TECHNICAL
REPORT

0633

NONPARAMETRIC REGRESSION WITH DEPENDENT CENSORED DATA

EL GHOUCH, A. and I. VAN KEILEGOM

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Nonparametric regression with dependent censored data

Anouar El Ghouch[1]
Institute of Statistics
Université catholique de Louvain

Ingrid Van Keilegom[1]
Institute of Statistics
Université catholique de Louvain

September 18, 2006

Abstract

Let (X_i, Y_i) $(i = 1, \ldots, n)$ be n replications of a random vector (X, Y), where Y is supposed to be subject to random right censoring. The data (X_i, Y_i) are assumed to come from a stationary α-mixing process. We consider the problem of estimating the function $m(x) = \mathbb{E}(\phi(Y)|X = x)$, for some known transformation ϕ. Particular choices of ϕ lead to the conditional moment function of Y given X, or the conditional distribution of Y given X. This problem is approached in the following way : first, we introduce a transformed variable Y_i^*, that is not subject to censoring and satisfies the relation $\mathbb{E}(\phi(Y_i)|X_i = x) = \mathbb{E}(Y_i^*|X_i = x)$, and then we estimate $m(x)$ by applying local linear regression techniques to the pseudo-data $(X_i, \hat{Y}_i^*)$, where $\hat{Y}_i^*$ is a certain estimator of Y_i^*. The asymptotic properties of the proposed estimator are established. We investigate the performance of the estimator for small samples through a simulation study, and we discuss the optimal choice of the transformation Y_i^*. As a by-product, we obtain a general result on the uniform rate of convergence of kernel type estimators of functionals of an unknown distribution function, under strong mixing assumptions. This result is of independent interest, and can be applied in a wide variety of contexts.

KEY WORDS: Censoring, kernel smoothing, local linear smoothing, mixing sequences, nonparametric regression, strong mixing, survival analysis.

[1] Financial support from the IAP research network nr. P5/24 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

1 Introduction

A crucial point in a variety of statistical problems is the study of a relation between a variable of interest Y and some covariate X. This can be done via estimating the function

$$m(x) = \mathbb{E}\left(\phi(Y)|X = x\right),$$

where the (known) transformation ϕ is introduced to include various functions of interest. For example taking $\phi(y) = y^r$ gives the rth conditional moment and if we take $\phi(y) = I(y \leq t)$ then m becomes the conditional distribution function (CDF) of Y given $X = x$ at t. Suppose that we have a set of n replications (X_i, Y_i) of (X, Y). Many kernel smoothing techniques consist in estimating $m(x)$ by calculating a weighted local average of the $\phi(Y_i)$'s. This can be written as

$$\sum_{i=1}^{n} \tilde{w}_i(x)\phi(Y_i), \tag{1.1}$$

where $\tilde{w}_i(x)$ is a given weight function describing the degree of smoothing. Special cases of (1.1) include the Nadaraya-Watson (NW) and local linear (LL) estimator. For a review about the statistical properties of these two estimators and many other related topics for independent data, we refer the reader to the book of Fan and Gijbels (1996).

For dependent observations, there is a large literature about the NW estimator under different kinds of associations, like mixing processes and Markovian chains. For more details see Györfi et al. (1989) and Bosq (1998) and the references therein. Masry and Fan (1997) consider estimating the conditional mean for mixing sequences using the LL estimator. They demonstrated the asymptotic normality for both strongly mixing and ρ-mixing processes. For more references about nonparametric regression techniques with dependent data see, for example, the bibliographical notes given in Fan and Yao (2003).

In this paper we consider the problem of nonparametrically estimating $m(x)$, when the data are spatially or temporally correlated, and when in addition the variable of interest is subject to censoring. To the best of our knowledge, this problem has not been studied in the literature before. However, in many practical applications this type of data is encountered. Consider for example economic duration data, in which event times are often correlated, and the observation of the event may be prevented by the occurrence of an earlier competing event (censoring). Observations on duration of unemployment e.g., may be right censored and are typically correlated. Such dependent censored data occur, for example, when study participants belong to clusters (e.g., month of unemployment, job type, neighborhood, school), with members of the same cluster having correlated risk of the event of interest. In all these cases, instead of observing Y (the survival time), we only observe the pair $(Z, \delta) = (\min(Y, C), I(Y \leq C))$, where C is another variable, known as the censoring variable. The available data are supposed to come from an α-mixing process.

Neither the NW nor the LL method can be directly applied with censored data and an adaptation of these techniques is therefore needed. One simple way to do inference in this context is first to transform the data in an unbiased manner, and then to apply the standard techniques to the transformed data as if they were uncensored. A variety of such transformations has been proposed and studied in the literature in the case of i.i.d. data. See for example, Buckley and James (1979), Koul et al. (1981), Doksum and Yandell (1983), Zheng (1984, 1987), Leurgans (1987), Zhou (1992), Srinivasan and Zhou (1994) and Lai et al. (1995). In all those papers, inference was done for a linear regression function with uncorrelated data. Inspired by those works, Fan and Gijbels (1994) proposed a more general transformation and used the LL method on the transformed data to estimate the regression relationship without any assumption made on its form. By using the Leurgans transformation, Singh and Lu (1999) also studied the nonparametric case but with the NW instead of the LL smoother, and in a multivariate context. All the transformations cited above need a prior estimation of $G_x(t)$, i.e. the CDF of C given $X = x$, since the transformed data involve this unknown quantity. Many authors cited above used the somewhat strong assumption that the censoring and the explanatory variables are independent, so that G, i.e. the unconditional DF, can be approximated using the well-known Kaplan-Meier estimator. This condition is reasonable whenever the censoring is not associated to the characteristic of the individuals under study. This is the case for example when censoring is caused by the termination of the study. But in many other situations, this hypothesis is not met. In this paper we will not make this assumption, so censoring is allowed to depend on X.

