)}^\star$ by (5.6) and (5.7) and set*

$$H_{N,(k,\lambda)}^\star := \left| \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_{N_\ell,(k,\lambda)}^\star(X_i^\star) - m^\star(X_i^\star)|^2 - \mathbf{E}\left[|m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] \right|,$$

where $\bar{\mathcal{D}}_{N_\ell}$ is given by (5.2).

If the conditions (5.10) and (5.11) hold, then we have

$$\lim_{N\to\infty} \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[2\, H_{N,(k,\lambda)}^\star > \mathbf{E}\left[|m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] + \frac{(\ln N)^2}{N} \right] = 0.$$

The proof of Lemma 5.1 is given in Section 5.5.

Let $(k_N, \lambda_N) \in K_N^\star \times \Lambda_N^\star$ be arbitrary (but deterministic). Theorem 5.1 and Lemma 5.1 yield that under the conditions of Theorem 5.1, the rate of convergence of the estimate $m_N^\star$ is given by

$$\int_{\mathbb{R}^d} \left| m_{N_\ell,(k_N,\lambda_N)}^\star(x) - m^\star(x) \right|^2 \mu(dx) + \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \frac{(\ln N)^2}{N}. \tag{5.13}$$

Now assume that in addition, $m^\star \in W_p([0,1]^d)$ with $0 < J_p^2(m^\star) < \infty$ for some $p \in \mathbb{N}$ with $2p > d$ and that there exists a constant $b_2 > 0$ such that (4.2) is fulfilled. For $k = p$ and λ chosen according to (4.3) (with n replaced by N_ℓ), one can then conclude from Corollary 4.1 that

$$\int_{\mathbb{R}^d} \left| m_{N_\ell,(k,\lambda)}^\star(x) - m^\star(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \left(\frac{(\ln N_\ell)^2}{N_\ell} \right)^{\frac{2p}{2p+d}} \right).$$

Obviously, one can choose $K_N^\star$ with $|K_N^\star| < \infty$ such that $p \in K_N^\star$ for all sufficiently large N (vide (5.4)). However, the choice of λ in (4.3) also depends on $p \in \mathbb{N}$, whereas condition (5.10) requires that $|\Lambda_N^\star| < \infty$. Since we assume that the smoothness p of $m^\star$ is unknown in an application, it is not possible to choose $\Lambda_N^\star$ with $|\Lambda_N^\star| < \infty$ such that there always exists a $N_0 \in \mathbb{N}$ with $\lambda \in \Lambda_N^\star$ for all $N \geq N_0$. Hence, one cannot simply apply Corollary 4.1 and (5.13) to derive a rate of convergence of $m_N^\star$ which does not depend on the rate of the estimate (5.7). Instead, we will combine the results of Theorem 4.1, Theorem 5.1, and Lemma 5.1 in order to prove the following corollary. Here, it is shown that for smooth $m^\star$ and $K_N^\star \times \Lambda_N^\star$ chosen according to (2.41) and (2.42), the rate of convergence of $m_N^\star$ is determined by three factors: the rates of the mean squared generalized transformation errors on the learning and testing data and the rate of the MSSE known from usual nonparametric regression (cf. Corollary 4.1).

Corollary 5.1. *Let d, $N \in \mathbb{N}$ with $N \geq 2$ and set $N_\ell := \lceil \frac{N}{2} \rceil$ and $N_{\mathrm{T}} := N - N_\ell$. Moreover, let $1 \leq \beta \leq \beta^\star < \infty$ and let the set of parameters $K_N^\star \times \Lambda_N^\star$ be given by*

$$K_N^\star := \left\{ \left\lfloor \frac{d}{2} \right\rfloor + 1, \left\lfloor \frac{d}{2} \right\rfloor + 2, \ldots, \left\lfloor \frac{d}{2} \right\rfloor + \lceil (\ln N)^2 \rceil \right\} \tag{5.14}$$

and

$$\Lambda_N^\star := \left\{ \frac{\ln N}{2^N}, \frac{\ln N}{2^{N-1}}, \ldots, \frac{\ln N}{1} \right\}. \tag{5.15}$$

Define the estimate $m_N^\star$ by (5.6) – (5.9). Assume that there exists a constant $b_2 > 0$ such that

$$\mathbf{P}\left[\max_{i=1,\ldots,N_\ell} |Y_i^\star - \bar{Y}_i^{(N)}|^2 > b_2 \right] \to 0 \quad (N_\ell \to \infty). \tag{5.16}$$

Let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. If the conditions (5.11) and (5.12) are fulfilled, then one gets

$$\begin{aligned} &\int_{\mathbb{R}^d} \left| m_N^\star(x) - m^\star(x) \right|^2 \mu(dx) \\ &\quad = \mathcal{O}_{\mathbf{P}} \left(\frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \left(\frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}} \right) \end{aligned}$$

for every distribution of $(X, Y^\star)$ with $(X, Y^\star) \in [0,1]^d \times [-\beta^\star, \beta^\star]$ a.s., $m^\star(X) \in [0, \beta]$ a.s., and $m^\star \in W_p([0,1]^d)$ with $0 < J_p^2(m^\star) < \infty$.

PROOF OF COROLLARY 5.1. Assume that the conditions (5.11) and (5.12) hold. First note that (5.14) and (5.15) imply (5.10), since

$$\frac{|K_N^\star \times \Lambda_N^\star|}{N^2} = \frac{\lceil (\ln N)^2 \rceil \cdot (N+1)}{N^2} \to 0 \quad (N \to \infty). \tag{5.17}$$

Furthermore, one can conclude for all p, d, $N \in \mathbb{N}$ that

$$\frac{(\ln N)^2}{N} \leq \left(\frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}}. \tag{5.18}$$

For each $(k, \lambda) \in K_N^\star \times \Lambda_N^\star$, define the estimate $m^\star_{N_\ell,(k,\lambda)}$ by (5.6) and (5.7). According to Theorem 5.1, (5.17), and (5.18), it remains to show

$$\begin{aligned} &\min_{(k,\lambda) \in K_N^\star \times \Lambda_N^\star} \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(k,\lambda)}(X_i^\star) - m^\star(X_i^\star)|^2 \\ &\quad = \mathcal{O}_{\mathbf{P}} \left(\frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \left(\frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}} \right). \end{aligned} \tag{5.19}$$

In the following, Theorem 4.1 will be applied. Observe that for sufficiently large N, there exist $(\check{k}, \check{\lambda}) \in K_N^\star \times \Lambda_N^\star$ and a constant $B_1 > 0$ such that $\check{k} = p$ and

$$B_1 \left(\frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}} \leq \check{\lambda} \leq 2 \cdot B_1 \left(\frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}} \tag{5.20}$$

(cf. Kohler, Krzyżak, and Schäfer (2002)). Here, we used that for all sufficiently large N, one gets

$$1 \leq h_N := \frac{1}{B_1} \ln N \left(\frac{N}{(\ln N)^2} \right)^{\frac{2p}{2p+d}} \leq 2^N$$

and therefore $j_N := \lfloor \log_2 h_N \rfloor \in \{0, 1, \ldots, N\}$. Indeed, this yields that (5.20) holds with $\check{\lambda} := 2^{-j_N} \cdot \ln N$.

Now note that $\frac{N}{2} \leq N_\ell = \lceil \frac{N}{2} \rceil \leq N$ implies

$$\left(\frac{(\ln N_\ell)^2}{N_\ell} \right)^{\frac{2p}{2p+d}} \leq \left(2 \, \frac{(\ln N)^2}{N} \right)^{\frac{2p}{2p+d}}. \tag{5.21}$$

From (5.20) and (5.21), we deduce that

$$\left(\frac{N_\ell}{\ln N_\ell} \right)^{\frac{2p}{2p+d}} \check{\lambda} \geq B_1 \left(\frac{\ln N_\ell}{2} \right)^{\frac{2p}{2p+d}} \to \infty \quad (N_\ell \to \infty). \tag{5.22}$$

Hence, Theorem 4.1 (where we set $n = N_\ell$, $k = \check{k}$, and $\lambda_n = \check{\lambda}$), (5.16), (5.22), and $\check{k} = p$ yield

$$\int_{\mathbb{R}^d} \left| m^\star_{N_\ell,(\check{k},\check{\lambda})}(x) - m^\star(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}} \left(\frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |Y_i^\star - \bar{Y}_i^{(N)}|^2 + \check{\lambda} + \left(\frac{(\ln N_\ell)^2}{N_\ell} \right)^{\frac{2p}{2p+d}} \right).$$

From this together with (5.18), (5.19), (5.21), and (5.20), one can conclude that it suffices to show

$$\begin{aligned} &\frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(\check{k},\check{\lambda})}(X_i^\star) - m^\star(X_i^\star)|^2 \\ &\qquad = \mathcal{O}_{\mathbf{P}} \left(\int_{\mathbb{R}^d} \left| m^\star_{N_\ell,(\check{k},\check{\lambda})}(x) - m^\star(x) \right|^2 \mu(dx) + \frac{(\ln N)^2}{N} \right). \end{aligned} \tag{5.23}$$

Below, we apply Lemma 5.1. For all $(k, \lambda) \in K_N^\star \times \Lambda_N^\star$, set

$$H^\star_{N,(k,\lambda)} := \left| \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(k,\lambda)}(X_i^\star) - m^\star(X_i^\star)|^2 - \mathbf{E}\left[|m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] \right|,$$

where $\bar{\mathcal{D}}_{N_\ell}$ is given by (5.2). Then one gets

$$\begin{aligned} &\mathbf{P}\left[\frac{2}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(\check{k},\check{\lambda})}(X_i^\star) - m^\star(X_i^\star)|^2 - 3 \int_{\mathbb{R}^d} \left| m^\star_{N_\ell,(\check{k},\check{\lambda})}(x) - m^\star(x) \right|^2 \mu(dx) > \frac{(\ln N)^2}{N} \right] \\ &= \mathbf{P}\left[\frac{2}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(\check{k},\check{\lambda})}(X_i^\star) - m^\star(X_i^\star)|^2 - 2\,\mathbf{E}\left[|m^\star_{N_\ell,(\check{k},\check{\lambda})}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] \right. \\ &\qquad\qquad \left. > \mathbf{E}\left[|m^\star_{N_\ell,(\check{k},\check{\lambda})}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] + \frac{(\ln N)^2}{N} \right] \\ &\le \mathbf{P}\left[2\, H^\star_{N,(\check{k},\check{\lambda})} > \mathbf{E}\left[|m^\star_{N_\ell,(\check{k},\check{\lambda})}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] + \frac{(\ln N)^2}{N} \right] \\ &\le \mathbf{P}\left[\exists\, (k,\lambda) \in K_N^\star \times \Lambda_N^\star : 2\, H^\star_{N,(k,\lambda)} > \mathbf{E}\left[|m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] + \frac{(\ln N)^2}{N} \right] \\ &\le \sum_{(k,\lambda) \in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[2\, H^\star_{N,(k,\lambda)} > \mathbf{E}\left[|m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \right] + \frac{(\ln N)^2}{N} \right]. \end{aligned} \tag{5.24}$$

This together with Lemma 5.1, (5.11), and (5.17) implies (5.23) and therefore the assertion of Corollary 5.1.

□

5.2 Adaptive MSSE of the regression function

In the following Theorem, we analyze the rate of convergence of the MSSE (2.45), which chooses the parameters k and λ in a data-dependent way via the splitting of the sample technique (cf. Section 2.4).

Theorem 5.2. (Adaptation via splitting of the sample) *Let $d, n \in \mathbb{N}$ with $n \geq 2$. Set $n_1 := \lceil \frac{n}{2} \rceil$ and $n_t := n - n_1$. Let the set of parameters $K_n \times \Lambda_n$ be defined by (2.41) and (2.42). Let $L \geq 1$, $\alpha_1 \in \mathbb{R}$, and the estimate m_n be given by (2.43) – (2.46). For any $p \in \mathbb{N}$ with $2p > d$, we have*

$$\int_{\mathbb{R}^d} \left| m_n(x) - m(x) \right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}} \left(\left(\frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}} \right)$$

for every distribution of (X, Y, C) which satisfies **(RA1)** – **(RA4)**, *$m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$, and (4.12).*

Remark 5.1. The rate of convergence in Theorem 5.2 is identical to the rate in Theorem 4.2, although the definition of the MSSE m_n does not depend on p or $J_p^2(m)$. In this sense, m_n is able to adapt automatically to the unknown smoothness of the regression function, which is measured by p and $J_p^2(m)$. Particularly, observe that in Theorem 5.2, the same conditions on the distribution of (X, Y, C) as in Theorem 4.2 are required.

Remark 5.2. It follows from Lemma 4.2 and the proof of Theorem 5.2 that the assertion of Theorem 5.2 also hold if the MSSE m_n is replaced by $\ddot{m}_n(\cdot) := m_{n_1,(\ddot{k}^{(1)},\ddot{\lambda}^{(1)})}(\cdot)$, where

$$\left(\ddot{k}^{(1)}, \ddot{\lambda}^{(1)}\right) := \operatorname*{arg\,min}_{(k,\lambda) \in K_n \times \Lambda_n} \left(\frac{1}{n_t} \sum_{i=n_1+1}^{n} |m_{n_1,(k,\lambda)}(X_i) - \hat{U}_i^{(1)}|^2 \right),$$

i.e., if we use $\hat{U}_i^{(1)}$ instead of $\hat{U}_{i,n_t}^{(1)}$ $(i = n_1 + 1, \ldots, n)$ in (2.46) to define our estimate. Here, $\hat{U}_i^{(1)}$ and $m_{n_1,(k,\lambda)}$ are defined by (2.25) and (2.44). Observe that $\hat{U}_i^{(1)}$ depends on the whole sample, while $\hat{U}_{i,n_t}^{(1)}$ is calculated only on the basis of the testing data.

Remark 5.3. In order to define an estimate of m in the presence of censored data, Máthé (2006) suggested to split the transformed data (2.26) in a learning and testing set in place of first splitting the data (1.19) and then transforming these learning and testing data seperately. Though, one can show that in general, the learning data defined by Máthé

and $(X_{n_1+1}, U^{(1)}_{n_1+1})$, …, $(X_n, U^{(1)}_n)$ are correlated (cf. (5.11)). Hence, it is not possible to directly apply standard exponential probability inequalities in order to derive a nearly optimal rate of convergence of such regression estimates.

Proof of Theorem 5.2. Define $U^{(1)}$, $U^{(1)}_i$ $(i = 1, \ldots, n)$, $\hat{U}^{(1)}_{i,n_1}$ $(i = 1, \ldots, n_1)$, and $\hat{U}^{(1)}_{i,n_t}$ $(n_1 + 1, \ldots, n)$ by (2.21), (2.22), (2.36), and (2.37). Moreover, let the transformed learning data $\hat{\mathcal{D}}^{(1)}_{n_1}$ and the whole transformed data $\hat{\mathcal{D}}^{(1)}_{n_1,n_t}$ be given by (2.38) and (2.40).

In the following, it is shown that the assumptions of Corollary 5.1 hold. As in the proof of Theorem 4.2, we first note that **(RA2)** and (2.24) imply $m(X) \in [0, L]$ a.s. and $|U^{(1)}| \leq L^\star_1 < \infty$ a.s., where

$$L^\star_1 = (1 + 2|\alpha_1|)\frac{L}{G(L)} \geq L.$$

If we set $N = n$, $N_\ell = n_1$, $N_T = n_t$, $\beta^\star = L^\star_1$, $\beta = L$, $K^\star_N = K_n$, $\Lambda^\star_N = \Lambda_n$, $Y^\star = U^{(1)}$, $(X^\star_i, Y^\star_i) = (X_i, U^{(1)}_i)$ $(i = 1, \ldots, N)$, and

$$\bar{Y}^{(N)}_i = \begin{cases} \hat{U}^{(1)}_{i,n_1} & \text{if} \quad i \in \{1, \ldots, N_\ell\} \\ \hat{U}^{(1)}_{i,n_t} & \text{if} \quad i \in \{N_\ell + 1, \ldots, N\} \end{cases}$$

in Section 5.1, then one can conclude from **(RA2)**, **(RA3)**, (2.23), (2.43) – (2.46), and (5.6) – (5.9) that $m^\star$ equals m and $m^\star_N$ equals m_n. Moreover, one gets $\bar{\mathcal{D}}_{N_\ell} = \hat{\mathcal{D}}^{(1)}_{n_1}$ and $\bar{\mathcal{D}}_N = \hat{\mathcal{D}}^{(1)}_{n_1,n_t}$.

Now, observe that (2.34) yields that the estimates (2.36) coincide with the estimates (2.25) if the latter are only calculated on the learning data (2.31) instead of the whole sample. Thus, we have from Lemma 3.1 and **(RA2)** – **(RA4)**

$$\max_{i=1,\ldots,n_1} |U^{(1)}_i - \hat{U}^{(1)}_{i,n_1}|^2 \to 0 \quad (n_1 \to \infty) \quad \text{a.s.} \tag{5.25}$$

Since (X, Y, C), (X_1, Y_1, C_1), …, (X_n, Y_n, C_n) are i.i.d. random vectors, (2.21), (2.22), and (2.34) – (2.37) imply that $\hat{\mathcal{D}}^{(1)}_{n_1}$ and $(X, U^{(1)})$, $(X_{n_1+1}, U^{(1)}_{n_1+1})$, …, $(X_n, U^{(1)}_n)$ are independent and moreover that $\hat{\mathcal{D}}^{(1)}_{n_1,n_t}$ and $(X, U^{(1)})$ are independent (cf. (5.11) and (5.12)). Hence, Corollary 5.1 and (5.25) yield

$$\begin{aligned} &\int_{\mathbb{R}^d} \left|m_n(x) - m(x)\right|^2 \mu(dx) \\ &\quad = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} |U^{(1)}_i - \hat{U}^{(1)}_{i,n_1}|^2 + \frac{1}{n_t}\sum_{i=n_1+1}^{n} |U^{(1)}_i - \hat{U}^{(1)}_{i,n_t}|^2 + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right). \end{aligned}$$

Therefore, it suffices to show

$$\frac{1}{n_1}\sum_{i=1}^{n_1}|U_i^{(1)}-\hat{U}_{i,n_1}^{(1)}|^2=\mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p}{2p+d}}\right) \tag{5.26}$$

and

$$\frac{1}{n_t}\sum_{i=n_1+1}^{n}|U_i^{(1)}-\hat{U}_{i,n_t}^{(1)}|^2=\mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p}{2p+d}}\right). \tag{5.27}$$

Set $\gamma := \frac{2p}{2p+d}$. Since $p, d \in \mathbb{N}$, one gets $\gamma \in (0,1)$. Similar to above, we deduce from (2.35) that the estimates (2.37) coincide with the estimates (2.25) if the latter are only calculated on the testing data (2.32). Thus, Lemma 4.2, **(RA2)** – **(RA4)**, (4.12), and (4.17) imply that there exists a constant $b_3 \in (0,\infty)$ such that

$$\limsup_{n_1\to\infty} n_1^{\frac{2p}{2p+d}}\frac{1}{n_1}\sum_{i=1}^{n_1}|U_i^{(1)}-\hat{U}_{i,n_1}^{(1)}|^2 \le \limsup_{n_1\to\infty} n_1^{\frac{2p}{2p+d}}\max_{i=1,\dots,n_1}|U_i^{(1)}-\hat{U}_{i,n_1}^{(1)}|^2 \le b_3 \quad \text{a.s.}$$

and

$$\limsup_{n_t\to\infty} n_t^{\frac{2p}{2p+d}}\frac{1}{n_t}\sum_{i=n_1+1}^{n}|U_i^{(1)}-\hat{U}_{i,n_t}^{(1)}|^2 \le \limsup_{n_t\to\infty} n_t^{\frac{2p}{2p+d}}\max_{i=n_1+1,\dots,n}|U_i^{(1)}-\hat{U}_{i,n_t}^{(1)}|^2 \le b_3 \quad \text{a.s.}$$

This together with

$$n \ge n_1 \ge n_t = n - n_1 = n - \left\lceil \frac{n}{2} \right\rceil \ge \frac{n}{3} \tag{5.28}$$

implies (5.26) and (5.27) and hence the assertion of Theorem 5.2.

□

In the following section, the rate of convergence of the MSSE σ_n^2, whose definition depends on the estimate m_{N_1} of the regression function, is analyzed. For this purpose, we now examine the rate of the empirical $\mathcal{L}_2$ error of m_{N_1} with respect to the testing data (cf. Corollary 4.2).

Lemma 5.2. *Let $d, n \in \mathbb{N}$ with $n \ge 3$. Set $N_1 := \lceil \frac{2n}{3} \rceil$, $n_1 := \lceil \frac{n}{3} \rceil$, and $N_r := N_1 - n_1$. Define the set of parameters $K_n \times \Lambda_n$ by (2.41) and (2.42). Let $L \ge 1$, $\alpha_1 \in \mathbb{R}$, and let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Moreover, define the estimate m_{N_1} by (2.43), (2.44), (2.66), and (2.67). For every distribution of (X, Y, C) satisfying* **(RA1)** – **(RA4)**, *$m \in W_p([0,1]^d)$ with $0 < J_p^2(m) < \infty$, and (4.12), it holds that*

$$\frac{1}{n-n_1}\sum_{i=n_1+1}^{n}|m_{N_1}(X_i)-m(X_i)|^2=\mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right).$$

PROOF OF LEMMA 5.2. First note that $N_1 = \lceil \frac{2n}{3} \rceil$ and $n_1 = \lceil \frac{n}{3} \rceil$ yield for all $n \in \mathbb{N}$

$$n_1 = \left\lceil \frac{N_1}{2} \right\rceil \tag{5.29}$$

and

$$\left(\frac{(\ln N_1)^2}{N_1} \right)^{\frac{2p}{2p+d}} \leq \left(\frac{3}{2} \frac{(\ln n)^2}{n} \right)^{\frac{2p}{2p+d}}. \tag{5.30}$$

For each $(k, \lambda) \in K_n \times \Lambda_n$, define $m_{n_1,(k,\lambda)}$ by (2.44). Let m_n, $\hat{U}^{(1)}_{i,n_\mathtt{t}}$ $(i = n_1 + 1, \ldots, n)$ and $\hat{U}^{(1)}_{i,N_\mathtt{r}}$ $(i = n_1 + 1, \ldots, N_1)$ be given by (2.45), (2.37), and (2.63). Observe that (2.45), (2.46), and (2.66) imply that m_{N_1} coincides with m_n, if we replace N_1 with n, $N_\mathtt{r}$ with $n_\mathtt{t}$, and $\hat{U}^{(1)}_{i,N_\mathtt{r}}$ with $\hat{U}^{(1)}_{i,n_\mathtt{t}}$ in (2.67), except that the minimum in (2.67) is calculated over $K_n \times \Lambda_n$ instead of $K_{N_1} \times \Lambda_{N_1}$. Hence, one can conclude from the proofs of Corollary 5.1 and Theorem 5.2, (5.29), $N_1 = \lceil \frac{2n}{3} \rceil \geq 2$, and $N_\mathtt{r} = N_1 - n_1$ that

$$\int_{\mathbb{R}^d} \left| m_{N_1}(x) - m(x) \right|^2 \mu(dx) = \mathcal{O}_\mathbf{P} \left(\left(\frac{(\ln N_1)^2}{N_1} \right)^{\frac{2p}{2p+d}} \right). \tag{5.31}$$

This together with

$$\left(\frac{(\ln n)^2}{n} \right)^{\frac{2a}{2a+d}} \geq \frac{(\ln n)^2}{n} \quad \forall a \in \mathbb{N}_0 \tag{5.32}$$

and (5.30) yields that it suffices to show

$$\frac{1}{n - n_1} \sum_{i=n_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 = \mathcal{O}_\mathbf{P} \left(\int_{\mathbb{R}^d} \left| m_{N_1}(x) - m(x) \right|^2 \mu(dx) + \frac{(\ln n)^2}{n} \right). \tag{5.33}$$

Below, Lemma 5.1 will be applied. For each $(k, \lambda) \in K_n \times \Lambda_n$, set

$$H^{(1)}_{n,(k,\lambda)} := \left| \frac{1}{n - n_1} \sum_{i=n_1+1}^{n} \left| m_{n_1,(k,\lambda)}(X_i) - m(X_i) \right|^2 - \int_{\mathbb{R}^d} \left| m_{n_1,(k,\lambda)}(x) - m(x) \right|^2 \mu(dx) \right|$$

and note that (cf. (5.24))

$$\mathbf{P} \left[\frac{2}{n - n_1} \sum_{i=n_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 - 3 \int_{\mathbb{R}^d} \left| m_{N_1}(x) - m(x) \right|^2 \mu(dx) > \frac{(\ln n)^2}{n} \right]$$

$$\leq \sum_{(k,\lambda) \in K_n \times \Lambda_n} \mathbf{P} \left[2H^{(1)}_{n,(k,\lambda)} > \int_{\mathbb{R}^d} \left| m_{n_1,(k,\lambda)}(x) - m(x) \right|^2 \mu(dx) + \frac{(\ln n)^2}{n} \right]. \tag{5.34}$$

In analogy to (5.17), we deduce from (2.41) and (2.42)

$$\frac{|K_n \times \Lambda_n|}{n^2} = \frac{\lceil (\ln n)^2 \rceil \cdot (n+1)}{n^2} \to 0 \quad (n \to \infty). \tag{5.35}$$

Let $U^{(1)}$, $U_i^{(1)}$ $(i = 1, \dots, n)$, and $\hat{U}_{i,n_1}^{(1)}$ $(i = 1, \dots, n_1)$ be given by (2.21), (2.22), and (2.36), respectively. Now, set $N = n$, $N_\ell = n_1$, $N_{\mathtt{T}} = n - n_1$, $\beta = L$, $K_N^\star = K_n$, $\Lambda_N^\star = \Lambda_n$, $Y^\star = U^{(1)}$, $Y_i^\star = U_i^{(1)}$ $(i = 1, \dots, N)$, $X_i^\star = X_i$ $(i = 1, \dots, N)$, and $\bar{Y}_i^{(N)} = \hat{U}_{i,n_1}^{(1)}$ $(i = 1, \dots, N_\ell)$ in Lemma 5.1. Similar to the proof of Theorem 5.2, one can conclude from (2.23), (2.44), and (5.7) that $m^\star$ then equals m and $m_{N_\ell,(k,\lambda)}^\star$ equals $m_{n_1,(k,\lambda)}$. Furthermore, we deduce that $\bar{\mathcal{D}}_{N_\ell} = \hat{\mathcal{D}}_{n_1}^{(1)}$ and moreover that $H_{N,(k,\lambda)}^\star = H_{n,(k,\lambda)}^{(1)}$ for all $(k, \lambda) \in K_N^\star \times \Lambda_N^\star$. In addition, since $(X, Y, C), (X_1, Y_1, C_1), \dots, (X_n, Y_n, C_n)$ are i.i.d. random vectors, (2.21), (2.22), (2.34), and (2.38) imply that $\hat{\mathcal{D}}_{n_1}^{(1)}$ and $(X, U^{(1)})$, $(X_{n_1+1}, U_{n_1+1}^{(1)})$, $\dots$, $(X_n, U_n^{(1)})$ are independent (cf. (5.11)). Hence, Lemma 5.1, (2.24), **(RA1)**, **(RA2)**, (5.34), (5.35), and $n_1 = \lceil \frac{n}{3} \rceil \leq \lceil \frac{n}{2} \rceil$ yield (5.33) and therefore the assertion of Lemma 5.2.

□

5.3 Adaptive MSSE of the conditional variance

This section investigates the rate of convergence of the MSSE (2.73). Theorem 4.3 indicates that this rate also depends on the smoothness of the underlying estimate of the regression function.

Theorem 5.3. (Adaptation via splitting of the sample) *Let $d, n \in \mathbb{N}$ with $n \geq 3$. Set $N_1 := \lceil \frac{2n}{3} \rceil$, $n_1 := \lceil \frac{n}{3} \rceil$, $N_{\mathtt{t}} := n - N_1$, and $N_{\mathtt{r}} := N_1 - n_1$. Define the set of parameters $K_n \times \Lambda_n$ by (2.41) and (2.42). Let $L \geq 1$, $\alpha_1, \alpha_2 \in \mathbb{R}$, and the estimate σ_n^2 be given by (2.71) – (2.74). Let $p_1, p_2 \in \mathbb{N}$ with $2p_1 > d$, $2p_2 > d$ be arbitrary. Set $p_{min} := \min\{p_1, p_2\}$ and $p_{max} := \max\{p_1, p_2\}$. Then we have*

$$\int_{\mathbb{R}^d} \left|\sigma_n^2(x) - \sigma^2(x)\right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_{min}}{2p_{min}+d}}\right)$$

for every distribution of (X, Y, C) which satisfies **(RA1)** – **(RA4)**, *(4.23), $m \in W_{p_1}([0,1]^d)$ with $0 < J_{p_1}^2(m) < \infty$, and $\sigma^2 \in W_{p_2}([0,1]^d)$ with $0 < J_{p_2}^2(\sigma^2) < \infty$.*

Remark 5.4. The rate of convergence in Theorem 5.3 is identical to that of Theorem 4.3. As mentioned in Remark 4.5, this rate may be far from the optimal nonparametric rate in case that $p_1 < p_2$. If, in contrast, $p_2 \geq p_1$ then we can conclude similar to Remark 4.5

that σ_n^2 achieves the optimal rate of convergence up to some logarithmic factor. In this case, σ_n^2 adapts automatically to the unknown smoothness of σ^2 (cf. Remark 5.1).

Remark 5.5. Let $\hat{G}^{(KM)}$, $\hat{U}_i^{(2)}$ $(i = 1, \ldots, n)$, and m_{N_1} be given by (2.33), (2.51) and (2.66). Define $G_{N_1}(\cdot) := \hat{G}^{(KM)}(\cdot, \{(Z_1, \delta_1), \ldots, (Z_{N_1}, \delta_{N_1})\})$ and, for $\alpha_2 \in \mathbb{R}$,

$$\hat{U}_{i,N_1}^{(2)} := (1 + \alpha_2) \int_0^{Z_i} \frac{2\,t}{G_{N_1}(t)}\, dt - \alpha_2 \frac{\delta_i Z_i^2}{G_{N_1}(Z_i)} \quad (i = 1, \ldots, N_1).$$

Moreover, set

$$\bar{U}_i^{(2)} := \hat{U}_i^{(2)} - m_{N_1}(X_i)^2 \quad (i = 1, \ldots, n) \tag{5.36}$$

and

$$\bar{U}_{i,N_1}^{(2)} := \hat{U}_{i,N_1}^{(2)} - m_{N_1}(X_i)^2 \quad (i = 1, \ldots, N_1). \tag{5.37}$$

Similar to Remark 5.2, we deduce that Theorem 5.2 holds for $\ddot{\sigma}_n^2(\cdot) := \ddot{\sigma}^2_{N_1,(\ddot{k}^{(2)},\ddot{\lambda}^{(2)})}(\cdot)$ with

$$\left(\ddot{k}^{(2)}, \ddot{\lambda}^{(2)}\right) := \operatorname*{arg\,min}_{(k,\lambda)\in K_n\times\Lambda_n} \left(\frac{1}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |\ddot{\sigma}^2_{N_1,(k,\lambda)}(X_i) - \bar{U}_i^{(2)}|^2 \right),$$

too. Here, $\ddot{\sigma}^2_{N_1,(k,\lambda)}(\cdot) := T_{[0,L^2]}\ddot{\tilde{\sigma}}^2_{N_1,(k,\lambda)}(\cdot)$, where

$$\ddot{\tilde{\sigma}}^2_{N_1,(k,\lambda)}(\cdot) := \operatorname*{arg\,min}_{f\in W_k([0,1]^d)} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} |f(X_i) - \bar{U}_{i,N_1}^{(2)}|^2 + \lambda J_k^2(f) \right).$$

In addition, Theorem 5.2 is also fulfilled if we replace m_{N_1} in (5.36) and (5.37) by the MSSE $\ddot{m}_n$, which is defined in Remark 5.2.

Proof of Theorem 5.3. The proof is divided into five steps.

Step 1. For each $(k, \lambda) \in K_n \times \Lambda_n$, let $\sigma^2_{N_1,(k,\lambda)}$ be given by (2.72) and set

$$V_{n,(k,\lambda)} := \frac{1}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |\sigma^2_{N_1,(k,\lambda)}(X_i) - \sigma^2(X_i)|^2.$$

In the first step of the proof, we use Theorem 5.1 in order to show

$$\begin{aligned} &\int_{\mathbb{R}^d} \left|\sigma_n^2(x) - \sigma^2(x)\right|^2 \mu(dx) \\ &= \mathcal{O}_{\mathbf{P}}\left(\min_{(k,\lambda)\in K_n\times\Lambda_n} V_{n,(k,\lambda)} + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2 + \frac{(\ln n)^2}{n} \right). \end{aligned} \tag{5.38}$$

Here, $U_i^{(2)}$ $(i = 1, \ldots, n)$ and $\bar{U}_{i,N_{\mathtt{t}}}$ $(i = N_1 + 1, \ldots, n)$ are defined by (2.50) and (2.69), respectively.

