Prediction from randomly right censored data *

Michael Kohler[1], Kinga Máthé[1] and Márta Pintér[2]

[1] *Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany, email: kohler@mathematik.uni-stuttgart.de, matheka@mathematik.uni-stuttgart.de.*

[2] *Department of Computer Science and Information Theory, Technical University of Budapest, Stoczek u. 2, H-1521 Budapest, Hungary, email: marti@szit.bme.hu.*

Abstract

Let X be a random vector taking values in $\mathbb{R}^d$, let Y be a bounded random variable, and let C be a right censoring random variable operating on Y. It is assumed that C is independent of (X, Y), the distribution function of C is continuous and the support of the distribution of Y is a proper subset of the support of the distribution of C. Given a sample $\{X_i, \min\{Y_i, C_i\}, I_{[Y_i \leq C_i]}\}$ and a vector of covariates X, we want to construct an estimate of Y such that the mean squared error is minimized.

Without censoring, i.e. for $C = \infty$ almost surely, it is well–known that the mean squared error of suitably defined kernel, partitioning, nearest neighbor, least squares and smoothing spline estimates converges for every distribution of (X, Y) to the optimal value almost surely, if the sample size tends to infinity. In this paper, we modify the above estimates and show that in the right random censoring model described above the same is true for the modified estimates.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: censored data, prediction, regression estimate, universal consistency, local averaging estimates, least squares estimates, smoothing splines.

*Research supported by DAAD and MÖB.

1 Introduction

Let X be a random vector taking values in $\mathbb{R}^d$ and let Y be a real random variable with $\mathbf{E}Y^2 < \infty$. The coordinates of X may have various types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. Hence, we do not assume anything on the distribution of X. In regression analysis one wants to estimate Y after having observed X, i.e. one wishes to find a function f with $f(X)$ "close" to Y. If the main goal of the analysis is the minimization of the mean squared error, then one wants to find a function f^* with

$$\mathbf{E}\,|f^*(X) - Y|^2 = \min_f \mathbf{E}\,|f(X) - Y|^2 . \tag{1}$$

Let $m(x) := \mathbf{E}\{Y|X = x\}$ be the regression function and denote the distribution of X by μ. It is well known that for each measurable function f the relation

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

holds. Therefore m is the solution of the minimization problem (1) and the mean squared error $\mathbf{E}|f(X) - Y|^2$ is close to its optimal value $\mathbf{E}|m(X) - Y|^2$ if and only if

$$\int_{\mathbb{R}^d} |\, f(x) - m(x)\,|^2 \, \mu(dx)$$

(in the following denoted as L_2 error) is close to zero.

For the regression estimation problem the distribution of (X, Y) (and therefore also m) is unknown. Given only a sample (X_1, Y_1),...,(X_n, Y_n) of the distribution of (X, Y) one has to construct an estimate $m_n(x)$=$m_n(x, (X_1, Y_1), ..., (X_n, Y_n))$ of $m(x)$ such that $\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx)$ is small.

In many medical studies it is not possible to observe a sample of (X, Y). There Y is the survival time of a patient, and often some of the patients are still alive at termination of the study, or withdraw alive during the study or die from other causes than those under study. Hence, instead of an observation of the survival time of a patient one observes only the minimum of the survival time and a censoring time. In this article we consider the problem of estimating the regression function from such incomplete data.

To be more precise, let (X, Y, C), (X_1, Y_1, C_1), (X_2, Y_2, C_2), ... be independent identically distributed $\mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}_+$–valued random variables. Set $Z = \min\{Y, C\}$, $\delta = I_{[Y \le C]}$

and $Z_i = \min\{Y_i, C_i\}$, $\delta_i = I_{[Y_i \le C_i]}$ $(i = 1, \ldots, n)$. We consider the problem of estimating the regression function from the data $\mathcal{D}_n = \{(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)\}$. In the sequel we assume that Y is a nonnegative bounded real random variable. This is a reasonable assumption because in applications Y represents the survival time of a patient which is always nonnegative and bounded.

The existing statistical approaches in the survival analysis can be basically classified in regression based and hazard risk based methods. In the latter, one studies the contribution of the risk factors by modelling the conditional hazard rate function under specific assumptions on the conditional distribution of Y, see, e.g., Cox (1972, 1975). An overview of existing bibliography concerning survival analysis can be found, e.g., in Fan and Gijbels (1995).

An estimation of the slope in a linear regression model via the proportional hazard model was proposed by Cox (1972). Based on the Kaplan-Meier estimator, Buckley and James (1979) intoduced an estimator, which estimates a linear regression function directly from the censored data. The consistency of this estimator was investigated by James and Smith (1984). For a slight modification of the estimator Ritov (1990) and Lai and Ying (1991) established the asymptotic normality.

In practical applications it is normally not known whether or not the model is linear. Important diagnostic tools for uncensored regression as scatterplots or residualplots are not directly applicable in the censored case. Therefore we are interested in nonparametric regression estimates for the randomly right censored case.

Without censoring, i.e. for $C = \infty$ almost surely, it is well–known that there exist estimates which are weakly and strongly universally consistent, i.e. for which the L_2 error tends to 0 in probability or almost surely for every distribution of (X, Y) with $\mathbf{E}|Y|^2 < \infty$. The existence of weakly universally consistent estimates was shown for the first time in Stone (1977). There nearest neighbor estimates were considered. Later this result was extended for various (kernel, partitioning, least squares, smoothing spline) regression estimates, see e.g. Devroye and Wagner (1980), Spiegelman and Sacks (1980), Györfi (1991), Walk (1997), Devroye et al. (1994), Lugosi and Zeger (1995), Kohler (1997, 1999), Kohler and Krzyżak (1998), Györfi and Walk (1996,1997) and Györfi, Kohler and Walk (1997).

Concerning the censored case, Beran (1981) introduced a nonparametric estimate of the conditional survival function and proved some consistency results which were later extended by Dabrowska (1987, 1989). They assumed that Y and C are conditionally independent given X. As explained in Section 3 of Carbonez, Györfi and van der Meulen (1995), one can use this estimate to derive a consistent estimate of the regression function in the censored case, but this estimate will be difficult to compute in an application. We provide consistent regression estimates under the somewhat stronger condition that C and (X, Y) are independent. This condition is reasonable whenever the censoring is independent of the characteristics of the individuum under study. Furthermore we assume, that the distribution function of C is continuous and that the support of the distribution of Y is a proper subset of the support of the distribution of C.

We present versions of kernel, partitioning, nearest neighbor, least squares and smoothing spline estimates that can be used for prediction from randomly right censored data. Under the above conditions we show the strong consistency of the estimates for every distribution of (X, Y, C) with Y nonnegative and almost surely bounded. We want to stress that the results hold without any assumption on the distribution of X or on the smoothness of the underlying regression function.

In Carbonez (1992) and Carbonez, Györfi and van der Meulen (1995) the same model of censored data is considered as in our paper, partitioning estimates are defined and it is shown that they are strongly consistent for bounded Y. Our results are based on a simplification of their proof; furthermore we show the strong consistency in the randomly right censored case not only of partitioning, but also of kernel, nearest neighbor, least squares and penalized least squares estimates. As in the above cited results concerning universally consistent estimation of the regression function in the uncensored case, our results require no restriction on the distribution of X.