To do so we need to control the error induced by estimating $G_x(t)$ uniformly in t. In order to bound this error, and motivated by the work of Härdle et al. (1988), we show a general result that can be applied in a large number of applications related to inference with correlated data. In fact, for completely observed data, we provide a uniform rate of convergence of NW type estimators of functionals of an unknown CDF L_x, i.e. $\int \beta_t(y)dL_x(y)$, under strong mixing assumptions. This result is established in the Appendix and it can be read and used independently of the rest of the paper. Using this result we prove then the asymptotic normality and weak consistency of a LL estimator of $m(x)$ based on transformed pre-estimated data from α-mixing censored processes.

Note that this approach can also be used as a tool to study nonparametrically the relationship between future and past values in the presence of censoring. For example one may predict the future values of a censored process $\{Y_t\}_t$ via kernel estimation of $\mathbb{E}\left(Y_{t+1}|Y_t\right)$.

2 Transformation of the data

Let (X_i, Z_i, δ_i), $i = 1, \ldots, n$, be a sample of dependent r.v. each having the same distribution as (X, Z, δ) considered in Section 1. The process (X_t, Y_t, C_t), $t = 0, \pm 1, \ldots, \pm\infty$, has the same distribution as (X, Y, C) and is assumed to be stationary α-mixing (or strong mixing). By this we mean that if $\mathcal{F}_J^L$ denotes the σ-field generated by the family $\{(X_t, Y_t, C_t),\ J \leq t \leq L\}$, then the mixing coefficients

$$\alpha(t) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_t^\infty} |P(A \cap B) - P(A)P(B)|$$

satisfy $\lim_{t\to\infty} \alpha(t) = 0$. For the properties of this and other mixing conditions we refer to Bradley (1986) and Doukhan (1994). Among all the strong mixing conditions available in the literature, α-mixing is the weakest and many time series models are α-mixing under mild conditions. See, for example, Pham and Tran (1985), Bougerol and Picard (1992) and Masry and Tjøstheim (1995).

In this work, the mixing coefficient $\alpha(t)$ is assumed to be $O(t^{-\nu})$ for some $\nu > 3.5$. The random variables Y and C are nonnegative random variables with continuous marginal DFs and they are independent given X. We denote, respectively, by $f_0(x)$, $F_x(t)$ and $G_x(t)$ the marginal density of X, the CDF of Y given X and the CDF of C given X. For a given conditional (sub)distribution function $L_x(t)$ we will use the notation $\bar{L}_x(t)$ for the corresponding survival function, i.e. $\bar{L}_x(t) = 1 - L_x(t)$, and $\dot{L}_x(t)$ for the partial derivative of $L_x(t)$ with respect to x. Define $H_x(t) = P\left(Z \leq t | x\right) = 1 - \bar{F}_x(t)\bar{G}_x(t)$, the CDF of the observed survival times, $H_x^0(t) = P\left(Z \leq t, \delta = 0 | x\right) = \int_0^t \bar{F}_x(s) dG_x(s)$, the sub-CDF of censored observations, and $\mathcal{T}_x = \sup\left\{t : \ H_x(t) < 1\right\}$, the right endpoint of the support of H_x for a given x. Also, let J be the support of X, which is an interval in $\mathbb{R}$ that may be infinite. We say that a real function f is ulL(J) if f is uniformly locally Lipschitz on J, that is,

$$\sup_{x, x' \in J, |x - x'| \leq \epsilon} |f(x) - f(x')| \leq M \left|x - x'\right|,$$

for some $\epsilon > 0$ and $M > 0$.

As we explained in the Introduction, the idea we follow here is to transform the triplet (X, Z, δ) to a new vector (X, Y^*) in such a way that, for a given x,

$$\mathbb{E}(Y^* | X = x) = \mathbb{E}(\phi(Y) | X = x).$$

Once this transformation is found, we estimate $m(x) = \mathbb{E}(\phi(Y)|X = x)$ by applying a local linear smoother to the transformed data (X_i, Y_i^*), that are not subject to censoring unlike the original data. Put

$$Y^* = \delta \varphi_x^1(Z) + (1 - \delta)\varphi_x^2(Z). \tag{2.1}$$

A general transformation is obtained by solving the differential equation

$$\varphi_x^1(t)\bar{G}_x(t) + \int_0^t \varphi_x^2(s) dG_x(s) = \phi(t). \tag{2.2}$$

Let $d_x(t) = \varphi_x^2(t) - \varphi_x^1(t)$. A general class of possible solutions of (2.2) is

$$\begin{cases} \varphi_x^1(t) = \phi(0) + \displaystyle\int_0^t \frac{d\phi(s)}{\bar{G}_x(s)} - \int_0^t \frac{d_x(s)}{\bar{G}_x(s)} dG_x(s) \\ \varphi_x^2(t) = \varphi_x^1(t) + d_x(t). \end{cases}$$

An interesting case is obtained by choosing $d_x(t) = \eta(t) - (1-\alpha)\phi(t)/\bar{G}_x(t)$, where $\alpha \in \mathbb{R}$ and $\eta(t)$ is a real valued function. This leads to

$$\begin{cases} \varphi_x^1(t) = \alpha \left(\phi(0) + \displaystyle\int_0^t \frac{d\phi(s)}{\bar{G}_x(s)} \right) + (1-\alpha)\frac{\phi(t)}{\bar{G}_x(t)} - \int_0^t \frac{\eta(s)}{\bar{G}_x(s)} dG_x(s) \\ \varphi_x^2(t) = \alpha \left(\phi(0) + \displaystyle\int_0^t \frac{d\phi(s)}{\bar{G}_x(s)} \right) + \eta(t) - \int_0^t \frac{\eta(s)}{\bar{G}_x(s)} dG_x(s). \end{cases}$$