In the following, we show that the conditions of Theorem 5.1 hold. Set $n_{\mathtt{t}} := n - n_{\mathtt{l}}$. Since $n_{\mathtt{l}} = \lceil \frac{n}{3} \rceil$, one has for all $n \geq 3$

$$n \geq n_{\mathtt{t}} = n - n_{\mathtt{l}} = n - \left\lceil \frac{n}{3} \right\rceil \geq \frac{n}{2}. \tag{5.39}$$

This implies

$$\frac{|K_n \times \Lambda_n|}{n_{\mathtt{t}}^2} = \frac{\lceil (\ln n)^2 \rceil \cdot (n+1)}{n_{\mathtt{t}}^2} \leq \frac{((\ln(2\,n_{\mathtt{t}}))^2 + 1) \cdot (2n_{\mathtt{t}} + 1)}{n_{\mathtt{t}}^2} \to 0 \quad (n_{\mathtt{t}} \to \infty) \tag{5.40}$$

and

$$\frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \leq 2\,\frac{(\ln n)^2}{n}. \tag{5.41}$$

Moreover, $n_{\mathtt{l}} = \lceil \frac{n}{3} \rceil$ and $N_{\mathtt{l}} = \lceil \frac{2n}{3} \rceil$ yield

$$N_{\mathtt{r}} = N_{\mathtt{l}} - n_{\mathtt{l}} = \left\lceil \frac{n_{\mathtt{t}}}{2} \right\rceil \quad \forall n \in \mathbb{N}. \tag{5.42}$$

Let $U^{(2)}$, $m_{N_{\mathtt{l}}}$, and $\bar{U}_{i,N_{\mathtt{r}}}$ $(i = n_{\mathtt{l}} + 1, \ldots, N_{\mathtt{l}})$ be given by (2.49), (2.66), and (2.68), respectively. Set

$$\hat{\mathcal{D}}_{N_{\mathtt{r}}}^{(2)} := \left\{ (X_{n_{\mathtt{l}}+1}, \bar{U}_{n_{\mathtt{l}}+1,N_{\mathtt{r}}}), \ldots, (X_{N_{\mathtt{l}}}, \bar{U}_{N_{\mathtt{l}},N_{\mathtt{r}}}) \right\}$$

and

$$\hat{\mathcal{D}}_{N_{\mathtt{r}},N_{\mathtt{t}}}^{(2)} := \left\{ (X_{n_{\mathtt{l}}+1}, \bar{U}_{n_{\mathtt{l}}+1,N_{\mathtt{r}}}), \ldots, (X_{N_{\mathtt{l}}}, \bar{U}_{N_{\mathtt{l}},N_{\mathtt{r}}}), (X_{N_{\mathtt{l}}+1}, \bar{U}_{N_{\mathtt{l}}+1,N_{\mathtt{t}}}), \ldots, (X_n, \bar{U}_{n,N_{\mathtt{t}}} \right\}.$$

Observe that (4.25) implies that there exists some constant $\bar{L}_2 \geq L^2$ such that we have with probability one $|U^{(2)} - m(X)^2| \leq \bar{L}_2$.

If we set $N = n_{\mathtt{t}}$, $N_\ell = N_{\mathtt{r}}$, $N_{\mathtt{T}} = N_{\mathtt{t}}$, $\beta^\star = \bar{L}_2$, $\beta = L^2$, $K_N^\star = K_n$, $\Lambda_N^\star = \Lambda_n$, $Y^\star = U^{(2)} - m(X)^2$, $Y_i^\star = U_{i+n_{\mathtt{l}}}^{(2)} - m(X_{i+n_{\mathtt{l}}})^2$ $(i = 1, \ldots, N)$, $X_i^\star = X_{i+n_{\mathtt{l}}}$ $(i = 1, \ldots, N)$, and

$$\bar{Y}_i^{(N)} = \begin{cases} \bar{U}_{i+n_{\mathtt{l}},N_{\mathtt{r}}} & \text{if} \quad i \in \{1, \ldots, N_\ell\} \\ \bar{U}_{i+n_{\mathtt{l}},N_{\mathtt{t}}} & \text{if} \quad i \in \{N_\ell + 1, \ldots, N\} \end{cases}$$

in Section 5.1, then one can deduce similar to the proof of Theorem 4.3 that $m^\star$ equals σ^2 and $m_N^\star$ equals σ_n^2. Moreover, one gets $\bar{\mathcal{D}}_{N_\ell} = \hat{\mathcal{D}}_{N_{\mathtt{r}}}^{(2)}$, $\bar{\mathcal{D}}_N = \hat{\mathcal{D}}_{N_{\mathtt{r}},N_{\mathtt{t}}}^{(2)}$,

$$\min_{(k,\lambda) \in K_N^\star \times \Lambda_N^\star} \frac{1}{N_{\mathtt{T}}} \sum_{i=N_\ell+1}^{N} |m_{N_\ell,(k,\lambda)}^\star(X_i^\star) - m^\star(X_i^\star)|^2 = \min_{(k,\lambda) \in K_n \times \Lambda_n} V_{n,(k,\lambda)}, \tag{5.43}$$

and

$$\frac{1}{N_{\mathtt{T}}} \sum_{i=N_\ell+1}^{N} |Y_i^\star - \bar{Y}_i^{(N)}|^2 = \frac{1}{N_{\mathtt{t}}} \sum_{i=N_{\mathtt{l}}+1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2. \tag{5.44}$$

Here, we used that $N_{\mathtt{r}} + n_{\mathtt{l}} = N_{\mathtt{l}}$ and $n_{\mathtt{l}} + n_{\mathtt{t}} = n$.

Observe that (X, Y, C), (X_1, Y_1, C_1), …, (X_n, Y_n, C_n) are i.i.d. random vectors. Hence, one can conclude from (2.34), (2.36), (2.43), (2.44), and (2.61) – (2.69) that $\hat{\mathcal{D}}^{(2)}_{N_{\mathtt{r}}}$ and $(X, U^{(2)} - m(X)^2)$, $(X_{N_{\mathtt{l}}+1}, U^{(2)}_{N_{\mathtt{l}}+1} - m(X_{N_{\mathtt{l}}+1})^2)$, …, $(X_n, U^{(2)}_n - m(X_n)^2)$ are independent (cf. (5.11)). Furthermore, we may deduce that $\hat{\mathcal{D}}^{(2)}_{N_{\mathtt{r}},N_{\mathtt{t}}}$ and $(X, U^{(2)} - m(X)^2)$ are independent (vide (5.12)). Therefore, one can conclude from Theorem 5.1, $\sigma^2(X) \in [0, L^2]$ a.s. (cf. **(RA2)**), (4.25), (5.40), and (5.42) – (5.44) that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left|\sigma^2_n(x) - \sigma^2(x)\right|^2 \mu(dx) \\
&= \mathcal{O}_{\mathbf{P}}\left(\min_{(k,\lambda)\in K_n\times\Lambda_n} V_{n,(k,\lambda)} + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_{\mathtt{l}}+1}^{n} |U^{(2)}_i - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2 + \frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \right). \quad (5.45)
\end{aligned}$$

Now, (5.38) follows from (5.41) and (5.45).

Step 2. Similar to the proof of Corollary 5.1, one can conclude that for n sufficiently large, there exists a pair of parameters $(\breve{k}, \breve{\lambda}) \in K_n \times \Lambda_n$ and a constant $B_1 > 0$ such that $\breve{k} = p_2$ and

$$B_1 \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}} \le \breve{\lambda} \le 2 \cdot B_1 \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}. \quad (5.46)$$

For all $(k, \lambda) \in K_n \times \Lambda_n$, define

$$H^{(2)}_{n,(k,\lambda)} := \left| V_{n,(k,\lambda)} - \int_{\mathbb{R}^d} \left|\sigma^2_{N_{\mathtt{l}},(k,\lambda)}(x) - \sigma^2(x)\right|^2 \mu(dx) \right|.$$

In analogy to (5.24), we have

$$\begin{aligned}
&\mathbf{P}\left[2\, V_{n,(\breve{k},\breve{\lambda})} - 3 \int_{\mathbb{R}^d} \left|\sigma^2_{N_{\mathtt{l}},(\breve{k},\breve{\lambda})}(x) - \sigma^2(x)\right|^2 \mu(dx) > \frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \right] \\
&\le \mathbf{P}\left[2\, H^{(2)}_{n,(\breve{k},\breve{\lambda})} > \int_{\mathbb{R}^d} \left|\sigma^2_{N_{\mathtt{l}},(\breve{k},\breve{\lambda})}(x) - \sigma^2(x)\right|^2 \mu(dx) + \frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \right] \\
&\le \mathbf{P}\left[\exists (k,\lambda) \in K_n \times \Lambda_n : 2\, H^{(2)}_{n,(k,\lambda)} > \int_{\mathbb{R}^d} \left|\sigma^2_{N_{\mathtt{l}},(k,\lambda)}(x) - \sigma^2(x)\right|^2 \mu(dx) + \frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \right] \\
&\le \sum_{(k,\lambda)\in K_n\times\Lambda_n} \mathbf{P}\left[2\, H^{(2)}_{n,(k,\lambda)} > \int_{\mathbb{R}^d} \left|\sigma^2_{N_{\mathtt{l}},(k,\lambda)}(x) - \sigma^2(x)\right|^2 \mu(dx) + \frac{(\ln n_{\mathtt{t}})^2}{n_{\mathtt{t}}} \right]. \quad (5.47)
\end{aligned}$$

If we set $N = n_{\mathtt{t}}$, $N_\ell = N_{\mathtt{r}}$, $N_{\mathtt{T}} = N_{\mathtt{t}}$, $\beta = L^2$, $K^\star_N = K_n$, $\Lambda^\star_N = \Lambda_n$, $Y^\star = U^{(2)} - m(X)^2$, $Y^\star_i = U^{(2)}_{i+n_{\mathtt{l}}} - m(X_{i+n_{\mathtt{l}}})^2$ $(i = 1, \ldots, N)$, $X^\star_i = X_{i+n_{\mathtt{l}}}$ $(i = 1, \ldots, N)$, and $\bar{Y}^{(N)}_i = \bar{U}_{i+n_{\mathtt{l}},N_{\mathtt{r}}}$

$(i = 1, \dots, N_\ell)$ in Section 5.1, then one can conclude in analogy to the first step of this proof that $m^\star$ equals σ^2 and $m^\star_{N_\ell,(k,\lambda)}$ equals $\sigma^2_{N_1,(k,\lambda)}$. Moreover, it holds that $\bar{\mathcal{D}}_{N_\ell} = \hat{\mathcal{D}}^{(2)}_{N_\mathtt{r}}$ and $H^\star_{N,(k,\lambda)} = H^{(2)}_{n,(k,\lambda)}$.

Therefore, Lemma 5.1, $\sigma^2(X) \in [0, L^2]$ a.s., (4.25), (5.32), (5.39) – (5.41), and (5.47) imply

$$\min_{(k,\lambda)\in K_n\times\Lambda_n} V_{n,(k,\lambda)} = \mathcal{O}_{\mathbf{P}}\left(\int_{\mathbb{R}^d} \left|\sigma^2_{N_1,(\check{k},\check{\lambda})}(x) - \sigma^2(x)\right|^2 \mu(dx) + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}\right). \tag{5.48}$$

Here, we used that the data $\hat{\mathcal{D}}^{(2)}_{N_\mathtt{r}}$ and the sequence of random vectors $(X, U^{(2)} - m(X)^2)$, $(X_{N_1+1}, U^{(2)}_{N_1+1} - m(X_{N_1+1})^2)$, …, $(X_n, U^{(2)}_n - m(X_n)^2)$ are independent (vide Step 1).

Step 3. In the third step of the proof, Lemma 3.1 and Theorem 4.1 will be applied in order to prove

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|\sigma^2_{N_1,(\check{k},\check{\lambda})}(x) - \sigma^2(x)\right|^2 \mu(dx)\\ &\qquad = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{N_\mathtt{r}} \sum_{i=n_1+1}^{N_1} |U^{(2)}_i - m(X_i)^2 - \bar{U}_{i,N_\mathtt{r}}|^2 + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}\right).\end{aligned} \tag{5.49}$$

Let $\epsilon > 0$ be arbitrary and set $B_2 := \epsilon + 2L^4$. First note that one can conclude similar to (4.27) from $m_{N_1}(X) \in [0, L]$ a.s., $m(X) \in [0, L]$ a.s. (see **(RA2)**), (2.68), and $(a+b)^2 \le 2a^2 + 2b^2$ $(a, b \in \mathbb{R})$ that

$$\begin{aligned}&\mathbf{P}\left[\max_{i=n_1+1,\dots,N_1} \left| U^{(2)}_i - m(X_i)^2 - \bar{U}_{i,N_\mathtt{r}} \right|^2 > B_2\right]\\ &= \mathbf{P}\left[\max_{i=n_1+1,\dots,N_1} \left| U^{(2)}_i - m(X_i)^2 - \left(\hat{U}^{(2)}_{i,N_\mathtt{r}} - m_{N_1}(X_i)^2\right) \right|^2 > B_2\right]\\ &\le \mathbf{P}\left[2 \max_{i=n_1+1,\dots,N_1} \left| U^{(2)}_i - \hat{U}^{(2)}_{i,N_\mathtt{r}} \right|^2 + 2 \max_{i=n_1+1,\dots,N_1} \left| m_{N_1}(X_i)^2 - m(X_i)^2 \right|^2 > \epsilon + 2L^4\right]\\ &\le \mathbf{P}\left[2 \max_{i=n_1+1,\dots,N_1} \left| U^{(2)}_i - \hat{U}^{(2)}_{i,N_\mathtt{r}} \right|^2 > \epsilon\right].\end{aligned} \tag{5.50}$$

Here, $\hat{U}^{(2)}_{n_1+1,N_\mathtt{r}}, \dots, \hat{U}^{(2)}_{N_1,N_\mathtt{r}}$ are defined by (2.64).

Now, let $\hat{U}^{(2)}_i$ $(i = 1, \dots, n)$ be given by (2.51). Since the random variables $\hat{U}^{(2)}_{i,N_\mathtt{r}}$ coincide with the random variables $\hat{U}^{(2)}_i$ if the latter ones are only computed on the data set (2.59) instead of the whole sample, Lemma 3.1, **(RA2)** – **(RA4)**, and (5.50) yield

$$\mathbf{P}\left[\max_{i=n_1+1,\dots,N_1} \left| U^{(2)}_i - m(X_i)^2 - \bar{U}_{i,N_\mathtt{r}} \right|^2 > B_2\right] \to 0 \quad (N_\mathtt{r} \to \infty). \tag{5.51}$$

Observe that for all $n \geq 3$, we have from (5.39) and (5.42)

$$n \geq N_{\mathtt{r}} \geq \frac{n}{4}, \tag{5.52}$$

which, in turn, implies

$$\left(\frac{(\ln N_{\mathtt{r}})^2}{N_{\mathtt{r}}}\right)^{\frac{2p_2}{2p_2+d}} \leq \left(4\,\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}. \tag{5.53}$$

From (5.46) and (5.53), one can conclude

$$\left(\frac{N_{\mathtt{r}}}{\ln N_{\mathtt{r}}}\right)^{\frac{2p_2}{2p_2+d}} \breve{\lambda} \geq B_1 \left(\frac{\ln N_{\mathtt{r}}}{4}\right)^{\frac{2p_2}{2p_2+d}} \to \infty \quad (N_{\mathtt{r}} \to \infty). \tag{5.54}$$

(cf. (5.22)).

Now note that Theorem 4.1 (where we set $n = N_{\mathtt{r}}$, $k = \breve{k}$, $\lambda_n = \breve{\lambda}$, $p = p_2$, $\beta^\star = \bar{L}_2$, $\beta = L^2$, $Y^\star = U^{(2)} - m(X)^2$, $Y_i^\star = U_{i+n_1}^{(2)} - m(X_{i+n_1})^2$ and $\bar{Y}_i^{(N)} = \bar{U}_{i+n_1,N_{\mathtt{r}}}$ $(i = 1, \ldots, n)$), $\sigma^2 \in W_{p_2}([0,1]^d)$ with $0 < J_{p_2}^2(\sigma^2) < \infty$, $\breve{k} = p_2$, $N_{\mathtt{r}} = N_{\mathtt{l}} - n_{\mathtt{l}}$, $L \geq 1$, **(RA1)**, **(RA2)**, (2.53), (4.25), (5.46), and (5.51) – (5.54) imply (5.49).

Step 4. In the fourth step of the proof, it is shown that

$$\begin{aligned}
&\frac{1}{N_{\mathtt{r}}} \sum_{i=n_1+1}^{N_1} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{r}}}|^2 + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2 \\
&\qquad = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 + n^{-\frac{2p_{max}}{2p_{max}+d}}\right).
\end{aligned} \tag{5.55}$$

Similar to (4.30), one can conclude from (2.68) and (2.69) with probability one

$$\begin{aligned}
&\frac{1}{N_{\mathtt{r}}} \sum_{i=n_1+1}^{N_1} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{r}}}|^2 + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2 \\
&\leq \frac{2}{N_{\mathtt{r}}} \sum_{i=n_1+1}^{N_1} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{r}}}^{(2)}|^2 + \frac{2}{N_{\mathtt{r}}} \sum_{i=n_1+1}^{N_1} |m_{N_1}(X_i)^2 - m(X_i)^2|^2 \\
&\quad + \frac{2}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{t}}}^{(2)}|^2 + \frac{2}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |m_{N_1}(X_i)^2 - m(X_i)^2|^2 \\
&\leq (2L)^2 \cdot \frac{2}{N_{\mathtt{r}}} \sum_{i=n_1+1}^{N_1} |m_{N_1}(X_i) - m(X_i)|^2 + (2L)^2 \cdot \frac{2}{N_{\mathtt{t}}} \sum_{i=N_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 \\
&\quad + 2 \max_{i=n_1+1,\ldots,N_1} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{r}}}^{(2)}|^2 + 2 \max_{i=N_1+1,\ldots,n} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{t}}}^{(2)}|^2.
\end{aligned} \tag{5.56}$$

Now note that $n_{\mathtt{t}} = n - n_{\mathtt{l}} \geq 2$ and (5.42) yield $N_{\mathtt{r}} \geq \frac{n_{\mathtt{t}}}{2}$ and

$$n \geq N_{\mathtt{t}} = n - N_{\mathtt{l}} = n_{\mathtt{t}} - N_{\mathtt{r}} = n_{\mathtt{t}} - \left\lceil \frac{n_{\mathtt{t}}}{2} \right\rceil \geq \frac{n_{\mathtt{t}}}{3}. \tag{5.57}$$

This, in turn, implies

$$\begin{aligned} &\frac{1}{N_{\mathtt{r}}} \sum_{i=n_{\mathtt{l}}+1}^{N_{\mathtt{l}}} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2 + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_{\mathtt{l}}+1}^{n} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2 \\ &\leq \frac{2}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{N_{\mathtt{l}}} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2 + \frac{3}{n_{\mathtt{t}}} \sum_{i=N_{\mathtt{l}}+1}^{n} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2 \\ &\leq \frac{3}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2. \end{aligned} \tag{5.58}$$

From (5.56) and (5.58), one gets with probability one

$$\begin{aligned} &\frac{1}{N_{\mathtt{r}}} \sum_{i=n_{\mathtt{l}}+1}^{N_{\mathtt{l}}} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{r}}}|^2 + \frac{1}{N_{\mathtt{t}}} \sum_{i=N_{\mathtt{l}}+1}^{n} |U_i^{(2)} - m(X_i)^2 - \bar{U}_{i,N_{\mathtt{t}}}|^2 \\ &\leq 2 \max_{i=n_{\mathtt{l}}+1,\ldots,N_{\mathtt{l}}} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{r}}}^{(2)}|^2 + 2 \max_{i=N_{\mathtt{l}}+1,\ldots,n} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{t}}}^{(2)}|^2 \\ &\quad + \frac{24L^2}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |m_{N_{\mathtt{l}}}(X_i) - m(X_i)|^2. \end{aligned} \tag{5.59}$$

Set $\gamma := \frac{2p_{max}}{2p_{max}+d}$. In the third step of this proof, we observed that the random variables $\hat{U}_{i,N_{\mathtt{r}}}^{(2)}$ coincide with the random variables $\hat{U}_i^{(2)}$ if the latter ones are only calculated on the data set (2.59) instead of the whole sample. Hence, Lemma 4.2, **(RA2)** – **(RA4)**, $\gamma \in (0,1)$, (4.23), and (4.31) imply that there exists some constant $b_3 \in (0,\infty)$ such that

$$\limsup_{N_{\mathtt{r}} \to \infty} N_{\mathtt{r}}^{\frac{2p_{max}}{2p_{max}+d}} \max_{i=n_{\mathtt{l}}+1,\ldots,N_{\mathtt{l}}} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{r}}}^{(2)}|^2 \leq b_3 \quad \text{a.s.}$$

This together with (5.52) yields

$$\max_{i=n_{\mathtt{l}}+1,\ldots,N_{\mathtt{l}}} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{r}}}^{(2)}|^2 = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p_{max}}{2p_{max}+d}}\right). \tag{5.60}$$

In analogy to (5.60), we deduce from Lemma 4.2, **(RA2)** – **(RA4)**, (2.51), (2.65), (4.23), (4.31), (5.39), and (5.57) that

$$\max_{i=N_{\mathtt{l}}+1,\ldots,n} |U_i^{(2)} - \hat{U}_{i,N_{\mathtt{t}}}^{(2)}|^2 = \mathcal{O}_{\mathbf{P}}\left(n^{-\frac{2p_{max}}{2p_{max}+d}}\right). \tag{5.61}$$

Now, (5.55) follows from (5.59) – (5.61).

Step 5. In the fifth step, we finish the proof of Theorem 5.3. From $p_2 \leq p_{max}$, (5.32), (5.38), (5.48), (5.49), and (5.55), one can conclude

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|\sigma_n^2(x) - \sigma^2(x)\right|^2 \mu(dx) \\ &= \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 + \left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_2}{2p_2+d}}\right) \end{aligned} \tag{5.62}$$

Since $m \in W_{p_1}([0,1]^d)$ with $0 < J_{p_1}^2(m) < \infty$, Lemma 5.2, **(RA1)** – **(RA4)**, (4.23), (4.35), $L \geq 1$, and $n_{\mathtt{t}} = n - n_{\mathtt{l}}$ imply

$$\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m_{N_1}(X_i) - m(X_i)|^2 = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p_1}{2p_1+d}}\right).$$

This together with (4.28) and (5.62) yields the assertion of Theorem 5.3.

□

5.4 Adaptive MSSE of the conditional survival function

In this section, we derive the rate of convergence of the MSSE (2.91), which is defined via the splitting of the sample technique (cf. Section 2.4).

Theorem 5.4. (Adaptation via splitting of the sample) *Let $d, n \in \mathbb{N}$ with $n \geq 2$ and let $\tau \in \mathbb{R}$ be arbitrary, but fixed. Set $n_{\mathtt{l}} := \lceil \frac{n}{2} \rceil$ and $n_{\mathtt{t}} := n - n_{\mathtt{l}}$. Moreover, let the set of parameters $K_n \times \Lambda_n$ be given by (2.41) and (2.42). Define the estimate $F_n(\tau\,|\,\cdot)$ by (2.89) – (2.92). For any $p \in \mathbb{N}$ with $2p > d$, we have*

$$\int_{\mathbb{R}^d} \left|F_n(\tau\,|\,x) - F(\tau\,|\,x)\right|^2 \mu(dx) = \mathcal{O}_{\mathbf{P}}\left(\left(\frac{(\ln n)^2}{n}\right)^{\frac{2p}{2p+d}}\right)$$

for every distribution of (X, Y, C) satisfying **(RA1)** – **(RA4)**, *$F(\tau\,|\,\cdot) \in W_p([0,1]^d)$ with $0 < J_p^2(F(\tau\,|\,\cdot)) < \infty$, and (4.12).*

Remark 5.6. Observe that the MSSE $F_n(\tau\,|\,\cdot)$ does not depend on p or $J_p^2(F(\tau\,|\,\cdot))$. Nevertheless, it achieves the same rate of convergence as the estimate in Theorem 4.4. Hence, $F_n(\tau\,|\,\cdot)$ adapts automatically to the unknown smoothness of $F(\tau\,|\,\cdot)$, which is measured by p and $J_p^2(F(\tau\,|\,\cdot))$.

Remark 5.7. In analogy to Remark 5.2, one can conclude that the assertion of Theorem 5.4 is also fulfilled for the estimate $\ddot{F}_n(\tau\,|\,\cdot) := F_{n_1,(\ddot{k}^{(3)},\ddot{\lambda}^{(3)})}(\tau\,|\,\cdot)$ with

$$\left(\ddot{k}^{(3)},\ddot{\lambda}^{(3)}\right) := \operatorname*{arg\,min}_{(k,\lambda)\in K_n\times\Lambda_n}\left(\frac{1}{n_{\mathtt{t}}}\sum_{i=n_1+1}^{n}|F_{n_1,(k,\lambda)}(\tau\,|\,X_i)-\hat{U}_i^{(3)}|^2\right).$$

Here $\hat{U}_i^{(3)}$ $(i=1,\ldots,n)$ and $F_{n_1,(k,\lambda)}(\tau\,|\,\cdot)$ are given by (2.51) and (2.90).

Proof of Theorem 5.4. In the following, we mimic the proof of Theorem 5.2. First, it is shown that the conditions of Corollary 5.1 hold.

Fix $\tau\in\mathbb{R}$. Furthermore, let $U^{(3)}$, $U_i^{(3)}$ $(i=1,\ldots,n)$, $\hat{U}_{i,n_1}^{(3)}$ $(i=1,\ldots,n_1)$, and $\hat{U}_{i,n_{\mathtt{t}}}^{(3)}$ $(i=n_1+1,\ldots,n)$ be defined by (2.77), (2.78), (2.84), and (2.85). Moreover, let $\hat{\mathcal{D}}_{n_1}^{(3)}$ and $\hat{\mathcal{D}}_{n_1,n_{\mathtt{t}}}^{(3)}$ be given by (2.86) and (2.88).

Obviously, we have with probability one that $F(\tau\,|\,X)=\mathbf{P}\,[\,Y>\tau\,|\,X\,]\in[0,1]$. In addition, note that (2.81) implies $|U^{(3)}|\le L_3^\star<\infty$ a.s., where $L_3^\star=\frac{1}{G(L)}\ge 1$.

Now set $N=n$, $N_\ell=n_1$, $N_{\mathtt{T}}=n_{\mathtt{t}}$, $\beta^\star=L_3^\star$, $\beta=1$, $K_N^\star=K_n$, $\Lambda_N^\star=\Lambda_n$, $Y^\star=U^{(3)}$, $(X_i^\star,Y_i^\star)=(X_i,U_i^{(3)})$ $(i=1,\ldots,N)$, and

$$\bar{Y}_i^{(N)}=\begin{cases}\hat{U}_{i,n_1}^{(3)} & \text{if}\quad i\in\{1,\ldots,N_\ell\}\\ \hat{U}_{i,n_{\mathtt{t}}}^{(3)} & \text{if}\quad i\in\{N_\ell+1,\ldots,N\}\end{cases}$$

in Section 5.1. From **(RA2)**, **(RA3)**, (2.80), (2.89) – (2.92), and (5.6) – (5.9), one can conclude similar to the proof of Theorem 4.2 that this implies that $m^\star(\cdot)$ equals $F\,(\,\tau\,|\,\cdot\,)$ and $m_N^\star(\cdot)$ equals $F_n\,(\,\tau\,|\,\cdot\,)$. Moreover, one gets $\bar{\mathcal{D}}_{N_\ell}=\hat{\mathcal{D}}_{n_1}^{(3)}$ and $\bar{\mathcal{D}}_N=\hat{\mathcal{D}}_{n_1,n_{\mathtt{t}}}^{(3)}$.

Note that (2.34) yields that the estimates (2.84) coincide with the estimates (2.79) if the latter are only calculated on the learning data (2.31) instead of the whole sample. Therefore, one gets from Lemma 3.1

$$\max_{i=1,\ldots,n_1}|U_i^{(1)}-\hat{U}_{i,n_1}^{(3)}|^2\to 0\quad(n_1\to\infty)\quad\text{a.s.}$$

and thus (5.16).

Furthermore, since (X,Y,C), (X_1,Y_1,C_1), …, (X_n,Y_n,C_n) are i.i.d. random vectors, one can conclude from (2.34), (2.35), (2.77), (2.78), (2.84) – (2.87) that $\hat{\mathcal{D}}_{n_1}^{(3)}$ and $(X,U^{(3)})$, $(X_{n_1+1},U_{n_1+1}^{(3)})$, …, $(X_n,U_n^{(3)})$ are independent as well as that $\hat{\mathcal{D}}_{n_1,n_{\mathtt{t}}}^{(3)}$ and $(X,U^{(3)})$ are independent. This, in turn, yields (5.11) and (5.12).

Hence, we have verified that the conditions of Corollary 5.1 are fulfilled. This together with (5.28) implies that in order to prove Theorem 5.4, it remains to show

$$\frac{1}{n_1}\sum_{i=1}^{n_1}|U_i^{(3)}-\hat{U}_{i,n_1}^{(3)}|^2=\mathcal{O}_{\mathbf{P}}\left(n_1^{-\frac{2p}{2p+d}}\right) \tag{5.63}$$

and

$$\frac{1}{n_t}\sum_{i=n_1+1}^{n}|U_i^{(3)}-\hat{U}_{i,n_t}^{(3)}|^2=\mathcal{O}_{\mathbf{P}}\left(n_t^{-\frac{2p}{2p+d}}\right). \tag{5.64}$$

Set $\gamma := \frac{2p}{2p+d}$. Since $p, d \in \mathbb{N}$, it holds that $\gamma \in (0,1)$. In analogy to above, one can conclude from (2.35) that the estimates (2.85) coincide with the estimates (2.79) if the latter are only calculated on the testing data (2.32). Thus, Lemma 4.2, (4.12), and (4.17) yield that there exists a constant $b_3 \in (0,\infty)$ such that

$$\limsup_{n_1\to\infty} n_1^{\frac{2p}{2p+d}}\frac{1}{n_1}\sum_{i=1}^{n_1}|U_i^{(3)}-\hat{U}_{i,n_1}^{(3)}|^2 \le \limsup_{n_1\to\infty} n_1^{\frac{2p}{2p+d}}\max_{i=1,\ldots,n_1}|U_i^{(3)}-\hat{U}_{i,n_1}^{(3)}|^2 \le b_3 \quad \text{a.s.}$$

and

$$\limsup_{n_t\to\infty} n_t^{\frac{2p}{2p+d}}\frac{1}{n_t}\sum_{i=n_1+1}^{n}|U_i^{(3)}-\hat{U}_{i,n_t}^{(3)}|^2 \le \limsup_{n_t\to\infty} n_t^{\frac{2p}{2p+d}}\max_{i=n_1+1,\ldots,n}|U_i^{(3)}-\hat{U}_{i,n_t}^{(3)}|^2 \le b_3 \quad \text{a.s.}$$

This together with (5.28) and (5.39) implies (5.63) and (5.64) and therefore the assertion of Theorem 5.4.

□

5.5 Proofs of Theorem 5.1 and Lemma 5.1

In this section, we show that the assertions of Theorem 5.1 and Lemma 5.1 hold. First, the proof of Lemma 5.1 is given. Subsequently, we apply this result in order verify Theorem 5.1.

Proof of Lemma 5.1. Assume that the conditions (5.10) and (5.11) are fulfilled and define $\bar{\mathcal{D}}_{N_\ell}$ by (5.2). For all $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$ set

$$g_{N_\ell,(k,\lambda)}(x) := |m^\star_{N_\ell,(k,\lambda)}(x) - m^\star(x)|^2 \quad (x \in [0,1]^d)$$

and $\varsigma^2_{(k,\lambda)} := \mathbf{Var}\left[g_{N_\ell,(k,\lambda)}(X)\middle|\bar{\mathcal{D}}_{N_\ell}\right]$. Observe that $0 \le m^\star(X) \le \beta$ a.s. and (5.7) imply with probability one

$$0 \le g_{N_\ell,(k,\lambda)}(X) \le \beta^2 \quad \forall\,(k,\lambda) \in K_N^\star \times \Lambda_N^\star. \tag{5.65}$$

Hence, one can conclude for all $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$

$$\varsigma_{(k,\lambda)}^2 \le \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X)^2 \middle| \bar{\mathcal{D}}_{N_\ell} \right] \le \beta^2 \, \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X) \middle| \bar{\mathcal{D}}_{N_\ell} \right] \quad \text{a.s.} \tag{5.66}$$

Since $(X, Y^\star)$, $(X_1^\star, Y_1^\star)$, …, $(X_N^\star, Y_N^\star)$ are i.i.d. random vectors, (5.6), (5.7), and assumption (5.11) imply for all $(k,\lambda) \in K_N^\star \times \Lambda_N^\star$ that $g_{N_\ell,(k,\lambda)}(X_{N_\ell+1}^\star)$, …, $g_{N_\ell,(k,\lambda)}(X_N^\star)$ are conditionally independent given $\bar{\mathcal{D}}_{N_\ell}$,

$$\mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X_i^\star) \middle| \bar{\mathcal{D}}_{N_\ell} \right] = \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X) \middle| \bar{\mathcal{D}}_{N_\ell} \right] \quad (i = N_\ell + 1, \dots, N),$$

and

$$\frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} \mathbf{Var}\left[g_{N_\ell,(k,\lambda)}(X_i^\star) \middle| \bar{\mathcal{D}}_{N_\ell} \right] = \varsigma_{(k,\lambda)}^2.$$

This together with (5.65), (5.66), and Bernstein's inequality (Lemma D.4) yields for all $t > 0$ with probability one

$$\begin{aligned}
&\sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[2\, H_{N,(k,\lambda)}^\star > \mathbf{E}\left[\left| m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X) \right|^2 \middle| \bar{\mathcal{D}}_{N_\ell} \right] + t \middle| \bar{\mathcal{D}}_{N_\ell} \right] \\
&= \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[\left| \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} g_{N_\ell,(k,\lambda)}(X_i^\star) - \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X) \middle| \bar{\mathcal{D}}_{N_\ell} \right] \right| \right. \\
&\qquad\qquad \left. > \frac{1}{2}\, \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X) \middle| \bar{\mathcal{D}}_{N_\ell} \right] + \frac{t}{2} \middle| \bar{\mathcal{D}}_{N_\ell} \right] \\
&\le \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[\left| \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} g_{N_\ell,(k,\lambda)}(X_i^\star) - \mathbf{E}\left[g_{N_\ell,(k,\lambda)}(X) \middle| \bar{\mathcal{D}}_{N_\ell} \right] \right| > \frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \middle| \bar{\mathcal{D}}_{N_\ell} \right] \\
&\le \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} 2 \exp\left(- \frac{3N_{\mathrm{T}} \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)^2}{6\varsigma_{(k,\lambda)}^2 + 2\beta^2 \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)} \right) \\
&\le 2\, |K_N^\star \times \Lambda_N^\star| \max_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \exp\left(- \frac{3N_{\mathrm{T}} \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)^2}{6\varsigma_{(k,\lambda)}^2 + 2\beta^2 \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)} \right) \\
&\le 2\, |K_N^\star \times \Lambda_N^\star| \max_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \exp\left(- \frac{3N_{\mathrm{T}} \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)^2}{(12\beta^2 + 2\beta^2) \left(\frac{\varsigma_{(k,\lambda)}^2}{2\beta^2} + \frac{t}{2} \right)} \right).
\end{aligned}$$

Hence, we may deduce for all $t > 0$

$$\sum_{(k,\lambda)\in K_N^\star\times\Lambda_N^\star} \mathbf{P}\left[2\, H^\star_{N,(k,\lambda)} > \mathbf{E}\left[\left| m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)\right|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell}\right] + t \,\middle|\, \bar{\mathcal{D}}_{N_\ell}\right]$$

$$\leq 2\, |K_N^\star \times \Lambda_N^\star| \max_{(k,\lambda)\in K_N^\star\times\Lambda_N^\star} \exp\left(- \frac{3N_{\mathtt{T}} \left(\frac{\varsigma^2_{(k,\lambda)}}{2\beta^2} + \frac{t}{2}\right)}{14\beta^2}\right)$$

$$\leq 2\, |K_N^\star \times \Lambda_N^\star| \exp\left(-3\, \frac{N_{\mathtt{T}}\, t}{28\beta^2}\right) \quad \text{a.s.} \tag{5.67}$$

Since $N_\ell \leq \lceil \frac{N}{2} \rceil$, we have for all $N \geq 2$

$$N_{\mathtt{T}} = N - N_\ell \geq N - \left\lceil \frac{N}{2}\right\rceil \geq \frac{N}{3} \tag{5.68}$$

(cf. (5.28)), which, in turn, implies for all sufficiently large N that

$$\exp\left(3\, \frac{N_{\mathtt{T}}}{28\beta^2} \cdot \frac{(\ln N)^2}{N}\right) \geq \exp\left(\frac{(\ln N)^2}{28\beta^2}\right) \geq \exp(2 \ln N) = N^2. \tag{5.69}$$

From (5.67) and (5.69), one can conclude

$$\sum_{(k,\lambda)\in K_N^\star\times\Lambda_N^\star} \mathbf{P}\left[2\, H^\star_{N,(k,\lambda)} > \mathbf{E}\left[\left| m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)\right|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell}\right] + \frac{(\ln N)^2}{N}\right]$$

$$\leq 2\, \frac{|K_N^\star \times \Lambda_N^\star|}{N^2}$$

for all sufficiently large N.