In Section 2 we describe estimates of regression function from uncensored data. The basic ideas behind the definition of the new estimates are presented in Section 3. The main results are formulated in Section 4 and proven in Section 5. In the appendix a list of some results of the so–called Vapnik–Chervonenkis theory is given. These results are used in the consistency proofs of the least squares and penalized least squares estimates.

2 Estimation of a regression function from uncensored data

In this section we give an overview of estimates of the unknown regression function that can be used if censoring is not present, i.e. for $C = \infty$ almost surely.

The value of the regression function at a point x is the conditional expectation of Y given $X = x$. It can be estimated by local averaging, i.e., by averaging those Y_i's where X_i is close to x. Such an estimate can be represented by

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) \cdot Y_i,$$

where $W_{n,i}(x) = W_{n,i}(x, X_1, \dots, X_n)$ are weights which depend on x and on $X_1, \dots, X_n$. Well-known examples of this type of estimates are k_n–*nearest neighbor estimates*, where $k_n \in \mathbb{N}$ is a parameter of the estimate and one uses weights defined (modulo some suitable tie–breaking strategy – see Devroye et al. (1994) for details) by

$$W_{n,i}(x) = \begin{cases} \frac{1}{k_n} & \text{if } X_i \text{ is among the } k_n \text{ nearest neighbors of } x \text{ in } \{X_1, \dots, X_n\}, \\ 0 & \text{else,} \end{cases} \tag{3}$$

kernel estimates, which use as parameters a positive number h_n (so–called bandwidth) and a function $K : \mathbb{R}^d \to \mathbb{R}$ (so–called kernel function) and where one uses weights

$$W_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)} \quad \left(\frac{0}{0} := 0\right), \tag{4}$$

and *partitioning estimates*, which use a partition $\pi_n = \{A_{n,j} \ : \ j\}$ as a parameter and weights defined by

$$W_{n,i}(x) = \sum_{j} \frac{I_{A_{n,j}}(X_i)}{\sum_{k=1}^{n} I_{A_{n,j}}(X_k)} I_{A_{n,j}}(x) \quad \left(\frac{0}{0} := 0\right). \tag{5}$$

Here I_A denotes the indicator function of the set A.

Next we introduce *least squares estimates.* Recall that the regression function minimizes the so–called L_2 risk $\mathbf{E}|f(X) - Y|^2$ over all (measurable) functions $f : \mathbb{R}^d \to \mathbb{R}$. This motivates to construct an estimate of m by minimizing an estimate of the L_2 risk, the so–called empirical L_2 risk

$$\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{6}$$

If one minimizes (6) over all (measurable) functions $f : \mathbb{R}^d \to \mathbb{R}$, then one gets (at least if the X_1 , ..., X_n are all distinct) a function which interpolates the data (X_1, Y_1), ..., (X_n, Y_n). Obviously, this is not a reasonable estimate of the regression function. Therefore one chooses sets $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, which depend on the sample size n and which get more and more complex for n tending to infinity, and minimizes (6) only over $\mathcal{F}_n$, i.e., one defines least squares estimates m_n by

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{7}$$

Here we assume for simplicity that the minima in (7) exists, but we do not assume that it is unique.

Instead of restricting the class of function over which one minimizes the empirical L_2 risk, one can also add a penalty term to the empirical L_2 risk. More precisely, let $k \in \mathbb{N}$ with $2k > d$ and denote by $W^k(\mathbb{R}^d)$ the Sobolev space containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ whose derivatives of total order k are in $L_2(\mathbb{R}^d)$. The condition $2k > d$ implies that the functions in $W^k(\mathbb{R}^d)$ are continuous and hence the evaluation of a function at a point is well defined. Set

$$J_k^2(f) = \sum_{\alpha_1, \ldots, \alpha_d \in \mathbb{N},\, \alpha_1 + \ldots + \alpha_d = k} \frac{k!}{\alpha_1! \cdot \ldots \cdot \alpha_d!} \int_{\mathbb{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) \right|^2 dx. \tag{8}$$

Then the *penalized least squares* (or *smoothing spline*) *estimate* m_n is defined by

$$m_n(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda_n J_k^2(f) \right), \tag{9}$$

where $\lambda_n > 0$ is a parameter of the estimate. Again we do not require for the minimum to be unique. A function which achieves the minima in (9) can be computed by solving a linear equation system (see, e.g., Duchon (1976), Section V).

3 Basic ideas

3.1 Estimation of a mean from censored data

A crucial point in the definition of the above estimates is estimation of a mean from i.i.d. data. In the definition of (penalized) least squares estimates the L_2 risk is estimated by (the sum of a penalty term and the) empirical L_2 risk (6), and in the definition of local averaging estimates the local average

$$\sum_{i=1}^{n} W_{n,i}(X_i)Y_i$$

is used as an approximation of the conditional expectation of Y given $X = x$. We show next how these estimates can be adapted to the case of censored data. For this we use an idea introduced in Carbonez, Györfi and van der Meulen (1995).

Let

$$T_F = \sup\{t : F(t) > 0\},$$

$$T_G = \sup\{y : G(y) > 0\},$$

$$T_K = \min\{T_F, T_G\},$$

where $F(t) = \mathbf{P}\{Y > t\}$ and $G(t) = \mathbf{P}\{C > t\}$.

Throughout the paper we will assume

$$C \text{ and } (X,Y) \text{ are independent,}$$

$$T_F < \infty, G \text{ is continuous and } G(T_F) > 0. \tag{10}$$

The condition $G(T_F) > 0$ implies that $T_F < T_G$ and hence $T_K = T_F$ by definition of T_K.

Let h be a function $h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ and consider the problem of estimating the mean $\mathbf{E}\{h(X,Y)\}$ from the censored data $\mathcal{D}_n$. Then an unbiased estimate of $\mathbf{E}\{h(X,Y)\}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i h(X_i, Z_i)}{G(Z_i)}. \tag{11}$$

Indead, using the independence of (X,Y) and C and properties of conditional expectations one gets

$$\begin{aligned}
\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(X_i,Z_i)}{G(Z_i)}\right\} &= \mathbf{E}\left\{\frac{I_{[Y_1 \le C_1]} h(X_1,Y_1)}{G(Y_1)}\right\} \\
&= \mathbf{E}\left\{\frac{h(X_1,Y_1)}{G(Y_1)}\mathbf{E}\left\{I_{[Y_1 \le C_1]}|X_1,Y_1\right\}\right\} \\
&= \mathbf{E}\left\{h(X_1,Y_1)\right\}.
\end{aligned}$$

In an application, G will usually be unknown, hence it is not possible to use the estimate (11). The basic idea in the definition of the estimates in this paper is to replace G in (11) by an estimate G_n of G. The way how to construct such an estimate G_n will be described in the next subsection.

3.2 The Kaplan–Meier estimator

In this section we present the definition of the Kaplan-Meier estimator (KME) together with some notations and properties. For more details see, e.g., Miller (1981).