So our theoretically transformed data (2.1) can be written as

$$Y^* = (1-\alpha)\zeta^1 + \alpha\zeta^2 + \zeta_\eta, \tag{2.3}$$

with

$$\zeta^1 = \frac{\delta\phi(Z)}{\bar{G}_x(Z)}, \;\; \zeta^2 = \phi(0) + \int_0^Z \frac{d\phi(s)}{\bar{G}_x(s)} \text{ and } \zeta_\eta = (1-\delta)\eta(Z) - \int_0^Z \frac{\eta(s)}{\bar{G}_x(s)} dG_x(s).$$

Remark 2.1
Another way to prove the validity of this transformation is by showing that $\mathbb{E}(\zeta^1|X = x) = \mathbb{E}(\zeta^2|X = x) = m(x)$ and $\mathbb{E}(\zeta_\eta|X = x) = 0$. This means that the transformation (2.3) is a linear combination of the two transformations ζ^1 and ζ^2 adjusted by the factor ζ_η. Both ζ^1 and ζ^2 are largely used in the censored data literature, the first one was originally proposed by Koul et al. (1981) and the second one is due to Leurgans (1987). By allowing the tuning parameter α to range from 0 to 1, we control the balance between these two methods. η is another user chosen parameter (real function). Taking $\eta \equiv 0$, (2.3) becomes the NC (New Class) transformation proposed by Fan and Gijbels (1994). By choosing a non-vanishing function η we hope to improve the quality of our transformation. Ideally, α and η have to be chosen to minimize the variation in the transformed data. However, it is hard to obtain an analytic formula for such an optimal theoretical choice. From practical point of view, a data-driven procedure is needed to make a reasonable choice of these two parameters. This will be discussed in more detail in Sections 5 and 6 .

From now on we only consider classes of functions ϕ that satisfy the following conditions:

Assumption $(\mathcal{H})$.

$(\mathcal{H}1)$ ϕ vanishes outside the interval $[0, \tau_x]$, for some $0 < \tau_x < \mathcal{T}_x$.

$(\mathcal{H}2)$ ϕ is a bounded non-decreasing function on $[0, \tau_x]$.

The function η is also assumed to satisfy these conditions.

Assumption $(\mathcal{H}1)$ is needed to address the identifiability issue due to censoring. It means that instead of estimating, for example, the mean regression function, $\mathbb{E}(Y|X = x)$, we will only estimate the truncated conditional mean $\mathbb{E}\left(YI(Y \leq \tau_x)|X = x\right)$. Condition $(\mathcal{H}2)$ is a technical assumption needed in the proof of Lemma 4.2 below. The 'non-decreasing' assumption is not required whenever, in (2.3), we take $\alpha = 0$ and $\eta \equiv 0$.

3 Estimation of $G_x(t)$

The transformation Y^* given in (2.3) depends on the unknown distribution G_x, which needs to be estimated, before we can apply this transformation in practice. In the independent data case, the problem of estimating G_x has been widely studied in the literature. Beran (1981) proposed to estimate $1 - G_x(t)$ by

$$1 - \hat{G}_x(t) = \prod_{i=1}^{n} \left(1 - \frac{(1-\delta_i)I(Z_i \leq t)\tilde{w}_{0i}(x)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\tilde{w}_{0j}(x)}\right),$$

where

$$\tilde{w}_{0i}(x) = \frac{K_0\left((x - X_i)/h_0\right)}{\sum_{j=1}^{n} K_0\left((x - X_j)/h_0\right)} \tag{3.1}$$

are Nadaraya-Watson (NW) weights, K_0 is a symmetric density function (kernel) with bounded support, say $[-1, 1]$, and with bounded first derivative, and $0 < h_0 \equiv h_{0n} \to 0$ is a bandwidth sequence. Note that this estimator reduces to the Kaplan-Meier estimator when all weights $\tilde{w}_{0i}(x)$ are equal to n^{-1}. Under the i.i.d. assumption, the asymptotic properties of this estimator have been further studied by Dabrowska (1987), González-Manteiga and Cadarso-Suarez (1994), Van Keilegom and Veraverbeke (1997), among others. We show below that in the present setup of strongly mixing processes, the estimator $\hat{G}_x(t)$ remains uniformly consistent.

Theorem 3.1 *Assume (A1) and (A2), given in the Appendix. Let $0 < \tau_x < \mathcal{T}_x$, and suppose $n^{-2\nu+7}h_0^{-3(2\nu+7)}(\log n)^{2\nu-3} = o(1)$.*
(i) If $H_x(t)$ and $H_x^0(t)$ are ulL(J) uniformly in $t \geq 0$ and f_0 is ulL(J), then

$$\sup_{x\in J}\sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)| = O_p(\Delta_n^{-1/2} + h_0).$$

(ii) If $\dot{H}_x(t)$ and $\dot{H}_x^0(t)$ exist and they are ulL(J) uniformly in $t \geq 0$ and f_0' exists and is ulL(J), then

$$\sup_{x\in J}\sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)| = O_p(\Delta_n^{-1/2} + h_0^2),$$

where $\Delta_n = nh_0/\log n$.