This together with (5.10) implies the assertion of Lemma 5.1.

□

Now we are in the position to prove Theorem 5.1. Here, Lemma A.2 will be applied, which is formulated and verified in Appendix A.

Proof of Theorem 5.1. Assume that the conditions (5.11) and (5.12) hold. Let $N \geq 2$ and set

$$\mathcal{F}_{K_N^\star\times\Lambda_N^\star} := \left\{ m^\star_{N_\ell,(k,\lambda)} : (k,\lambda) \in K_N^\star \times \Lambda_N^\star \right\}.$$

First observe that (5.8) and (5.9) imply

$$m_N^\star(\cdot) = \operatorname*{arg\,min}_{f\in\mathcal{F}_{K_N^\star\times\Lambda_N^\star}} \frac{1}{N_{\mathtt{T}}} \sum_{i=N_\ell+1}^{N} |f(X_i^\star) - \bar{Y}_i^{(N)}|^2. \tag{5.70}$$

Define

$$A_N^\star := 18 \min_{(k,\lambda)\in K_N^\star\times\Lambda_N^\star} \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(k,\lambda)}(X_i^\star) - m^\star(X_i^\star)|^2 + \frac{512}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |Y_i^\star - \bar{Y}_i^{(N)}|^2.$$

Assumption (5.11), Lemma A.2, and (5.70) yield

$$\begin{aligned}
&\mathbf{P}\left[\int_{\mathbb{R}^d} \left|m_N^\star(x) - m^\star(x)\right|^2 \mu(dx) > 2A_N^\star + 3\frac{(\ln N)^2}{N}\right] \\
&\quad\le \mathbf{P}\left[\int_{\mathbb{R}^d} \left|m_N^\star(x) - m^\star(x)\right|^2 \mu(dx) - \frac{2}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_N^\star(X_i^\star) - m^\star(X_i^\star)|^2 > \frac{(\ln N)^2}{N}\right] \\
&\qquad + \mathbf{P}\left[\frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_N^\star(X_i^\star) - m^\star(X_i^\star)|^2 > A_N^\star + \frac{(\ln N)^2}{N}\right] \\
&\quad\le \mathbf{P}\left[\int_{\mathbb{R}^d} \left|m_N^\star(x) - m^\star(x)\right|^2 \mu(dx) - \frac{2}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_N^\star(X_i^\star) - m^\star(X_i^\star)|^2 > \frac{(\ln N)^2}{N}\right] \\
&\qquad + \frac{2|K_N^\star\times\Lambda_N^\star|}{\exp\left(b_7 N_{\mathrm{T}} \frac{(\ln N)^2}{N}\right) - 1} \\
&\quad=: q_N + \frac{2|K_N^\star\times\Lambda_N^\star|}{\exp\left(b_7 N_{\mathrm{T}} \frac{(\ln N)^2}{N}\right) - 1}.
\end{aligned} \tag{5.71}$$

Here, $b_7 > 0$ is the constant in Lemma A.2 which only depends on $\beta^\star$.

For all sufficiently large N, one can conclude from (5.68) for the last term on the right hand side of (5.71)

$$\frac{2\,|K_N^\star\times\Lambda_N^\star|}{\exp\left(b_7 N_{\mathrm{T}} \frac{(\ln N)^2}{N}\right) - 1} \le \frac{2\,|K_N^\star\times\Lambda_N^\star|}{\exp\left(2\ln N + \ln 2\right) - 1} = \frac{2\,|K_N^\star\times\Lambda_N^\star|}{2N^2 - 1} \le \frac{2\,|K_N^\star\times\Lambda_N^\star|}{N^2}.$$

This together with (5.10) and (5.71) implies that it remains to show

$$q_N \to 0 \quad (N\to\infty). \tag{5.72}$$

In the following, Lemma 5.1 will be applied. Let the data sets $\bar{\mathcal{D}}_N$, $\bar{\mathcal{D}}_{N_\ell}$, and $\bar{\mathcal{D}}_{N_{\mathrm{T}}}$ be given by (5.1) – (5.3). For each $(k,\lambda)\in K_N^\star\times\Lambda_N^\star$ define

$$H^\star_{N,(k,\lambda)} := \left| V^\star_{N,(k,\lambda)} - \mathbf{E}\left[\, |m^\star_{N_\ell,(k,\lambda)}(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell}\right]\right|,$$

where

$$V^\star_{N,(k,\lambda)} := \frac{1}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m^\star_{N_\ell,(k,\lambda)}(X_i^\star) - m^\star(X_i^\star)|^2.$$

From (5.11) and (5.12), we deduce that $\bar{\mathcal{D}}_{N_{\mathrm{T}}}$ and $(X, Y^\star)$ are conditionally independent given $\bar{\mathcal{D}}_{N_\ell}$. This together with (5.8) and properties of conditional expectation yields

$$
\begin{aligned}
q_N &= \mathbf{P}\left[\int_{\mathbb{R}^d} \left|m_N^\star(x) - m^\star(x)\right|^2 \mu(dx) - \frac{2}{N_{\mathrm{T}}} \sum_{i=N_\ell+1}^{N} |m_N^\star(X_i^\star) - m^\star(X_i^\star)|^2 > \frac{(\ln N)^2}{N}\right] \\
&= \mathbf{P}\left[\mathbf{E}\left[\, |m_N^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_N \,\right] - 2V_{N,(k^\star,\lambda^\star)}^\star > \frac{(\ln N)^2}{N}\right] \\
&\le \mathbf{P}\left[\exists (k,\lambda) \in K_N^\star \times \Lambda_N^\star : \mathbf{E}\left[\, |m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_N \,\right] - 2V_{N,(k,\lambda)}^\star > \frac{(\ln N)^2}{N}\right] \\
&\le \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[\mathbf{E}\left[\, |m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_N \,\right] - 2V_{N,(k,\lambda)}^\star > \frac{(\ln N)^2}{N}\right] \\
&= \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[\mathbf{E}\left[\, |m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \,\right] - 2V_{N,(k,\lambda)}^\star > \frac{(\ln N)^2}{N}\right] \\
&\le \sum_{(k,\lambda)\in K_N^\star \times \Lambda_N^\star} \mathbf{P}\left[2H_{N,(k,\lambda)}^\star > \mathbf{E}\left[\, |m_{N_\ell,(k,\lambda)}^\star(X) - m^\star(X)|^2 \,\middle|\, \bar{\mathcal{D}}_{N_\ell} \,\right] + \frac{(\ln N)^2}{N}\right].
\end{aligned}
$$

Finally, the last inequality and Lemma 5.1 imply (5.72).

□

Chapter 6

Applications to simulated data

In this chapter, we analyze the performances of the MSSE (2.45), (2.73), and (2.91) in a simulation study. Section 6.1 introduces our choice of the distribution of (X, Y, C). Here, it is shown that this choice assures that the assumptions of Theorems 5.2 – 5.4 on the distribution of (X, Y, C) are fulfilled, and the suitably defined MSSE (2.45), (2.73), and (2.91) hence achieve their optimal rate of convergence up to some logarithmic factor. In the three subsequent sections, the results of the simulation study are discussed, which suggest that our MSSE are reliable estimates, which perform well for moderate sample sizes. Finally, Section 6.5 contains the proofs of two lemmata presented in Section 6.1.

All results of the simulation study (including pictures) presented in this chapter were performed by means of statistical software R-2.6.2 (www.r-project.org), in which simulations were based on self-written functions.

6.1 Simulation model

This section establishes the specific choice of the distribution of $(X, Y, C) \in \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}_+$ which is applied in the simulation studies presented in Sections 6.2 – 6.4. In particular, for the lifetime Y with $\mathbf{E}Y^4 < \infty$ it is assumed that the *heteroscedastic model*

$$Y = m(X) + \sigma(X) \cdot \epsilon \tag{6.1}$$

holds, where the so-called error term ϵ is a real-valued random variable with $\mathbf{E}\epsilon^4 < \infty$. In order to ensure that the functions m and σ in (6.1) are actually identical to the regression

function of (X, Y) and the conditional standard deviation of Y given X, respectively, we require that $\mathbf{E}[\,\epsilon \,|\, X\,] = 0$ and $\mathbf{Var}[\,\epsilon \,|\, X\,] = 1$.

Indeed, (6.1) and the last two conditions imply that with probability one

$$\mathbf{E}[\,Y \,|\, X\,] = \mathbf{E}[\,m(X) + \sigma(X) \cdot \epsilon \,|\, X\,] = m(X) + \sigma(X) \cdot \mathbf{E}[\,\epsilon \,|\, X\,] = m(X)$$

and

$$\begin{aligned} \mathbf{Var}[\,Y \,|\, X\,] &= \mathbf{E}\left[\, |\, Y - m(X)\,|^2 \,\middle|\, X\,\right] = \mathbf{E}\left[\, |\, \sigma(X) \cdot \epsilon\,|^2 \,\middle|\, X\,\right] \\ &= \sigma^2(X) \cdot \mathbf{E}\left[\, \epsilon^2 \,\middle|\, X\,\right] = \sigma^2(X) \cdot \mathbf{Var}[\,\epsilon \,|\, X\,] \\ &= \sigma^2(X). \end{aligned}$$

The term heteroscedastic model refers to all models allowing for a non-constant conditional variance of the dependent variable. A broad survey of different types of such models, in particular in the context of parametric regression, can be found in Carroll and Ruppert (1988) and the literature cited therein. The construction of nonparametric estimates of the regression function or the conditional variance within heteroscedastic models is, e.g., described in Müller and Stadtmüller (1987), Carroll and Hall (1989), Neumann (1994), Stadtmüller and Tsybakov (1995), Wang, Brown, Cai, and Levine (2008), and Cai, Levine, and Wang (2009).

For our simulation study, we now choose a specific form of m and σ in (6.1) as well as the distributions of X, C, and the error term ϵ. To be more precise, in this chapter, it is assumed that the following conditions on the distribution of (X, Y, C) are fulfilled:

(SM1) The heteroscedastic model (6.1) holds, where

- **(SM1a)** $m(X) = \frac{7}{8} - \frac{1}{3}\,(2X-1)^4 - 48\,(X\,(1-X))^{\frac{7}{2}}$,
- **(SM1b)** $\sigma(X) = \frac{\sqrt{3}}{24} + 16\,\sqrt{3}\,(X\,(1-X))^{\frac{7}{2}}$,
- **(SM1c)** ϵ is uniformly distributed on $\left[\,-\sqrt{3}, \sqrt{3}\,\right]$.

(SM2) X is uniformly distributed on $[\,0, 1\,]$.

(SM3) C is exponentially distributed with mean $\frac{4}{3}$.

(SM4) C, X, and ϵ are mutually independent.

From **(SM1c)** and **(SM4)**, one can conclude that

$$\mathbf{E}\,[\,\epsilon\,|\,X\,] = \mathbf{E}\,\epsilon = \frac{1}{2}\left(\sqrt{3}-\sqrt{3}\right) = 0 \quad \text{and} \quad \mathbf{Var}\,[\,\epsilon\,|\,X\,] = \mathbf{Var}\,\epsilon = \frac{1}{12}\left|\sqrt{3}+\sqrt{3}\right|^2 = 1.$$

I.e., our conditions on the first and second conditional moment of ϵ are satisfied. Moreover, note that **(SM1)** – **(SM4)** imply that in our simulation study, on average approximately 37.7% of the data is censored.

In the next proposition, it is shown the distribution of (X, Y, C) specified above meets the regularity assumptions **(RA1)** – **(RA4)**.

Proposition 6.1. *Let the conditions* **(SM1)** – **(SM4)** *hold. Then the distribution of* (X, Y, C) *fulfills the regularity conditions* **(RA1)** – **(RA4)**.

Proof of Proposition 6.1. Obviously, **(RA1)** is implied by **(SM2)** (with $d = 1$). Furthermore, **(RA3)** directly follows from (6.1) and **(SM4)**. Moreover, **(SM3)** yields **(RA4)**. In order to show regularity assumption **(RA2)**, first observe that for all $x \in [0,1]$, we have from **(SM1a)** and **(SM1b)**

$$\frac{2}{3} \leq M^+(x) := m(x) + \sqrt{3}\cdot\sigma(x) = 1 - \frac{1}{3}\,(2x-1)^4 \leq 1 \tag{6.2}$$

and

$$0 \leq M^-(x) := m(x) - \sqrt{3}\cdot\sigma(x) \leq M^-_{max} := \max_{x\in[0,1]} M^-(x) \approx 0.603. \tag{6.3}$$

Now, (6.1) – (6.3), **(SM1c)**, and **(SM2)** imply with probability one

$$0 \leq M^-(X) \leq Y \leq M^+(X) \leq 1. \tag{6.4}$$

This together with **(SM3)** yields that **(RA2)** is fulfilled with any $L \in [1, \infty)$.

□

Let $\alpha_1, \alpha_2 \in \mathbb{R}$ and $k, k_1, k_2 \in \mathbb{N}$. with $2k, 2k_1, 2k_2 > 1$. For $n \in \mathbb{N}$ choose the smoothing parameters $\lambda_n, \lambda_{1,n}, \lambda_{2,n} > 0$ such that the conditions (3.19), (3.20), and (3.35) are satisfied. Moreover, let $\tau \in \mathbb{R}$ be arbitrary, but fixed. Then one can conclude from Theorem 3.2, Theorem 3.3, and Theorem 3.4 that for the distribution of (X, Y, C) defined by **(SM1)** – **(SM4)**, the MSSE (2.28), (2.57), and (2.83) are strongly consistent.

In the following, we derive the explicit form of the conditional survival function $F\,(t|x)$ $(t \in \mathbb{R},\, x \in [0,1])$. First observe that **(SM1b)** yields $\sigma(x) > 0\ \forall x \in [0,1]$. Set

$$R(t,x) := \frac{t - m(x)}{\sigma(x)} \quad (t \in \mathbb{R}, x \in [0,1]). \tag{6.5}$$

Now, **(SM4)**, (6.1), (6.2), and (6.5) imply for all $t \in \mathbb{R}$ and all $x \in [0,1]$

$$\begin{aligned} F(t|x) &= \mathbf{P}[\, Y > t \mid X = x \,] = \mathbf{P}[\, m(X) + \sigma(X) \cdot \epsilon > t \mid X = x \,] \\ &= \mathbf{P}[\, \epsilon > R(t,X) \mid X = x \,] = \int_{R(t,x)}^{\sqrt{3}} \frac{1}{2\sqrt{3}} I_{[-\sqrt{3} \le s \le \sqrt{3}]}\, ds \\ &= I_{[R(t,x) < -\sqrt{3}]} + I_{[-\sqrt{3} \le R(t,x) \le \sqrt{3}]} \cdot \frac{1}{2\sqrt{3}} \left(\sqrt{3} - R(t,x)\right) \\ &= I_{[R(t,x) < -\sqrt{3}]} + I_{[-\sqrt{3} \le R(t,x) \le \sqrt{3}]} \cdot \frac{M^+(x) - t}{2\sqrt{3}\,\sigma(x)}. \end{aligned} \tag{6.6}$$

Now note that from (6.2) and (6.3), we have for all $t \in \left[M_{max}^-, \frac{2}{3}\right]$ and all $x \in [0,1]$

$$m(x) - \sqrt{3} \cdot \sigma(x) \le t \le m(x) + \sqrt{3} \cdot \sigma(x)$$

and therefore $-\sqrt{3} \le R(t,x) \le \sqrt{3}$. This together with (6.6) yields

$$F(t|x) = \frac{M^+(x) - t}{2\sqrt{3}\,\sigma(x)} \quad \left(M_{max}^- \le t \le \frac{2}{3}, x \in [0,1]\right). \tag{6.7}$$

In the next two lemmata, we show that in the scenario of our simulation study, the conditions of Theorems 5.2 – 5.4 on the distribution of (X,Y,C) are fulfilled with $p = 3$ (or $p_1 = p_2 = 3$) and $d = 1$. The first lemma states that m, σ^2, and $F(\tau|\cdot)$ ($\tau \in \left[M_{max}^-, \frac{2}{3}\right)$ fixed) are functions in the Sobolev space $W_3([0,1])$.

Lemma 6.1. *Let M_{max}^- be given by (6.3). Furthermore, let $\tau \in \left[M_{max}^-, \frac{2}{3}\right)$ be arbitrary, but fixed. If the conditions* **(SM1)** – **(SM4)** *hold, then we have*

1. $m \in W_3([0,1])$ *with* $0 < J_3^2(m) < \infty$, *but* $m \notin W_k([0,1])$ $\forall k \ge 4$.
2. $\sigma^2 \in W_3([0,1])$ *with* $0 < J_3^2(\sigma^2) < \infty$, *but* $\sigma^2 \notin W_k([0,1])$ $\forall k \ge 4$.
3. $F(\tau|\cdot) \in W_3([0,1])$ *with* $0 < J_3^2(F(\tau|\cdot)) < \infty$, *but* $F(\tau|\cdot) \notin W_k([0,1])$ $\forall k \ge 4$.

The proof of Lemma 6.1 is given in Section 6.5.

Let $k \in \mathbb{N}$ and let $\tau \in \left[M_{max}^-, \frac{2}{3}\right)$ be arbitrary, but fixed. The Sobolev space $W_k([0,1])$ contains all functions whose weak derivatives up to order k are contained in $\mathcal{L}_2([0,1])$ (see (2.3)). This implies that $W_k([0,1]) \subseteq W_\kappa([0,1])$ for all $\kappa \in \mathbb{N}$ with $\kappa \le k$. Therefore, one can conclude from Lemma 6.1 that if the conditions **(SM1)** – **(SM4)** are fulfilled, $m, \sigma^2, F(\tau|\cdot) \in W_\kappa([0,1])$ for all $\kappa \in \{1,2,3\}$.

In the second lemma of this section, we investigate the conditions (4.12) and (4.23), which determine the rate of convergence of the mean squared transformation errors (cf. Lemma 4.2).

Lemma 6.2. *If the conditions* **(SM1)** – **(SM4)** *hold, then one gets* $\tau_F = 1$ *and*

$$-\int_0^1 F(t)^{-\frac{3}{4}}\, dG(t) < \infty.$$

The proof of Lemma 6.2 is given in Section 6.5.

Let $n \in \mathbb{N}$ with $n \geq 3$, $L \geq 1$, $\alpha_1, \alpha_2 \in \mathbb{R}$, and $\tau \in \left[M_{max}^-, \frac{2}{3}\right)$. Theorems 5.2 – 5.4, Lemma 6.1, and Lemma 6.2 imply that for the distribution of (X, Y, C) defined by **(SM1)** – **(SM4)**, the rate of stochastic convergence of the MSSE (2.45), (2.73), and (2.91) (with $n_{\mathtt{l}}$, $n_{\mathtt{t}}$, $N_{\mathtt{l}}$, $N_{\mathtt{r}}$, and $N_{\mathtt{t}}$ chosen according to Theorems 5.2 – 5.4) is given by

$$\left(\frac{(\ln n)^2}{n}\right)^{\frac{6}{7}}, \tag{6.8}$$

which is optimal up to the logarithmic factor.

However, in statistical applications, such asymptotic results are often secondary, since in general, it is unknown whether they are valid for or they do not apply to a given finite sample size. In order to assess the accuracy and the precision of our MSSE on data with a moderate sample size n, the estimates (2.45), (2.73), and (2.91) are therefore analyzed below on simulated data sets with $n = 200$.

6.2 Results for MSSE of the regression function

Throughout this and the two subsequent sections of this chapter, we will assume that (X, Y, C), (X_1, Y_1, C_1), …, (X_n, Y_n, C_n) are i.i.d. random vectors, whose distribution is given by **(SM1)** – **(SM4)**. For a single simulation run and for each (X_i, Y_i, C_i), a realization (x_i, y_i, c_i) $(i = 1, \ldots, n)$ is generated. According to (1.1) and (1.2), we then compute the realizations (Δ_i, z_i) of (δ_i, Z_i), where $\Delta_i := I_{[y_i < c_i]}$ and $z_i := \min\{y_i, c_i\}$ $(i = 1, \ldots, n)$.

As mentioned above, in each simulation run a sample size of $n = 200$ is chosen. Furthermore, we set the "known upper bound" on the distribution of Y in assumption **(RA2)** to $L = 1$. However, for all estimates presented in this chapter, the effect of the choice of L (with $L \geq 1$) on the results is negligible.

In order to calculate the MSSE m_n, the 200 data points of a single simulation run are first randomly split into a learning data set and a testing data set with sample sizes $n_{\mathtt{l}} = n_{\mathtt{t}} = 100$ (cf. Section 2.4). For each of these sets, we then compute the estimates of the transformed data according to (2.36) and (2.37).

Observe that these data still depend on $\alpha_1 \in \mathbb{R}$. This parameter provides the possibility to improve the quality of the transformation and therefore the estimation of m. Let the transformed random variables $U_i^{(1)}$ $(i = 1, \ldots, n)$ be given by (2.22). As pointed out by Fan and Gijbels (1994, 1996), the choice $\alpha_1 > 0$ is more intuitive, because it rather focuses on the censored than on the uncensored observations. On the other hand, if α_1 is too large, then we have $U_i^{(1)} \geq Z_i$ a.s. if $\delta_i = 0$ but $U_i^{(1)} \leq Z_i$ a.s. if $\delta_i = 1$ $(i = 1, \ldots, n)$. Hence, this would lead to an increased variability of the transformed random variables. Likely, this would deteriorate the performance of the estimation of m by any regression estimate which is based on this sample.

Thus, Fan and Gijbels (1994, 1996) suggested to take the largest α_1 in a data-dependent way such that the observed lifetimes do not exceed the corresponding estimates of the transformed times. Since (2.36) and (2.37) are calculated separately, we adapt their proposition to our situation by computing their choice of α_1 only on the learning data, i.e., set

$$\alpha_1^{FG} := \alpha_1^{FG}(\mathcal{D}_{n_1}^{(1)}) := \min_{\substack{i=1,\ldots,n_1: \\ \delta_i = 1}} \frac{\displaystyle\int_0^{Y_i} \frac{1}{G_{n_1}(t)}\, dt - Y_i}{\displaystyle\frac{Y_i}{G_{n_1}(Y_i)} - \int_0^{Y_i} \frac{1}{G_{n_1}(t)}\, dt}. \tag{6.9}$$

Here, $\mathcal{D}_{n_1}^{(1)}$ and G_{n_1} are given by (2.31) and (2.34). In order to guarantee that α_1^{FG} is always calculable, we require that the learning data contains at least one uncensored observation.

Observe that (2.34) implies $G_{n_1}(a) \leq G_{n_1}(t) \leq 1$ a.s. for all $0 \leq t \leq a$ and therefore

$$Y_i \leq \int_0^{Y_i} \frac{1}{G_{n_1}(t)}\, dt \leq \frac{Y_i}{G_{n_1}(Y_i)} \quad \text{a.s.} \quad (i = 1, \ldots, n).$$

This yields with probability one that $\alpha_1^{FG} \geq 0$ and hence that the estimates of the transformed times do not fall below the corresponding censoring times.

Once the value of α_1^{FG} is determined, we apply this choice to the computation of the transformed random variables (2.37). Note that while (6.9) ensures that $\hat{U}_{i,n_1}^{(1)} \geq Y_i$ for all $i = 1, \ldots, n_1$ with $\delta_i = 1$, it may happen that there exists some $i \in \{n_1 + 1, \ldots, n\}$ with $\delta_i = 1$ such that $\hat{U}_{i,n_{\mathtt{t}}}^{(1)} < Y_i$. As mentioned above, this may somewhat increase the

variability of the transformed observations in the testing data. One way to avoid this would be to calculate α_1^{FG} according to Fan and Gijbels (1994, 1996) on the whole data set. But in this case, one would use information from the testing set (2.32) in order to calculate the random variables (2.36) in the learning set (2.38). To examine the performance of our regression estimate of m, we rather like to transform the learning data (2.31) as if (2.32) would be a new sample of (X, δ, Z) which is not available in this step (cf. Section 2.4).

In order to ensure that (2.38) and (2.39) do not depend on each other, one could choose $\alpha_1 = \alpha_1^{FG}(\mathcal{D}_{n_1}^{(1)})$ in (2.36) and $\alpha_1 = \alpha_1^{FG}(\mathcal{D}_{n_t}^{(1)})$ in (2.37), where $\alpha_1^{FG}(\mathcal{D}_{n_t}^{(1)})$ is given similar to (6.9) with $\mathcal{D}_{n_1}^{(1)}$ replaced by $\mathcal{D}_{n_t}^{(1)}$. Since we randomly split the data set in (2.31) and (2.32), $\alpha_1^{FG}(\mathcal{D}_{n_1}^{(1)})$ and $\alpha_1^{FG}(\mathcal{D}_{n_t}^{(1)})$ should not differ very much. However, especially for small sample sizes, it may nevertheless happen that this difference is so large that this would considerably diminish the accuracy of any regression estimate.

Other suggestions in order to choose α_1, which do not depend on the data can, e.g., be found in Leurgans (1987) or Koul, Susarla, and Van Ryzin (1981) which propose to use $\alpha_1 = 0$ and $\alpha_1 = -1$, respectively. Note that $\alpha_1 = 0$ implies that uncensored and censored observations are treated equally. In contrast, the choice $\alpha_1 = -1$ yields that all censored observations are set to zero, while all the uncensored observations are increased.

Once the value of the transformation parameter is determined, we compute the MSSE m_n according to Section 2.4 on the realizations of the transformed learning and testing data. In order to analyze the effect of the censoring mechanism on the estimation of m, a second MSSE m_n^{UD} is calculated on the realizations of the (in an statistical application not observable) uncensored data. This estimate is derived by simply replacing $\mathcal{D}_{n_1}^{(1)}$ and $\mathcal{D}_{n_t}^{(1)}$ in the definition of m_n with $\{(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})\}$ and $\{(X_{n_1+1}, Y_{n_1+1}), \ldots, (X_n, Y_n)\}$, respectively. I.e., m_n is identical to m_n^{UD} if no censoring arises. Observe that m_n^{UD} does not depend on the transformation parameter α_1.

For a single simulated data set, Figure 6.1 (a) shows the (in an application) unobserved data together with the regression function m (vide **(SM1a)**), Figure 6.1 (b) the observed censored data with censored data points marked by "+", Figure 6.1 (c) the transformed data with $\alpha_1 = \alpha_1^{FG} \approx 0.44$, and Figure 6.1 (d) m (green solid line), the MSSE m_n of m based on the transformed data (blue dotted line), and the in an application not calculable MSSE m_n^{UD} of m based on the unobserved data (red dashed line).

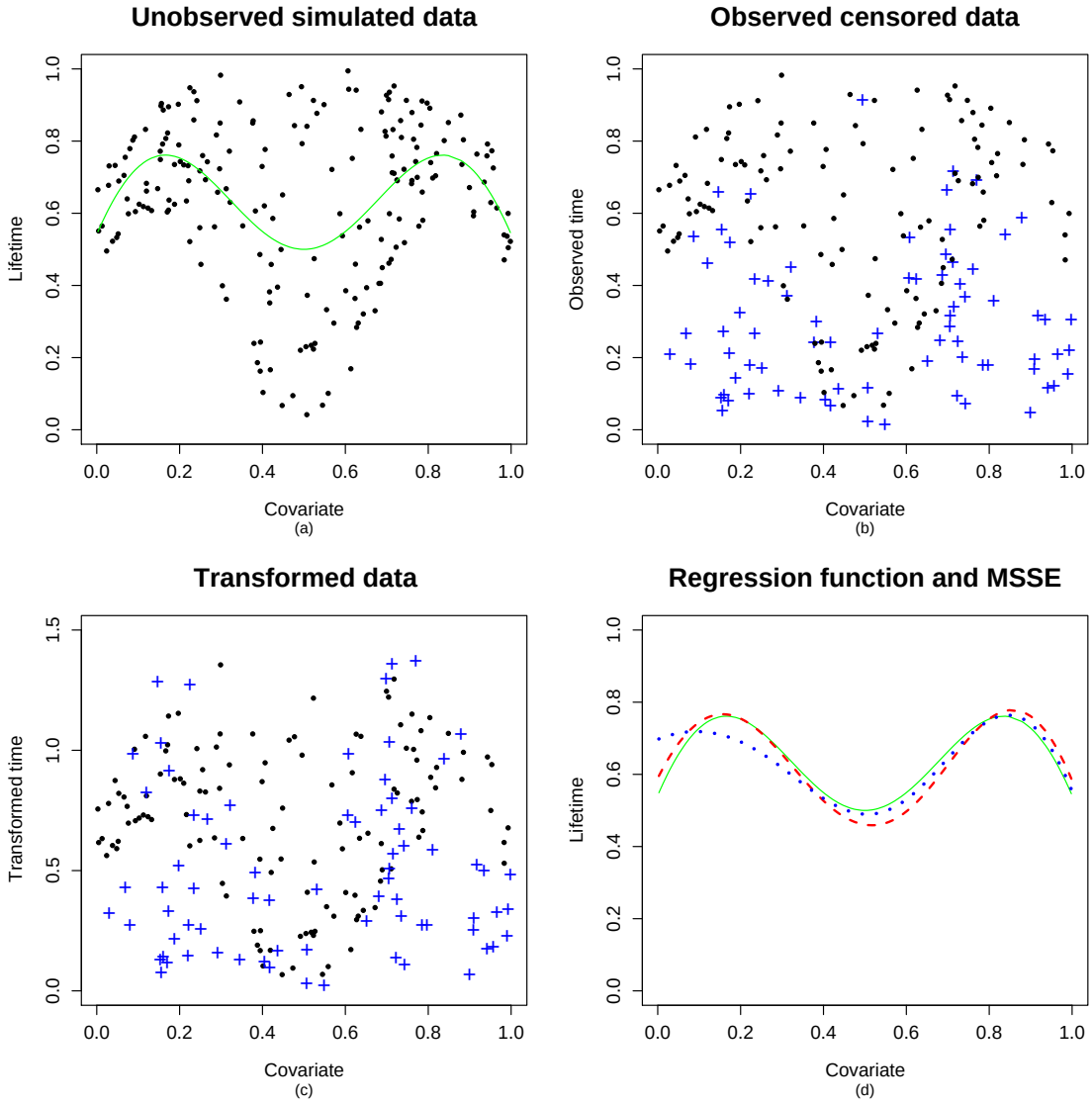

Figure 6.1: Data set for a single simulation run. The character "•" indicates the uncensored observations, while the censored observations are presented by "+". The same characters are used for the transformed data. The individual panels show (a) the unobserved data together with m (green solid line), (b) the observed simulated data, (c) the transformed data with $\alpha_1 = \alpha_1^{FG} \approx 0.44$, and (d) m (green solid line) with the MSSE m_n (blue dotted line) and m_n^{UD} (red dashed line).

Figure 6.1 (d) indicates that even for the small to moderate sample size of $n = 200$ chosen, both nonparametric regression estimates perform very well in our simulation study.

For the simulated data set displayed in Figure 6.1, the empirical $\mathcal{L}_2$ error of m_n^{UD} is approximately $1.5 \cdot 10^{-3}$, while the empirical $\mathcal{L}_2$ error of m_n (with $\alpha_1 = \alpha_1^{FG} \approx 0.44$) is about $5.2 \cdot 10^{-3}$. It is evident that the MSSE m_n^{UD}, which is solely computed on the uncensored observations, will commonly be an better estimate of m (in terms of the $\mathcal{L}_2$ error) than m_n. However, in an application where censored data occurs, m_n^{UD} is not calculable.

According to Lemma 6.1, m, σ^2, and $F(\tau|\cdot)$ ($M_{max}^- \leq \tau < \frac{2}{3}$ fixed) are functions in the Sobolev space $W_3([0,1])$. All the MSSE presented in this and the two subsequent sections (for the MSSE of $F(\tau|\cdot)$ provided that $M_{max}^- \leq \tau < \frac{2}{3}$) typically choose the value of the parameter k (or k_1 and k_2 for the estimation of σ^2) via the splitting of the sample technique to $k = 3$ (or $k_1 = k_2 = 3$). I.e., in general, our estimates adapt automatically to the (in an application unknown) smoothness of m, σ^2, and $F(\tau|\cdot)$, respectively, as indicated by Theorems 5.2 – 5.4.

In order to determine the influence of the parameter α_1 on the estimation of the regression function m, we generated 50 independent samples (with sample size $n = 200$) of the distribution of (X, Y, C) given by **(SM1)** – **(SM4)**. For each sample, the empirical $\mathcal{L}_2$ error of m_n for six different choices of α_1 was calulated. For our first version of m_n, $\alpha_1 = \alpha_1^{FG}$ was considered, where α_1^{FG} is given by (6.9). Secondly, we chose $\alpha_1 = \frac{14}{25}$, which corresponds to the median of the α_1^{FG}'s calculated on the learning data (2.31) for the 50 independently repeated simulation runs. Thirdly, $\alpha_1 = 0$ was taken (cf. Leurgans (1987)). In order to cover a wider range, these versions of m_n were compared with three others, where we considered values of α_1 which lay somewhere between ($\alpha_1 = \frac{1}{3}$) or above ($\alpha_1 = \frac{2}{3}$ and $\alpha_1 = \frac{3}{4}$) the choices mentioned before. Furthermore, for each of the 50 samples, the empirical $\mathcal{L}_2$ error of the MSSE m_n^{UD} based on the unobserved data was computed.