Let G_n be the Kaplan-Meier estimator of G introduced by Kaplan and Meier (1958). As often in the literature we treat the largest observation as uncensored. In this case G_n is defined by

$$G_n(t) = \begin{cases} \prod_{i=1}^{n} \left[1 - \frac{1-\delta_{(i)}}{n-i+1}\right]^{I_{[Z_{(i)} \le t]}} & \text{if} \quad t < T_{K,n}, \\ \lim_{s \to T_{K,n},\, s < T_{K,n}} G_n(t) & \text{if} \quad t \ge T_{K,n}, \end{cases} \tag{12}$$

where $T_{K,n} := \max\{Z_1, \ldots, Z_n\}$ and the pairs $(Z_{(i)}, \delta_{(i)}), i = 1, \ldots, n$ are the n pairs of observed (Z_i, δ_i) ordered on the $Z_{(i)}$, i.e.

$$Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)} = T_{K,n}. \tag{13}$$

For G arbitrary choosen, some of the Z_i may be identical, so that the ordering is not unique. However, the KME is unique (see Peterson (1977)).

A Glivenko-Cantelli theorem for the censored case was obtained e.g. by Stute and Wang (1993). Their results imply that for continuous G one has

$$\sup_{t \le T_K} |G_n(t) - G(t)| \longrightarrow 0 \quad (n \to \infty) \quad a.s., \tag{14}$$

a result which will be extremely useful in this paper.

3.3 Definition of the new estimates

Next we use the ideas described above to adapt the definiton of local averaging, least squares and penalized least squares estimates to the case of censored data.

Let G_n be the Kaplan Meier estimate of G defined by (12). We define local averaging estimates for censored data by

$$\tilde{m}_{n,1}(x) = \tilde{m}_{n,1}(x, \mathcal{D}_n) = \sum_{i=1}^{n} W_{n,i}(x) \cdot \frac{\delta_i \cdot Z_i}{G_n(Z_i)}, \tag{15}$$

least squares estimates for censored data by

$$\tilde{m}_{n,2}(\cdot) = \tilde{m}_{n,2}(\cdot, \mathcal{D}_n) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \cdot |f(X_i) - Z_i|^2}{G_n(Z_i)} \tag{16}$$

and penalized least squares estimates for censored data by

$$\tilde{m}_{n,3}(\cdot) = \tilde{m}_{n,3}(\cdot, \mathcal{D}_n) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \cdot |f(X_i) - Z_i|^2}{G_n(Z_i)} + \lambda_n J_k^2(f) \right), \tag{17}$$

where $J_k^2(f)$ is as defined in (8).

The fact $0 \leq Y \leq T_F < \infty$ *a.s.* implies that $0 \leq m(x) \leq T_F$ $(x \in \mathbb{R}^d)$. In order to ensure that the estimates are bounded in the same way, set

$$m_{n,j}(x) = m_{n,j}(x, \mathcal{D}_n) = \begin{cases} T_{K,n} & \text{if} \quad \tilde{m}_{n,j}(x) > T_{K,n}, \\ \tilde{m}_{n,j}(x) & \text{if} \quad 0 \leq \tilde{m}_{n,j}(x) \leq T_{K,n}, \\ 0 & \text{if} \quad \tilde{m}_{n,j}(x) < 0, \end{cases} \tag{18}$$

$(j = 1, 2, 3)$, with $T_{K,n}$ as in (13). Observe $0 \leq T_{K,n} \leq T_K = T_F < \infty$.

4 Main results

4.1 Local averaging estimates

Our first result concerns the strong consistency of local averaging estimates for censored data.

Theorem 1 *Assume that the weights $\{W_{n,i}\}$ are nonnegative, that the sum of the weights is at most one and that for all distributions of (X, Y) with $|Y|$ bounded almost surely*

$$\int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} W_{n,i}(x) Y_i - m(x) \right)^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s. \tag{19}$$

Then the estimate $m_{n,1}$ defined by (15) and (18) satisfies

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

for all distributions of (X, Y, C) which satisfy (10).

By Theorem 1 and Devroye and Györfi (1983) we get the following result concerning the consistency of partitioning estimates.

Corollary 1 *(Carbonez, Györfi and van der Meulen (1995)). Let $\pi_n = \{A_{n,j} : j = 1, 2, \ldots\}$ be partitions of $\mathbb{R}^d$ such that for each sphere S centered at the origin*

$$\lim_{n \to \infty} \max_{j : A_{n,j} \cap S \neq 0} diam(A_{n,j}) = 0$$

and

$$\lim_{n\to\infty} \frac{|\{j : A_{n,j} \cap S \neq 0\}|}{n} = 0.$$

Define the partitioning estimate $m_{n,1}$ by (15), (18) and (5). Then

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

for all distributions of (X, Y, C) which satisfy (10).

By Theorem 1 and Devroye and Krzyżak (1989) we get the following result concerning the consistency of kernel estimates.

Corollary 2 *Let K be a regular kernel, i.e. it is nonnegative, and there is a ball $S_{0,r}$ of radius $r > 0$ centered at the origin, and constant $b > 0$ such that*

$$K(x) \geq bI_{\{x \in S_{0,r}\}}$$

and

$$\int_{\mathbb{R}^d} \sup_{u \in x + S_{0,r}} K(u) dx < \infty.$$

If $h_n \to 0$ and $nh_n^d \to \infty$, then for the kernel regression estimate $m_{n,1}$ defined by (15), (18) and (4),

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

for all distributions of (X, Y, C) which satisfy (10).

By Theorem 1 and Devroye et al. (1994) we get the following result concerning the consistency of nearest neighbor estimates.

Corollary 3 *If $k_n \to \infty$ and $k_n/n \to 0$, then for the nearest neighbor regression estimate $m_{n,1}$ defined by (15), (18) and (3),*

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

for all distributions of (X, Y, C) which satisfy (10) and for which for each $x \in \mathbb{R}^d$ the random variable $\|X - x\|$ is absolutely continuous.

4.2 Least squares estimates

In the next theorem we use the notion $\mathcal{V}_{\mathcal{F}_n^+}$ for the VC dimension of the set of graphs of functions in $\mathcal{F}_n$ (see appendix for exact definition).

Theorem 2 *Assume that the sets $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfy*

$$\frac{\mathcal{V}_{\mathcal{F}_n^+}}{n} \to 0 \quad (n \to \infty) \tag{20}$$

and

$$\inf_{\substack{f \in \mathcal{F}_n, \\ 0 \le f(x) \le T_F \ (x \in \mathbb{R}^d)}} \sup_{x \in [-A,A]^d} |f(x) - g(x)| \to 0 \quad (n \to \infty) \tag{21}$$

for every $A \in \mathbb{R}_+$ and for every continuous function g with $0 \le g(x) \le T_F$ $(x \in \mathbb{R}^d)$ and $g(x) = 0$ for all $x \in \mathbb{R}^d \setminus [-A, A]^d$. Let $m_{n,2}$ be defined by (16) and (18). Then

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

for all distributions of (X, Y, C) which satisfy (10).

We now give an example for sets of functions which satisfy the conditions of Theorem 2.
Example. Let $M \in \mathbb{N}_0$, $A_n \in \mathbb{R}_+$, $K_n \in \mathbb{N}$, such that

$$\frac{K_n^d}{n} \to 0 \quad (n \to \infty), \tag{22}$$

$$A_n \to \infty \quad (n \to \infty) \quad \text{and} \quad \frac{A_n}{K_n} \to 0 \quad (n \to \infty). \tag{23}$$

Let $\pi_n = \{B_{n,1}, B_{n,2}, \ldots B_{n,K_n^d}\}$ be the partition of $[-A_n, A_n]^d$ into K_n^d equidistant cubes. Let $\mathcal{F}_n$ be the set of all (multivariate) piecewise polynomials of degree M (or less, in each coordinate) with respect to π_n.