Proof. Using similar arguments as in the proof of Proposition 2.2 in Dabrowska (1987), one can easily demonstrate that, uniformly in $x \in J$ and $t \in [0, \tau_x]$,

$$|\hat{G}_x(t) - G_x(t)| = O(1) \left[\sup_{x \in J} \sup_{t \geq 0} |\hat{H}_x(t) - H_x(t)| + \sup_{x \in J} \sup_{t \geq 0} |\hat{H}_x^0(t) - H_x^0(t)| \right],$$

where $\hat{H}_x(t) = \sum_{i=1}^n I(Z_i \leq t)\tilde{w}_{0i}$ and $\hat{H}_x^0(t) = \sum_{i=1}^n (1 - \delta_i) I(Z_i \leq t)\tilde{w}_{0i}$. The result follows as a direct application to $\hat{H}_x(t)$ and $\hat{H}_x^0(t)$ of Theorem 7.4 in the Appendix. □

Note that, under similar assumptions, the same result holds for the Beran estimator of F_x and also for the conditional hazard function estimator.

4 Estimation of $m(x)$

Let us start with the case where G_x is known and denote by $m^G(x)$ the conditional mean of Y^* given $X = x$. The LL estimator of $m^G(x)$ is given by

$$\hat{m}^G(x) = \sum_{i=1}^n \tilde{w}_{1i}(x) Y_i^*, \tag{4.1}$$

where

$$\tilde{w}_{1i}(x) = \frac{K_1\left((x - X_i)/h_1\right)\left[S_{n,2}(x) - (x - X_i)S_{n,1}(x)\right]}{\sum_{j=1}^n K_1\left((x - X_j)/h_1\right)\left[S_{n,2}(x) - (x - X_j)S_{n,1}(x)\right]} \tag{4.2}$$

are local linear (LL) weights, with $S_{n,l}(x) = \sum_{i=1}^n K_1\left((x - X_i)/h_1\right)(x - X_i)^l$, $l = 0, 1, 2$ and where $0 < h_1 \equiv h_{1n} \to 0$ is a bandwidth and K_1 is a kernel, assumed to be a bounded function with bounded support. Unlike the local constant approach, which cannot adapt to unbalanced design situations and which has adverse boundary effects that require boundary correction, LL regression is known to have many good statistical properties that are detailed in the book of Fan and Gijbels (1996).

Since $m(x) = m^G(x)$, (4.1) is also a LL estimator for $m(x)$ based on the transformed data (X_i, Y_i^*), $i = 1, \ldots, n$. Put $u_j = \int u^j K(u)du$ and $v_j = \int v^j K^2(v)dv$ and, suppose that $u_0 = 1$ and $u_1 = 0$. Let $G_x^*(t)$ and $\sigma_*^2(x)$ denote respectively the CDF and the conditional variance of Y^* given $X = x$. The random sequence (X_t, Y_t^*) is strongly mixing with mixing coefficient $\alpha^*(t) \leq \alpha(t)$, see e.g. Eberlein and Taqqu (1986). So by applying Theorem 5 in Masry and Fan (1997) we have the following result.

Lemma 4.1 *Assume (A1), (A2) and condition ($\mathcal{H}$) for ϕ and η. Let $h_1 = C_1 n^{-\gamma_1}$, for some $C_1 > 0$ and $1/5 \leq \gamma_1 < (\nu - 1)/(\nu + 1)$. If $f_0(.)$, $G_.^*(t)$ and $\sigma_*^2(.)$ are continuous on J, then*

$$\sqrt{nh_1}\left(\hat{m}^G(x) - m(x) - u_2 h_1^2 m''(x)/2\right) \xrightarrow{d} \mathcal{N}\left(0, v_0 \sigma_*^2(x)/f_0(x)\right),$$

for each x in J, provided that m'' exists and is continuous on J.

Remark 4.1
Assumption (A2) is weaker than condition 2(ii) required in Masry and Fan (1997). However, using similar techniques as in the proof of Lemma 7.1 in the Appendix, it can be shown that their result continues to hold under this weaker condition. Note also that the finite moment conditions in Masry and Fan (1997) are fulfilled in our case since Y^* is bounded.

We next examine the limiting distribution of the regression estimator based on the estimated transformation. More precisely, we propose to plug-in Beran's estimator $\hat{G}_x$ in the formula of Y^*, see (2.3). We denote by $\hat{Y}^*$ the resulting transformation and by $\hat{m}^{\hat{G}}$ the corresponding LL estimator, i.e.

$$\hat{m}^{\hat{G}}(x) = \sum_{i=1}^{n} \tilde{w}_{1i}(x)\hat{Y}_i^*. \tag{4.3}$$

Note that if in (4.3) we take $\alpha = 0$ and $\eta \equiv 0$ and instead of the LL we use the NW estimator for the regression function (with bandwidth h_0 and kernel K_0), the resulting estimator $\sum_{i=1}^n \tilde{w}_{0i}(x)\hat{Y}_i^*$, can be written as

$$\int \phi(t) d\hat{F}_x(t), \tag{4.4}$$

where $\hat{F}_x$ is the Beran estimator of F_x. This is due to the fact that the jumps of $\hat{F}_x$ at the uncensored points Z_i are exactly $\tilde{w}_{0i}(x)/\hat{\bar{G}}_x(Z_i)$. This means that our estimator (4.3) improves the 'naive' estimator (4.4) from three points of view: (1) The LL weights are used instead of the classical NW weights, (2) A more general transformation is allowed, and (3) The second bandwidth (kernel) used for the regression function does not need to be the same as the first one used for estimating G_x. This last point is especially interesting because as we will see in the simulation section, the best results are typically obtained for $h_1 << h_0$.

To state the asymptotic normality of $\hat{m}^{\hat{G}}(x)$, we first need to bound the error induced by approximating the true DF G_x by its Beran estimator $\hat{G}_x$.