Figure 6.2 (a) displays the boxplots for the empirical $\mathcal{L}_2$ errors of these estimates, Figure 6.2 (b) for each choice of α_1 the boxplots for the ratios of the empirical $\mathcal{L}_2$ errors of the MSSE m_n to the empirical $\mathcal{L}_2$ errors of m_n^{UD}.

Figure 6.2 (b) shows that the empirical $\mathcal{L}_2$ errors of the MSSE based on the transformed data is for all six different choices of α_1 in 50% of the simulation runs between approximately 1.8 and 6.4 times larger than the empirical $\mathcal{L}_2$ error of m_n^{UD}. In this example, choosing $\alpha_1 = \frac{14}{25}$ leads to the estimate with the lowest median of the empirical

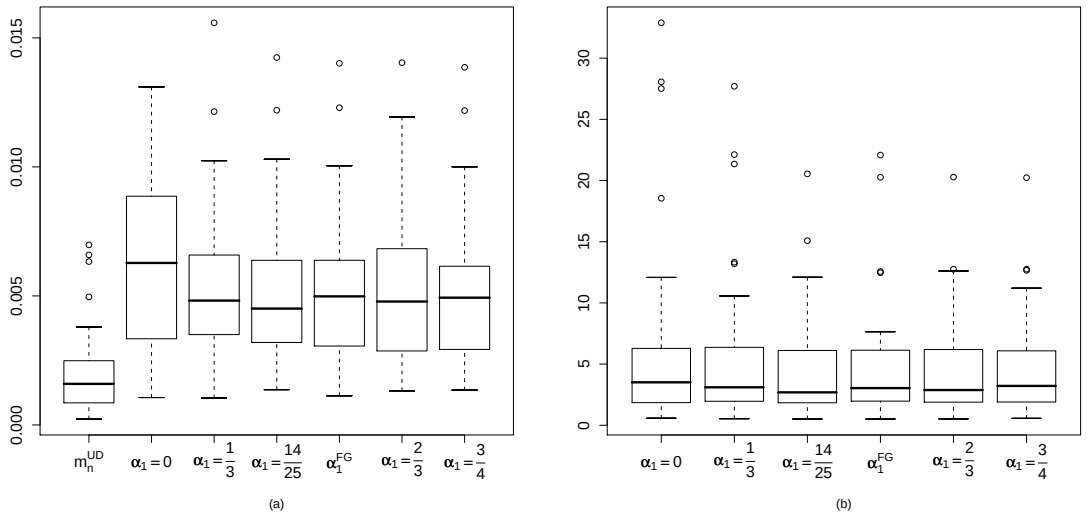

Figure 6.2: Boxplots of the empirical $\mathcal{L}_2$ errors and the ratio of the empirical $\mathcal{L}_2$ errors of m_n and m_n^{UD} for different values of α_1 in 50 independently repeated simulation runs. Panel (a) displays (from left to right) the boxplot of the empirical $\mathcal{L}_2$ errors of m_n^{UD} and the boxplots of the $\mathcal{L}_2$ errors of m_n for six different choices of α_1, panel (b) for each value of α_1 the boxplots for the ratios of the empirical $\mathcal{L}_2$ errors of m_n to the corresponding empirical $\mathcal{L}_2$ error of m_n^{UD}.

$\mathcal{L}_2$ errors (about $4.5 \cdot 10^{-3}$, vide Figure 6.2 (a)) as well as the lowest median of the ratios of the empirical $\mathcal{L}_2$ errors (about 2.7, vide Figure 6.2 (b)) of all six MSSE for censored data. This estimate also achieves the lowest 75%–quantile of the ratios of the empirical $\mathcal{L}_2$ errors for the 50 independentyl repeated simulation runs compared to the other five MSSE based on the transformed data. As described above, the choice $\alpha_1 = \frac{14}{25}$ corresponds to the median of the α_1^{FG}'s which were calculated for the 50 samples. As expected, the MSSE based on the transformed data with $\alpha_1 = \alpha_1^{FG}$ performs quite as well as the estimate with $\alpha_1 = \frac{14}{25}$, with the first MSSE having a slightly higher median of the empirical $\mathcal{L}_2$ errors and ratios of the empirical $\mathcal{L}_2$ errors, respectively, than the second. In our example, only the MSSE with $\alpha_1 = 0$ shows a reduced accuracy and precision of the estimation of m compared to these two estimates (vide Figure 6.2 (a)), while the three remaining MSSE show a more or less comparable performance. Here, in 50% of the independently repeated simulation runs, the empirical $\mathcal{L}_2$ error of the estimate with $\alpha_1 = 0$ is between 1.0 and 1.6 times larger than the empirical $\mathcal{L}_2$ error of the estimate with $\alpha_1 = \frac{14}{25}$.

The best choice of the parameter α_1 in an application is still an open question. From a theoretical point of view, α_1 should be determined in such a way that it minimizes the variability of the transformed data. However, the derivation of an analytic formula for this selection is complicated (cf. Fan and Gijbels (1994, 1996) and El Ghouch and Van Keilegom (2008)).

As indicated by the difference between the empirical $\mathcal{L}_2$ errors for MSSE based on the transformed data with $\alpha_1 = \frac{14}{25}$ and $\alpha_1 = 0$, choosing α_1 properly might be important. But on the other hand, this difference is much smaller than the difference between the empirical $\mathcal{L}_2$ errors of the estimates based on the censored data for our six choices of α_1 and the empirical $\mathcal{L}_2$ errors of MSSE m_n^{UD} based on the uncensored data. Therefore, for the MSSE in our simulation study, the different choices of the transformation parameter α_1 have a much smaller impact on the accuracy of the estimation of m than the effect of the censoring itself.

6.3 Results for MSSE of the conditional variance

Below, we present our results of the estimate (2.73). In order to compute the MSSE σ_n^2 of the conditional variance σ^2, the data of a single simulation run with sample size $n = 200$ is split into three parts. As mentioned in Section 2.5, the first part with sample size $n_1 = 67$ and the second part with sample size $N_r = 66$ are used to calculate the underlying estimate of m. The second part and the third part with sample size $N_t = 67$ are then treated as learning data and testing data in the computation of (2.73). Note that this estimate depends on two transformation parameters, α_1 and α_2. According to the results of Section 6.2, the parameter of the underlying estimate of m is in this section set to $\alpha_1 = \frac{14}{25}$ or $\alpha_1 = \alpha_1^{FG}$. Furthermore, we consider two similar choices for α_2 :

$$\alpha_2^{FG} := \alpha_2^{FG}\left(\mathcal{D}_{N_r}^{(2)}\right) := \min_{\substack{i=n_1+1,\ldots,N_1:\\ \delta_i=1}} \frac{\displaystyle\int_0^{Y_i} \frac{2\,t}{G_{N_r}(t)}\,dt - Y_i^2}{\displaystyle\frac{Y_i^2}{G_{N_r}(Y_i)} - \int_0^{Y_i} \frac{2\,t}{G_{N_r}(t)}\,dt}$$

(cf. (6.9)) and $\alpha_2 = 1$, which agrees with the median of α_2^{FG} for 50 independently repeated simulation runs.

Let $\bar{U}_{i,N_r}$ $(i = 68, \ldots, 133)$ and $\bar{U}_{i,N_t}$ $(i = 134, \ldots, 200)$ be given by (2.68) and (2.69).

For a single simulated data set, Figure 6.3 displays (a) the (realizations of the) unobserved data points $(X_i, Y_i^2 - m(X_i)^2)$ $(i = 1, \ldots, 200)$ together with the conditional variance σ^2 (see **(SM1b)**) and (b) the transformed data points $(X_i, \bar{U}_{i,N_{\mathrm{r}}})$ $(i = 68, \ldots, 133)$ and $(X_i, \bar{U}_{i,N_{\mathrm{t}}})$ $(i = 134, \ldots, 200)$ with transformed censored data points marked by "×", σ^2 (green solid line), the MSSE σ_n^2 based on the censored data (blue dotted line), and the in an application not calculable MSSE $\sigma_n^{2,UD}$ of σ^2 based on the uncensored data (red dashed line). In this example, we have $\alpha_1 = \alpha_1^{FG} \approx 0.25$ and $\alpha_2 = \alpha_2^{FG} \approx 1.18$.

The estimates σ_n^2 and $\sigma_n^{2,UD}$ depend on the MSSE m_{N_1} and $m_{N_1}^{UD}$, respectively. Hence, the performances of m_{N_1} and $m_{N_1}^{UD}$ may have a big influence on the precision of our estimation of σ^2. On the left hand side of Figure 6.3 (b), one can identify a larger domain, where the MSSE σ_n^2 tends to overestimate the conditional variance σ^2. For the displayed simulated data set, this is due to the fact that in this area, the underlying MSSE m_{N_1} takes smalller values than the regression function m, and the transformed random variables $\bar{U}_{i,N_{\mathrm{r}}}$ $(i = 68, \ldots, 133)$ and $\bar{U}_{i,N_{\mathrm{t}}}$ $(i = 134, \ldots, 200)$ are therefore more likely to exceed the unobserved random variables $Y_i^2 - m(X_i)^2$ $(i = 68, \ldots, 200)$ (cf. (2.68) and (2.69)). In the opposite scenario, one can similarly expect that σ_n^2 inclines to underestimate σ^2. The same remarks apply to $m_{N_1}^{UD}$ and $\sigma_n^{2,UD}$.

The MSSE σ_n^2 depends on two transformation parameters, α_1 and α_2. In analogy to our analysis in Section 6.2, we created 50 independent data sets according to the setting of our simulation study. For each of these samples and for the four different choices of (α_1, α_2) mentioned above, the estimate σ_n^2 was calculated. Furthermore, we computed $\sigma_n^{2,UD}$, the MSSE based on the uncensored data, which does not depend on α_1 or α_2.

Table 6.1 presents the minimum, the 25%–quantile, the median, the 75%–quantile, and the maximum of the empirical $\mathcal{L}_2$ errors of σ_n^2 on the 50 independently repeated simulation runs for each choice of the two transformation parameters. Moreover the interquartile range (abbreviated as IQR) is given, which is defined as the difference between the 75%–quantile and the 25%–quantile. Here, the IQR serves as a measure of the variability of the empirical $\mathcal{L}_2$ errors in the 50 simulated data sets.

From Table 6.1, one can conclude that for the 50 samples and out of the four different choices of α_1 and α_2, the choice $\alpha_1 = \frac{14}{25}$ and $\alpha_2 = \alpha_2^{FG}$ results in the estimate with the lowest median and the lowest 25%–quantile of the empirical $\mathcal{L}_2$ errors. However, the MSSE

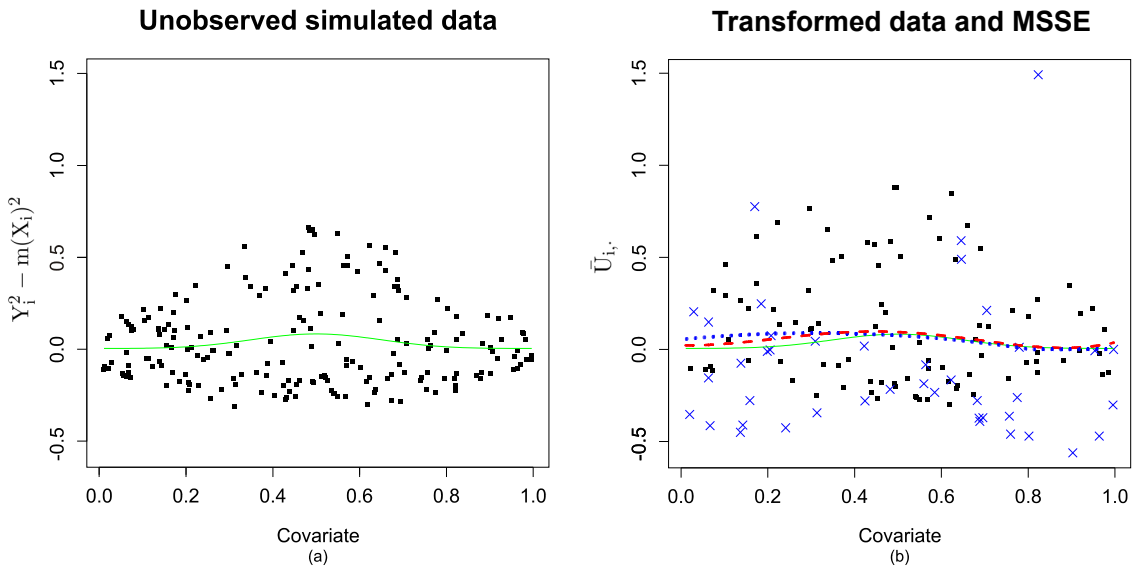

Figure 6.3: Data set for a single simulation run. The character "■" indicates the (transformed) uncensored observations, while the (transformed) censored observations are presented by "×". The individual panels show (a) the (realizations of the) unobserved data $(X_i, Y_i^2 - m(X_i)^2)$ together with σ^2 (green solid line) and (b) the (realizations of the) transformed data $(X_i, \bar{U}_{i,\cdot})$ — where $\bar{U}_{i,\cdot} := \bar{U}_{i,N_{\mathrm{r}}}$ for $i \in \{68, \ldots, 133\}$ and $\bar{U}_{i,\cdot} := \bar{U}_{i,N_{\mathrm{t}}}$ for $i \in \{134, \ldots, 200\}$ — together with σ^2 (green solid line), MSSE σ_n^2 (blue dotted line) and $\sigma_n^{2,UD}$ (red dashed line).

with the lowest 75%–quantile and the lowest IQR is derived for $\alpha_1 = \frac{14}{25}$ and $\alpha_2 = 1$. A comparison of the first and the third column of Table 6.1 shows that choosing $\alpha_1 = \alpha_1^{FG}$ instead of $\alpha_1 = \frac{14}{25}$ while keeping $\alpha_2 = \alpha_2^{FG}$, leads to a slightly higher median and IQR of the empirical $\mathcal{L}_2$ errors. This conclusion also holds true with $\alpha_2 = 1$ replaced by $\alpha_2 = \alpha_2^{FG}$, as indicated by the second and the fourth column.

In Section 6.2, we already observed that for the MSSE m_n, the choice $\alpha_1 = \frac{14}{25}$ is slightly better in terms of the median of the empirical $\mathcal{L}_2$ errors than taking $\alpha_1 = \alpha_1^{FG}$. Since the estimate σ_n^2 depends on m_{N_1}, choosing $\alpha_1 = \frac{14}{25}$ instead of $\alpha_1 = \alpha_1^{FG}$ also leads to a smaller median and IQR of the empirical $\mathcal{L}_2$ errors of σ_n^2, as mentioned above. Moreover, Figure 6.2 and Table 6.1 imply that for $\alpha_1 = \frac{14}{25}$, the median of the empirical $\mathcal{L}_2$ errors of the two MSSE of σ^2 are equal to or slightly smaller than the median of the empirical $\mathcal{L}_2$ error of the MSSE of m. However, this relation does not hold for the estimates with $\alpha_1 = \alpha_1^{FG}$, and in both cases, the 75%–quantiles and the IQR of the empirical $\mathcal{L}_2$ errors

	α_1^{FG}, α_2^{FG}	α_1^{FG}, $\alpha_2 = 1$	$\alpha_1 = \frac{14}{25}$, α_2^{FG}	$\alpha_1 = \frac{14}{25}$, $\alpha_2 = 1$
Minimum	$0.81 \cdot 10^{-3}$	$0.56 \cdot 10^{-3}$	$1.09 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$
25%–Quantile	$2.01 \cdot 10^{-3}$	$2.12 \cdot 10^{-3}$	$1.93 \cdot 10^{-3}$	$2.16 \cdot 10^{-3}$
Median	$5.61 \cdot 10^{-3}$	$5.13 \cdot 10^{-3}$	$4.27 \cdot 10^{-3}$	$4.5 \cdot 10^{-3}$
75%–Quantile	$9.52 \cdot 10^{-3}$	$9.08 \cdot 10^{-3}$	$9.09 \cdot 10^{-3}$	$7.67 \cdot 10^{-3}$
Maximum	$23.99 \cdot 10^{-3}$	$25.27 \cdot 10^{-3}$	$25.83 \cdot 10^{-3}$	$27.95 \cdot 10^{-3}$
IQR	$7.51 \cdot 10^{-3}$	$6.97 \cdot 10^{-3}$	$7.16 \cdot 10^{-3}$	$5.51 \cdot 10^{-3}$

Table 6.1: Summary statistics for the empirical $\mathcal{L}_2$ errors of σ_n^2 on 50 independently generated simulated data sets for four different choices of (α_1, α_2).

of the estimates of σ^2 exceed those of the estimates of m. Since the MSSE of σ^2 depend on the MSSE of m, the empirical $\mathcal{L}_2$ error of the latter ones are in our setting expected to be on average smaller than (or at least equal to) those of the first ones.

6.4 Results for MSSE of the conditional survival function

In the following, the results of the simulation study for our MSSE of the conditional survival function $F(\tau|\cdot)$ ($\tau \in \mathbb{R}$ fixed) are presented.

For each simulation run, we first create 200 data points which are split at random into a learning and a testing data set, each of sample size $n_{\mathtt{l}} = n_{\mathtt{t}} = 100$ (cf. Section 6.2). Then the MSSE $F_n(\tau \,|\, \cdot)$ is calculated on the basis of the random variables (2.84) and (2.85). In addition, we compute $F_n^{UD}(\tau \,|\, \cdot)$ by applying the estimate $F_n(\tau \,|\, \cdot)$ to the (in an statistical application not observable) uncensored data. Note that $F_n(\tau \,|\, \cdot)$ and $F_n^{UD}(\tau \,|\, \cdot)$ do not depend on the truncation parameter L or any transformation parameter.

For $\tau = \frac{2}{3}$ and a single simulated data set, Figure 6.4 shows (a) the unobserved simulated data, the conditional survival function $F(\tau|\cdot)$ (green solid line), and a horizontal line at level $\frac{2}{3}$ (orange dot and dash line), (b) the observed simulated data with censored data points marked by "+" and the horizontal line from panel (a), (c) the transformed data, computed according to (2.84) and (2.85), and (d) $F(\tau|\cdot)$ (green solid line) together with the MSSE $F_n(\tau \,|\, \cdot)$ (red dashed line) and $F_n^{UD}(\tau \,|\, \cdot)$ (blue dotted line).

The transformed times displayed in Figure 6.4 (c) are calculated from the observed

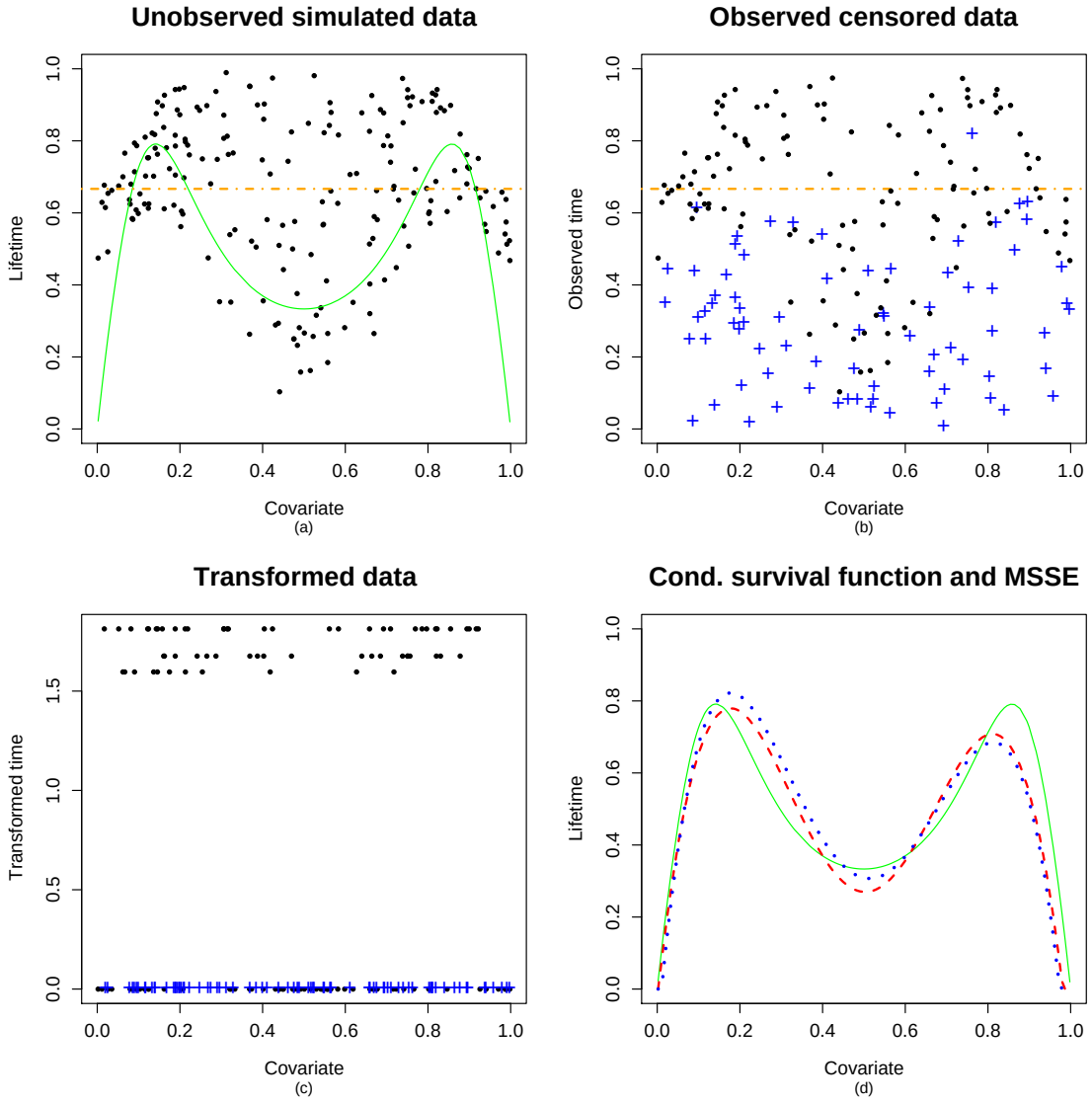

Figure 6.4: Data set for a single simulation run. The character "•" indicates the (transformed) uncensored observations, while the (transformed) censored observations are presented by "+". The individual panels show (a) the unobserved data together with $F(\tau \,|\, \cdot)$ (green solid line) and a horizontal line at level $\tau = \frac{2}{3}$ (orange dot and dash line), (b) the observed simulated data with the horizontal line from panel (a), (c) the transformed data, and (d) $F(\tau \,|\, \cdot)$ (green solid line) with MSSE $F_n(\tau \,|\, \cdot)$ (red dashed line) and $F_n^{UD}(\tau \,|\, \cdot)$ (blue dotted line).

times of panel (b) as follows (vide (2.84) and (2.85)). All uncensored data points below the orange horizontal line at level $\tau = \frac{2}{3}$ and all censored data points and are set to zero. All

uncensored observations which exceed $\frac{2}{3}$ are replaced by $G_{n_1}(Y_i)^{-1}$ $(i = 1, \ldots, 100)$ if they are contained in the learning data or by $G_{n_t}(Y_i)^{-1}$ $(i = 101, \ldots, 200)$ if they belong to the testing data, respectively. Therefore, one part of the transformed data points takes values larger than or identical to one, while the data points in the other part are all identical to zero. Note that if we compute the MSSE $F_n^{UD}(\tau\,|\,\cdot)$ or if no censoring arises, then the transformed random variables are simply given by $I_{[Z_i>\tau]}$ $(i = 1, \ldots, 200)$.

But then, it holds that $F(t|x) = \mathbf{P}[Y > t \,|\, X = x] \in [0,1]$ for all $(x,t) \in [0,1]^d \times \mathbb{R}$. This means that we seek to estimate a function which takes only values between zero and one on the basis of some data, where the dependent random variables are in $(0,1)$ with probability zero.

It is not evident that an estimate based on such data performs well. However, Figure 6.4 (d) indicates that $F_n(\tau\,|\,\cdot)$ and $F_n^{UD}(\tau\,|\,\cdot)$ are quite reliable estimates of the conditional survival function, especially if one recalls the small sample size we considered. Though, the performance of $F_n(\tau\,|\,\cdot)$ may be poor if our assumption that C and X are independent does not hold. In this case, it may happen that the censored observations are much more likely to occur for particular intervals of the corresponding covariates. Since these observations are all transformed to zero, this can cause $F_n(\tau\,|\,\cdot)$ to underestimate $F(\tau\,|\,\cdot)$ in these domains, while it may overestimate in the others. Yet, if the censoring is not to heavy, then this effect is compensated by the smoothing parameter of $F_n(\tau\,|\,\cdot)$, at least to some extent.

Figure 6.5 (a) and (b) display the empirical $\mathcal{L}_2$ errors of $F_n^{UD}(\tau\,|\,\cdot)$ and $F_n(\tau\,|\,\cdot)$ for seven different values of τ on 50 independently generated samples of the chosen distribution of (X, Y, C). Note that since $0 \leq Y \leq 1$ a.s. (cf. Section 6.1), we only consider $\tau \in [0,1]$.

Obviously, the performances of $F_n(\tau\,|\,\cdot)$ and $F_n^{UD}(\tau\,|\,\cdot)$ depend on the particular choice of τ. In our simulation study, the conditional survival function is given by (6.6).

From (6.3) and (6.5), we deduce that if $\tau \leq M_{max}^- \approx 0.603$ then there exist $x \in [0,1]$ with $F(\tau\,|\,x) = 1$. In case that $\tau \in \left(M_{max}^-, \frac{2}{3}\right)$, then (6.7) yields $F(\tau\,|\,x) \in (0,1)$ for all $x \in [0,1]$. And if $\tau \geq \frac{2}{3}$, then one can conclude from (6.2) and (6.5) that there exist $x \in [0,1]$ with $F(\tau\,|\,x) = 0$. In addition, note that for $\tau = \frac{2}{3}$, $F(\tau\,|\,x) = 0$ holds if and only if $x = 0$ or $x = 1$, and that for $\tau = M_{max}^- \approx 0.603$, there exist exactly one value $x_{max}^- \approx 0.13$ with $F(\tau\,|\,x_{max}^-) = F(\tau\,|\,1 - x_{max}^-) = 1$. Furthermore, we have $F(1\,|\,x) = 0$

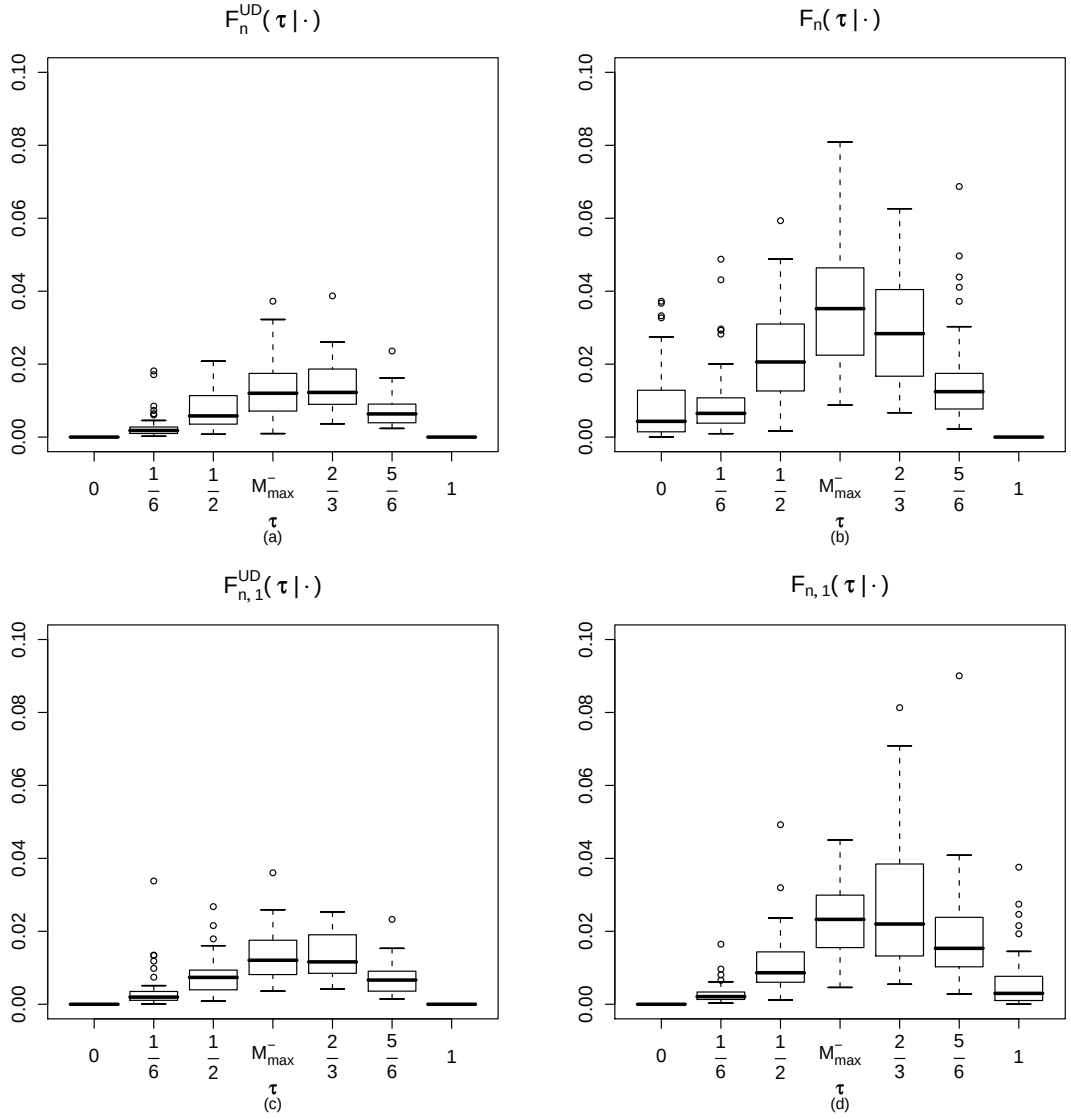

Figure 6.5: Boxplots of the empirical $\mathcal{L}_2$ errors for seven different values of τ in 50 independently repeated simulation runs. The individual panels show the empirical $\mathcal{L}_2$ errors of the MSSE (a) $F_n^{UD}(\tau\,|\,\cdot)$, (b) $F_n(\tau\,|\,\cdot)$, (c) $F_{n,1}^{UD}(\tau\,|\,\cdot)$, and (d) $F_{n,1}(\tau\,|\,\cdot)$. The estimates $F_n^{UD}(\tau\,|\,\cdot)$ and $F_{n,1}^{UD}(\tau\,|\,\cdot)$ are based on the unobserved simulated data, while $F_n(\tau\,|\,\cdot)$ and $F_{n,1}(\tau\,|\,\cdot)$ are calculated on the censored data.

and $F(\,0\,|\,x\,) = 1$ for all $x \in [0,1]$. Since our MSSE $F_n(\tau\,|\,\cdot)$ and $F_n^{UD}(\tau\,|\,\cdot)$ are truncated at zero and one, this implies that in case that $\tau \in \left[M_{max}^-, \frac{2}{3}\right]$, these estimates are likely to have larger empirical $\mathcal{L}_2$ errors than if $\tau \notin \left[M_{max}^-, \frac{2}{3}\right]$ (vide Figure 6.5 (a) and (b)).

In addition, the empirical $\mathcal{L}_2$ errors of $F_n^{UD}(\,0\,|\,\cdot\,)$, $F_n^{UD}(\,1\,|\,\cdot\,)$, and $F_n(\,1\,|\,\cdot\,)$ are zero a.s., because with probability one, the transformed random variables then all equal 1 (for $\tau = 0$) or 0 (for $\tau = 1$). This is indicated by Figure 6.5 (a) and (b). But if censoring arises, then with a non-zero probability, the empirical $\mathcal{L}_2$ error of $F_n(\,0\,|\,\cdot\,)$ is not identical to zero, since in this case, with probability one, the censored observations are set to zero, while the transformed uncensored observations take values larger than or equal to 1.

Let the assumptions of Section 2.6 hold. Next, a MSSE $F_{n,1}(\,\tau\,|\,\cdot\,)$ of $F(\,\tau\,|\,\cdot\,)$ is derived, which performs for small values of τ (and a finite sample size) better than $F_n(\,\tau\,|\,\cdot\,)$ in terms of the $\mathcal{L}_2$ error (see below). Let therefore $\tau \in \mathbb{R}$ be arbitrary, but fixed and set

$$\bar{U}^{(3)} := \frac{\delta I_{[Z \le \tau]}}{G(Z)} \tag{6.10}$$

and

$$\bar{U}_i^{(3)} = \frac{\delta_i I_{[Z_i \le \tau]}}{G(Z_i)} \quad (i = 1, \ldots, n). \tag{6.11}$$

Note that (6.10) and (6.11) are given similar to (2.77) and (2.78), but with $I_{[Z>\tau]}$ and $I_{[Z_i>\tau]}$ replaced by $I_{[Z\le\tau]}$ or $I_{[Z_i\le\tau]}$.