$\mathcal{F}_n$ is a $K_n^d \cdot (M+1)^d$-dimensional vector space, so that by Lemma 5 of the appendix it follows that $V_{\mathcal{F}_n^+} \le K_n^d \cdot (M+1)^d + 1$. Condition (20) is then satisfied by (22). In order to show (21) let g be a continuous function with $0 \le g(x) \le T_F$ $(x \in \mathbb{R}^d)$ and $g(x) = 0$ for all $x \in \mathbb{R}^d \setminus [-A_n, A_n]^d$. Choose fixed $x_i \in B_{n,i}$ $(i = 1, \ldots, K_n^d)$ and define $f_g \in \mathcal{F}_n$ by

$$f_g(x) = \sum_i g(x_i) I_{B_{n,i}}(x) \quad (x \in \mathbb{R}^d).$$

For arbitrary points $u_1, u_2 \in B_{n,i}$ $(i = 1, \ldots, K_n^d)$ it holds that $||u_1 - u_2|| \leq \sqrt{d} \cdot \frac{2A_n}{K_n}$. Furthermore for given x there exists a cube $B_{n,i}$ by the definition of f_g, so that we have $f_g(x) = g(x_i)$. It follows

$$\begin{aligned}
\inf_{\substack{f \in \mathcal{F}_n, \\ 0 \leq f(x) \leq T_F \ (x \in \mathbb{R}^d)}} \sup_{x \in [-A_n, A_n]^d} |f(x) - g(x)| &\leq \sup_{x \in [-A_n, A_n]^d} |f_g(x) - g(x)| \\
&\leq \sup_{u_1, u_2 \in B_{n,i}} |g(u_1) - g(u_2)| \\
&\leq \sup_{||u_1 - u_2|| \leq \sqrt{d} \cdot \frac{2A_n}{K_n}} |g(u_1) - g(u_2)| \\
&\longrightarrow 0
\end{aligned}$$

by (23) and by uniform continuity of g, so that condition (21) is satisfied.

4.3 Smoothing spline estimates

Theorem 3 *Let $k \in \mathbb{N}$ with $2k > d$, and for $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n \to \infty) \tag{24}$$

and

$$n \cdot \lambda_n \to \infty \quad (n \to \infty). \tag{25}$$

Let $m_{n,3}$ be defined by (17) and (18). Then

$$\int_{\mathbb{R}^d} |m_{n,3}(x) - m(x)|^2 \mu(dx) \to 0 \quad a.s. \quad (n \to \infty)$$

for all distributions of (X, Y, C) which satisfy (10) and $|X|$ bounded a.s.

Remark. The assumption $|X|$ bounded $a.s.$ may be dropped if one modifies the estimate as described in Remark 3 in Kohler and Krzyżak (1998).

5 Proofs

We use following notations: For any $0 \leq t < \infty$ and $x \in \mathbb{R}$ define

$$T_{[0,t]}(x) = \begin{cases} t & \text{if} \quad x > t, \\ x & \text{if} \quad 0 \leq x \leq t, \\ 0 & \text{if} \quad x < 0. \end{cases} \tag{26}$$

For $f : \mathbb{R}^d \to \mathbb{R}$ define $T_{[0,t]} f : \mathbb{R}^d \to \mathbb{R}$ by $(T_{[0,t]} f)(x) := T_{[0,t]}(f(x))$ $(x \in \mathbb{R}^d)$.

In analogy to (18) set

$$m^\star_{n,j}(x) = \begin{cases} T_F & \text{if} \quad \tilde{m}_{n,j}(x) > T_F, \\ \tilde{m}_{n,j}(x) & \text{if} \quad 0 \le \tilde{m}_{n,j}(x) \le T_F, \\ 0 & \text{if} \quad \tilde{m}_{n,j}(x) < 0, \end{cases} \tag{27}$$

$(j = 1, 2, 3)$. Furthermore, define

$$\mathcal{B}^\star_n \mathcal{F}_n := \{ f \in \mathcal{F}_n : \; 0 \le f(x) \le T_F \; (x \in \mathbb{R}^d) \} \tag{28}$$

and

$$\mathcal{T}^\star_n \mathcal{F}_n := \{ g : \mathbb{R}^d \to \mathbb{R}_+, \; g = T_{[0,T_F]} f \text{ for some } f \in \mathcal{F}_n \}. \tag{29}$$

In order to prove our main results we need the following lemma:

Lemma 1 *Let $j \in \{1, 2, 3\}$. Then*

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

if and only if

$$\int_{\mathbb{R}^d} |m^\star_{n,j}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

Proof. It suffices to show

$$\int_{\mathbb{R}^d} |m^\star_{n,j}(x) - m_{n,j}(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s. \tag{30}$$

Since $T_{K,n} \le T_K = T_F$, it is easy to see that

$$|m^\star_{n,j}(x) - m_{n,j}(x)| = |T_{[0,T_F]} \tilde{m}_{n,j}(x) - T_{[0,T_{K,n}]} \tilde{m}_{n,j}(x)| \le T_F - T_{K,n} \quad (x \in \mathbb{R}^d).$$

Thus

$$\int_{\mathbb{R}^d} |m^\star_{n,j}(x) - m_{n,j}(x)|^2 \mu(dx) \le \int_{\mathbb{R}^d} (T_F - T_{K,n})^2 \mu(dx) = (T_F - T_{K,n})^2.$$

Since $T_{K,n} \to T_K$ $(n \to \infty)$ $a.s.$ and $T_K = T_F$, (30) follows. $\square$

5.1 Proof of Theorem 1

Because of Lemma 1 it suffices to show

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

Set

$$\bar{m}_n(x) = T_{[0,T_F]}\left(\sum_{i=1}^n W_{n,i}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)} \right).$$

Then

$$\begin{aligned}
&\int_{\mathbb{R}^d} |m^\star_{n,1}(x) - m(x)|^2 \mu(dx) \\
&\leq 2 \int_{\mathbb{R}^d} |m^\star_{n,1}(x) - \bar{m}_n(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx). \qquad (31)
\end{aligned}$$

Because of $0 \leq \frac{\delta_1 Z_1}{G(Z_1)} \leq \frac{T_F}{G(T_F)}$ $a.s.$ and

$$\mathbf{E}\left\{ \frac{\delta_1 Z_1}{G(Z_1)} \middle| X_1 \right\} = \mathbf{E}\left\{ \frac{I_{[Y_1 \leq C_1]} Y_1}{G(Y_1)} \middle| X_1 \right\} = \mathbf{E}\left\{ \frac{Y_1}{G(Y_1)} \mathbf{E}\{I_{[Y_1 \leq C_1]} | X_1, Y_1\} \middle| X_1 \right\} = m(X_1),$$

where the last equation follows from the independence of (X, Y) and C, (19) and $0 \leq m(x) \leq T_F$ $(x \in \mathbb{R}^d)$ imply

$$\int_{\mathbb{R}^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx) \leq \int_{\mathbb{R}^d} \left| \sum_{i=1}^n W_{n,i}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)} - m(x) \right|^2 \mu(dx) \to 0 \quad (32)$$