Lemma 4.2 *If the conditions of Theorem 3.1(i) hold, the functions ϕ and η satisfy condition ($\mathcal{H}$), $nh_0/\log n \to \infty$ and $nh_1 \to \infty$, then*

$$\hat{m}^{\hat{G}}(x) - \hat{m}^{G}(x) = O_p\left(\sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)|\right),$$

for all $x \in J$.

Proof. First note that

$$\left|\hat{m}^{\hat{G}}(x) - \hat{m}^{G}(x)\right| \leq \sup_i |\hat{Y}_i^* - Y_i^*| \sum_{i=1}^{n} |\tilde{w}_{1i}(x)|.$$

From the definition of $\tilde{w}_{1i}(x)$, see (4.2), using Theorem 1 in Masry and Fan (1997),

$$\sum_{i=1}^{n} |\tilde{w}_{1i}(x)| \leq \frac{2S_{n,2}(x)S_{n,0}(x)}{S_{n,2}(x)S_{n,0}(x) - S_{n,1}^2(x)} = \frac{2u_2 + o_p(1)}{u_2 + o_p(1)} = O_p(1).$$

On the other hand,

$$|\hat{Y}_i^* - Y_i^*| \leq |\hat{\zeta}_i^1 - \zeta_i^1| + |\hat{\zeta}_i^2 - \zeta_i^2| + |\hat{\zeta}_{\eta i} - \zeta_{\eta i}|. \tag{4.5}$$

We will only show the derivation for the third term on the right hand side of (4.5), since for the two other terms the development is similar.

$$\begin{aligned}|\hat{\zeta}_{\eta i} - \zeta_{\eta i}| &\leq \left| \int_0^{Z_i} \left(\frac{\eta(s)}{\bar{\hat{G}}_x(s)} - \frac{\eta(s)}{\bar{G}_x(s)} \right) d\hat{G}_x(s) \right| + \left| \int_0^{Z_i} \frac{\eta(s)}{\bar{G}_x(s)} d(G_x(s) - \hat{G}_x(s)) \right| \\ &= I_1 + I_2 \qquad \text{(say)}.\end{aligned}$$

Clearly,

$$I_1 \leq \sup_{t\in[0,\tau_x]} |\eta(t)| \sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)| \int_0^{\tau_x} \frac{d\hat{G}_x(t)}{\bar{\hat{G}}_x(t)\bar{G}_x(t)}.$$

By Theorem 3.1(i) we have that $\sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)| = o_p(1)$ and since $\bar{G}_x(t) \geq \bar{G}_x(\tau_x) > 0$, for all $t \in [0, \tau_x]$, it follows that $I_1 = O_p\left(\sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)|\right)$. For I_2, using integration by parts and after some easy algebra, we obtain

$$I_2 \leq 4\frac{\sup_{t\in[0,\tau_x]} |\eta(t)|}{\bar{G}_x^2(\tau_x)} \sup_{t\in[0,\tau_x]} |\hat{G}_x(t) - G_x(t)|,$$

which completes the proof. □

Remark 4.2 (uniform rate)
Let $0 < \tau < \inf\{\mathcal{T}_x : x \in J\}$. If ϕ and η satisfy condition $(\mathcal{H})$ with τ instead of τ_x, then it follows from Theorem 3.1(ii) that $\sup_{x\in J, t\in[0,\tau]} |\hat{G}_x(t) - G_x(t)| = O_p((nh_0/\log n)^{-1/2} + h_0^2)$. Now from Corollary 1 in Masry (1996) we have that $S_{n,j}(x) \to f_0(x)u_j$ uniformly on J. Hence, it follows from the proof of Lemma 4.2 that

$$\sup_{x\in J} |\hat{m}^{\hat{G}}(x) - \hat{m}^G(x)| = O_p((nh_0/\log n)^{-1/2} + h_0^2).$$

Moreover, by Theorem 6 in Masry (1996), $\sup_{x\in J} |\hat{m}^G(x) - m(x)| = O_p((nh_1/\log n)^{-1/2} + h_1^2)$. We conclude that

$$\sup_{x\in J} |\hat{m}^{\hat{G}}(x) - m(x)| = O_p((n\underline{h}/\log n)^{-1/2} + \overline{h}^2),$$

where $\underline{h} = \min(h_0, h_1)$ and $\overline{h} = \max(h_0, h_1)$, whenever the required assumptions are fulfilled.

The following theorem is a direct consequence of Lemma 4.1 and Lemma 4.2.

Theorem 4.1 *Assume the conditions of Lemma 4.1 hold. If $\dot{H}_x(t)$ and $\dot{H}^0_x(t)$ exist and are ulL(J) uniformly in $t \geq 0$, if f_0' exists and is ulL(J), and if* $\log n/(nh_0^5) = O(1)$ *and $n^{-2\nu+7}h_0^{-3(2\nu+7)}(\log n)^{2\nu-3} = o(1)$, then for any x in J,*

$$\sqrt{nh_1}\left(\hat{m}^{\hat{G}}(x) - m(x) - u_2h_1^2m''(x)/2 + O_p(h_0^2)\right) \xrightarrow{d} \mathcal{N}\left(0, v(x)\right), \qquad (4.6)$$

with $v(x) = v_0\sigma_*^2(x)/f_0(x)$.