Next, we define estimates of (6.11) in analogy to (2.84) and (2.85). Let G_{n_1} and G_{n_t} be given by (2.34) and (2.35). Set

$$\bar{U}_{i,n_1}^{(3)} := \frac{\delta_i I_{[Z_i \le \tau]}}{G_{n_1}(Z_i)} \quad (i = 1, \ldots, n_1) \tag{6.12}$$

and

$$\bar{U}_{i,n_t}^{(3)} := \frac{\delta_i I_{[Z_i \le \tau]}}{G_{n_t}(Z_i)} \quad (i = n_1 + 1, \ldots, n) \tag{6.13}$$

$\left(\frac{0}{0} := 0\right)$. If $\tau = 0$ and $\mathbf{P}\,[\,Y = 0\,] = 0$, then **(RA2)** implies that with probability one

$$\bar{U}_{i,n_1}^{(3)} = 0 \quad (i = 1, \ldots, n_1) \quad \text{and} \quad \bar{U}_{i,n_t}^{(3)} = 0 \quad (n_1 + 1, \ldots, n). \tag{6.14}$$

Let the MSSE $\bar{F}_n(\,\tau\,|\,\cdot\,)$ be defined by (2.89) – (2.92) with the only difference that $U_{i,n_1}^{(3)}$ is replaced by $\bar{U}_{i,n_1}^{(3)}$ $(i = 1, \ldots, n_1)$ and $U_{i,n_t}^{(3)}$ by $\bar{U}_{i,n_t}^{(3)}$ $(i = n_1 + 1, \ldots, n)$. Observe that one can conclude similar to (2.80)

$$\mathbf{E}\left[\,\bar{U}^{(3)}\,\middle|\,X\,\right] = \mathbf{P}\,[\,Y \le \tau\,|\,X\,] = 1 - F\,(\,\tau\,|\,X\,)\,.$$

Therefore, $\bar{F}_n(\,\tau\,|\,\cdot\,)$ may be considered as a regression estimate of $1 - F(\,\tau\,|\,\cdot\,)$, the conditional distribution function of the lifetime Y at fixed point $\tau \in \mathbb{R}$ (cf. Section 2.6).

Hence, our MSSE $F_{n,1}(\tau\,|\,\cdot\,)$ of $F(\tau\,|\,\cdot\,)$ is now given by

$$F_{n,1}(\tau\,|\,\cdot\,) := 1 - \bar{F}_n(\tau\,|\,\cdot\,). \tag{6.15}$$

Assume that the conditions of Theorem 5.4 hold. From the proofs of Lemma 4.2 and Theorem 5.4, one can conclude that in this case, $F_{n,1}(\tau\,|\,\cdot\,)$ and $\bar{F}_n(\tau\,|\,\cdot\,)$ achieve their optimal rate of convergence up to some logarithmic factor. Here, we used the fact that (6.15) implies

$$|\,F_{n,1}(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,| = \big|\,\bar{F}_n(\tau\,|\,x\,) - (1 - F(\tau\,|\,x\,))\,\big| \quad \forall x \in [0,1]^d.$$

If $\mathbf{P}\,[\,Y = 0\,] = 0$, then (6.14) and (6.15) yield that the $\mathcal{L}_2$ error of $F_{n,1}(\,0\,|\,\cdot\,)$ is zero a.s. for all $n \geq 2$. Moreover, in case that τ is close to zero, it is likely that this $\mathcal{L}_2$ error is smaller than the $\mathcal{L}_2$ error of $F_n(\tau\,|\,\cdot\,)$. On the other hand, if τ is close to L and censoring arises, then one can conclude similar to above that $F_n(\tau\,|\,\cdot\,)$ is a more reliable estimate of $F(\tau\,|\,\cdot\,)$ in terms of the $\mathcal{L}_2$ error than $F_{n,1}(\tau\,|\,\cdot\,)$ (note that we use $L = 1$ in our simulation study).

Therefore, an obvious idea is to construct an estimate $F_{n,2}(\tau\,|\,\cdot\,)$ of $F(\tau\,|\,\cdot\,)$, where $F_{n,2}(\tau\,|\,\cdot\,)$ equals $F_{n,1}(\tau\,|\,\cdot\,)$ if $\tau \leq 0$ and $F_n(\tau\,|\,\cdot\,)$ if $\tau \geq L$, respectively. In order to account for values of $\tau \in (0, L)$, one can introduce a weight function $w : \mathbb{R} \to [0,1]$ with $w(\tau) = 1$ if $\tau \geq L$ and $w(\tau) = 0$ if $\tau \leq 0$ and set

$$F_{n,2}(\tau\,|\,\cdot\,) := w(\tau) \cdot F_n(\tau\,|\,\cdot\,) + (1 - w(\tau)) \cdot F_{n,1}(\tau\,|\,\cdot\,). \tag{6.16}$$

Observe that $(a+b)^2 \leq 2 \cdot a^2 + 2 \cdot b^2$ and $0 \leq w(\tau) \leq 1$ $(\tau \in \mathbb{R})$ yield for all $x \in [0,1]^d$

$$\begin{aligned}
&|\,F_{n,2}(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,|^2 \\
&\quad = \quad |\,w(\tau) \cdot F_n(\tau\,|\,x\,) - w(\tau) \cdot F(\tau\,|\,x\,) + (1 - w(\tau)) \cdot F_{n,1}(\tau\,|\,x\,) - (1 - w(\tau)) \cdot F(\tau\,|\,x\,)\,|^2 \\
&\quad \leq \quad 2 \cdot w(\tau)^2 \cdot |\,F_n(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,|^2 + 2 \cdot (1 - w(\tau))^2 \cdot |\,F_{n,1}(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,|^2 \\
&\quad \leq \quad 2 \cdot |\,F_n(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,|^2 + 2 \cdot |\,F_{n,1}(\tau\,|\,x\,) - F(\tau\,|\,x\,)\,|^2 .
\end{aligned}$$

Similar to above, this implies that the assertion of Theorem 5.4 still holds if we replace $F_n(\tau\,|\,\cdot\,)$ by $F_{n,2}(\tau\,|\,\cdot\,)$. The proper choice of the weight function w in a statistical application may be a non-trivial task, since it depends on the underlying censoring mechanism,

and should in general be done via some data-driven method (e.g., be based on the empirical $\mathcal{L}_2$ risks of $F_n(\tau\,|\,\cdot)$ and $F_{n,1}(\tau\,|\,\cdot)$).

Figure 6.5 (b) and (d) show the boxplots of the empirical $\mathcal{L}_2$ errors of $F_n(\tau\,|\,\cdot)$ and $F_{n,1}(\tau\,|\,\cdot)$ on the 50 independently repeated simulation runs for seven different values of τ. In our example, $F_n(\tau\,|\,\cdot)$ is a more reliable estimate of $F(\tau\,|\,\cdot)$ than $F_{n,1}(\tau\,|\,\cdot)$ if $\tau > \frac{2}{3}$. This is due to several reasons. From the proof of Lemma 6.2, we deduce that for $\tau > \frac{2}{3}$, there exist $x_t^-, x_t^+ \in (0,1)$ such that $F(\tau\,|\,x) = 0$ for all $x \notin [x_t^-, x_t^+]$. Therefore, we have $\mathbf{P}[\,Z \leq \tau\,|\,X = x\,] = 1$ for all $\tau > \frac{2}{3}$ and all $x \notin [x_t^-, x_t^+]$. And this in turn implies that outside of the interval $[x_t^-, x_t^+]$, the transformed random variables, which we use in order to compute $F_n(\tau\,|\,\cdot)$, are all a.s. identical to zero. But then, for uncensored observations, (6.12) and (6.13) are in this case larger than or equal to one. Moreover, for $x \in [x_t^-, x_t^+]$ and $\tau > \frac{2}{3}$, (6.6) yields that $F_n(\tau\,|\,\cdot)$ is likely to be closer to $F(\tau\,|\,\cdot)$ than $F_{n,1}(\tau\,|\,\cdot)$.

In addition, for $\tau = 1$ we have $x_t^- = x_t^+ = \frac{1}{2}$. As mentioned above, the empirical $\mathcal{L}_2$ error of $F_n(1\,|\,\cdot)$ is zero a.s., while with a non-zero probability the empirical $\mathcal{L}_2$ error of $F_{n,1}(1\,|\,\cdot)$ does not equal zero. Similar, one can conclude that for $\frac{2}{3} < \tau < 1$, the former is likely to be close to zero, whereas this does not apply for the latter one. This implies that the empirical $\mathcal{L}_2$ error of $F_n(\tau\,|\,\cdot)$ will in general be smaller than the empirical $\mathcal{L}_2$ error of $F_{n,1}(\tau\,|\,\cdot)$ if $\tau > \frac{2}{3}$. In analogous way, one can argue that for $\tau < M_{max}^-$ the reverse is true.

As a consequence, for $\tau = \frac{5}{6}$, the median of the empirical $\mathcal{L}_2$ errors displayed in Figure 6.5 (d) is approximately 1.21 times larger than that in panel (b). Furthermore, this also holds true for the 25%–quantile, the 75%–quantile, and the IQR. The median of the ratios of the empirical $\mathcal{L}_2$ errors of $F_{n,1}(\frac{5}{6}\,|\,\cdot)$ to the corresponding empirical $\mathcal{L}_2$ errors of $F_n(\frac{5}{6}\,|\,\cdot)$ is approximately 1.21.

On the other hand, Figure 6.5 (b) and (d) also demonstrate that if $\tau < M_{max}^-$, then $F_{n,1}(\tau\,|\,\cdot)$ is a more accurate and preciser estimate of the conditional survival function of Y than $F_n(\tau\,|\,\cdot)$. For $\tau = \frac{1}{6}$ and $\tau = \frac{1}{2}$, the median of the ratios of the empirical $\mathcal{L}_2$ errors of $F_{n,1}(\tau\,|\,\cdot)$ to the corresponding empirical $\mathcal{L}_2$ errors of $F_n(\tau\,|\,\cdot)$ is about 0.34 and 0.42, respectively.

Additionally, as discussed above, one can observe that the empirical $\mathcal{L}_2$ errors of $F_n(\tau\,|\,\cdot)$ and $F_{n,1}(\tau\,|\,\cdot)$ are generally larger in case that $\tau \in \left[M_{max}^-, \frac{2}{3}\right]$ than if τ is not

contained in this interval. And the closer τ is chosen to zero or one, the smaller is the median of the empirical $\mathcal{L}_2$ errors of both MSSE evaluated on the 50 independently repeated simulation runs.

These characteristics are also visible in Figure 6.5 (a) and (c) for the estimates $F_n^{UD}(\tau\,|\,\cdot\,)$ and $F_{n,1}^{UD}(\tau\,|\,\cdot\,)$, which are calculated on the uncensored data. But in contrast to the MSSE for censored regression, panels (a) and (c) indicate that $F_n^{UD}(\tau\,|\,\cdot\,)$ and $F_{n,1}^{UD}(\tau\,|\,\cdot\,)$ perform similar for all $\tau \in \mathbb{R}$. This is due to the fact that if no censoring arises, the transformed random variables (2.78) are in this case identical to $I_{[Y_i>\tau]}$, whereas (6.11) equal $I_{[Y_i\leq\tau]} = 1 - I_{[Y_i>\tau]}$ $(i = 1, \ldots, n)$. As a consequence, $\bar{F}_n^{UD}(\tau\,|\,\cdot\,)$ is close to $1 - F_n^{UD}(\tau\,|\,\cdot\,)$ and therefore $F_{n,1}^{UD}(\tau\,|\,\cdot\,)$ to $F_n^{UD}(\tau\,|\,\cdot\,)$. For the 50 independently generated simulated data sets and all seven choices of τ considered, 90% of the absolute differences between the empirical $\mathcal{L}_2$ errors of $F_n^{UD}(\tau\,|\,\cdot\,)$ and $F_{n,1}^{UD}(\tau\,|\,\cdot\,)$ are smaller than $8.9 \cdot 10^{-3}$.

6.5 Proofs of Lemmata 6.1 and 6.2

In this section, it is shown that the assumptions of Lemma 6.1 and Lemma 6.2 hold. First, we verify that m, σ^2, and $F(\tau|\cdot)$ ($M_{max}^- \leq \tau < \frac{2}{3}$ fixed) are functions in the Sobolev space $W_3([0,1])$.

Proof of Lemma 6.1. First note that **(SM1b)** yields for all $x \in [0,1]$

$$\frac{d}{dx}\,\sigma(x) = 56\sqrt{3}\,(1-2x)\,(x\,(1-x))^{\frac{5}{2}}, \tag{6.17}$$

$$\frac{d^2}{dx^2}\,\sigma(x) = 28\sqrt{3}\,(5-24x+24x^2)\,(x\,(1-x))^{\frac{3}{2}}, \tag{6.18}$$

and

$$\frac{d^3}{dx^3}\,\sigma(x) = 210\sqrt{3}\,(1-10x+24x^2-16x^3)\,(x\,(1-x))^{\frac{1}{2}}. \tag{6.19}$$

Since

$$0 \leq x\,(1-x) \leq \frac{1}{4} \quad \forall x \in [0,1], \tag{6.20}$$

(SM1b) and (6.17) – (6.19) imply for all $k \in \{0,1,2,3\}$

$$\left|\frac{d^k}{dx^k}\,\sigma(x)\right| \leq \frac{\sqrt{3}}{24} + 210\sqrt{3}\cdot 53\cdot\frac{1}{2} < 1 + 210\cdot 53 = 11131 \quad (x \in [0,1]). \tag{6.21}$$

Furthermore, one can conclude from (6.19) for all $x \in (0,1)$ that

$$\frac{d^4}{dx^4}\,\sigma(x) \;=\; 105\sqrt{3}\,(1-32x+160x^2-256x^3+128x^4)\,(x\,(1-x))^{-\frac{1}{2}}. \tag{6.22}$$

From (6.22), one gets for all $0 < x \leq \frac{7}{500}$

$$\frac{d^4}{dx^4}\,\sigma(x) \geq 105\,\sqrt{3}\,\frac{1}{\sqrt{3}}\,(x\,(1-x))^{-\frac{1}{2}} = 105\,(x\,(1-x))^{-\frac{1}{2}}. \tag{6.23}$$

Here, we used that

$$1-32x+160x^2-256x^3+128x^4 \geq \frac{1}{\sqrt{3}} \quad \left(0 < x \leq \frac{7}{500}\right). \tag{6.24}$$

Now, (6.21) yields for all $k \in \{0,1,2,3\}$

$$\int_0^1 \left|\frac{d^k}{dx^k}\,\sigma(x)\right|^2 dx < \infty. \tag{6.25}$$

In contrast, (6.23) implies

$$\int_0^{\frac{7}{500}} \left|\frac{d^4}{dx^4}\,\sigma(x)\right|^2 dx \geq \int_0^{\frac{7}{500}} 105^2 (x\,(1-x))^{-1}\,dx = 105^2 \ln\left(\frac{x}{1-x}\right)\Bigg|_0^{\frac{7}{500}},$$

i.e.,

$$\int_0^{\frac{7}{500}} \left|\frac{d^4}{dx^4}\,\sigma(x)\right|^2 dx = \infty. \tag{6.26}$$

Furthermore, we deduce from (6.2)

$$\int_0^1 \left|\frac{d^k}{dx^k}\,M^+(x)\right|^2 dx < \infty \quad \forall k \in \mathbb{N}_0.$$

The last inequality together with (6.2), (6.25), and $(a+b)^2 \leq 2a^2+2b^2$ $(a,b \in \mathbb{R})$ yields for all $k \in \{0,1,2,3\}$

$$\int_0^1 \left|\frac{d^k}{dx^k}\,m(x)\right|^2 dx \leq 2 \cdot \int_0^1 \left|\frac{d^k}{dx^k}\,M^+(x)\right|^2 dx + 6 \cdot \int_0^1 \left|\frac{d^k}{dx^k}\,\sigma(x)\right|^2 dx < \infty. \tag{6.27}$$

This proves that $m \in W_3([0,1])$ (cf. (2.3)). On the other hand, (6.2), (6.23), and (6.26) imply

$$\begin{aligned}\int_0^1 \left|\frac{d^4}{dx^4}\,m(x)\right|^2 dx \;&=\; \int_0^1 \left|\sqrt{3}\cdot\frac{d^4}{dx^4}\,\sigma(x)+128\right|^2 dx \\ &\geq\; 3\cdot\int_0^{\frac{7}{500}} \left|\frac{d^4}{dx^4}\,\sigma(x)\right|^2 dx = \infty,\end{aligned} \tag{6.28}$$

i.e., $m \notin W_k([0,1])$ for all $k \in \mathbb{N} \setminus \{1,2,3\}$ (vide (2.3)). In addition, one can conclude from (2.5), **(SM1a)**, and (6.27), that

$$0 < J_3^2(m) = \int_0^1 \left| \frac{d^3}{dx^3}\, m(x) \right|^2 dx < \infty \tag{6.29}$$

and therefore the first part of Lemma 6.1.

Next, we show the second part of Lemma 6.1. First observe that (6.21) implies

$$\int_0^1 \left|\sigma^2(x)\right|^2 dx = \int_0^1 |\sigma(x)|^4\, dx < \infty.$$

Moreover, we have from (6.21)

$$\begin{aligned}
\int_0^1 \left| \frac{d}{dx}\, \sigma^2(x) \right|^2 dx &= \int_0^1 \left| 2\,\sigma(x) \cdot \frac{d}{dx}\, \sigma(x) \right|^2 dx < \infty, \\
\int_0^1 \left| \frac{d^2}{dx^2}\, \sigma^2(x) \right|^2 dx &= \int_0^1 \left| 2\,\sigma(x) \cdot \frac{d^2}{dx^2}\, \sigma(x) + 2 \left(\frac{d}{dx}\, \sigma(x) \right)^2 \right|^2 dx < \infty,
\end{aligned}$$

and

$$\int_0^1 \left| \frac{d^3}{dx^3}\, \sigma^2(x) \right|^2 dx = \int_0^1 \left| 2\,\sigma(x) \cdot \frac{d^3}{dx^3}\, \sigma(x) + 6\, \frac{d}{dx}\, \sigma(x) \cdot \frac{d^2}{dx^2}\, \sigma(x) \right|^2 dx < \infty. \tag{6.30}$$

On the other hand, $\sigma(x) \geq \frac{\sqrt{3}}{24}$ $(x \in [0,1])$ and (6.26) yield

$$\begin{aligned}
\int_0^1 \left| \frac{d^4}{dx^4}\, \sigma^2(x) \right|^2 dx &= \int_0^1 \left| 2\sigma(x) \cdot \frac{d^4}{dx^4}\, \sigma(x) + 8 \frac{d}{dx}\, \sigma(x) \cdot \frac{d^3}{dx^3}\, \sigma(x) + 6 \left(\frac{d^2}{dx^2}\, \sigma(x) \right)^2 \right|^2 dx \\
&\geq \int_0^{\frac{7}{500}} \left| 2\sigma(x) \cdot \frac{d^4}{dx^4}\, \sigma(x) \right|^2 dx \\
&\geq \frac{1}{48} \int_0^{\frac{7}{500}} \left| \frac{d^4}{dx^4}\, \sigma(x) \right|^2 dx \\
&= \infty.
\end{aligned}$$

Here, the first inequality follows from (6.23) and

$$8\, \frac{d}{dx}\, \sigma(x) \cdot \frac{d^3}{dx^3}\, \sigma(x) + 6 \left(\frac{d^2}{dx^2}\, \sigma(x) \right)^2 \geq 0 \quad \left(0 \leq x \leq \frac{7}{500} \right).$$

Finally, one can conclude from (2.5), **(SM1b)**, and (6.30)

$$0 < J_3^2(\sigma^2) = \int_0^1 \left| \frac{d^3}{dx^3}\, \sigma^2(x) \right|^2 dx < \infty.$$

This proves the second part of Lemma 6.1.

Now, let M_{max}^- be given by (6.3) and let $\tau \in \left[M_{max}^-, \frac{2}{3}\right)$ be arbitrary, but fixed. From (6.2), (6.7), and $\sigma(x) > 0$ $(x \in [0,1])$, one gets

$$\int_0^1 |F(\tau|x)|^2 \, dx = \frac{1}{12} \int_0^1 \left| \frac{M^+(x) - \tau}{\sigma(x)} \right|^2 dx < \infty. \tag{6.31}$$

Furthermore, (6.2), (6.7), (6.17) – (6.19), and

$$\max_{x\in[0,1]} \left| \frac{d^i}{dx^i} \left(M^+(x) - \tau\right) \right| \leq \max_{x\in[0,1]} \left| \frac{d^3}{dx^3} M^+(x) \right| = 64 \quad (i = 0,1,2,3)$$

yield for all $k \in \{1,2,3\}$

$$\begin{aligned} \int_0^1 \left| \frac{d^k}{dx^k} F(\tau|x) \right|^2 dx &= \frac{1}{12} \int_0^1 \left| \frac{d^k}{dx^k} \frac{M^+(x) - \tau}{\sigma(x)} \right|^2 dx \\ &\leq \frac{1}{12} \cdot 3^2 \cdot 64^2 \sum_{i=0}^k \sum_{j=0}^k \int_0^1 \left| \frac{d^i}{dx^i} \sigma^{-1}(x) \right| \cdot \left| \frac{d^j}{dx^j} \sigma^{-1}(x) \right| dx \\ &= 3072 \sum_{i=0}^k \sum_{j=0}^k \int_0^1 \left| \frac{d^i}{dx^i} \sigma^{-1}(x) \right| \cdot \left| \frac{d^j}{dx^j} \sigma^{-1}(x) \right| dx, \end{aligned} \tag{6.32}$$

where

$$\sigma^{-a}(x) := \left(\frac{1}{\sigma(x)} \right)^a \quad (a \in \mathbb{N}, x \in [0,1]).$$

Next, we express the first three derivatives of σ^{-1} by the derivatives of σ, i.e.,

$$\frac{d}{dx} \sigma^{-1}(x) = -\sigma^{-2}(x) \cdot \frac{d}{dx} \sigma(x), \tag{6.33}$$

$$\frac{d^2}{dx^2} \sigma^{-1}(x) = 2\sigma^{-3}(x) \cdot \left(\frac{d}{dx} \sigma(x) \right)^2 - \sigma^{-2}(x) \cdot \frac{d^2}{dx^2} \sigma(x), \tag{6.34}$$

and

$$\begin{aligned} \frac{d^3}{dx^3} \sigma^{-1}(x) &= -6\sigma^{-4}(x) \cdot \left(\frac{d}{dx} \sigma(x) \right)^3 + 6\sigma^{-3}(x) \cdot \frac{d}{dx} \sigma(x) \cdot \frac{d^2}{dx^2} \sigma(x) \\ &\quad -\sigma^{-2}(x) \cdot \frac{d^3}{dx^3} \sigma(x) \end{aligned} \tag{6.35}$$

$(0 \leq x \leq 1)$.

For all $k \in \{1,2,3\}$, (6.21), (6.32) – (6.35), and $\sigma(x) > 0$ $(x \in [0,1])$ imply

$$\int_0^1 \left| \frac{d^k}{dx^k} F(\tau|x) \right|^2 dx < \infty. \tag{6.36}$$

This together with (6.31) shows that $F(\tau|\cdot) \in W_3([0,1])$.

On the other hand, one can conclude from (6.35) for all $x \in (0,1)$

$$\begin{aligned}\frac{d^4}{dx^4}\,\sigma^{-1}(x) &= 24\,\sigma^{-5}(x)\cdot\left(\frac{d}{dx}\,\sigma(x)\right)^4 - 36\,\sigma^{-4}(x)\cdot\left(\frac{d}{dx}\,\sigma(x)\right)^2\cdot\frac{d^2}{dx^2}\,\sigma(x)\\ &\quad +6\,\sigma^{-3}(x)\cdot\left(\frac{d^2}{dx^2}\,\sigma(x)\right)^2 + 8\,\sigma^{-3}(x)\cdot\frac{d}{dx}\,\sigma(x)\cdot\frac{d^3}{dx^3}\,\sigma(x)\\ &\quad -\sigma^{-2}(x)\cdot\frac{d^4}{dx^4}\,\sigma(x). \qquad (6.37)\end{aligned}$$

Now (6.2), (6.7), (6.26), and (6.37) yield

$$\begin{aligned}\int_0^1\left|\frac{d^4}{dx^4}\,F\left(\tau|x\right)\right|^2 dx &= \frac{1}{12}\int_0^1\left|\sum_{i=0}^{4}\frac{24}{i!\cdot(4-i)!}\cdot\frac{d^{4-i}}{dx^{4-i}}\,\left(M^+(x)-\tau\right)\cdot\frac{d^i}{dx^i}\,\sigma^{-1}(x)\right|^2 dx\\ &\geq \frac{1}{12}\int_0^{\frac{7}{500}}\left|\left(M^+(x)-\tau\right)\cdot\frac{d^4}{dx^4}\,\sigma^{-1}(x)\right|^2 dx\\ &\geq \frac{1}{12}\left(\frac{2}{3}-\tau\right)^2\cdot\int_0^{\frac{7}{500}}\left|\frac{d^4}{dx^4}\,\sigma(x)\right|^2 dx\\ &= \infty. \qquad (6.38)\end{aligned}$$

Here, we used that for all $x \in (0, \frac{7}{500}]$

$$\begin{aligned}\sum_{i=0}^{3}\frac{24}{i!\cdot(4-i)!}\cdot\frac{d^{4-i}}{dx^{4-i}}\,\left(M^+(x)-\tau\right)\cdot\frac{d^i}{dx^i}\,\sigma^{-1}(x) &\leq 0,\\ \left(M^+(x)-\tau\right)\cdot\frac{d^4}{dx^4}\,\sigma^{-1}(x) &\leq 0,\\ \left|\frac{d^4}{dx^4}\,\sigma^{-1}(x)\right| &\geq \left|\frac{d^4}{dx^4}\,\sigma(x)\right|,\end{aligned}$$

and $M^+(x) > M^+(0) = \frac{2}{3} > \tau$.

Finally, $0 < J_3^2(F\left(\tau|\cdot\right)) < \infty$ follows from (6.36) and

$$\begin{aligned}J_3^2(F\left(\tau|\cdot\right)) &= \frac{1}{12}\int_0^1\left|\sum_{i=0}^{3}\frac{6}{i!\cdot(3-i)!}\cdot\frac{d^{3-i}}{dx^{3-i}}\,\left(M^+(x)-\tau\right)\cdot\frac{d^i}{dx^i}\,\sigma^{-1}(x)\right|^2 dx\\ &\geq \frac{1}{12}\int_0^1\left|\left(M^+(x)-\tau\right)\cdot\frac{d^3}{dx^3}\,\sigma^{-1}(x)\right|^2 dx\\ &\geq \frac{1}{12}\left(\frac{2}{3}-\tau\right)^2\int_0^{\frac{1}{200}}\left|\frac{d^3}{dx^3}\,\sigma(x)\right|^2 dx > 0\end{aligned}$$

(cf. (6.29) and (6.38)). This proves the third part of lemma 6.1.

□

In the remaining part of this section we show that for the distribution of (X, Y, C) chosen in our simulation study, regularity assumption (4.10), which controls the rate of convergence of the maximum squared transformation errors in Lemma 4.2, is fulfilled for $\gamma = \frac{6}{7}$ (cf. (6.8)).

Proof of Lemma 6.2. First note that (6.1), (6.2), **(SM1c)**, and **(SM2)** imply $\tau_F = 1$. Let $M^+(x)$, M^-_{max}, and $R(t,x)$ be given by (6.2), (6.3), and (6.5), respectively ($t \in \mathbb{R}$, $x \in [0,1]$). Similar to (6.7), one can conclude for all $t \geq \frac{2}{3}$ and all $x \in [0,1]$ from (6.2), (6.3), and (6.6)

$$F\left(t|x\right) = I_{\left[\, R(t,x) \leq \sqrt{3} \,\right]} \cdot \frac{M^+(x) - t}{2\sqrt{3}\,\sigma(x)}. \tag{6.39}$$

Now, observe that $R(t,x) \leq \sqrt{3}$ is equivalent to $t \leq M^+(x) = 1 - \frac{1}{3}(2x-1)^4$ and this, in turn, to

$$x_t^- := \frac{1}{2} - \frac{1}{2}\left(3 - 3t\right)^{\frac{1}{4}} \leq x \leq \frac{1}{2} + \frac{1}{2}\left(3 - 3t\right)^{\frac{1}{4}} =: x_t^+ \quad \left(\frac{2}{3} \leq t \leq 1\right).$$

Hence, one gets for all $\frac{2}{3} \leq t \leq 1$ from (6.39)

$$F\left(t|x\right) = \begin{cases} \frac{M^+(x)-t}{2\sqrt{3}\,\sigma(x)} & \text{if} \quad x \in \left[x_t^-, x_t^+\right] \\ 0 & \text{if} \quad x \notin \left[x_t^-, x_t^+\right]. \end{cases} \tag{6.40}$$

Next, we apply (6.40) in order to derive a lower bound on the survival function $F(t)$ for $\frac{2}{3} \leq t \leq 1$. From **(SM1b)**, one can conclude that

$$0 < \sigma(x) \leq \sigma\left(\frac{1}{2}\right) = \frac{1}{2\sqrt{3}} \quad \forall x \in [0,1]$$

This together with **(SM2)** and (6.40) implies for all $\frac{2}{3} \leq t \leq 1$

$$\begin{aligned} F(t) &= \mathbf{E}\left[\, F\left(t|X\right)\right] = \int_{x_t^-}^{x_t^+} \frac{M^+(x) - t}{2\sqrt{3}\,\sigma(x)}\, dx \\ &\geq \int_{x_t^-}^{x_t^+} \left(M^+(x) - t\right)\, dx = \int_{x_t^-}^{x_t^+} \left(1 - \frac{1}{3}\left(2x-1\right)^4 - t\right) dx \\ &= x \cdot (1-t) - \frac{1}{30}\left(2x-1\right)^5 \Big|_{x_t^-}^{x_t^+} = \frac{4}{5}\, 3^{\frac{1}{4}}\, (1-t)^{\frac{5}{4}}. \end{aligned} \tag{6.41}$$

Here, we used that $0 \leq x_t^- \leq x_t^+ \leq 1$ for all $\frac{2}{3} \leq t \leq 1$.

On the other hand, we have from **(SM3)**

$$G(t) = \mathbf{P}\left[\, C > t\,\right] = \begin{cases} \exp\left(-\frac{3}{4}\, t\right) & \text{if} \quad t \geq 0 \\ 1 & \text{if} \quad t < 0. \end{cases} \tag{6.42}$$

Since F is monotonically decreasing on $[0,1]$, (6.41) and (6.42) yield

$$\begin{aligned}
-\int_0^1 F(t)^{-\frac{3}{4}}\,dG(t) &= \frac{3}{4}\int_0^1 F(t)^{-\frac{3}{4}}\exp\left(-\frac{3}{4}t\right)\,dt \\
&\le \frac{3}{4}\int_0^1 F(t)^{-\frac{3}{4}}\,dt = \frac{3}{4}\int_0^{\frac{2}{3}} F(t)^{-\frac{3}{4}}\,dt + \frac{3}{4}\int_{\frac{2}{3}}^1 F(t)^{-\frac{3}{4}}\,dt \\
&\le \frac{3}{4}\cdot\frac{2}{3}\cdot\left(\frac{4}{5}\cdot 3^{\frac{1}{4}}\cdot 3^{-\frac{5}{4}}\right)^{-\frac{3}{4}} + \frac{3}{4}\int_{\frac{2}{3}}^1 \left(\frac{4}{5}\cdot 3^{\frac{1}{4}}\cdot(1-t)^{\frac{5}{4}}\right)^{-\frac{3}{4}}\,dt \\
&\le 2 + \int_{\frac{2}{3}}^1 (1-t)^{-\frac{15}{16}}\,dt = 2 + (-16)\cdot(1-t)^{\frac{1}{16}}\Big|_{\frac{2}{3}}^1 \\
&= 2 + 16\cdot 3^{-\frac{1}{16}} \\
&< \infty.
\end{aligned}$$

This implies the assertion of Lemma 6.2.

□

Chapter 7

Applications to real data

In Chapter 7, we apply our MSSE of the regression function to two publicly available data sets. Section 7.1 contains an analysis of the well-known Stanford heart transplant data, which has already been investigated by Miller and Halpern (1982), Escobar and Meeker (1992), and Fan and Gijbels (1994, 1996) among many others. Here, different versions of our MSSE are compared with various estimates of the conditional mean or median lifetime time defined in literature. In Section 7.2, we discuss interesting features related to the breast cancer data set described in Van de Vijver, He, Van't Veer et al. (2002).

All pictures and computations of the estimates presented below were performed using R-2.6.2 (www.r-project.org), including libraries Design 2.1-1, MASS 7.2-40, scatterplot3d 0.3-25, and survival 2.33 as well as self-written functions for the computation of the MSSE.

7.1 Stanford heart transplant data

The Stanford heart transplantation program, described in more detail in Crowley and Hu (1977) and Miller and Halpern (1982), started in October 1967. Until February 1980, 249 patients were admitted to the study after a medical inspection. For each participant, a donor heart, matched on the blood type, was sought. Out of the 249 patients, 184 received a heart transplantation, a few having multiple transplants. The other participants were lost due to follow up, died or improved sufficiently while waiting for a donor heart.

For the 184 recipients, Miller and Halpern (1982) reported two covariates, the age at time of first transplant (ranging from 12 to 64 years) and the T5 mismatch score (ranging

from 0 to 3.05). The latter one is a measure for the degree of tissue incompatibility between the donor and the patients heart. Because the tissue typing was never completed, the T5 mismatch scores of 27 recipients are not available. The time of interest, i.e., the "lifetime", in the analysis of the 184 patients is the (logarithm of the) lifetime since transplantation. Of the recipients, 113 (61.4%) died before February 1980, i.e., were uncensored, while 71 (38.6%) were still alive and a censored observation was listed.

In their study of the heart transplant data set, Miller and Halpern (1982) and Escobar and Meeker (1992) both reported that there is no evidence that the T5 mismatch score is a useful explanatory variable for survival. Furthermore, Escobar and Meeker (1992) observed that this conclusion is stable, since it is not seriously influenced by a limited perturbation of the observed times or if the most influential data points were dropped. Consequently, in Miller and Halpern (1982), Escobar and Meeker (1992), Fan and Gijbels (1994, 1996), and this thesis, the T5 mismatch score is not used in the further analysis.

Observe that Miller and Halpern (1982) and Fan and Gijbels (1994, 1996) initially only considered the 157 patients with complete tissue typing. However, after discarding T5 mismatch score, this is not necessary and we use the data of all 184 recipients. Note that the differences between the estimates computed for the 157 and the 184 patients are minor and do not affect our interpretation of the results.

Miller and Halpern (1982) applied four different semiparametric methods to the Stanford heart transplant data in order to estimate the regression function m and the conditional median lifetime, respectively. Their first approach is based on the proportional hazard model with linear predictor, introduced by Cox (1972, 1975). The other three methods, which are due to Miller (1976), Buckley and James (1979), and Koul, Susarla, and Van Ryzin (1981), assume that m is a linear function.