$a.s.$ as $n \to \infty$. Furthermore, the assumptions that the weights are nonnegative and that their sum is at most one imply

$$\begin{aligned}
&\int_{\mathbb{R}^d} |m^\star_{n,1}(x) - \bar{m}_n(x)|^2 \mu(dx) \\
&\leq T_F \int_{\mathbb{R}^d} \left| \sum_{i=1}^n W_{n,i}(x) \cdot \frac{\delta_i Z_i}{G_n(Z_i)} - \sum_{i=1}^n W_{n,i}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)} \right| \mu(dx) \\
&\leq T_F \int_{\mathbb{R}^d} \sum_{i=1}^n W_{n,i}(x) \cdot T_F \cdot \left| \frac{1}{G_n(Z_i)} - \frac{1}{G(Z_i)} \right| \mu(dx) \\
&\leq T_F^2 \frac{1}{G(T_F) \cdot G_n(T_F)} \sup_{t \leq T_F} |G(t) - G_n(t)| \to 0 \quad (n \to \infty) \quad a.s.,
\end{aligned}$$

where we have used $G(T_F) > 0$ and (14). This together with (31) and (32) implies the assertion. □

5.2 Proof of Theorem 2

By Lemma 1 it suffices to show

$$\int_{\mathbb{R}^d} |m^\star_{n,2}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

The proof of this relation is divided into three steps.

In the first step we show

$$\begin{aligned}\int_{\mathbb{R}^d} |m^\star_{n,2}(x) - m(x)|^2 \mu(dx) \quad &\leq \quad 2 \sup_{f \in \mathcal{T}^\star_n \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Y_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \\ &\quad + \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx). \end{aligned} \tag{33}$$

Because of (2) we have the following decomposition of the L_2 risk:

$$\begin{aligned}\int_{\mathbb{R}^d} |m^\star_{n,2}(x) - m(x)|^2 \mu(dx) \quad &= \quad \left\{ \mathbf{E}\left(|m^\star_{n,2}(X) - Y|^2 \, | \mathcal{D}_n \right) - \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 \right\} \\ &\quad + \left\{ \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 \right\}. \end{aligned}$$

Again by (2)

$$\inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 = \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx).$$

Thus we just need an upper bound for the first term. It holds that

$$\begin{aligned} &\mathbf{E}\left(|m^\star_{n,2}(X) - Y|^2 \, | \mathcal{D}_n \right) - \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 \\ &= \sup_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \left[\mathbf{E}\left(|m^\star_{n,2}(X) - Y|^2 \, | \mathcal{D}_n \right) - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |m^\star_{n,2}(X_i) - Z_i|^2}{G_n(Z_i)} \right. \\ &\quad + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |m^\star_{n,2}(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |\tilde{m}_{n,2}(X_i) - Z_i|^2}{G_n(Z_i)} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |\tilde{m}_{n,2}(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} \\ &\quad \left. + \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right]. \end{aligned}$$

Now we give upper bounds for the terms on the right-hand side of the last equality.

The third term is by (16) and $f \in \mathcal{B}^\star_n \mathcal{F}_n \subset \mathcal{F}_n$ less than or equal to 0.

The same is true for the second term on the right-hand side, because of (27) and $0 \le Z_i \le T_{K,n} \le T_K = T_F$ $(i = 1, \ldots n)$ $a.s.$ we have for all $i \in \{1, \ldots, n\}$

$$|\tilde{m}_{n,2}(X_i) - Z_i| = |m^\star_{n,2}(X_i) - Z_i| + |m^\star_{n,2}(X_i) - \tilde{m}_{n,2}(X_i)| \ge |m^\star_{n,2}(X_i) - Z_i|.$$

Since $m^\star_{n,2} \in \mathcal{T}^\star_n \mathcal{F}_n$ and $f \in \mathcal{B}^\star_n \mathcal{F}_n \subset \mathcal{T}^\star_n \mathcal{F}_n$ we can bound the remaining two terms by

$$\sup_{f \in \mathcal{T}^\star_n \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Y_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right|,$$

which proves (33).

In the second step of the proof we show

$$\inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx) \longrightarrow 0 \ (n \to \infty) \ a.s. \tag{34}$$

Let $\epsilon > 0$ be arbitrary. Denote the set of all continuous functions with compact support by $C_0(\mathbb{R}^d)$. Since $m \in L_2(\mu)$ and $C_0(\mathbb{R}^d)$ dense in $L_2(\mu)$ for any distribution μ of X, there exists a function $h \in C_0(\mathbb{R}^d)$ with

$$\int_{\mathbb{R}^d} |h(x) - m(x)|^2 \mu(dx) < \epsilon.$$

Choose $A > 0$ with

$$h(x) = 0 \ (x \in \mathbb{R}^d \setminus [-A, A]^d) \ \text{ and } \ \mu(\mathbb{R}^d \setminus [-A, A]^d) \le \frac{\epsilon}{T_F^2}. \tag{35}$$

Set

$$\bar{h}(x) := T_{[0,T_F]} h(x) \ \ (x \in \mathbb{R}^d).$$

Then $\bar{h} \in C_0(\mathbb{R}^d)$, $0 \le \bar{h}(x) \le T_F$ $(x \in \mathbb{R}^d)$, and $0 \le m(x) \le T_F$ $(x \in \mathbb{R}^d)$ implies

$$\int_{\mathbb{R}^d} |\bar{h}(x) - m(x)|^2 \mu(dx) \ \ \le \ \ \int_{\mathbb{R}^d} |h(x) - m(x)|^2 \mu(dx) \le \epsilon.$$

From this, $|f(x) - m(x)|^2 \le T_F^2$ $(x \in \mathbb{R}^d)$ for all $f \in \mathcal{B}^\star_n \mathcal{F}_n$ and (35) we conclude

$$\begin{aligned}
&\inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx) \\
&\le \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \int_{[-A,A]^d} |f(x) - m(x)|^2 \mu(dx) + T_F^2 \cdot \mu\left(\mathbb{R}^d \setminus [-A, A]^d\right) \\
&\le \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} 2 \int_{[-A,A]^d} |f(x) - \bar{h}(x)|^2 \mu(dx) + 2 \int_{[-A,A]^d} |\bar{h}(x) - m(x)|^2 \mu(dx) + \epsilon \\
&\le 2 \inf_{f \in \mathcal{B}^\star_n \mathcal{F}_n} \sup_{x \in [-A,A]^d} |f(x) - \bar{h}(x)|^2 + 3\epsilon.
\end{aligned}$$

With (21) and $\epsilon \to 0$ we have proved (34).