As a consequence of this theorem, $\hat{m}^{\hat{G}}(x)$ is a consistent estimator for $m(x)$ with the asymptotic bias and variance given respectively by $h_1^2m''(x)u_2/2 + O_p(h_0^2)$ and $v(x)/(nh_1)$. The extra error term $O_p(h_0^2)$ comes from the bias of $\hat{G}_x$. The asymptotic variance is also larger than in the familiar case, since $\sigma_*^2(x) \geq \mathbb{V}ar(\phi(Y)|X = x) \equiv \sigma^2(x)$. Without censoring, Y^* becomes $\phi(Y)$ and so the asymptotic variance reduces to $v_0\sigma^2(x)/f_0(x)$, which is the asymptotic variance for uncensored data. Note also that our assumptions on h_0 and h_1 imply that $nh_0^5 \to \infty$ and $nh_1^5 = O(1)$ which means that $h_1/h_0 = o(1)$. Therefore, the asymptotic bias of $\hat{m}^{\hat{G}}(x)$ is dominated by $O_p(h_0^2)$. By ignoring the bias term, i.e. by assuming that $nh_1h_0^4 = (h_1/h_0)nh_0^5 \to 0$, (4.6) becomes

$$\sqrt{nh_1}\left(\hat{m}^{\hat{G}}(x) - m(x)\right) \xrightarrow{d} \mathcal{N}\left(0, v(x)\right).$$

This result may be used to construct an asymptotic confidence interval for $m(x)$. To do so, $\sigma_*^2(x) = \mathbb{V}ar(Y^*|X = x)$ needs to be estimated. A simple estimator of $\sigma_*^2(x)$ is given by $\sum_{i=1}^n \tilde{w}_{1'i}(x)(\hat{Y}_i^* - \hat{m}^{\hat{G}}(X_i))^2$, where $\tilde{w}_{1'i}(x)$ is given by (4.2) but with another bandwidth $h_1^{'}$ instead of h_1. Using similar arguments as in the proof of Lemma 4.2, it can be easily shown that this estimator is asymptotically equivalent to $\sum_{i=1}^n \tilde{w}_{1'i}(x)(Y_i^* - \hat{m}^G(X_i))^2$, which is the classical LL estimator for the conditional variance for completely observed data. Finally, the results stated above may also be extended to construct a simultaneous confidence band for $m(x)$. In fact, as we have done in Lemma 4.1, using some known results from the literature, see for example Xia (1998), it can easily be shown that, under some regularity conditions, $\sqrt{nh_1}\left(\hat{m}^G(x) - m(x)\right) \xrightarrow{d} Y_n(x, h_1)\sigma_*(x)/\sqrt{f_0(x)}$ uniformly in $x \in [0, \tau]$, where $Y_n(x, h_1) = h_1^{-1/2}\int_0^1 K_1((z-x)h_1^{-1})dW_n(z)$ and $W_n(z)$ is a sequence of standard Wiener processes. Given our Remark 4.2, it is clear that the same result is also available for $\hat{m}^{\hat{G}}(x)$.

5 Numerical study

In this section we present the results of a simulation study, in which the finite sample performance of the proposed method is investigated. Let X_t have a uniform distribution on $[0, 3]$, and let $Y_t = r(X_t) + \sigma(X_t)\epsilon_t$, where $r(x) = 12.5 + 3x - 4x^2 + x^3$, $\sigma(x) = (x - 1.5)^2a_0 + a_1$ and ϵ_t is a standard normal random variable. Also, define $C_t =$

$\tilde{r}(X_t) + \sigma(X_t)\tilde{\epsilon}_t$, with $\tilde{r}(x) = r(x) + \beta(x)\sigma(x)$, $\beta(x) = (x - 1.5)^2 b_0 + b_1$, and $\tilde{\epsilon}_t$ is also standard normal. The variables X_t, ϵ_t and $\tilde{\epsilon}_t$ are mutually independent. The parameters b_0 and b_1 allow to control the percentage of censoring (PC) which is given by $PC(x) = P(Y_t > C_t | X_t = x) = 1 - P(\epsilon_t \leq \beta(x) + \tilde{\epsilon}_t) = 1 - \Phi(\beta(x)/\sqrt{2})$, where Φ is the distribution function of a standard normal random variable. Our objective is to estimate the truncated conditional mean function $m(x) = \int_0^{12.39} t dF_x(t)$. This corresponds to $\phi(t) = tI(t \leq T)$, with $T = 12.39$ which is the 0.98 upper quantile of the DF H_x for $x = 1.5$. Different values of x were investigated but we only show here the results for $x = 1.5$. Four cases are studied :
(1) $b_1 = 0.95$, $b_0 = 0$: PC is constant and is equal to 25%.
(2) $b_1 = 0.95$, $b_0 = -0.27$: PC is convex with minimum, 25%, at $x = 1.5$.
(3) $b_1 = 0$, $b_0 = 0$: PC is constant and is equal to 50%.
(4) $b_1 = 0$, $b_0 = -0.238$: PC is convex with minimum, 50%, at $x = 1.5$.
The parameters a_0 and a_1 allow to control the variation in the generated data. Three values for a_0 are investigated : $a_0 = 0$, $a_0 = -0.25$ and $a_0 = 0.25$. The first one corresponds to a homoscedastic regression model. In the second (third) case, $\sigma(x)$ is concave (convex) with maximum (minimum) at $x = 1.5$. Finally, we chose two values for a_1 : $a_1 = 0.5$ and $a_1 = 1$.