For the latter one, the observed times are first transformed according to (2.22) with $\alpha_1 = -1$. Then a least squares approach in order to estimate m is applied to the new, virtually uncensored data. The method of Buckley and James (1979) is quite similar, but uses a somewhat different transformation which involves the unknown regression function and therefore leads to an iterative scheme. Miller (1976) suggested to estimate the slope and the intercept of m by minimizing a weighted sum of squares of residuals. Here, the weights depend on a modified Kaplan-Meier estimate of F. Note that if the lifetimes

are measured on a logarithmic scale, then the linear regression model corresponds to an accelerated failure time model (vide, e.g., Martinussen and Scheike (2006)).

As mentioned above, Miller and Halpern (1982) also presented a hazard risk approach in order to analyze censored data. As suggested by Cox (1972, 1975), they assumed that the failure rate for the lifetime Y of an individual given a d-dimensional vector of covariates $X \in \mathbb{R}^d$ follows a log-linear model. Here, the existence of the conditional density of Y given X is required. In this model, one can derive an explicit form of the conditional survival function $F(t|x)$ of Y ($t \in \mathbb{R}, x \in \mathbb{R}^d$). Note that for arbitrary, but fixed $(t, x) \in \mathbb{R} \times \mathbb{R}^d$, this form is completely determined by two unknown parameters. The first one is the vector of regression coefficients, which is independent of t and x, and the second one is the baseline hazard rate, which is a univariate function in t. Hence, an estimate $\hat{F}_n^{Cox}(t|x)$ of $F(t|x)$ ($t \in \mathbb{R}, x \in \mathbb{R}^d$) may simply be obtained by replacing these unknown parameters with appropriate estimators. Miller and Halpern (1982) proposed to calculate the latter ones with the techniques established by Cox (1972, 1975) and Breslow (1974). Based on these estimates, they introduced an estimate med_n^{Cox} of the conditional median lifetime, which is derived by calculating at each $x \in \mathbb{R}^d$ the value of $t \in \mathbb{R}$, where $\hat{F}_n^{Cox}(t|x) = \frac{1}{2}$.

In the following, we compare the performance of the four methods presented above on the Stanford heart transplant data, where all 184 recipients are considered. As already mentioned a similar analysis for the 157 patients with complete tissue typing can be found in Miller and Halpern (1982). Observe that the three regression based approaches are applied in order to estimate m, while med_n^{Cox} estimates the conditional median lifetime. Miller and Halpern (1982) argued that within the hazard risk approach, it is reasonable to select med_n^{Cox} because it is much easier to compute and, since the time scale is irrelevant to this method, more appropriate than an estimate of m.

Figure 7.1 (a) displays the scatterplot of the observed times (in days) on a logarithmic scale (base 10) versus age at transplant (in years) for the 184 Stanford heart transplant patients. Furthermore, the linear fits to these data with the four estimates described in Miller and Halpern (1982) are given. Here, the three methods based on parametric regression are abbreviated as BJ (Buckley and James (1979)), Miller (Miller (1982)), and KSV (Koul, Susarla, and Van Ryzin (1981)).

Note that for one of the 184 patients, death at time $t = 0$ was reported. Since in the

log-linear model, it is impossible to account for such an observation, it was recoded as $t = 1$ according to Miller and Halpern (1982) and Fan and Gijbels (1994, 1996).

In the comparison of the four methods mentioned above, Miller and Halpern (1982) concluded that the estimate of Buckley and James (1979) and med_n^{Cox} performed best on the heart transplant data. Here, the method of Koul, Susarla, and Van Ryzin (1981) differs substantially from the other three approaches (vide Figure 7.1 (a)). This is due to the fact that within the Stanford heart transplant data, there occur proportionately more censored observations in younger recipients than in older patients. Transforming the sample according to (2.22) with $\alpha_1 = -1$ sets all censored observations to zero, while the uncensored observations are enlarged (cf. Section 6.4). Hence, the linear regression estimate of Koul, Susarla, and Van Ryzin (1981) predicts a positive slope. Yet intuitively, the lifetime since transplantation should rather diminish with increasing recipient age and the slope should be negative, as for all other linear estimates. Applying the linear fit of Miller (1976) to the heart transplant data also seems to be quite inappropriate (cf. Figure 7.1 (a)), since it is nearly constant over the whole range, implying that the age of a patient is not a risk factor for survival.

However, the linear fits to the data are poor for all of the four methods. In particular, the estimate of the conditional median log-lifetime, med_n^{Cox}, exceeds all observed lifetimes (and censoring times) for young patients. Similar to Miller and Halpern (1982), we therefore computed a quadratic fit with med_n^{Cox} and the estimate of Buckley and James (1979). The results of these analyses are illustrated in Figure 7.1 (b) and compared to two different versions of the parametric estimate described in Escobar and Meeker (1992). Here, it is assumed that an accelerated failure time model holds, where the lifetime is either lognormal or Weibull distributed. For both versions, a quadratic fit to the data was computed (abbreviated as EM_ln and EM_W, respectively).

Figure 7.1 (b) shows that all four estimates perform quite similar for older recipients, whereas there is some difference for younger patients. Compared to Figure 7.1 (a), the quadratic model fits the data better than the linear model. In particular, it reveals that for the extreme young recipients, the lifetime is shorter than for middle-aged patients. Anyway, the quadratic fits all have a big drawback: they suggest that the lifetime increases from young to middle-aged recipients in the same way as it diminishes from middle-aged

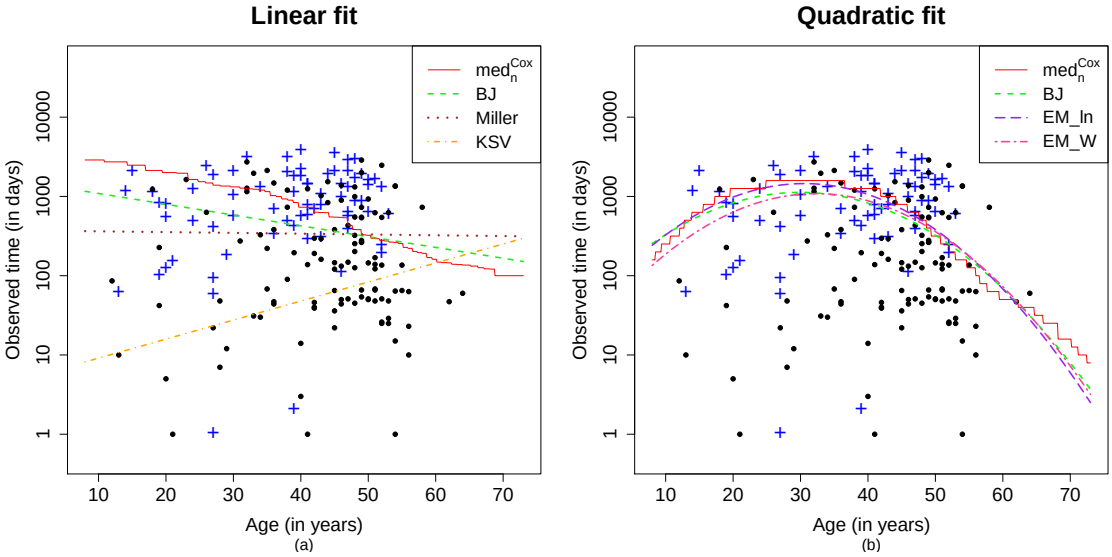

Figure 7.1: Stanford heart transplant data set for the 184 recipients and the estimates described in Miller and Halpern (1982). The character "•" indicates the uncensored observations, while the censored observations are presented by "+". The individual panels show (a) Stanford heart transplant data with observed time (in days) on logarithmic scale (base 10) plotted against age at transplant (in years) and the four different estimates (linear fit) described in Miller and Halpern (1982) and (b) as in Figure 7.1 (a) but for quadratic regression and the two best performing estimates due to Miller and Halpern (1982) as well as the estimates of Escobar and Meeker (1992). Details are given in the text.

to old patients, which seems to be fairly unrealistic. Beyond, this feature essentially cannot be improved by using more complex parametric fits with ordinary polynomials (i.e., with natural exponents exceeding two). This is indicated by an analysis of the cubic model with the four estimates of Figure 7.1 (b) which shows that there is only a very small difference to the corresponding quadratic fit. Hence, we conclude that the presented global parametric and semiparametric methods miss important features of the data, and a suitable nonparametric approach may be more reasonable.

As mentioned in Section 2.2, Fan and Gijbels (1994, 1996) also investigated the idea of the censoring unbiased transformation in order to estimate the regression function in the presence of censored data. But in contrast to Buckley and James (1979) and Koul, Susarla, and Van Ryzin (1981), a nonparametric regression technique was chosen. To be

more precise, Fan and Gijbels (1994, 1996) suggested to transform the data according to (2.22) and then to apply a locally weighted least squares estimate. Here, the parameter α_1 was calculated similar to (6.9) with the only difference that the whole sample instead of a learning data set was used.

Furthermore, in the univariate situation, Fan and Gijbels (1994, 1996) proposed an alternative transformation, which incorporates the values of the covariate via the same kernel function which is used in order to compute the locally weighted least squares estimate. This transformation has the advantage that it leaves the uncensored observations unchanged and does not require that the censoring times and the covariates are independent, but that Y and C are conditionally independent given X. In their applications to real and simulated data, the kernel was chosen as the standard normal density function. The parameter k^{FG} of the bandwidth of the estimate was determined via a n-fold cross-validation (also known as leave-one out crossvalidation). However, the disadvantage of this alternative transformation is that it is only applicable in univariate regression.

Figure 7.2 displays the Stanford heart transplant data in analogy to Figure 7.1 together with the estimates described in Fan and Gijbels (1994, 1996) based (a) on the transformation in the univariate case and (b) on (2.22) with $\alpha_1 \approx 0.056$, respectively. The optimal choice of the bandwidth parameter due to the cross-validation procedure is $k^{FG} = 24$ in the former and $k^{FG} = 34$ in the latter case. Similar to Fan and Gijbels (1994), we additionally choose smaller bandwith parameters $k^{FG} = 8$ and $k^{FG} = 9$, thereby trying to detect a finer structure.

Observe that in panel (a) the curves are nearly constant for the 17 to 48 year old recipients. In addition, for patients older than 48 years, the estimate with $k^{FG} = 24$ predicts a nearly linear slope. As a consequence, Fan and Gijbels (1994, 1996) conjectured that for younger recipients, the log-lifetime is constant, while it decreases linearly for older patients. On the other hand, the estimates in panel (b) predict that the log-lifetime is diminished for young patients compared to middle-aged recipients. This is mainly due to the different treatment of the censored observations by (2.22) in contrast to the univariate transformation given in Fan and Gijbels (1994, 1996), which generally results in a smaller variability of the random variables of the latter transformation. Yet, the whole sample is used to compute $\hat{U}_1^{(1)}, \ldots, \hat{U}_n^{(1)}$. In contrast, for the conversion of a

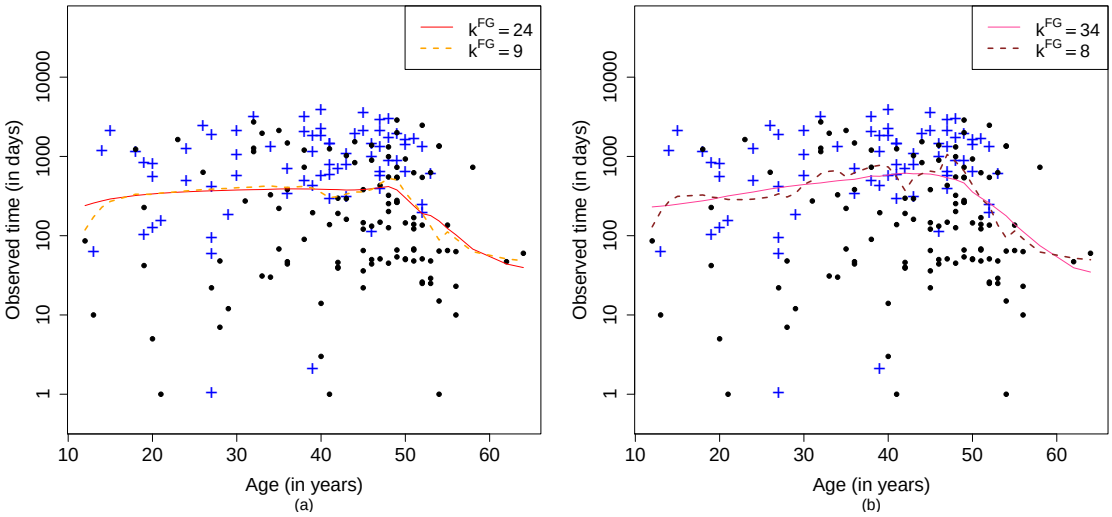

Figure 7.2: Stanford heart transplant data set for the 184 patients similar to Figure 7.1 together with the estimates of Fan and Gijbels (1994, 1996). The individual panels show the Stanford heart transplant data with observed time (in days) on logarithmic scale (base 10) plotted against age at transplant (in years) and the estimates of Fan and Gijbels (1994, 1996) based on the transformation (a) in the univariate case with bandwidth parameter $k^{FG} = 24$ and $k^{FG} = 9$ and (b) according to (2.22) with transformation parameter $\alpha_1 \approx 0.056$, and bandwidth parameter $k^{FG} = 34$ and $k^{FG} = 8$, respectively..

censored observation C_i by the univariate transformation, only those uncensored data points (X_j, Y_j) with $Y_j > C_i$ and X_j being in a neighborhood of X_i (defined via k^{FG}) are incorporated ($i \neq j; i, j \in \{1, \ldots, n\}$). If no such datum point exists, then C_i is not transformed.

As for the Stanford heart transplant data, the majority of the censored observations exceed the most uncensored ones. Moreover, for 38–52 year old recipients the amount of the censored data points with this property is larger than for the other patients. Therefore, for most censored observations only few uncensored data points are used in order to estimate the random variables of the univariate transformation. Hence, calculating $\hat{U}_1^{(1)}, \ldots, \hat{U}_n^{(1)}$ instead may be more appropriate. But then, by transforming the data according to (2.22), one assumes that C and X are independent, which obviously is inappropriate for the heart transplant data.

To sum it up, the different behavior of the estimates in Figure 7.2 (a) and (b) for young to middle-aged patients is mainly caused by the following facts. The univariate transformation causes the censored observations to be located at about the same level. This results in the curves in panel (a) which are nearly constant over a large fraction of the plot. On the other hand, the transformation according to (2.22) with $\alpha_1 \approx 0.056$ operates mainly on the censored observations of the 38–52 year old patients, while the data points for recipients younger than 38 years are nearly left unchanged. Hence, the estimates in panel (b) predict that the lifetime increases from young to middle-aged patients.

Finally, we applied our MSSE (2.45) to the Stanford heart transplant data. Note that m_n depends on the particular choice of the learning and the testing data (cf. Section 2.4). In order to determine its effect on our estimation of m, we randomly selected 50 different learning and testing data sets, each of sample size $n_{\mathtt{l}} = n_{\mathtt{t}} = 92$. For each of these sets, an MSSE was calculated according to (2.45). For each recipient age, we computed the median $m_{n,med}$, the 10%–quantile $m_{n,10\%}$, and the 90%–quantile $m_{n,90\%}$ of the log-lifetimes predicted by the 50 estimates. Beyond, $m^{UT}_{n,med}$, $m^{UT}_{n,10\%}$, and $m^{UT}_{n,90\%}$ were derived in an analogous way by replacing (2.36) and (2.37) with the transformation suggested by Fan and Gijbels (1994, 1996) for the univariate case where $k^{FG} = 24$.

In a statistical application, one may prefer to avoid the dependency of m_n on the particular choice of the learning and testing data set in order to get a unique (and at best a more accurate) result. Here, a n-fold cross-validation similar to Fan and Gijbels (1994, 1996) instead of the splitting of the sample technique can be applied. Note that in this case, it is not advisable to calculate (2.36) and (2.37) separately (at least if censored data occurs). This is due to the fact that for n-fold cross-validation, each of the n testing data sets has sample size $n_{\mathtt{t}} = 1$, i.e., we would have just one observation to compute the Kaplan-Meier estimate of G on the testing data. Therefore, we rather define a cross-validation version m^{CV}_n of m_n as follows (cf. Fan and Gijbels (1994, 1996)).

First, the censored data is transformed according to (2.25). Then, for all $i = 1, \ldots, n$ and all $(k, \lambda) \in K_n \times \Lambda_n$, we compute a MSSE $m_{n-1,i,(k,\lambda)}$ in analogy to (2.28), where we replace $\hat{\mathcal{D}}^{(1)}_n$ with $\hat{\mathcal{D}}^{(1)}_n \setminus \{(X_i, \hat{U}^{(1)}_i)\}$ and λ_n with λ. Here, $\hat{\mathcal{D}}^{(1)}_n$, K_n, and Λ_n are given by (2.26), (2.41), and (2.42), respectively. In the next step, we choose that pair of parameters out of $K_n \times \Lambda_n$ which minimizes the empirical $\mathcal{L}_2$ risk of $m_{n-1,i,(k,\lambda)}$ $(i = 1, \ldots, n)$. To be

more precise, let

$$\left(k^{CV}, \lambda^{CV}\right) := \underset{(k,\lambda)\in K_n\times\Lambda_n}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^{n} |m_{n-1,i,(k,\lambda)}(X_i) - \hat{U}_i^{(1)}|^2 \right),$$

where $\hat{U}_i^{(1)}$ $(i = 1, \ldots, n)$ is given by (2.25). Finally, we use this pair of parameters in order to construct a MSSE according to (2.28) from the whole transformed data $\hat{\mathcal{D}}_n^{(1)}$, i.e, we define $m_n^{CV}(\cdot) := m_{n,(k^{CV},\lambda^{CV})}(\cdot)$. Moreover, let $m_n^{CV,UT}$ be given by replacing (2.26) in the estimation procedure described above with the data from the univariate transformation of Fan and Gijbels (1994, 1996).

In general, it is more complicated to show that a regression estimate based on a cross-validation procedure achieves the optimal rate of convergence than in case that a splitting of the sample technique is applied (cf., e.g., Györfi, Kohler, Krzyżak, and Walk (2002)), Chapters 7 and 8). If censored data occurs, this problem gets even worse. In our case, we have to deal with the fact that $m_{n-1,i,(k,\lambda)}$ and $(X_i, \hat{U}_i^{(1)})$ both depend on (X_i, Z_i, δ_i) $(i \in \{1, \ldots, n\})$. This is because for each of the random variables (2.25), we calculate the Kaplan-Meier estimate of G on the whole sample

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

Hence, in contrast to the proofs of Theorem 5.1 and Lemma 5.1, we cannot use standard techniques from usual nonparametric regression in order to reduce the problem mentioned above to the problem of deriving a fast stochastic rate of convergence of $m_{n-1,i,(k,\lambda)}$. In contrast, the mode of dependency of $m_{n-1,i,(k,\lambda)}$ and $\hat{\mathcal{D}}_n^{(1)} \setminus \{(X_i, \hat{U}_i^{(1)})\}$ on $(X_i, \hat{U}_i^{(1)})$ has to be established and one has to impose additional assumptions on the distribution of (X, Y, C). Similar to above, one can conclude that the same problem occurs for $m_n^{CV,UT}$.

Figure 7.3 displays the Stanford heart transplant data as in Figure 7.1 together with (a) $m_n^{CV,UT}$ (green solid line), $m_{n,med}^{UT}$ (red solid line), $m_{n,10\%}^{UT}$ (lower red dotted line), and $m_{n,90\%}^{UT}$ (upper red dotted line) with $k^{FG} = 24$, and (b) m_n^{CV} (green solid line) $m_{n,med}$ (red solid line), $m_{n,10\%}$, and $m_{n,90\%}$ (lower and upper red dotted line) with $\alpha_1 \approx 0.056$.

Similar to Figure 7.2 (a), one observes that $m_n^{CV,UT}$ and $m_{n,med}^{UT}$ show a nearly constant shape for the 17-48 year old recipients. However, for patients older than 55 years, there is a difference between the estimates of Fan and Gijbels (1994, 1996) and the curves in Figure 7.3 (a). While the first ones show a similar linear slope as for 48–55 year old

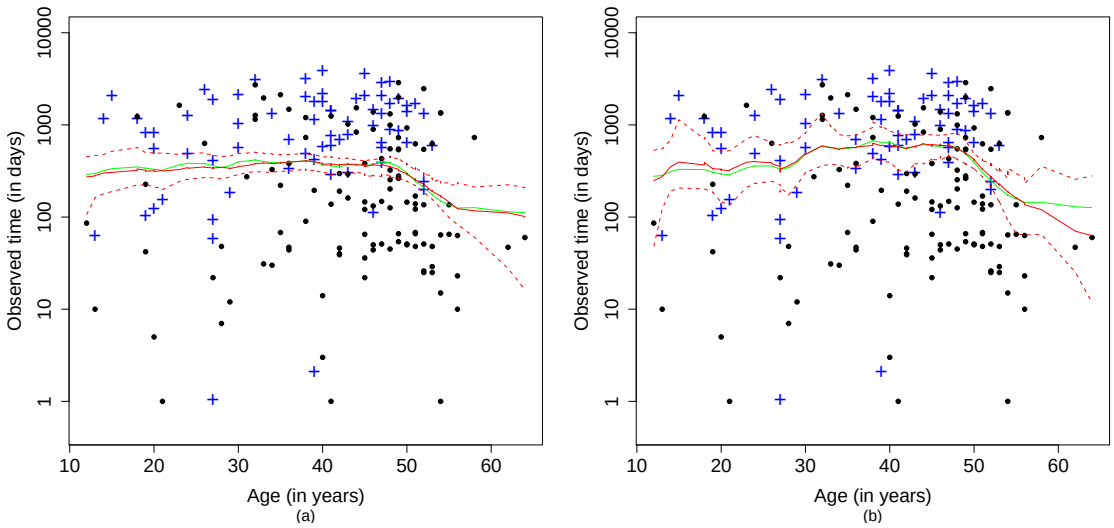

Figure 7.3: Stanford heart transplant data set for the 184 patients similar to Figure 7.1 with observed time (in days) on logarithmic scale (base 10) plotted against age at transplant (in years) and (a) $m_n^{CV,UT}$ (green solid line), $m_{n,med}^{UT}$ (red solid line), $m_{n,10\%}^{UT}$ (lower red dotted line), and $m_{n,90\%}^{UT}$ (upper red dotted line) with $k^{FG} = 24$ and (b) m_n^{CV} (green solid line), $m_{n,med}$ (red solid line), $m_{n,10\%}$ (lower red dotted line) and $m_{n,90\%}$ (upper red dotted line) with $\alpha_1 \approx 0.056$.

recipients, the latter ones are almost constant. The curve with $k^{FG} = 24$ even falls below all the data points on the right hand side of Figure 7.2 (a). Hence, one may argue that $m_n^{CV,UT}$ and $m_{n,med}^{UT}$ are more appropriate estimates of the regression function on this interval. However, there are only few patients being older than 55 years at time of first transplantation.

In analogy to the estimates in Figure 7.2 (b), m_n^{CV} and $m_{n,med}$ increase from young to middle-aged recipients. Here, we observe that m_n^{CV} and $m_{n,med}$ do not predict a linear slope as the curve with $k^{FG} = 34$ in Figure 7.2 (b), but rather show a somewhat similar behavior to the estimate with $k^{FG} = 8$. Both estimates are nearly constant for up to 30 year old recipients as well as for 32 to 50 year old patients with a small increase in between. Note that the mean level of the curves on these intervals is at about the same level as of the estimates in panel (a).

As mentioned above, the variability of $\hat{U}_1^{(1)}, \ldots, \hat{U}_n^{(1)}$ is in general larger than that of

the random variables of univariate transformation. Consequently, for the estimates based on the latter ones the difference between the 90%–quantiles and the 10%–quantiles of the log-lifetimes predicted by the 50 estimates is smaller than for those based on the first ones. Moreover, $m_{n,10\%}^{UT}$ and $m_{n,90\%}^{UT}$ are smoother than $m_{n,10\%}$ and $m_{n,90\%}$.

Overall, the curves in Figure 7.3 show a similar behavior to those in Figure 7.2. Table 7.1 displays the median empirical L_2 risks of m_n for the 50 different choices of the learning and the traing data set together with the empirical L_2 risk of m_n^{CV} (or $m_n^{CV,UT}$ for the univariate transformation) and m_n^{FG} with respect to the transformed censored data. The first column lists these values for the transformation according to (2.25) with $\alpha_1 \approx 0.056$ (cf. Figure 7.2 (b) and Figure 7.3 (b)). For the estimates in the second column, $\alpha_1 \approx 0.013$ was chosen which corresponds to the median of the α_1^{FG}'s which were calculated on the 50 learning data sets, where α_1^{FG} is given by (6.9). The third column shows the median empirical L_2 risk and the empirical L_2 risks with respect to the univariate transformation, where $k^{FG} = 24$.

From Table 7.1, we deduce that at large, the MSSE fit the transformed heart transplant data for all three transformations slightly better than the estimates of Fan and Gijbels (1994, 1996). However, the difference between both estimation procedures is considerably smaller than the gap between the univariate transformation and the transformation according to (2.25). This is due to the above-mentioned difference in the variability of the converted random variables when applying the two transformations. Thus, for this data set where the censoring mechanism is not independent of the covariate under study,

	Transformation (2.36)		Univariate Transformation
	$\alpha_1 \approx 0.056$	$\alpha_1 \approx 0.013$	$k^{FG} = 24$
m_n	0.979	0.955	0.577
$m_n^{CV}, m_n^{CV,UT}$	1.016	0.974	0.568
m_n^{FG}	1.038	0.998	0.574

Table 7.1: Median empirical L_2 risk of m_n and empirical L_2 risk of m_n^{CV} (or $m_n^{CV,UT}$ for the univariate transformation) and m_n^{FG} with respect to the transformed censored data of three different transformations.

choosing the univariate transformation instead of the transformation according to (2.25) is much more important than the choice of either of the two estimation procedures.

Note that for the MSSE based on the transformation according to (2.25), further improvements may be possible by choosing a different value of the transformation parameter. In particular, one can deduce from the proof of Lemma 4.2 that Theorem 5.2 still holds, if we replace α_1 in (2.22) and (2.25) for each $i \in \{1, \ldots, n\}$ by $\bar{\alpha}_1(X_i)$. Here, $\bar{\alpha}_1 : \mathbb{R}^d \to \mathbb{R} : x \mapsto \bar{\alpha}_1(x)$ with $|\bar{\alpha}_1(x)| < \infty$ for all $x \in \mathbb{R}^d$. This provides the possibility to transform the data for different values of the covariates in a different way and hence reduce the variability of the transformed random variables.

7.2 Breast cancer data set of Van de Vijver et al. (2002)

The breast cancer data set published by Van de Vijver, He, Van't Veer et al. (2002) comprises tumors of 295 women selected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute according to several predefined criteria, which should assure a homogeneous patient collective. Among other things it was required that the tumor was a so-called *primary invasive breast carcinoma*, which has the ability to metastasize, but has yet not spread beyond the breast. Furthermore, only tumors with less than 5 cm in diameter were considered. Moreover, it was demanded that there was no prior history of cancer (except for a non-aggressive form of skin cancer), the apical axillary lymph nodes were tumor-negative, the year of diagnosis was between 1984 and 1995, and the age at diagnosis was below 53 years.

For a period of at least five years, all women were assessed at least annually. The median duration of follow-up since surgery was 6.7 years. Out of the 295 patients, 130 received a so-called *adjuvant systemic therapy*, consisting of chemotherapy (90 patients), hormonal therapy (20 patients) or both (20 patients). Overall survival was defined as the time from date of surgery to date of death from any cause (uncensored observations) or to date of the last follow up visit (censored observation). Distant metastasis-free survival time was specified as the time from date of surgery to date of first occurrence of a distant metastasis (uncensored observations) or to date of last follow up visit, non-breast cancer related death, recurrence of local or regional disease or second primary cancer (censored

observation).

Several clinical and histopathological characteristics of the tumors were reported. These comprehend widely used prognostic factors in medical praxis such as the estrogen receptors (ER) expression level, the number of metastases, the lymph node status, the tumor diameter, and the histological grade of the tumors. Moreover, by histopathological methods, it was determined whether a patient was a carrier of a germline mutation in the BRCA1 or the BRCA2 gene. The wild type alleles of these so-called *tumor suppressor genes* are known to lower the risk of getting breast cancer and are assumed to be incorporated in the DNA repair in the cells. In contrast, carriers of a mutation in BRCA1 or BRCA2 are more likely to develop tumors.

However, breast cancer is a very complex disease of the mammary gland with many subgroups, partially differing seriously in their clinical courses. It thus cannot be explained satisfactorily by the common clinical or histopathological characteristics. Today, it is generally agreed that breast cancer, like other types of cancer, is the final outcome of multiple environmental, hereditary, and genetic factors. Carcinogenesis is often considered as a multistage process, in which normal cells are transformed to tumor cells.

Among women worldwide, breast cancer is the most common cause of cancer death with about $460,000$ disease-related deaths in 2008. Every eighth to tenth women is at least once in her lifetime affected by breast cancer (Ferlay et al., GLOBOCAN 2008 v2.0).

At present, treatment of breast cancer may include a surgery (breast-conserving or mastectomy, possibly with axillary lymph node dissection), radiotherapy, chemotherapy, hormonal treatment, and/or antibody therapy. In clinical practice, the decision which therapies should be applied is based on the various histopathological or clinical characteristics. Here, often a systemic treatment, consisting of a chemotherapy, hormonal treatment and/or antibody therapy is ordered. However, it is known that $70\% - 80\%$ of the patients receiving this treatment do not benefit from it and may in addition suffer from the strong side effects. This is due to the fact that disappointingly, just on the basis of the histopathological and clinical variables, these patients cannot be distinguished from the remaining $20\% - 30\%$. Up to now, systemic therapy is therefore in general given to all patients with certain histopathological or clinical characteristics, although only a minor part of them benefits from this treatment.

Since genetic factors play a major role in carcinogenesis, measuring gene expression levels has become a standard tool in cancer research. Here, the hope is that the heterogeneity and complexity of tumors can be better modeled by adding these factors than with histopathological or clinical variables alone. Indeed, it has been shown in several studies that this improves tumor diagnosis and classification as well as the prediction of prognosis and the response to therapy and hence offers the potential to refine treatment for cancer patients. Today, one important tool for monitoring gene expression levels are the so-called *DNA microarrays*. They usually consist of a solid glass, plastic or silicon surface, where microscopic DNA spots, each representing a single gene, are fixed in a rectangular grid. Depending on the subject of a study, DNA microarrays may, e.g., cover just some interesting genes up to the complete genome of an organism or several organisms.

For each of the 295 women in their study, Van de Vijver, He, Van't Veer et al. (2002) measured the gene expression levels of the tumor cells with microarray containing approximately $25,000$ human genes and therefore covering nearly the whole human genome. Their goal was to demonstrate that solely on the basis of these data, one may better predict prognosis and response to therapy of breast cancer patients than with classical histopathological and clinical variables. To be more precise, for each of the 295 tumors, the expression levels of the genes on the microarray were used in order to calculate a single measure, which was then applied in order to classify the patients in a low-risk and a high-risk group.

In a previous study with the same microarrays and 78 breast cancer patients, Van't Veer, Dai, Van de Vijver et al. (2002) selected 70 genes based on the association of the expression of each gene with the probability of developing a distant metastasis within 5 years. For the 44 patients which remained disease free for at least 5 years, the mean of the gene expression levels of each the 70 genes (in the following denoted as average profile) were calculated. Then, for each tumor, Pearson's correlation coefficient between this average profile and the expression levels of the 70 genes was computed. A patient with a correlation coefficient of more than 0.4 was then assigned to the low-risk group, while all other patients were classified into the high-risk group. Here, this threshold was chosen by a cross-validation procedure and caused a 10% rate of false negative results.

Van de Vijver, He, Van't Veer et al. (2002) used the average profile of Van't Veer,

Dai, Van de Vijver et al. (2002) in order to determine the patients with low-risk and high-risk in their study. For each of the 295 women, the correlation coefficient with this profile was calculated and the same threshold in order to distinguish between the two groups was applied. Van de Vijver, He, Van't Veer et al. (2002) showed that within their dataset, this measure is a more powerful predictor of overall survival, development of distant metastasis within 5 years, and response to therapy than standard systems based on clinical and histopathological factors. In the following, we will refer to the correlation coefficients described above as VHV's.

Besides, Chang, Nuyten, Sneddon et al. (2005) observed that for the 295 patients of Van de Vijver, He, Van't Veer et al. (2002), overall survival and distant metastasis-free survival are also markedly diminished in women whose tumors are similar to a second, independent prognosis profile. In contrast to Van de Vijver, He, Van't Veer et al. (2002), the determination of this profile was knowledge-based. Here, only genes which are involved in processes of wound-healing of cells were considered. This choice was motivated by the observation of many similarities between the tumor microenvironment and wound-healing. Due to this fact, it has been proposed that carcinogenesis can at least partly be explained by malfunctions in the reparative processes of cells.

In a preceding study with a DNA microarray covering approximately $36,000$ genes, Chang, Sneddon, Alizadeh et al. (2004) identified 512 genes, which are associated with wound-healing processes in human cells. Note that the selection of these genes was based on their expression profiles in experiments with normal cells and cell lines, and no experiments with tumor cells were included in this step. Rather, Chang, Sneddon, Alizadeh et al. (2004) subsequently demonstrated by hierarchical cluster analyses of several cancer data sets, including the breast cancer data set of Van't Veer, Dai, Van de Vijver et al. (2002), that their gene set has the power to reveal a low-risk and a high-risk population. Furthermore, a prognosis profile similar to Van't Veer, Dai, Van de Vijver et al. (2002) was derived by computing for each of the of 512 genes the mean expression level for the patients in the latter group. Moreover, it was shown that 62 genes are sufficient to classify the patients in the data set of Van't Veer, Dai, Van de Vijver et al. (2002) into a group with poor and with good wound-healing properties, where the misclassification rate is approximately 6.4%. Chang, Sneddon, Alizadeh et al. (2004) report that this misclassifi-

cation rate could not be improved, even if all genes associated with wound-healing were considered.

Chang, Nuyten, Sneddon et al. (2005) applied this profile in order to analyze the breast cancer data set of Van de Vijver, He, Van't Veer et al. (2002). Out of the 512 genes which were selected in the study of Chang, Sneddon, Alizadeh et al. (2004), 459 were also measured by the microarray used in the study of Van de Vijver, He, Van't Veer et al. (2002). In analogy to Van't Veer, Dai, Van de Vijver et al. (2002), Pearson's correlation coefficient between the prognosis profile of Chang, Sneddon, Alizadeh et al. (2004), restricted to these genes, and the expression levels of the 459 genes was calculated for each of the 295 tumors. In the following, we term these coefficients as CNS.