In the third step of the proof we show

$$\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Y_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \longrightarrow 0 \ (n \to \infty) \ a.s. \tag{36}$$

We use the decomposition

$$\begin{aligned} & \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \\ \leq & \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} \right| \\ & + \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right|. \end{aligned}$$

Using $0 \leq f(x) \leq T_F$ ($x \in \mathbb{R}^d$) for all $f \in \mathcal{T}_n^\star \mathcal{F}_n$, $0 \leq Z_i \leq T_{K,n} \leq T_K = T_F$ ($i = 1, \ldots, n$) $a.s.$ and $G(T_F) > 0$, we have

$$\begin{aligned} & \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} \right| \\ = & \frac{1}{n} \sum_{i=1}^n \delta_i |f(X_i) - Z_i|^2 \left| \frac{1}{G(Z_i)} - \frac{1}{G_n(Z_i)} \right| \\ \leq & \frac{T_F^2}{G(T_F) \cdot G_n(T_F)} \sup_{t \leq T_F} |G(t) - G_n(t)|, \end{aligned}$$

which together with (14) implies

$$\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} \right| \longrightarrow 0 \quad (n \longrightarrow \infty) \quad a.s.$$

It remains to prove that

$$\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \longrightarrow 0 \quad (n \longrightarrow \infty) \quad a.s. \tag{37}$$

holds. For the proof of (37) we will use Lemma 3 and Lemma 4 of the appendix.

For $V := (X, Z, \delta)$, $V_1 := (X_1, Z_1, \delta_1), \ldots, V_n := (X_n, Z_n, \delta)$ i.i.d, define

$$\begin{aligned} \mathcal{H}_n \ := \ & \Big\{ h : \mathbb{R}^d \times [0, T_F] \times \{0,1\} \longrightarrow \mathbb{R}_+ : \exists f \in \mathcal{T}_n^\star \mathcal{F}_n \text{ with} \\ & \quad h(x, z, \delta) := \frac{\delta |f(x) - z|^2}{G(z)} \ \text{ for all } (x, z, \delta) \in \mathbb{R}^d \times [0, T_F] \times \{0,1\} \Big\}. \end{aligned}$$

Since $0 \le Y \le T_F \le \infty$ *a.s.* and $0 \le f(x) \le T_F$ for all $x \in \mathbb{R}^d$ and all $f \in \mathcal{T}_n^\star \mathcal{F}_n$ the functions in $\mathcal{H}_n$ are nonnegative and bounded by $T_F^2/G(T_F)$.

Furthermore we have

$$\mathbf{E}h(V) = \mathbf{E}\frac{\delta|f(X)-Z|^2}{G(Z)} = \mathbf{E}|f(X)-Z|^2$$

and therefore

$$\sup_{f\in\mathcal{T}_n^\star\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n \frac{\delta_i|f(X_i)-Z_i|^2}{G(Z_i)} - \mathbf{E}|f(X)-Y|^2\right| = \sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n h(V_i) - \mathbf{E}h(V)\right|.$$

If $h_1, h_2 \in \mathcal{H}_n$ and f_1, f_2 are the corresponding functions in $\mathcal{T}_n^\star\mathcal{F}_n$, then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n |h_1(V_i)-h_2(V_i)| &= \frac{1}{n}\sum_{i=1}^n\left|\frac{\delta_i|f_1(X_i)-Z_i|^2}{G(Z_i)} - \frac{\delta_i|f_2(X_i)-Z_i|^2}{G(Z_i)}\right| \\
&\le \frac{1}{G(T_F)}\frac{1}{n}\sum_{i=1}^n |(f_1(X_i)+f_2(X_i)-2Z_i)(f_1(X_i)-f_2(X_i))| \\
&\le \frac{2T_F}{G(T_F)}\frac{1}{n}\sum_{i=1}^n |f_1(X_i)-f_2(X_i)|,
\end{aligned}$$

which implies that

$$\mathcal{N}(\epsilon, \mathcal{H}_n, V_1^n) \le \mathcal{N}\left(\frac{\epsilon\cdot G(T_F)}{2T_F}, \mathcal{T}_n^\star\mathcal{F}_n, X_1^n\right), \tag{38}$$

where $\mathcal{N}(\epsilon, \mathcal{F}, z_1^n)$ is the covering number defined in the appendix. From this, Lemma 3 and Lemma 4 of the appendix we get for arbitrary $t > 0$ that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n h(V_i) - \mathbf{E}h(V)\right| > t\right\} \\
&\le 8\mathbf{E}\left\{\mathcal{N}\left(\frac{t\cdot G(T_F)}{16T_F}, \mathcal{T}_n^\star\mathcal{F}_n, X_1^n\right)\right\}\cdot\exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right) \\
&\le 8\left(\frac{64eT_F^3}{t\cdot G(T_F)}\right)^{2\mathcal{V}_{\mathcal{T}_n^\star\mathcal{F}_n^+}}\cdot\exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right).
\end{aligned}$$

Because of $\mathcal{V}_{\mathcal{T}_n^\star\mathcal{F}_n^+} \le \mathcal{V}_{\mathcal{F}_n^+}$ and of condition (20) of the theorem the above probability is summable. Hence by the Borel-Cantelli lemma we get (36).

The assertion follows from steps 1 to 3. $\square$

5.3 Proof of Theorem 3

We will need the following lemma.

Lemma 2 *Let $m^\star_{n,3}$ be defined by (17) and (27). Assume that $2k > d$ and that conditions (24) and (25) hold. Then*

$$\mathbf{E}\left(|m^\star_{n,3}(X) - Y|^2 \,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |m^\star_{n,3}(X_i) - Z_i|^2}{G(Z_i)} \to 0$$

a.s. as $n \to \infty$ for all distributions of (X, Y, C) which satisfy (10) with $X \in [0,1]^d$ a.s.

Proof. By the definition of the estimate

$$\frac{1}{n}\sum_{i=1}^n \frac{\delta_i \cdot |\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} + \lambda_n J_k^2(\tilde{m}_{n,3})$$

$$\leq \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \cdot |0 - Z_i|^2}{G_n(Z_i)} + \lambda_n \cdot 0$$

$$\leq \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \cdot Z_i^2}{G(Z_i)} + \frac{1}{n}\sum_{i=1}^n \left| \frac{\delta_i \cdot Z_i^2}{G_n(Z_i)} - \frac{\delta_i \cdot Z_i^2}{G(Z_i)} \right|$$

$$\leq \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \cdot Z_i^2}{G(Z_i)} + \frac{T_F^2}{G(T_F)G_n(T_F)} \sup_{t \leq T_F} |G(t) - G_n(t)|$$

$$\to \mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\} \quad a.s.$$

as $n \to \infty$ because of the strong law of large numbers and (14). This together with (25) implies that for sufficiently large n

$$J_k^2(\tilde{m}_{n,3}) \leq \frac{2}{\lambda_n}\mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\} = n \cdot \frac{2\mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\}}{n \cdot \lambda_n} \leq n,$$

hence for n sufficiently large

$$\tilde{m}_{n,3} \in \mathcal{F}_n := \left\{f : f \in W^k(\mathbb{R}^d) \text{ and } J_k^2(f) \leq n\right\},$$

thus

$$m^\star_{n,3} \in \mathcal{T}_n^\star \mathcal{F}_n.$$

Similarly as in the proof of Theorem 2, let

$$\mathcal{H}_n := \left\{h : \mathbb{R}^d \times [0, T_F] \times \{0,1\} \longrightarrow \mathbb{R}^+ : h(x,z,\delta) := \frac{\delta |f(x) - z|^2}{G(z)} \text{ for some } f \in \mathcal{T}_n^\star \mathcal{F}_n\right\}.$$