To generate a mixing process X_t with uniform distribution on $[0, 3]$, we first consider an ARMA time series of the form $E_t = \sum_i \delta_i E_{t-i} + \sum_i \gamma_i \upsilon_{t-i} + \upsilon_t$, where the υ_t are i.i.d. $\mathcal{N}(0, 1)$. By an appropriate choice of δ_i's and γ_i's, the resulting E_t is a strongly mixing Gaussian process, with $\alpha(n) \to 0$ at an exponential rate (see Pham and Tran (1985) and Bougerol and Picard (1992)). Then, in order to get an explanatory variable that is α-mixing and has the required distribution, we use the probability integral transform method (see Hoel et al. (1971)). Three situations are considered :

- **Model 1**: X_t is generated from an $AR(1)$, with $\gamma_1 = 0.5$, ϵ_t and $\tilde{\epsilon_t}$ are i.i.d.
- **Model 2**: X_t is generated from an $AR(1)$, with $\gamma_1 = -0.5$, ϵ_t and $\tilde{\epsilon_t}$ are i.i.d.
- **Model 3**: X_t, ϵ_t and $\tilde{\epsilon_t}$ are generated from an $AR(1)$, with γ_1 equal to 0.8, 0.5 and 0.5, respectively.

The mutual independence of X_t, ϵ_t and $\tilde{\epsilon}_t$, implies that $(X_t, \epsilon_t, \tilde{\epsilon_t})$ is a strongly mixing process, and hence this is also the case for the sequences (X_t, Y_t, C_t). The sample size is taken equal to $n = 350$. For all the data analyzed, the Epanechnikov kernel, which is known to have certain optimal properties, $K(x) = (3/4)(1 - x^2)I(-1 \leq x \leq 1)$, is used for both the Beran estimator of G_x and for the LL smoother of $m(x)$. To calculate the transformed data, we first need to choose the tuning parameter α and the 'adjustment' function η. In this study, five values of the parameter α are investigated, $\alpha = 0, 0.25, 0.5, 0.75, 1$, and two functionals η are considered, namely $\eta \equiv 0$ and $\eta = \phi$. For all scenarios, the results using the second choice are considerably better than those obtained with the zero adjustment function. Therefore, we restrict attention here to showing the results for $\eta = \phi$. To evaluate $\hat{m}^{\hat{G}}(x)$ we also need the two bandwidths

h_0 and h_1. In this study the value of h_0 and h_1 ranges from 0.2 to 3 by steps of 0.04. To avoid instability of the transformed data and following the idea of Fan and Gijbels (1994), we do not transform the data points for which $Z_i > T$. For each scenario, the bias, the empirical variance and the mean squared error are calculated over 1500 replications. The results are summarized in Tables 1-4. Each entry in the table represents the result for which the MSE is minimal, obtained over all possible values of h_0, h_1 and α. The tables also show the values of these parameters, for which the best result is obtained.

We first discuss the findings for the bandwidths h_0 and h_1. Almost in all situations the optimal value of h_0 is larger than the optimal value of h_1. This is not surprising, since intuitively, to correctly calculate the LL estimator one may need only a 'small' portion of the transformed data, which already contain some information from the neighborhood. This also confirms the theoretical fact that h_1/h_0 must converge to 0 (see Section 4). Interestingly, regarding the proportion of censoring in the simulated data, h_0 and h_1 behave differently. In fact as the PC increases, h_1 becomes larger but, globally, this is not the case for h_0. The behavior of h_1 can be attributed to the increase in the variation of the transformed data due to censoring. For h_0, remember that this bandwidth is only used to estimate the conditional distribution function G_x of the censoring variables. Estimating G_x becomes easier in the presence of highly censored data, so, in this case, h_0 tends to be smaller. It seems that the bandwidths are also influenced by the heteroscedasticity in the random samples.

The second finding is about the α parameter. Clearly for low censoring rate, the transformation ζ^1 corresponding to $\alpha = 0$ in (2.3), works better for all cases. But once censoring becomes higher, the value of α increases. One also needs a large value of α, at least 0.5, to obtain reasonable results when a high censoring proportion is combined with a large variance (see Table 4). Now, concerning the MSE, in general our method leads to satisfactory results even with heteroscedastic dependent residuals (see Table 3), but the finite sample performance gets worse as the degree of dependency in the data increases. Another factor that clearly acts on the quality of the resulting estimator is the variance of the residuals. Globally, better results are obtained when the variance remains constant ($a_0 = 0$). Also, when increasing the value of a_1, we find that the performance of $\hat{m}^{\hat{G}}$ decreases (compare Table 1 and Table 4). As it can be seen from Tables 1 and 2, the impact of the sign of the autocorrelation parameter γ_1 in the simulations is not clear. However, with high proportion of censoring, it seems that our estimator shows better performance with $\gamma_1 = -0.5$. It is also obvious from the tables that the MSE is mainly due to the variance component of the estimator. Finally, as we said before, we found that the adjustment function $\eta = \phi$ (see (2.3)) has a good effect on the resulting estimator. Actually, when we take $\eta \equiv 0$, the MSE increases for all simulations, typically in the range of 2% to 5%. In this case, we also noted that the optimal value of α is not the same as for the case $\eta = \phi$. In fact, in contrast to the results shown in Table 1-4, the optimal value of α is often larger than 0.5 in that case.

a_0	b_1	b_0	α	h_0	h_1	MSE	Bias	Var
-0.25	0.95	0	0	2.76	0.72	0.0072	0.0155	0.0070
		-0.27	0	2.96	0.80	0.0079	0.0226	0.0074
	0	0	0.25	1.68	0.96	0.0189	0.0118	0.0188
		-0.238	0.25	1.68	1.08	0.0192	0.0154	0.0190
0	0.95	0	0	2.92	0.72	0.0069	0.0120	0.0068
		-0.27	0	2.68	0.80	0.0081	0.0120	0.0080
	0	0	0.25	1.96	1.04	0.0182	0.0117	0.0181
		-0.238	0.25	2.24	1.28	0.0186	0.0121	0.0185
0.25	0.95	0	0	2.96	0.76	0.0070	0.0172	0.0067
		-0.27	0	3	0.84	0.0084	0.0145	0.0082
	0	0	0.25	2.04	1.36	0.0165	0.0106	0.0164
		-0.238	0.25	3	3	0.0289	-0.0900	0.0208