In order to assign the 295 women either to a low-risk or a high-risk group, a threshold for the CNS's was determined via a splitting of the sample technique (cf. Section 2.4). Here, the learning and testing data were matched for all available clinical variables and all known risk factors. Chang, Nuyten, Sneddon et al. (2005) demonstrated that a threshold of -0.15 gave approximately 90% sensitivity for predicting distant metastasis as the first recurrence event in both data sets. Note that in contrast to the procedure of Van de Vijver, He, Van't Veer et al. (2002), in the approach of Chang, Nuyten, Sneddon et al. (2005), patients are classified into the high-risk group if their CNS is below the established threshold.

According to the results of Chang, Nuyten, Sneddon et al. (2005), the approaches of Van't Veer, Dai, Van de Vijver et al. (2002) and Chang, Sneddon, Alizadeh et al. (2004) gave overlapping and generally consistent predictions of outcomes, although their derivation was quite different. Note that only 2 of the 512 genes in the prognosis profile of Chang, Nuyten, Sneddon et al. (2005) were also contained in the 70 genes selected by Van de Vijver, He, Van't Veer et al. (2002). Moreover, approximately 70.8% of the patients were classified into the high-risk group or low risk-group by both procedures. Of the remaining 86 women, 67 were classified as high-risk by the profile of Chang, Sneddon, Alizadeh et al. (2004), but as low-risk by that of Van't Veer, Dai, Van de Vijver et al. (2002). In contrast, 19 patients were grouped vice versa.

An analysis of overall survival and distant metastasis-free survival time by means of Kaplan-Meier estimates revealed that the profile of Van't Veer, Dai, Van de Vijver et al.

(2002) performs slightly better on the data set of the 295 breast cancer patients than the profile of Chang, Sneddon, Alizadeh et al. (2004). E.g., for the approach of Van de Vijver, the predicted value of the survival function $F(t)$ at time $t = 10$ years is 94.5% for the low-risk group (mean standard error: 2.6%) and 54.6% in the high-risk group (mean standard error: 4.4%). In contrast, for the low-risk group determined by Chang, Nuyten, Sneddon et al. (2005), the estimated probability is only 92.3% (standard error: 3.8%), while it is 63.7% (standard error: 3.8%) for the corresponding high-risk group. This demonstrates that of the two prognosis profiles described above, the profile of Van't Veer, Dai, Van de Vijver et al. (2002) more accurately allocates patients to a low- or high-risk group.

However, the choice of the number of groups and the selection of the correlation thresholds in the analyses of Van de Vijver, He, Van't Veer et al. (2002) and Chang, Nuyten, Sneddon et al. (2005) are somewhat arbitrary. Moreover, many features can be left unrevealed if one only applies a simple classification rule. Regression based approaches may give deeper insight into such data. Here, we calculated our MSSE (2.45) on the breast cancer data set of Van de Vijver, He, Van't Veer et al. (2002). In order to compare our results with those of Chang, Nuyten, Sneddon et al. (2005), only the VHV's and CNS's are used as covariates. Furthermore, the parameters of the estimates were determined by the splitting of the sample technique with the same learning and testing data set as in Chang, Nuyten, Sneddon et al. (2005). Note that in this example, it is not advisable to randomly choose learning and testing set, since then it can happen that risk factors are not equally distributed between these two samples. This could seriously bias any analysis based on these data sets.

Figure 7.4 displays the breast cancer data set of Van de Vijver, He, Van't Veer et al. (2002) together with the MSSE (2.45) of the regression function. In panel (a), the lifetimes of the 295 patients since surgery (in years) and the censoring times, respectively, are plotted against the VHV's and CNS's. Panel (b) shows the transformed data together with the MSSE m_n (red surface). Here, the censored data is transformed according to (2.36) and (2.37) with $\alpha_1 = \alpha_1^{FG} \approx 0.09$.

Note that in this example, m_n is approximately a linear function of the two covariates VHV and CNS with intercept 12.2 and slopes 12.1 (VHV) and -4.7 (CNS). The signs of the slopes coincide with the fact that larger values of VHV, but smaller values of CNS,

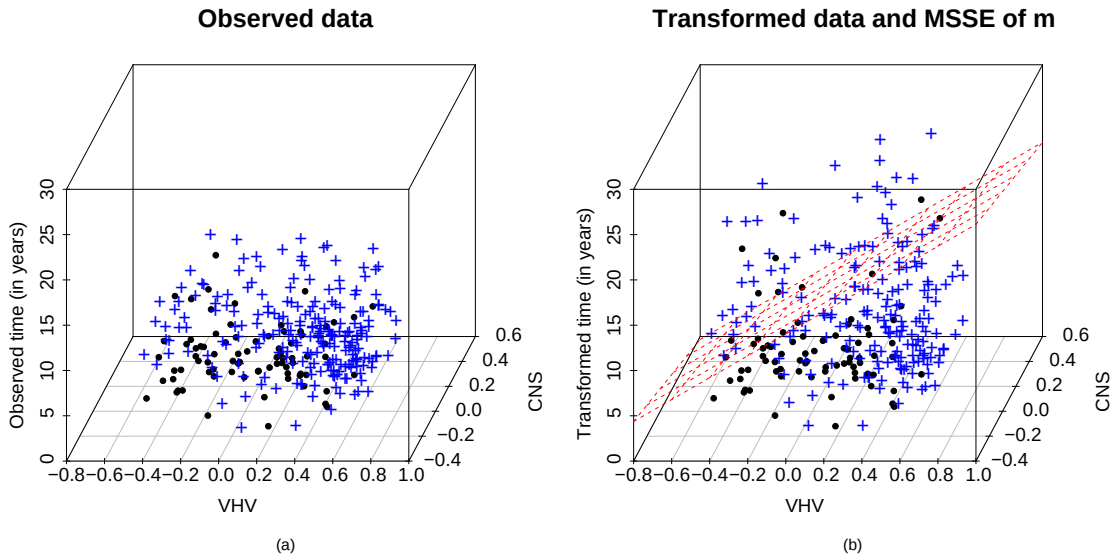

Figure 7.4: Breast cancer data set of Van de Vijver, He, Van't Veer et al. (2002) and the MSSE m_n of the regression function. The character "•" indicates the (transformed) uncensored observations, while the (transformed) censored observations are presented by "+". The individual panels show (a) the breast cancer data with observed time (in years) plotted against the VHV's and CNS's and (b) the transformed data together with m_n (red surface), where $\alpha_1 = \alpha_1^{FG} \approx 0.09$.

agree with a higher risk of women to develop distant metastases and hence having a shorter overall survival. Furthermore, the difference in magnitude of the absolute values of the slopes reflects the observation of Chang, Nuyten, Sneddon et al. (2005) that the VHV's better distinguish between a high-risk and low-risk group of the patients than the CNS's.

Chang, Nuyten, Sneddon et al. (2005) reported that by combining the VHV's and CNS's, one may distinguish between three risk-groups of the 295 patients by simply setting a threshold for each prognosis profile. They observed that the women with VHV's larger than 0.4 had a very good prognosis (cf. Figure 7.4 (b)). With a threshold of 0.05 for the CNS's, the remaining patients could then be divided into those with moderate and slightly worse outcomes. Here, this choice was motivated by the fact that only 10% of the patients in the good prognosis group had a CNS larger than 0.05, while a threshold of -0.15 similar to above would simply split this group in nearly two equally sized classes which show a minor difference in outcome.

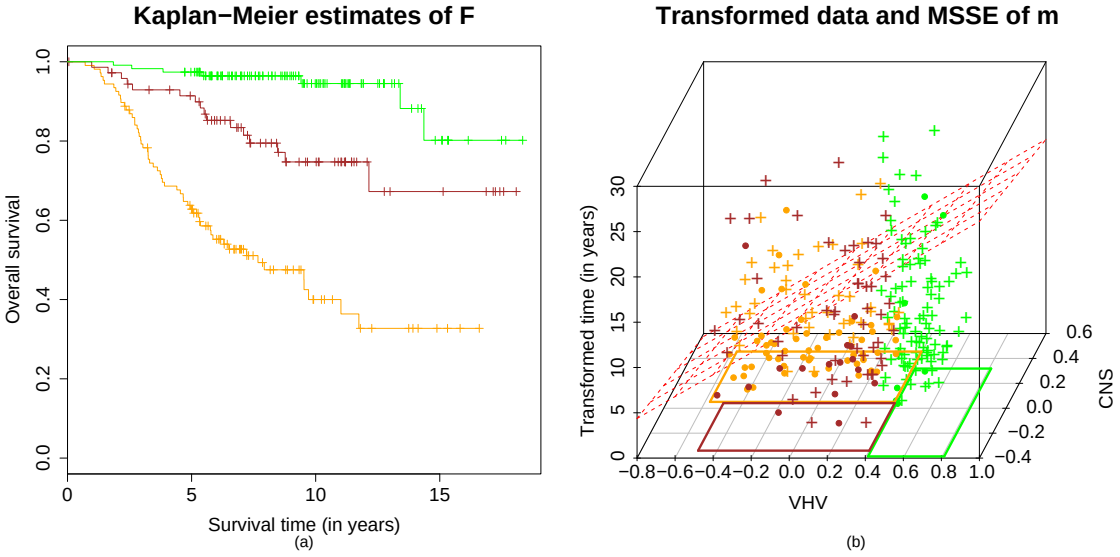

Figure 7.5: Breast cancer data set of Van de Vijver, He, Van't Veer et al. (2002) with three different subgroups defined by Chang, Nuyten, Sneddon et al. (2005). The individual panels display (a) the Kaplan-Meier estimates of the survival function F for the three subgroups (green, orange, and brown solid line) with censored observations marked by "+" and (b) the transformed breast cancer data with m_n (red surface) as in Figure 7.4, but with the data points of the different subgroups colored according to the corresponding Kaplan-Meier curves in panel (a). The colored rectangles in panel (b) display the range of the VHV's and CNS's in the three subgroups.

Figure 7.5 presents a comparison between this division of the 295 patients and m_n. Panel (a) displays the Kaplan-Meier estimates of the overall survival for patients with good, intermediate, and poor prognosis (green, brown, and orange solid line) according to Chang, Nuyten, Sneddon et al. (2005). Figure 7.5 (b) is identical to Figure 7.4 with the only difference that the data points of the different subgroups are colored as the corresponding Kaplan-Meier curves in Figure 7.5 (a). Moreover, the colored rectangles display the range of the VHV's and CNS's in the three groups.

The three prognosis groups marked in Figure 7.5 (b) clearly coincide with high, medium and low levels of the MSSE m_n. However, in contrast to the assumption of having two or three well-seperated risk groups of patients in the data, the shape of the MSSE rather suggest that there exist no clear boundaries between such groups, but the mean survival

rather alters smoothly for different values of the VHV's and CNS's. Hence, the group definitions of Chang, Nuyten, Sneddon et al. (2005) should be handled with care, since a slight change of the chosen thresholds would affect the assignment of more than just very few patients.

Appendix A

Results for fixed design regression

Below we formulate and prove two auxiliary results, which are used in the proofs of Lemma 4.1 and Theorem 5.1, in a fixed design regression model.

Let therefore $x_1, \dots, x_n \in [0,1]^d$ be arbitrary, but fixed and let $m^\star : [0,1]^d \to \mathbb{R}$ be a real-valued function. Assume that for all $i = 1, \dots, n$

$$Y_i^\star = m^\star(x_i) + \epsilon_i, \tag{A.1}$$

where $\epsilon_1, \dots, \epsilon_n$ are independent, real-valued random variables with expectation zero. Then the following result holds:

Lemma A.1. *Let $n, d \in \mathbb{N}$, $\lambda_n > 0$, $\beta^\star \geq 1$, and $b_4 > 0$. Set $l := \beta^\star + b_4$ and let $p \in \mathbb{N}$ with $2p > d$ be arbitrary. Assume that $Y_1^\star, \dots, Y_n^\star$ are given by (A.1) with $|Y_i^\star| \leq \beta^\star$ a.s. and $|m^\star(x_i)| \leq \beta^\star$ for all $i \in \{1, \dots, n\}$. Let $\bar{Y}_1, \dots, \bar{Y}_n$ be arbitrary real-valued random variables. Define estimates $\tilde{m}^\star_{n,(p,\lambda_n)}$ and $\hat{m}^\star_{n,(p,\lambda_n)}$ by*

$$\tilde{m}^\star_{n,(p,\lambda_n)}(\cdot) := \operatorname*{arg\,min}_{f \in W_p([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^n |f(x_i) - \bar{Y}_i|^2 + \lambda_n J_p^2(f) \right), \tag{A.2}$$

where $W_p([0,1]^d)$ and $J_p^2(\cdot)$ are given by (2.3) and (2.5), and

$$\hat{m}^\star_{n,(p,\lambda_n)}(\cdot) := T_{[-l,l]} \tilde{m}^\star_{n,(p,\lambda_n)}(\cdot). \tag{A.3}$$

Assume $m^\star \in W_p([0,1]^d)$ with $J_p^2(m^\star) < \infty$ and set

$$\hat{S}_n^\star := \frac{1}{n} \sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)|^2 + \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}) - \frac{64}{n} \sum_{i=1}^n |Y_i^\star - \bar{Y}_i|^2 - 2\lambda_n J_p^2(m^\star).$$

Then there exist constants $b_5, b_6 > 0$ which only depend on $\beta^\star$, such that for any $t_n > 0$ with

$$t_n \to 0 \quad (n \to \infty), \tag{A.4}$$

$$\frac{n\, t_n}{\ln n} \to \infty \quad (n \to \infty), \tag{A.5}$$

and

$$\frac{n\, t_n}{\ln n} \lambda_n^{\frac{d}{2p}} \to \infty \quad (n \to \infty), \tag{A.6}$$

we have for all $t \geq t_n$ and for all sufficiently large n

$$\mathbf{P}\left[\hat{S}_n^\star > t,\ \frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2 \leq \frac{1}{n}\sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2\right] \leq b_5 \exp\left(-b_6 n t\right).$$

Proof of Lemma A.1. First, we notice that $|Y_i^\star| \leq \beta^\star$ a.s. and $|m^\star(x_i)| \leq \beta^\star$ imply

$$\epsilon_i^2 = |Y_i^\star - m^\star(x_i)|^2 \leq 4(\beta^\star)^2 \quad \text{a.s.} \tag{A.7}$$

for all $i \in \{1, \ldots, n\}$. Set

$$\hat{V}^\star_{n,(p,\lambda_n)} := \frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)|^2 + \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}).$$

By an application of Lemma B.1 in combination with (A.7), we have for all $t > 0$

$$\begin{aligned}
&\mathbf{P}\left[\hat{S}_n^\star > t,\ \frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2 \leq \frac{1}{n}\sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2\right] \\
&= \mathbf{P}\left[\hat{V}^\star_{n,(p,\lambda_n)} > t + \frac{64}{n}\sum_{i=1}^n |Y_i^\star - \bar{Y}_i|^2 + 2\lambda_n J_p^2(m^\star),\right. \\
&\qquad\qquad \left.\frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2 \leq \frac{1}{n}\sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2\right] \\
&\leq \mathbf{P}\left[t < \hat{V}^\star_{n,(p,\lambda_n)} \leq \frac{8}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)) \cdot \epsilon_i\right] \\
&= \mathbf{P}\left[t < \hat{V}^\star_{n,(p,\lambda_n)} \leq \frac{8}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)) \cdot \epsilon_i,\ \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \leq 4(\beta^\star)^2\right] \\
&=: q_{1,n}.
\end{aligned} \tag{A.8}$$

In order to derive an upper bound on $q_{1,n}$, observe that the Cauchy–Schwarz inequality yields with probability one

$$\frac{8}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)) \cdot \epsilon_i \leq 8\sqrt{\frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)|^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2}.$$

Therefore, one can conclude that inside of $q_{1,n}$

$$\begin{aligned}
\hat{V}^{\star}_{n,(p,\lambda_n)} &= \frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2+\lambda_n J_p^2(\tilde{m}^{\star}_{n,(p,\lambda_n)}) \\
&\leq \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}+\frac{\lambda_n J_p^2(\tilde{m}^{\star}_{n,(p,\lambda_n)})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}\right)^2 \\
&= \left(\frac{\hat{V}^{\star}_{n,(p,\lambda_n)}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}\right)^2 \\
&\leq \left(\frac{\frac{8}{n}\sum_{i=1}^{n}(\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i))\cdot\epsilon_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}\right)^2 \\
&\leq \left(\frac{16\beta^{\star}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}\right)^2 \\
&= 256(\beta^{\star})^2. \qquad \text{(A.9)}
\end{aligned}$$

For arbitrary $t>0$ set

$$\bar{\jmath}_{min} := \min\left\{j\in\mathbb{N}: 2^j t \geq 256(\beta^{\star})^2\right\}.$$

An application of the peeling-technique (cf. (4.50)) together with (A.8) and (A.9) implies for all $t>0$

$$\begin{aligned}
&\mathbf{P}\left[\hat{S}^{\star}_n > t,\ \frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-\bar{Y}_i|^2 \leq \frac{1}{n}\sum_{i=1}^{n}|\tilde{m}^{\star}_{n,(p,\lambda_n)}(x_i)-\bar{Y}_i|^2\right] \\
&\leq \mathbf{P}\left[t<\hat{V}^{\star}_{n,(p,\lambda_n)}\leq 256(\beta^{\star})^2,\ \hat{V}^{\star}_{n,(p,\lambda_n)}\leq\frac{8}{n}\sum_{i=1}^{n}(\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i))\cdot\epsilon_i,\right. \\
&\qquad\qquad \left.\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\leq 4(\beta^{\star})^2\right] \\
&\leq \sum_{j=1}^{\bar{\jmath}_{min}}\mathbf{P}\left[\frac{2^j t}{2}<\hat{V}^{\star}_{n,(p,\lambda_n)}\leq 2^j t,\ \hat{V}^{\star}_{n,(p,\lambda_n)}\leq\frac{8}{n}\sum_{i=1}^{n}(\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i))\cdot\epsilon_i,\right. \\
&\qquad\qquad \left.\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\leq 4(\beta^{\star})^2\right].
\end{aligned}$$

Hence, we have for all $t > 0$

$$\begin{aligned}
&\mathbf{P}\left[\hat{S}_n^\star > t,\ \frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2 \le \frac{1}{n}\sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{Y}_i|^2\right] \\
&\quad\le \sum_{j=1}^{\bar{\jmath}_{min}} \mathbf{P}\left[\hat{V}^\star_{n,(p,\lambda_n)} \le 2^j t,\ \frac{1}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\cdot\epsilon_i > \frac{2^j t}{16},\ \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \le 4(\beta^\star)^2\right] \\
&\quad\le \sum_{j=1}^{\bar{\jmath}_{min}} \mathbf{P}\left[\frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)|^2 \le 2^j t,\ \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}) \le 2^j t,\right. \\
&\qquad\qquad \left.\frac{1}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\cdot\epsilon_i > \frac{2^j t}{16},\ \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \le 4(\beta^\star)^2\right] \\
&\quad\le \sum_{j=1}^{\bar{\jmath}_{min}} \mathbf{P}\left[\sup_{g\in\mathcal{G}_{2^j t/\lambda_n}} \left|\frac{1}{n}\sum_{i=1}^n g(x_i)\cdot\epsilon_i\right| \ge \frac{2^j t}{16},\ \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \le 4(\beta^\star)^2\right] \\
&\quad=: \sum_{j=1}^{\bar{\jmath}_{min}} q_{2,n,j}. \qquad\qquad (A.10)
\end{aligned}$$

Here, for all $j \in \{1,\dots,\bar{\jmath}_{min}\}$

$$\mathcal{G}_{2^j t/\lambda_n} := \left\{ f - m^\star : f \in \mathcal{F}_{2^j t/\lambda_n},\ \frac{1}{n}\sum_{i=1}^n |f(x_i) - m^\star(x_i)|^2 \le 2^j t \right\}$$

with

$$\mathcal{F}_{2^j t/\lambda_n} := \left\{ T_{[-l,l]} f : f \in W_p([0,1]^d),\ J_p^2(f) \le \frac{2^j t}{\lambda_n} \right\}.$$

For two arbitrary functions $g_1, g_2 \in \mathcal{G}_{2^j t/\lambda_n}$ with $g_1 = f_1 - m^\star$ and $g_2 = f_2 - m^\star$, where $f_1, f_2 \in \mathcal{F}_{2^j t/\lambda_n}$ ($j \in \{1,\dots,\bar{\jmath}_{min}\}$), one gets

$$\frac{1}{n}\sum_{i=1}^n |g_1(x_i) - g_2(x_i)|^2 = \frac{1}{n}\sum_{i=1}^n |f_1(x_i) - f_2(x_i)|^2$$

and therefore

$$\mathcal{N}_2\left(s, \mathcal{G}_{2^j t/\lambda_n}, x_1^n\right) = \mathcal{N}_2\left(s, \mathcal{F}_{2^j t/\lambda_n}, x_1^n\right) \quad \forall s > 0. \qquad (A.11)$$

In the following, Lemma C.5 will be applied to the probabilities $q_{2,n,j}$ ($j = 1,\dots,\bar{\jmath}_{min}$) on the right hand side of (A.10). For this purpose, we first check that the conditions (C.4), (C.5), and (C.6) are fulfilled.

Set $V_i = \epsilon_i$ ($i = 1,\dots,n$), $\xi = \frac{2^j t}{16}$, $\nu = K = 2\beta^\star$, $R = \sqrt{2^j t}$, $\nu_0 = 2\sqrt{2}\beta^\star$, and $\mathcal{G} = \mathcal{G}_{2^j t/\lambda_n}$ in Lemma C.5 ($j \in \{1,\dots,\bar{\jmath}_{min}\}$). Then (A.7) and the definition of $\mathcal{G}_{2^j t/\lambda_n}$ imply that (C.4) and (C.5) hold, respectively.

In order to show that (C.6) is fulfilled, first note that (A.5) yields

$$\sqrt{n}\,\xi = \frac{\sqrt{n}\,2^j t}{16} \geq \frac{\sqrt{n\,t_n}}{16}\sqrt{2^j t} \geq 2b_{14}\sqrt{2^j t} = 2b_{14}R \tag{A.12}$$

for all $t \geq t_n$, all $j \in \{1,\dots,\bar{\jmath}_{min}\}$, and all sufficiently large n. Here, $b_{14} > 0$ is the constant in Lemma C.5. Note that Lemma C.5 implies that b_{14} only depends on $\beta^\star$.

Furthermore, from the definition of $\bar{\jmath}_{min}$, we have that $2^j t = 2 \cdot 2^{j-1} t < 2 \cdot 256(\beta^\star)^2$, i.e., $\xi = \frac{2^j t}{16} < \sqrt{2^j t} \cdot 2\beta^\star = R\nu$ in Lemma C.5 ($j \in \{1,\dots,\bar{\jmath}_{min}\}$). For all $t \geq t_n$, all $j \in \{1,\dots,\bar{\jmath}_{min}\}$, and all sufficiently large n, (A.5) yields $R^2 = 2^j t \geq 2^j t_n \geq \frac{l^2}{n}$. Therefore, one can conclude similar to (4.55) that Lemma B.2 and (A.11) imply

$$\begin{aligned}
\int_{\frac{2^j t}{2^8\beta^\star}}^{\sqrt{2^j t}} \sqrt{\ln \mathcal{N}_2\left(s, \mathcal{G}_{2^j t/\lambda_n}, x_1^n\right)}\, ds &\leq \int_0^{\sqrt{2^j t}} \sqrt{\ln \mathcal{N}_2\left(s, \mathcal{F}_{2^j t/\lambda_n}, x_1^n\right)}\, ds \\
&\leq b_8\, \lambda_n^{-\frac{d}{4p}} \sqrt{2^j t}\sqrt{\ln n} + b_9\sqrt{2^j t}\sqrt{\ln n} \\
&= \sqrt{n} 2^j t \left(b_8 \sqrt{\frac{\ln n}{n 2^j t}} \lambda_n^{-\frac{d}{2p}} + b_9 \sqrt{\frac{\ln n}{n 2^j t}} \right) \\
&\leq \sqrt{n} 2^j t \left(b_8 \sqrt{\frac{\ln n}{n t_n}} \lambda_n^{-\frac{d}{2p}} + b_9 \sqrt{\frac{\ln n}{n t_n}} \right)
\end{aligned} \tag{A.13}$$

for all $t \geq t_n$, all $j \in \{1,\dots,\bar{\jmath}_{min}\}$, and all sufficiently large n. Here, $b_8, b_9 > 0$ are the constants from Lemma B.2 which only depend on p and d. Now, (A.13) together with (A.5), (A.6), and (A.12) yields that (C.6) holds for all $t \geq t_n$, all $j \in \{1,\dots,\bar{\jmath}_{min}\}$, and all sufficiently large n.

Hence, we can deduce from Lemma C.5 and (A.5) for all $t \geq t_n$ and all sufficiently large n

$$\begin{aligned}
\sum_{j=1}^{\bar{\jmath}_{min}} q_{2,n,j} &\leq \sum_{j=1}^{\bar{\jmath}_{min}} b_{14} \exp\left(-\frac{n}{4\, b_{14}^2 2^j t} \left(\frac{2^j t}{16} \right)^2 \right) = \sum_{j=1}^{\bar{\jmath}_{min}} b_{14} \exp\left(-\frac{n 2^j t}{1024\, b_{14}^2} \right) \\
&\leq \sum_{j=1}^{\bar{\jmath}_{min}} b_{14} \exp\left(-\frac{njt}{1024\, b_{14}^2} \right) \leq b_{14} \frac{\exp\left(-\frac{nt}{1024\, b_{14}^2}\right)}{1 - \exp\left(-\frac{nt}{1024\, b_{14}^2}\right)} \\
&\leq 2\, b_{14} \exp\left(-\frac{nt}{1024\, b_{14}^2} \right),
\end{aligned} \tag{A.14}$$

where we used that $2^j \geq j$ ($j \in \mathbb{N}$). Since in our case b_{14} only depends on $\beta^\star$, the assertion of Lemma A.1 follows from (A.10) and (A.14).

□

The next lemma is applied in the proof of Theorem 5.1 in order to bound the empirical $\mathcal{L}_2$ error of the MSSE (5.8) on the testing data.

Lemma A.2. *Let $d, n_{\mathtt{l}}, n_{\mathtt{t}} \in \mathbb{N}$ and set $n := n_{\mathtt{l}} + n_{\mathtt{t}}$. Let $0 < \beta \leq \beta^\star < \infty$. Define $Y_i^\star$ by (A.1) with $|Y_i^\star| \leq \beta^\star$ a.s. and $m^\star(x_i) \in [0, \beta]$ $(i \in \{1, \dots, n\})$. Let $\bar{Y}_1, \dots, \bar{Y}_n$ be real-valued random variables, where $\bar{Y}_1, \dots, \bar{Y}_{n_{\mathtt{l}}}$ and $Y^\star_{n_{\mathtt{l}}+1}, \dots, Y^\star_n$ are independent sequences of random variables.*

Let $K_n^\star \times \Lambda_n^\star$ be a finite, non-empty set of parameters with

$$K_n^\star \subset \left\{ \left\lfloor \frac{d}{2} \right\rfloor + 1, \left\lfloor \frac{d}{2} \right\rfloor + 2, \dots \right\} \quad \text{and} \quad \Lambda_n^\star \subset (0, \infty).$$

For each $(k, \lambda) \in K_n^\star \times \Lambda_n^\star$, define the estimates $\tilde{m}^\star_{n_{\mathtt{l}},(k,\lambda)}$ and $m^\star_{n_{\mathtt{l}},(k,\lambda)}$ via

$$\tilde{m}^\star_{n_{\mathtt{l}},(k,\lambda)}(\cdot) := \operatorname*{arg\,min}_{f \in W_k([0,1]^d)} \left(\frac{1}{n_{\mathtt{l}}} \sum_{i=1}^{n_{\mathtt{l}}} |f(x_i) - \bar{Y}_i|^2 + \lambda J_k^2(f) \right),$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are given by (2.3) and (2.5), and

$$m^\star_{n_{\mathtt{l}},(k,\lambda)}(\cdot) := T_{[0,\beta]} \tilde{m}^\star_{n_{\mathtt{l}},(k,\lambda)}(\cdot).$$

Now, let

$$m_n^\star(\cdot) := \operatorname*{arg\,min}_{f \in \mathcal{F}_{K_n^\star \times \Lambda_n^\star}} \frac{1}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |f(x_i) - \bar{Y}_i|^2 \tag{A.15}$$

with

$$\mathcal{F}_{K_n^\star \times \Lambda_n^\star} := \left\{ m^\star_{n_{\mathtt{l}},(k,\lambda)} : (k, \lambda) \in K_n^\star \times \Lambda_n^\star \right\}$$

and set

$$A_n^\star := 18 \min_{(k,\lambda) \in K_n^\star \times \Lambda_n^\star} \frac{1}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |m^\star_{n_{\mathtt{l}},(k,\lambda)}(x_i) - m^\star(x_i)|^2 + \frac{512}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} \left| Y_i^\star - \bar{Y}_i \right|^2.$$

Then there exists a constant $b_7 > 0$ which depends only on $\beta^\star$, such that for all $t > 0$

$$\mathbf{P}\left[\frac{1}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |m_n^\star(x_i) - m^\star(x_i)|^2 > A_n^\star + t \right] \leq \frac{2\,|K_n^\star \times \Lambda_n^\star|}{\exp\left(b_7\, n_{\mathtt{t}} t\right) - 1}.$$

Proof of Lemma A.2. Set

$$\hat{m}_n(\cdot) := \operatorname*{arg\,min}_{f \in \mathcal{F}_{K_n^\star \times \Lambda_n^\star}} \frac{1}{n_{\mathtt{t}}} \sum_{i=n_{\mathtt{l}}+1}^{n} |f(x_i) - m^\star(x_i)|^2. \tag{A.16}$$

In the proof, we will apply the following lemma.

Lemma A.3. *(Kohler (2006)) Let $t > 0$, $d \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{R}^d$, $y_1^\star, \bar{y}_1, \ldots, y_n^\star, \bar{y}_n \in \mathbb{R}$, $m^\star : \mathbb{R}^d \to \mathbb{R}$, and $\mathcal{F}$ a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. Set*

$$m_n^\star(\cdot) := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - \bar{y}_i|^2 , \tag{A.17}$$

$$\hat{m}_n(\cdot) := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - m^\star(x_i)|^2 , \tag{A.18}$$

and

$$a_n^\star := 18 \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - m^\star(x_i)|^2 + \frac{512}{n} \sum_{i=1}^n |y_i^\star - \bar{y}_i|^2 .$$

If both minima in (A.17) and (A.18) exist and

$$\frac{1}{n} \sum_{i=1}^n |m_n^\star(x_i) - m^\star(x_i)|^2 > a_n^\star + t, \tag{A.19}$$

then we have

$$\frac{t}{2} < \frac{1}{n} \sum_{i=1}^n |m_n^\star(x_i) - \hat{m}_n(x_i)|^2 \leq \frac{16}{n} \sum_{i=1}^n (m_n^\star(x_i) - \hat{m}_n(x_i)) (y_i^\star - m^\star(x_i)). \tag{A.20}$$

For the proof of Lemma A.3, see Kohler (2006), Lemma 1.

Since

$$\min_{(k,\lambda) \in K_n^\star \times \Lambda_n^\star} \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n |m_{n_1,(k,\lambda)}^\star(x_i) - m^\star(x_i)|^2 = \min_{f \in \mathcal{F}_{K_n^\star \times \Lambda_n^\star}} \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n |f(x_i) - m^\star(x_i)|^2 ,$$

one can conclude from Lemma A.3 for arbitrary $t > 0$

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n |m_n^\star(x_i) - m^\star(x_i)|^2 > A_n^\star + t \right] \\
&\leq \mathbf{P}\left[\frac{t}{2} < \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n |m_n^\star(x_i) - \hat{m}_n(x_i)|^2 \right. \\
&\qquad\qquad \left. \leq \frac{16}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n (m_n^\star(x_i) - \hat{m}_n(x_i)) (Y_i^\star - m^\star(x_i)) \right] \\
&\leq \mathbf{P}\left[\exists (k,\lambda) \in K_n^\star \times \Lambda_n^\star : \frac{t}{2} < \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n |m_{n_1,(k,\lambda)}^\star(x_i) - \hat{m}_n(x_i)|^2 \right. \\
&\qquad\qquad \left. \leq \frac{16}{n_{\mathtt{t}}} \sum_{i=n_1+1}^n (m_{n_1,(k,\lambda)}^\star(x_i) - \hat{m}_n(x_i)) (Y_i^\star - m^\star(x_i)) \right] \\
&\leq |K_n^\star \times \Lambda_n^\star| \max_{(k,\lambda) \in K_n^\star \times \Lambda_n^\star} \mathbf{P}\left[\frac{t}{2} < \hat{V}_{1,n,(k,\lambda)}^\star \leq 16\, \hat{V}_{2,n,(k,\lambda)}^\star \right].
\end{aligned} \tag{A.21}$$

Here,

$$\hat{V}^{\star}_{1,n,(k,\lambda)} := \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)|^2$$

and

$$\hat{V}^{\star}_{2,n,(k,\lambda)} := \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} (m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)) \, (Y^{\star}_i - m^{\star}(x_i)) \, .$$

An application of the peeling-technique (cf. (4.50)) to the right hand side of (A.21) yields for all $t > 0$

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m^{\star}_n(x_i) - m^{\star}(x_i)|^2 > A^{\star}_n + t \right] \\
&\leq \; |K^{\star}_n \times \Lambda^{\star}_n| \max_{(k,\lambda) \in K^{\star}_n \times \Lambda^{\star}_n} \sum_{s=0}^{\infty} \mathbf{P}\left[\frac{2^s t}{2} < \hat{V}^{\star}_{1,n,(k,\lambda)} \leq 2^s t, \hat{V}^{\star}_{1,n,(k,\lambda)} \leq 16 \, \hat{V}^{\star}_{2,n,(k,\lambda)} \right] \\
&\leq \; |K^{\star}_n \times \Lambda^{\star}_n| \max_{(k,\lambda) \in K^{\star}_n \times \Lambda^{\star}_n} \sum_{s=0}^{\infty} \mathbf{P}\left[\hat{V}^{\star}_{1,n,(k,\lambda)} \leq 2^s t, \hat{V}^{\star}_{2,n,(k,\lambda)} > \frac{2^s t}{32} \right] . \qquad \text{(A.22)}
\end{aligned}$$

Now fix $(k,\lambda) \in K^{\star}_n \times \Lambda^{\star}_n$ and $s \in \mathbb{N}_0 \cup \{\infty\}$. Set $\bar{Y}_1^{n_1} := \left\{ \bar{Y}_1, \ldots, \bar{Y}_{n_1} \right\}$. Next, we apply Hoeffding's inequality (Lemma D.3) in order to bound the conditional probability

$$\begin{aligned}
q_{n,(k,\lambda),s}(t) \;&:=\; \mathbf{P}\left[\hat{V}^{\star}_{1,n,(k,\lambda)} \leq 2^s t, \hat{V}^{\star}_{2,n,(k,\lambda)} > \frac{2^s t}{32} \middle| \bar{Y}_1^{n_1} \right] \\
&=\; \mathbf{P}\left[\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)|^2 \leq 2^s t, \right. \\
&\qquad \left. \frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} (m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)) \, (Y^{\star}_i - m^{\star}(x_i)) > \frac{2^s t}{32} \middle| \bar{Y}_1^{n_1} \right]
\end{aligned}$$

for all $t > 0$.