Hence using the notations $V := (X, Z, \delta)$ and $V_i := (X_i, Z_i, \delta_i)$ $(i = 1, \ldots, n)$, and the fact that

$$\mathbf{E}h(V) = \mathbf{E}\left\{\frac{\delta|f(X) - Z|^2}{G(Z)}\right\} = \mathbf{E}|f(X) - Z|^2,$$

it suffices to show that

$$\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n} h(V_i) - \mathbf{E}h(V)\right| \to 0 \quad (n \to \infty) \quad a.s.$$

From (38) we know that

$$\mathcal{N}(\epsilon, \mathcal{H}_n, V_1^n) \leq \mathcal{N}\left(\frac{\epsilon \cdot G(T_F)}{2T_F}, \mathcal{T}_n^\star \mathcal{F}_n, X_1^n\right),$$

where $\mathcal{N}(\epsilon, \mathcal{F}, z_1^n)$ is the covering number defined in the appendix. Now from this and Lemma 3 and 6 of the appendix we get for any $t > 0$ that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n} h(V_i) - \mathbf{E}h(V)\right| > t\right\} \\
\leq\ & 8\mathbf{E}\left\{\mathcal{N}\left(\frac{t \cdot G(T_F)}{16T_F}, \mathcal{T}_n^\star \mathcal{F}_n, X_1^n\right)\right\} \cdot \exp\left(-\frac{nt^2 G(T_F)^2}{128T_F^4}\right) \\
\leq\ & 8\left(c_1 \frac{T_F \cdot n}{\frac{t \cdot G(T_F)}{16T_F}}\right)^{c_2\left(\frac{\sqrt{n}\cdot 16T_F}{t\cdot G(T_F)}\right)^{\frac{d}{k}} + c_3} \exp\left(-\frac{nt^2 G(T_F)^2}{128T_F^4}\right).
\end{aligned}$$

which is summable because of $2k > d$. Hence by the Borel-Cantelli lemma we get the assertion. $\square$

Proof of Theorem 3.

We may assume w.l.o.g. that $X \in [0, 1]^d$ a.s. By Lemma 1 it suffices to show

$$\int_{\mathbb{R}^d} |m_{n,3}^\star(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

Let $\epsilon > 0$ be arbitrary, and choose a bounded infinitely often differentiable function $g_\epsilon : \mathbb{R}^d \to \mathbb{R}$ such that

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

We use the following error decomposition

$$\int_{\mathbb{R}^d} |m_{n,3}^\star(x) - m(x)|^2 \mu(dx)$$

$$
\begin{aligned}
&= \mathbf{E}\left(|m^\star_{n,3}(X)-Y|^2\,|\mathcal{D}_n\right) - \mathbf{E}|m(X)-Y|^2 \\
&= \mathbf{E}\left(|m^\star_{n,3}(X)-Y|^2\,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G(Z_i)} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|\tilde{m}_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|\tilde{m}_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|g_\epsilon(X_i)-Z_i|^2}{G_n(Z_i)} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|g_\epsilon(X_i)-Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|g_\epsilon(X_i)-Z_i|^2}{G(Z_i)} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|g_\epsilon(X_i)-Z_i|^2}{G(Z_i)} - \mathbf{E}|g_\epsilon(X)-Y|^2 \\
&\quad + \mathbf{E}|g_\epsilon(X)-Y|^2 - \mathbf{E}|m(X)-Y|^2.
\end{aligned}
$$

By Lemma 2

$$
\mathbf{E}\left(|m^\star_{n,3}(X)-Y|^2\,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G(Z_i)} \to 0 \quad (n\to\infty) \quad a.s.
$$

Since $0 \le m^\star_{n,3}(X_i) \le T_{K,n}$ and $0 \le Z_i \le T_{K,n} \le T_K = T_F$ $(i=1,\dots,n)$ $a.s.$, $G(T_F)>0$ one can conclude from (14)

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} \\
&\le \frac{1}{n}\sum_{i=1}^n \delta_i|m^\star_{n,3}(X_i)-Z_i|^2 \left|\frac{1}{G(Z_i)} - \frac{1}{G_n(Z_i)}\right| \\
&\le \frac{T_F^2}{G(T_F)G_n(T_F)} \sup_{t\le T_F} |G(t)-G_n(t)| \\
&\to 0 \quad a.s.
\end{aligned}
$$

If $a,b\in\mathbb{R}$ with $0\le b\le T_F$ and $c=T_{[0,T_F]}a$ then $|c-b|\le|a-b|$, which implies that

$$
\frac{1}{n}\sum_{i=1}^n \frac{\delta_i|m^\star_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|\tilde{m}_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} \le 0.
$$

It follows from the definition of the estimate and (24) that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{\delta_i|\tilde{m}_{n,3}(X_i)-Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i|g_\epsilon(X_i)-Z_i|^2}{G_n(Z_i)} &\le \lambda_n J_k^2(g_\epsilon) - \lambda_n J_k^2(\tilde{m}_{n,3}) \\
&\le \lambda_n J_k^2(g_\epsilon) \to 0 \quad (n\to\infty).
\end{aligned}
$$

Since $G(T_F) > 0$ one can conclude from (14)

$$\frac{1}{G(T_F)G_n(T_F)} \sup_{t \le T_F} |G(t) - G_n(t)| \to 0 \quad (n \to \infty),$$

thus

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} \\ \le \; & \frac{1}{G(T_F)G_n(T_F)} \sup_{t \le T_F} |G(t) - G_n(t)| \, \frac{1}{n}\sum_{i=1}^{n} \delta_i |g_\epsilon(X_i) - Z_i|^2 \\ \le \; & \frac{1}{G(T_F)G_n(T_F)} \sup_{t \le T_F} |G(t) - G_n(t)| \cdot (2T_F^2 + 2 \sup_{x \in \mathbb{R}^d} |g_\epsilon(x)|^2) \to 0 \quad (n \to \infty) \quad a.s. \end{aligned}$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|g_\epsilon(X) - Y|^2 \to 0 \quad (n \to \infty) \quad a.s.$$

Finally,

$$\mathbf{E}|g_\epsilon(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 = \int_{\mathbb{R}^d} |m(x) - g_\epsilon(x)|^2 \mu(dx) < \epsilon.$$

With $\epsilon \to 0$ the result follows. □

A Some results of the Vapnik–Chervonenkis theory

In this section we list the definitions and results of the Vapnik–Chervonenkis theory which we have used in section 5. An excellent introduction to most of these results can be found in Devroye, Györfi and Lugosi (1996).

We start with the definition of covering numbers of classes of functions.

Definition 1 *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, z_1^n)$ is defined for any $\epsilon > 0$ and $z_1^n = (z_1, ..., z_n) \in \mathbb{R}^{d \cdot n}$ as the smallest integer k such that there exist functions $g_1, ..., g_k : \mathbb{R}^d \to \mathbb{R}$ with*

$$\min_{1 \le i \le k} \frac{1}{n} \sum_{j=1}^{n} |f(z_j) - g_i(z_j)| \le \epsilon$$

for each $f \in \mathcal{F}$.

If $Z_1^n = (Z_1, ..., Z_n)$ is a sequence of $\mathbb{R}^d$–valued random variables, then $\mathcal{N}(\epsilon, \mathcal{F}, Z_1^n)$ is a random variable with expected value $\mathbf{E}\mathcal{N}(\epsilon, \mathcal{F}, Z_1^n)$. The next result due to Pollard is the main tool in the proofs of Theorem 2 and Theorem 3.

Lemma 3 *(Pollard (1984), section II.5, Th. 24) Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$, and let $Z_1^n = (Z_1, ..., Z_n)$ be $\mathbb{R}^d$–valued i.i.d. random variables. Then for any $\epsilon > 0$*

$$\mathbf{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbf{E}f(Z_1)\right| > \epsilon\right] \leq 8\mathbf{E}\left(\mathcal{N}(\frac{\epsilon}{8}, \mathcal{F}, Z_1^n)\right)\exp\left(-\frac{n\,\epsilon^2}{128\,B^2}\right).$$

To bound covering numbers, we use the following definition of the VC dimension.

Definition 2 *Let $\mathcal{D}$ be a class of subsets of $\mathbb{R}^d$ and let $F \subseteq \mathbb{R}^d$. One says that $\mathcal{D}$ shatters F if each subset of F has the form $D \cap F$ for some D in $\mathcal{D}$. The VC dimension $\mathcal{V}_{\mathcal{D}}$ of $\mathcal{D}$ is defined as the largest integer k for which a set of cardinality k exists which is shattered by $\mathcal{D}$.*

A connection between covering numbers and VC dimensions is given by the following lemma, which uses the notation $\mathcal{V}_{\mathcal{F}^+}$ for the VC dimension of the set

$$\mathcal{F}^+ := \left\{\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \;:\; t \leq f(x)\} \;:\; f \in \mathcal{F}\right\}$$

of all graphs of functions of $\mathcal{F}$.

Lemma 4 *(Haussler (1992), Th. 6) Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [-B, B]$. Then one has for any $z_1^n \in \mathbb{R}^{d\cdot n}$ and any $0 < \epsilon < B$*

$$\mathcal{N}(\epsilon, \mathcal{F}, z_1^n) \leq 2\left(\frac{4eB}{\epsilon}\log\left(\frac{4eB}{\epsilon}\right)\right)^{\mathcal{V}_{\mathcal{F}^+}}.$$

The following result is often useful for bounding the VC dimension.

Lemma 5 *(Steele (1975), Dudley (1978)) Let $\mathcal{F}$ be a k–dimensional vector space of functions $f : \mathbb{R}^d \to \mathbb{R}$. Then the class of sets of the form $\{x \in \mathbb{R}^d \;:\; f(x) \geq 0\}$, $f \in \mathcal{F}$, has VC dimension less than or equal to k.*

The following lemma gives an upper bound for the covering number which occurs in the proof of Theorem 3.

Lemma 6 *(Kohler, Krzyżak (1998)) Let $L, c > 0$ and set*

$$\mathcal{F} = \left\{T_{[0,L]}f : f \in W^k(\mathbb{R}^d) \;\; \textit{and} \;\; J_k^2(f) \leq c\right\},$$

where the definition of $J_k^2(f)$ *is the same as in (8). Then for any* $\epsilon > 0$ *and* $x_1, ..., x_n \in [0,1]^d$

$$\mathcal{N}(\epsilon, \mathcal{F}, x_1^n) \le \left(c_1 \frac{Ln}{\epsilon}\right)^{c_2\left(\frac{\sqrt{c}}{\epsilon}\right)^{\frac{d}{k}} + c_3},$$

where $c_1, c_2, c_3 \in \mathbb{R}_+$ *are constants which only depend on* k *and* d*.*

References

[1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

[2] Buckley, J., and James, I. (1979). Linear regression with censored data. *Biometrica* **66** 429–436.

[3] Carbonez, A. (1992). *Nonparametric functional estimation under random censoring and a new semiparametric model of random censorship.* PhD thesis, Kaholieke Universiteit Leuven.

[4] Carbonez, A., Györfi, L., and van der Meulen, E. C. (1995). Partition–estimates of a regression function under random censoring. *Satistics & Decisions* **13** 21-37.

[5] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Royal Statist. Soc. B* **34** 187–220.

[6] Cox, D. R. (1975). Partial likelihood. *Biometrica* **62** 269–276.

[7] Dabrowska, D. M. (1987). Nonparametric regression with censored survival data. *Scand. J. Statist.* **14** 181–197.

[8] Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Ann. Statist.* **17** 1157–1167.

[9] Devroye, L., and Györfi, L. (1983). Distribution–free exponential bounds on the L_1 error of partitioning estimates of a regression function. In: *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, F. Konecny, J. Mogyoródi and W. Wertz, eds., pp. 67–76, Akadémiai Kiadó, Budapest, Hungary.

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

[11] Devroye, L., Györfi. L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, Berlin.

[12] Devroye, L., and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimates. *J. Stat. Plann. Inference* **23** 71–82.

[13] Devroye, L. P., and Wagner, T. J. (1980). Distribution–free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

[14] Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique* **10** 5–12.

[15] Dudley, R. (1978). Central limit theorems for empirical measures. *Annals of Probability* **6** 899–929.

[16] Fan, J., and Gijbels, I. (1995). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

[17] Györfi, L. (1991). Universal consistency of a regression estimate for unbounded regression functions. In: *Nonparametric Functional Estimation*, G. Roussas, ed., pp. 329–338, NATO ASI Series, Springer, Berlin, 1991.

[18] Györfi, L., Kohler, M., and Walk, H. (1998). Weak and strong universal consistency of semi–recursive partitioning and kernel regression estimates. *Statistics & Decisions* **16** 1–18.

[19] Györfi, L., and Walk, H. (1996). On the strong universal consistency of a series type regression estimate. *Mathematical Methods of Statistics* **5** 332–342.

[20] Györfi, L., and Walk, H. (1997). On the strong universal consistency of a recursive regression estimate by Pál Révész. *Statistics and Probability Letters* **31** 177–183.

[21] Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* **100** 78–150.

[22] James, I. R., and Smith, P. J. (1984). Consistency results for linear regression with censored data. *Ann. Statist.* **19** 590–600.

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

[24] Kohler, M. (1997). On the universal consistency of a least squares spline regression estimator. *Math. Methods of Statistics* **6** 349–364.

[25] Kohler, M. (1999). Universally consistent regression function estimation using hierarchical B-splines. *Journal of Multivariate Analysis* **67** 138–164.

[26] Kohler, M., and Krzyżak, A. (1998). A Vapnik–Chervonenkis approach to penalized least squares estimates. Under revision by *Journal of Multivariate Analysis.*

[27] Lai, T. L., and Ying, Z. (1991). Large–sample theory of a modified Buckley–James estimator. *Ann. Statist.* **19** 1370–1902.

[28] Lugosi, G., and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory* **41** 677–687.

[29] Miller, R. G. Jr. (1981). *Survival Analysis.* Wiley, New York.

[30] Peterson, A. V. Jr. (1977). Expressing the Kaplan–Meier estimator as a function of empirical subsurvival functions. *J. Amer. Satist. Assoc.* **72** 854-858.

[31] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

[32] Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18** 303–328.

[33] Spiegelman, C., and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* **8** 240–246.

[34] Steele, J. (1975) *Combinatorial entropy and uniform limit laws.* PhD Thesis, Stanford University, Stanford, CA.

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

[36] Stute, W., and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.* **21** 1591-1607.

[37] Walk, H. (1997). Strong universal consistency of kernel and partitioning regression estimates. Preprint 97–1, Mathematisches Institut A, Universität Stuttgart.