Let

$$\hat{V}^{\star}_{3,n,(k,\lambda),i} := (m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)) \, (Y^{\star}_i - m^{\star}(x_i)) \quad (i = n_1 + 1, \ldots, n).$$

Observe that due to (A.1), we have $Y^{\star}_i = m^{\star}(x_i) + \epsilon_i$ $(i = 1, \ldots, n)$, where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with expectation zero. This together with the assumption that $\bar{Y}_1, \ldots, \bar{Y}_{n_1}$ and $Y^{\star}_{n_1+1}, \ldots, Y^{\star}_n$ are independent sequences of random variables implies

$$\mathbf{E}\left[Y^{\star}_i \,\middle|\, \bar{Y}_1^{n_1} \right] = \mathbf{E}\left[Y^{\star}_i \right] = m^{\star}(x_i) + \mathbf{E}\left[\epsilon_i \right] = m^{\star}(x_i) \quad (i = n_1 + 1, \ldots, n), \qquad \text{(A.23)}$$

and moreover that $\hat{V}^{\star}_{3,n,(k,\lambda),n_1+1}, \ldots, \hat{V}^{\star}_{3,n,(k,\lambda),n}$ are conditionally independent given $\bar{Y}_1^{n_1}$.

From (A.23), we deduce with probability one

$$\mathbf{E}\left[\,\hat{V}^{\star}_{3,n,(k,\lambda),i}\,\middle|\,\bar{Y}_1^{n_1}\,\right] = (m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)) \cdot \mathbf{E}\left[\,Y_i^{\star} - m^{\star}(x_i)\,\middle|\,\bar{Y}_1^{n_1}\,\right] = 0$$

$(i = n_1 + 1, \ldots, n)$. Furthermore, $|Y_i^{\star}| \le \beta^{\star}$ a.s. and $0 \le m^{\star}(x_i) \le \beta \le \beta^{\star}$ yield

$$|\,\hat{V}^{\star}_{3,n,(k,\lambda),i}\,| \le 2\,\beta^{\star} \cdot |m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)| =: B_i \quad \text{a.s.} \quad (i = n_1 + 1, \ldots, n). \tag{A.24}$$

This implies that inside of $q_{n,(k,\lambda),s}(t)$, one gets for all $t > 0$

$$\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |2\,B_i|^2 = \frac{16\,(\beta^{\star})^2}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m^{\star}_{n_1,(k,\lambda)}(x_i) - \hat{m}_n(x_i)|^2 \le 16\,(\beta^{\star})^2 \cdot 2^s t. \tag{A.25}$$

Therefore, we have shown that (conditioned on $\bar{Y}_1^{n_1}$) the assumptions of Lemma D.3 hold, and one can conclude from Lemma D.3, (A.24), and (A.25) that for all $t > 0$

$$q_{n,(k,\lambda),s}(t) \le 2\,\exp\left(-\frac{2\,n_{\mathtt{t}}\,\left(\frac{2^s\,t}{32}\right)^2}{16(\beta^{\star})^2 \cdot 2^s t}\right) = 2\,\exp\left(-b_7\,n_{\mathtt{t}}\,2^s\,t\right) \quad \text{a.s.},$$

where $b_7 := 2^{-13}(\beta^{\star})^{-2}$.

This together with (A.22) and $2^s \ge s + 1$ $(s \in \mathbb{N}_0)$ yields for all $t > 0$

$$\begin{aligned}
\mathbf{P}\left[\frac{1}{n_{\mathtt{t}}} \sum_{i=n_1+1}^{n} |m^{\star}_n(x_i) - m^{\star}(x_i)|^2 > A^{\star}_n + t\right] &\le |K^{\star}_n \times \Lambda^{\star}_n| \sum_{s=0}^{\infty} 2\,\exp\left(-b_7\,n_{\mathtt{t}}\,2^s\,t\right) \\
&\le 2\,|K^{\star}_n \times \Lambda^{\star}_n| \sum_{s=0}^{\infty} \exp\left(-b_7\,n_{\mathtt{t}}\,(s+1)\,t\right) \\
&= 2\,|K^{\star}_n \times \Lambda^{\star}_n|\,\frac{\exp\left(-b_7\,n_{\mathtt{t}} t\right)}{1 - \exp\left(-b_7\,n_{\mathtt{t}} t\right)} \\
&= \frac{2\,|K^{\star}_n \times \Lambda^{\star}_n|}{\exp\left(b_7\,n_{\mathtt{t}} t\right) - 1}.
\end{aligned}$$

This implies the assertion of Lemma A.2.

□

Appendix B

Two deterministic lemmata

This chapter contains two deterministic lemmata which are used in the proofs of Theorem 4.1 and Lemma A.1.

Lemma B.1. *Let $l > 0$, $n, d \in \mathbb{N}$, $x_1, \dots, x_n \in [0,1]^d$, and $y_1^\star, \bar{y}_1, \dots, y_n^\star, \bar{y}_n \in \mathbb{R}$. Let $p \in \mathbb{N}$ with $2p > d$ be arbitrary and let $\lambda_n > 0$. Define the estimates $\tilde{m}^\star_{n,(p,\lambda_n)}$ and $\hat{m}^\star_{n,(p,\lambda_n)}$ by*

$$\tilde{m}^\star_{n,(p,\lambda_n)}(\cdot) := \operatorname*{arg\,min}_{f \in W_p([0,1]^d)} \left(\frac{1}{n} \sum_{i=1}^n |f(x_i) - \bar{y}_i|^2 + \lambda_n J_p^2(f) \right), \tag{B.1}$$

where $W_p([0,1]^d)$ and $J_p^2(\cdot)$ are given by (2.3) and (2.5), and

$$\hat{m}^\star_{n,(p,\lambda_n)}(\cdot) := T_{[-l,l]} \tilde{m}^\star_{n,(p,\lambda_n)}(\cdot). \tag{B.2}$$

Assume that $m^\star \in W_p([0,1]^d)$ with $J_p^2(m^\star) < \infty$. Let $t > 0$ and set

$$\hat{v}^\star_{n,(p,\lambda_n)} := \frac{1}{n} \sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i)|^2 + \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}).$$

If

$$\hat{v}^\star_{n,(p,\lambda_n)} > t + \frac{64}{n} \sum_{i=1}^n |y_i^\star - \bar{y}_i|^2 + 2\lambda_n J_p^2(m^\star) \tag{B.3}$$

and

$$\frac{1}{n} \sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{y}_i|^2 \leq \frac{1}{n} \sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{y}_i|^2, \tag{B.4}$$

then we have

$$\hat{v}^\star_{n,(p,\lambda_n)} \leq \frac{8}{n} \sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))(y_i^\star - m^\star(x_i)). \tag{B.5}$$

Proof of Lemma B.1. Assume $m^\star \in W_p([0,1]^d)$ with $J_p^2(m^\star) < \infty$. Definition (B.1) and inequality (B.4) imply

$$\begin{aligned}
\hat{v}^\star_{n,(p,\lambda_n)} &= \frac{1}{n}\sum_{i=1}^n |\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{y}_i|^2 + \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}) - \frac{1}{n}\sum_{i=1}^n |m^\star(x_i) - \bar{y}_i|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i)) \\
&\le \frac{1}{n}\sum_{i=1}^n |\tilde{m}^\star_{n,(p,\lambda_n)}(x_i) - \bar{y}_i|^2 + \lambda_n J_p^2(\tilde{m}^\star_{n,(p,\lambda_n)}) - \frac{1}{n}\sum_{i=1}^n |m^\star(x_i) - \bar{y}_i|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i)) \\
&\le \frac{1}{n}\sum_{i=1}^n |m^\star(x_i) - \bar{y}_i|^2 + \lambda_n J_p^2(m^\star) - \frac{1}{n}\sum_{i=1}^n |m^\star(x_i) - \bar{y}_i|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i)) \\
&= \lambda_n J_p^2(m^\star) + \frac{2}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i))\,. \qquad \text{(B.6)}
\end{aligned}$$

Here, the first equality follows with

$$(a_1 - a_2)^2 = (a_1 - a_3)^2 - (a_2 - a_3)^2 + 2\,(a_1 - a_2)\,(a_3 - a_2) \quad \forall a_1, a_2, a_3 \in \mathbb{R}.$$

If

$$\frac{2}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i)) < \lambda_n J_p^2(m^\star),$$

then we can conclude from (B.3) and (B.6)

$$t + 2\lambda_n J_p^2(m^\star) < \hat{v}^\star_{n,(p,\lambda_n)} < 2\lambda_n J_p^2(m^\star),$$

in contradiction to $t > 0$. Therefore, we have shown that

$$\begin{aligned}
\hat{v}^\star_{n,(p,\lambda_n)} &\le \frac{4}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - m^\star(x_i)) \\
&= \frac{4}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(\bar{y}_i - y_i^\star) \\
&\quad + \frac{4}{n}\sum_{i=1}^n (\hat{m}^\star_{n,(p,\lambda_n)}(x_i) - m^\star(x_i))\,(y_i^\star - m^\star(x_i))\,. \qquad \text{(B.7)}
\end{aligned}$$

Now assume that the second term on the right hand side of (B.7) is smaller than the first one. In this case, (B.7) and the Cauchy–Schwarz inequality yield

$$\begin{aligned}\hat{v}^{\star}_{n,(p,\lambda_n)} &\leq \frac{8}{n}\sum_{i=1}^{n}(\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i))\,(\bar{y}_i-y_i^{\star})\\ &\leq 8\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i^{\star}-\bar{y}_i|^2}. \qquad \text{(B.8)}\end{aligned}$$

If

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2\neq 0,$$

then (B.3) together with (B.8) implies (cf. (A.9))

$$t+\frac{64}{n}\sum_{i=1}^{n}|y_i^{\star}-\bar{y}_i|^2<\hat{v}^{\star}_{n,(p,\lambda_n)}\leq\left(\frac{\hat{v}^{\star}_{n,(p,\lambda_n)}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2}}\right)^2\leq\frac{64}{n}\sum_{i=1}^{n}|y_i^{\star}-\bar{y}_i|^2,$$

in contradiction to $t>0$. And if

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{\star}_{n,(p,\lambda_n)}(x_i)-m^{\star}(x_i)|^2=0,$$

then we can conclude from (B.3), (B.8), and the definition of $\hat{v}^{\star}_{n,(p,\lambda_n)}$ that $t<\hat{v}^{\star}_{n,(p,\lambda_n)}=0$. From this together with (B.7), the assertion (B.5) of Lemma B.1 follows.

□

Lemma B.2. *Let $l,b>0$. Let $p,d,n\in\mathbb{N}$ with $2p>d$ and $n>1$. Set*

$$\mathcal{F}_b:=\left\{T_{[-l,l]}f:f\in W_p([0,1]^d),J_p^2(f)\leq b\right\},$$

where $W_p([0,1]^d)$ and $J_p^2(\cdot)$ are given by (2.3) and (2.5). There exist constants $b_8,b_9>0$ which only depend on p and d, such that for all $\zeta\geq\frac{l^2}{n}$ and all $x_1,\ldots,x_n\in[0,1]^d$

$$\int_0^{\sqrt{\zeta}}\sqrt{\ln\mathcal{N}_2(s,\mathcal{F}_b,x_1^n)}\,ds\leq b_8\left(\frac{b}{\zeta}\right)^{\frac{d}{4p}}\sqrt{\zeta}\sqrt{\ln n}+b_9\sqrt{\zeta}\sqrt{\ln n}. \qquad \text{(B.9)}$$

Proof of Lemma B.2. For any $\zeta>0$ and all $x_1,\ldots,x_n\in[0,1]^d$ set

$$\mathcal{I}_\zeta:=\int_0^{\sqrt{\zeta}}\sqrt{\ln\mathcal{N}_2(s,\mathcal{F}_b,x_1^n)}\,ds.$$

Lemma C.3 implies that there exist two constants $B_1, B_2 > 0$ which only depend on p and d, such that for all $\zeta \in (0, l^2]$ and all $x_1, \dots, x_n \in [0,1]^d$

$$\mathcal{I}_\zeta \le B_1 b^{\frac{d}{4p}} \int_0^{\sqrt{\zeta}} s^{-\frac{d}{2p}} \sqrt{\ln\left(\frac{64 l^2 e n}{s^2}\right)}\, ds + B_2 \int_0^{\sqrt{\zeta}} \sqrt{\ln\left(\frac{64 l^2 e n}{s^2}\right)}\, ds. \tag{B.10}$$

Here, we used that $\sqrt{a_1 + a_2} \le \sqrt{a_1} + \sqrt{a_2}$ for all $a_1, a_2 \ge 0$.

Substituting $t := \frac{\sqrt{\zeta}}{s}$ and applying Hölder's inequality, one can conclude from (B.10) for all $\zeta \in (0, l^2]$ and all $x_1, \dots, x_n \in [0,1]^d$

$$\begin{aligned}
\mathcal{I}_\zeta &= B_1 b^{\frac{d}{4p}} \zeta^{\frac{1}{2}-\frac{d}{4p}} \int_1^\infty t^{\frac{d}{2p}-2} \sqrt{\ln\left(\frac{64 l^2 e n}{\zeta}\, t^2\right)}\, dt + B_2 \sqrt{\zeta} \int_1^\infty t^{-2} \sqrt{\ln\left(\frac{64 l^2 e n}{\zeta}\, t^2\right)}\, dt \\
&\le B_1 b^{\frac{d}{4p}} \zeta^{\frac{1}{2}-\frac{d}{4p}} \sqrt{\int_1^\infty t^{\frac{d}{2p}-2}\, dt \cdot \int_1^\infty t^{\frac{d}{2p}-2} \ln\left(\frac{64 l^2 e n}{\zeta}\, t^2\right) dt} \\
&\quad + B_2 \sqrt{\zeta} \sqrt{\int_1^\infty t^{-2}\, dt \cdot \int_1^\infty t^{-2} \ln\left(\frac{64 l^2 e n}{\zeta}\, t^2\right) dt} \\
&= B_3 \left(\frac{b}{\zeta}\right)^{\frac{d}{4p}} \sqrt{\zeta} \sqrt{\ln\left(B_4 \frac{l^2 n}{\zeta}\right)} + B_2 \sqrt{\zeta} \sqrt{\ln\left(B_5 \frac{l^2 n}{\zeta}\right)}
\end{aligned} \tag{B.11}$$

with the constants $B_3 := \frac{B_1}{1-\frac{d}{2p}}$, $B_4 := 64 e^{1+\frac{2}{1-\frac{d}{2p}}}$, and $B_5 := 64 e^3$. In (B.11), the last equality follows from

$$\int t^a \ln(B\, t^2)\, dt = \frac{t^{a+1}}{a+1} \ln\left(B\, e^{-\frac{2}{a+1}}\, t^2\right) \quad (a \ne -1, B > 0).$$

Note that this formula is applicable for the first term on the right hand side of (B.11) since $2p > d$ implies $\frac{d}{2p} - 2 < -1$.

For all $\zeta > l^2$, we deduce from Lemma C.3 that $\mathcal{I}_\zeta = \mathcal{I}_{l^2}$. Finally, for all $\zeta \ge \frac{l^2}{n}$ and all $x_1, \dots, x_n \in [0,1]^d$, this together with (B.11) yields

$$\begin{aligned}
\mathcal{I}_\zeta &\le B_3 \left(\frac{b}{\zeta}\right)^{\frac{d}{4p}} \sqrt{\zeta} \sqrt{\ln(B_4 n^2)} + B_2 \sqrt{\zeta} \sqrt{\ln(B_5 n^2)} \\
&\le B_3 \left(\frac{b}{\zeta}\right)^{\frac{d}{4p}} \sqrt{\zeta} \sqrt{2(2 + \ln B_4)} \sqrt{\ln n} + B_2 \sqrt{\zeta} \sqrt{2(2 + \ln B_5)} \sqrt{\ln n},
\end{aligned}$$

where the last inequality follows from $\ln(a \cdot n^2) \le 2(2 + \ln a) \ln n$ for all $a \ge 1$ and $n \ge 2$.

□

Appendix C

Results from empirical process theory

Definition C.1. (Covering number) *Let $d \in \mathbb{N}$, $1 \leq r < \infty$, and $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. For any $\epsilon > 0$ and any $v_1^n = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$, the* $\mathcal{L}_r$*– ϵ–* **covering number** *$\mathcal{N}_r(\epsilon, \mathcal{F}, v_1^n)$ is defined as the smallest $N \in \mathbb{N}$ such that there exist functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with*

$$\min_{1 \leq j \leq N} \left(\frac{1}{n} \sum_{i=1}^{n} |f(v_i) - g_j(v_i)|^r \right)^{\frac{1}{r}} \leq \epsilon$$

for each $f \in \mathcal{F}$. If no such $N \in \mathbb{N}$ exists, then set $\mathcal{N}_r(\epsilon, \mathcal{F}, v_1^n) := \infty$.

If $V_1^n = (V_1, \ldots, V_n)$ is a vector of $\mathbb{R}^d$-valued random variables, then $\mathcal{N}_r(\epsilon, \mathcal{F}, V_1^n)$ is a random variable with expected value $\mathbf{E}\mathcal{N}_r(\epsilon, \mathcal{F}, V_1^n)$.

Lemma C.1. *(Pollard (1984)) Let $d \in \mathbb{N}$, $B > 0$, $V, V_1, \ldots, V_n$ be $\mathbb{R}^d$-valued i.i.d. random variables, and let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to [0, B]$. Then, for any $\epsilon > 0$,*

$$\mathbf{P}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(V_i) - \mathbf{E}g(V) \right| > \epsilon \right] \leq 8 \exp\left(-\frac{n\epsilon^2}{128 B^2} \right) \cdot \mathbf{E}\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, V_1^n \right).$$

Lemma C.2. *(Kohler and Krzyżak (2001)) Let $k, d \in \mathbb{N}$, $l, b > 0$ and set*

$$\mathcal{F}_b := \left\{ T_{[-l,l]} f : f \in W_k([0,1]^d), J_k^2(f) \leq b \right\},$$

where $W_k([0,1]^d)$ and $J_k^2(\cdot)$ are defined as in (2.3) and (2.5). For any $0 < \epsilon < l$ and any $x_1, \ldots, x_n \in [0,1]^d$ there exist constants $b_{10}, b_{11}, b_{12} > 0$, depending only on k and d, such

that

$$\mathcal{N}_1\left(\epsilon, \mathcal{F}_b, x_1^n\right) \leq \left(b_{10}\frac{l\,n}{\epsilon}\right)^{b_{11}\left(\frac{\sqrt{b}}{\epsilon}\right)^{\frac{d}{k}}+b_{12}}.$$

Lemma C.3. *(Kohler, Krzyżak, and Schäfer (2002)) Let $l, b > 0$, $p, d \in \mathbb{N}$, and set*

$$\mathcal{F}_b := \left\{T_{[-l,l]}f : f \in W_p([0,1]^d), J_p^2(f) \leq b\right\},$$

where $W_p([0,1]^d)$ and $J_p^2(\cdot)$ are given by (2.3) and (2.5). There exists a constant $b_{13} > 0$ which depends only on p and d such that for any $\epsilon > 0$ and all $x_1, \ldots, x_n \in [0,1]^d$

$$\ln \mathcal{N}_2\left(\epsilon, \mathcal{F}_b, x_1^n\right) \leq b_{13}\left(\left(\frac{\sqrt{b}}{\epsilon}\right)^{\frac{d}{p}} + 1\right) \cdot \ln\left(\frac{64el^2n}{\epsilon^2}\right) \cdot I_{[\epsilon \leq l]}.$$

Lemma C.4. *(Kohler (2000)) Let $V, V_1, \ldots, V_n$ be i.i.d. random variables with values in some set Θ. Let $K_1, K_2 \geq 1$ and let $\mathcal{G}$ be a permissible class of functions $g : \Theta \to [-K_1, K_1]$ satisfying*

$$\mathbf{E}\, g(V)^2 \leq K_2 \cdot \mathbf{E}\, g(V). \tag{C.1}$$

If for $0 < \epsilon < 1$ and $\nu > 0$

$$\sqrt{n}\epsilon\sqrt{1-\epsilon}\sqrt{\nu} \geq 288 \max\{2\,K_1, \sqrt{2\,K_2}\} \tag{C.2}$$

and for all $v_1, \ldots, v_n \in \Theta$ and all $\xi \geq \frac{\nu}{4}$

$$\sqrt{n}\,\epsilon\,(1-\epsilon)\,\xi \geq 288 \max\{K_1, 2\,K_2\} \int_0^{\sqrt{\xi}} \sqrt{\ln \mathcal{N}_2\left(s, \mathcal{G}^*, v_1^n\right)}\, ds, \tag{C.3}$$

where

$$\mathcal{G}^* = \left\{g \in \mathcal{G} : \frac{1}{n}\sum_{i=1}^n g(v_i)^2 \leq 4\xi\right\},$$

then

$$\mathbf{P}\left[\sup_{g \in \mathcal{G}} \frac{\left|\mathbf{E}\, g(V) - \frac{1}{n}\sum_{i=1}^n g(V_i)\right|}{\nu + \mathbf{E}\, g(V)} > \epsilon\right] \leq 50 \exp\left(-\frac{n\nu\epsilon^2(1-\epsilon)}{128 \cdot 2304 \max\{K_1^2, K_2\}}\right).$$

Lemma C.5. *(Van de Geer (2000)) Let $d \in \mathbb{N}$, $R > 0$, $x_1, \ldots, x_n \in \mathbb{R}^d$, and $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to \mathbb{R}$. Assume that*

$$\sup_{g \in \mathcal{G}} \sqrt{\frac{1}{n}\sum_{i=1}^n |g(x_i)|^2} \leq R. \tag{C.4}$$

Let $K > 0$, $\nu_0 > 0$. Suppose that $V_1, \ldots, V_n$ are independent, real-valued random variables with expectation zero, which fulfill the sub-Gaussian condition

$$\max_{i=1,\ldots,n} K^2 \, \mathbf{E}\left[\exp\left(\frac{|V_i|^2}{K^2}\right) - 1\right] \leq \nu_0^2. \tag{C.5}$$

Then for some constant $b_{14} > 0$ which depends only on K and ν_0, and for $\xi > 0$ and $\nu > 0$ satisfying $\xi < R\nu$ and

$$\sqrt{n}\xi \geq 2b_{14} \max\left\{\int_{\frac{\xi}{8\nu}}^{R} \sqrt{\ln \mathcal{N}_2(s, \mathcal{G}, x_1^n)}\, ds, R\right\}, \tag{C.6}$$

we have

$$\mathbf{P}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\cdot V_i\right| \geq \xi, \frac{1}{n}\sum_{i=1}^{n} V_i^2 \leq \nu^2\right] \leq b_{14}\cdot\exp\left(-\frac{n\xi^2}{4\, b_{14}^2 R^2}\right).$$

Appendix D

Auxiliary results

Definition D.1. *Let $p \in (0, \infty)$ and let Ω be an arbitrary measure space with positive measure μ. The $\mathcal{L}_p(\Omega)$ space is defined as*

$$\mathcal{L}_p(\Omega) := \left\{ f : \Omega \to \mathbb{C} : f \text{ measurable}, \int_\Omega |f(x)|^p \mu(dx) < \infty \right\}.$$

Definition D.2. *Let $\Omega \subseteq \mathbb{R}^d$ $(d \in \mathbb{N})$ and let $\mathcal{F}$ be an arbitrary set consisting of functions $f : \Omega \to \mathbb{R}$. For $1 \leq p < \infty$ and for an arbitrary probability measure μ on Ω, $\mathcal{F}$ is dense in $\mathcal{L}_p(\Omega)$ if and only if for any $g \in \mathcal{L}_p(\Omega)$ and any $\epsilon > 0$, there is a function $f \in \mathcal{F}$ such that*

$$\int_\Omega |g(x) - f(x)|^p \, \mu(dx) \leq \epsilon.$$

Lemma D.1. *(Rudin (1974)) Let $d \in \mathbb{N}$ and let $\Omega \subseteq \mathbb{R}^d$ be a locally compact Hausdorff space. For any $1 \leq p < \infty$ and any probability measure μ on Ω, the collection of all continuous real-valued functions on Ω whose support is compact is dense in $\mathcal{L}_p(\Omega)$.*

Lemma D.2. *(Borel-Cantelli lemma) Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Furthermore, let $A_1, A_2, \ldots \in \mathcal{A}$ be a sequence of events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. Then we have*

$$\mathbf{P}\left[\limsup_{n \to \infty} A_n \right] = 0.$$

I.e., if for a sequence V_1, V_2, $\ldots$, of real-valued random variables and arbitrary $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbf{P}\left[|V_n| > \epsilon \right] < \infty, \tag{D.1}$$

then

$$\mathbf{P}\left[\limsup_{n\to\infty}|V_n|>\epsilon\right]=0.$$

If D.1 even holds for all $\epsilon>0$, then one can conclude with probability one that $V_n\to 0$ $(n\to\infty)$.

Lemma D.3. *(Hoeffding's inequality, Hoeffding (1963)) Let $a_1,\ b_1,\ \ldots,\ a_n,\ b_n\in\mathbb{R}$. Assume that $V_1,\ \ldots,\ V_n$ are independent real-valued random variables with $V_i\in[a_i,b_i]$ a.s. for all $i=1,\ldots,n$. Then, for all $\epsilon>0$,*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(V_i-\mathbf{E}V_i)\right|>\epsilon\right]\leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}|b_i-a_i|^2}\right).$$

Lemma D.4. *(Bernstein's inequality, Bernstein (1946)) Let $a,b\in\mathbb{R}$ with $a<b$. Assume that $V_1,\ \ldots,\ V_n$ are independent real-valued random variables with $V_i\in[a,b]$ a.s. for all $i=1,\ldots,n$. Let*

$$\varsigma^2:=\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(V_i)>0.$$

Then it holds for all $\epsilon>0$ that

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(V_i-\mathbf{E}V_i)\right|>\epsilon\right]\leq 2\exp\left(-\frac{3n\epsilon^2}{6\,\varsigma^2+2\epsilon(b-a)}\right).$$

Bibliography

[1] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes.* Springer series in statistics, Springer-Verlag, New York.

[2] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

[3] Bernstein, S. N. (1946). *The theory of probabilities.* Gastehizdat Publishing House, Moscow.

[4] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

[5] Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

[6] Carbonez, A. (1992). *Nonparametric functional estimation under random censoring and a new semiparametric model of random censorship.* PhD thesis, Katholieke Universiteit Leuven.

[7] Cai, T. T., Levine, M., and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100**, 126–136.

[8] Carroll, R. J. and Hall, P. (1989). Variance function estimation in regression: The effect of estimating the mean. *J. Royal Statist. Soc. B* **51**, 3–14.

[9] Carroll, R. J. and Ruppert, D. (1988). *Transformation and weighting in regression.* Chapman & Hall, London.

[10] Chang, H. Y., Sneddon, J. B., Alizadeh, A. A. et al. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biology* **2**, 206–214.

[11] Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B. et al. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *PNAS* **10**, 3738–3743.

[12] Chen, K. and Lo, S.-H. (1997). On the rate of uniform convergence of the product-limit estimator: strong and weak laws. *Ann. Statist.* **25**, 1050–1087.

[13] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Royal Statist. Soc. B* **34**, 187–220.

[14] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

[15] Cox, D. R. and Oakes, D. (1984). *Analysis of survival data.* Chapman & Hall, London.

[16] Crowley, J. and Hu, M. (1977). Covariance Analysis of heart transplant survival data. *J. Amer. Statist. Assoc.* **72**, 27–36.

[17] Dabrowska, D. M. (1987). Nonparametric regression with censored lifetime data. *Scand. J. Statist.* **14**, 181–197.

[18] Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Ann. Statist.* **17**, 1157–1167.

[19] Deheuvels, P. and Mason, D.M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inf. for Stoch. Proc.* **7**, 225–277.

[20] Devroye, L. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattern Anal. Machine Intell.* **4**, 154–157.

[21] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer series for stochastic modelling and applied probability, Springer-Verlag, Berlin.

[22] Devroye, L. and Wagner, T. (1980). Distribution–free consistency results in non-parametric discrimination and regression function estimation. *Ann. Statist.* **8**, 231–239.

[23] Dippon, J. (2011). Kernel and support vector methods for censored survival data with covariates. Submitted for publication.

[24] Döhler, S. and Rüschendorf, L. (2002). Adaptive estimation of hazard functions. *Prob. and Math. Statist.* **22**, 355–379.

[25] Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces (in French). *R.A.I.R.O. Analyse Numérique* **10**, 5–12.

[26] El Ghouch, A. and Van Keilegom, I. (2008). Non-parametric regression with dependent censored data. *Scand. J. Statist.* **35**, 228–247.

[27] Escobar, L. A. and Meeker, W. Q. Jr. (1992). Assessing influence in regression analysis with censored data. *Biometrics* **48**, 507–528.

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

[29] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

[30] Ferlay J., Shin H. R., Bray F., Forman D., Mathers C. and Parkin D.M.. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]. Lyon, France: International Agency for Research on Cancer; 2010. Available from: http://globocan.iarc.fr, accessed on 17/02/2013.

[31] Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *J. Stat. Plann. and Inf.* **140**, 335-352.

[32] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis.* Wiley, New York.

[33] Gneyou, K.E. (2005). Vitesse der convergence de certains estimateurs de Kaplan-Meier de la régression (in French). *Afr. Statist.* **1**, 77-92.

[34] Grenander, U. (1981). *Abstract inference.* Wiley, New York.

[35] Guessoum, Z. and Ould-Saïd, E.(2008). On nonparametric estimation of the regression function under random censorship mode. *Stat. & Dec.* **26**, 159-177.

[36] Guessoum, Z. and Ould-Saïd, E. (2010). Kernel regression uniform rate estimation for censored data under α-mixing condition. *El. J. Statist.* **4**, 117-132.

[37] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer series in statistics, Springer-Verlag, New York.

[38] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

[39] Huang, J. Z. and Stone, C. J. (1998). The $\mathcal{L}_2$ rate of convergence for event history regression with time-dependent covariates. *Scand. J. Statist.* **25**, 603–620.

[40] James, I. R. and Smith, P. J. (1984). Consistency results for linear regression with censored data. *Ann. Statist.* **12**, 590–600.

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

[42] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Stat. Plann. and Inf.* **89**, 1–23.

[43] Kohler, M. (2006). Nonparametric regression with additional measurement errors in the dependent variable. *J. Stat. Plann. and Inf.* **136**, 3339–3361.

[44] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Trans. Inform. Theory* **47**, 3054–3058.

[45] Kohler, M., Krzyżak, A., and Schäfer, D. (2002). Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli* **8**, 475–489.

[46] Kooperberg, C., Stone, C. J., and Troung, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90**, 78–94.

[47] Kooperberg, C., Stone, C. J., and Troung, Y. K. (1995b). The $\mathcal{L}_2$ rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143–157.

[48] Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right–censored data. *Ann. Statist.* **9**, 1276–1288.

[49] Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370–1402.

[50] Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74**, 301–309.

[51] Maillot, B. and Viallon, V. (2009). Uniform limit laws of the logarithm for nonparametric estimators of the regression function in presence of censored data. *Math. Meth. Statist.* **18**, 159–184.

[52] Martinussen, T. and Scheike, T. H. (2006). *Dynamic regression models for survival data.* Springer series for biology and health, Springer-Verlag, New York.

[53] Máthé, K. (2006). *Regressionsanalyse mit zensierten Daten* (in German). PhD thesis, Faculty of Mathematics and Physics, University of Stuttgart.

[54] Miller, R. G. (1976). Least squares regression with censored data. *Biometrika* **63**, 449–464.

[55] Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

[56] Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–625.

[57] Neumann, M. H. (1994). Fully data–driven nonparametric variance estimators. *Statistics* **25**, 189–212.

[58] Park, J. (2004). Optimal global rate of convergence in nonparametric regression with left-truncated and right-censored data. *J. Multivariate Anal.* **89**, 70–86.

[59] Peterson, A. V. Jr. (1977). Expressing the Kaplan–Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.* **72**, 854–858.

[60] Pintér, M. (2001). *Consistency results in nonparametric regression and classification.* PhD thesis, Budapest University of Technology and Economics.

[61] Pollard, D. (1984). *Convergence of stochastic processes.* Springer series in statistics, Springer-Verlag, New York.

[62] The R Project for Statistical Computing [Internet]. Available from: http://www.r-project.org.

[63] Rudin, W. (1974). *Real and complex analysis.* Mc Graw-Hill series in higher mathematics, McGraw-Hill Book Co., New York.

[64] Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18**, 303–328.

[65] Stadtmüller, U. and Tsybakov, A. B. (1995). Nonparametric recursive variance estimation. *Statistics* **27**, 55–63.

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

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

[68] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.* **21**, 1591–1607.

[69] Van de Geer, S. (2000). *Empirical processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

[70] Van't Veer, L. J., Dai, H., Van de Vijver, M. J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

[71] Van de Vijver, M. J., He, Y. D., Van't Veer, L. J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009.

[72] Wahba, G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

[73] Wang L., Brown, L. D., Cai, T. T., and Levine, M. (2008). Effect of mean variance function estimation in nonparametric regression. *Ann. Statist.* **36**, 646–664.

[74] Zheng, Z. (1987). A class of estimators of the parameters in linear regression with censored data. *Acta Math. Appl. Sin. Engl. Ser.* **3**, 231–241.