Table 1: *Optimal results for Model* 1 *and for* $a_1 = 0.5$.

a_0	b_1	b_0	α	h_0	h_1	MSE	Bias	Var
-0.25	0.95	0	0	2.76	0.72	0.0076	0.0170	0.0073
		-0.27	0	2.84	0.80	0.0082	0.0232	0.0077
	0	0	0.25	2.88	0.72	0.0119	0.0119	0.0118
		-0.238	0.25	2.76	0.88	0.0117	0.0112	0.0116
0	0.95	0	0	2.72	0.72	0.0074	0.0121	0.0073
		-0.27	0	2.76	0.84	0.0084	0.0250	0.0078
	0	0	0.25	2.76	0.92	0.0117	0.0101	0.0116
		-0.238	0.25	2.80	1.20	0.0126	0.0046	0.0126
0.25	0.95	0	0	2.84	0.76	0.0075	0.0171	0.0072
		-0.27	0	2.24	0.88	0.0089	0.0234	0.0084
	0	0	0.25	3	1.24	0.0123	0.0101	0.0122
		-0.238	0.25	3	3	0.0264	-0.0916	0.0180

Table 2: *Optimal results for Model* 2 *and for* $a_1 = 0.5$.

a_0	b_1	b_0	α	h_0	h_1	MSE	Bias	Var
-0.25	0.95	0	0	2.88	0.76	0.0112	0.0249	0.0106
		-0.27	0	2.84	0.84	0.0123	0.0308	0.0114
	0	0	0.25	1.24	1.28	0.0338	0.0185	0.0335
		-0.238	0.25	1.6	1.16	0.0350	0.0234	0.0345
0	0.95	0	0	2.64	0.76	0.0110	0.0180	0.0107
		-0.27	0	2.84	0.84	0.0126	0.0195	0.0122
	0	0	0.25	1.56	1.24	0.0331	0.0154	0.0329
		-0.238	0.25	2	1.36	0.0367	0.0118	0.0366
0.25	0.95	0	0	2.84	0.80	0.0110	0.0221	0.0105
		-0.27	0	2.56	0.92	0.0130	0.0284	0.0122
	0	0	0.25	2	1.4	0.0322	0.0068	0.0322
		-0.238	0.25	3	3	0.0548	-0.0983	0.0451

Table 3: *Optimal results for Model* 3 *and for* $a_1 = 0.5$.

a_0	b_1	b_0	α	h_0	h_1	MSE	Bias	Var
-0.25	0.95	0	0	2.76	1.04	0.0137	0.0121	0.0136
		-0.27	0	1.4	1.16	0.0201	0.0278	0.0193
	0	0	0.5	1.28	1.84	0.0475	0.0757	0.0418
		-0.238	0.5	1.24	0.24	0.0994	0.0488	0.0970
0	0.95	0	0	2.84	1.08	0.0130	0.0092	0.0129
		-0.27	0	1.40	1.24	0.0177	0.0241	0.0171
	0	0	0.5	1.24	1.12	0.0408	0.0413	0.0391
		-0.238	0.5	1.52	1.04	0.0487	0.0511	0.0461
0.25	0.95	0	0	2.88	1.12	0.0122	0.0057	0.0122
		-0.27	0	1.44	1.4	0.0163	0.0247	0.0157
	0	0	0.5	1.24	1.56	0.0356	0.0441	0.0337
		-0.238	0.5	1.76	1.48	0.0357	0.0439	0.0338

Table 4: *Optimal results for Model* 1 *and for* $a_1 = 1$.

6 Parameters selection

In practice, η, α, h_0 and h_1 need to be chosen in some data driven way, in order to obtain satisfactory results. From our simulation study, it becomes clear that an appropriate value of those parameters is very important since they influence the behavior of the estimator. Especially the choice of the bandwidths h_0 and h_1 requires more attention as those parameters control the amount of smoothing inherent to the process. It is known that undersmoothing leads to a large variance and oversmoothing increases the bias. For this reason several methods (e.g. plug-in, cross-validation, bootstrap, ...) for selecting smoothing parameters, based on the observed data, have been proposed and studied by many researchers. Much effort in this area has been made, assuming the data are independent and completely observed. For dependent but uncensored observations, the results are sparser. See for example Härdle and Vieu (1992), Quintela del Río and Vilar Fernández (1992) and Hall et al. (1995). Because of the technical difficulties encountered when working with censored data, the bandwidth selection problem becomes really problematic in this case. To the best of our knowledge, no optimal rule has been proposed in the literature for this type of data. In this section we will discuss this problem from a practical point of view and we propose some guidelines that might help in selecting a reasonable value for the parameters needed to calculate $\hat{m}^{\hat{G}}$.

For its simplicity and consistency the cross-validation (CV) is one of the most used methods in the literature. It aims at minimizing the mean square of the prediction error which is, in our case, given by $n^{-1}\sum(\hat{m}^{\hat{G}}(X_i) - Y_i)^2$. Let us start by assuming that G_x is known. In this case, one may use the following local 'leave block out' CV criterion :

$$CV(x, h_1) = n_k^{-1} \sum_{j \in J_k} \left(\hat{m}_r^G(X_j) - Y_j^*\right)^2, \tag{6.1}$$

where, for some $0 < k \leq 1$, J_k is the set of the $n_k = \lfloor nk \rfloor$ nearest neighbor points to x and $\hat{m}_r^G(X_j)$ is the LL estimator of m at X_j without the observations (X_i, Y_i^*), $i = 1, \ldots, n$, for which $|i - j| \leq r$, i.e.

$$\hat{m}_r^G(X_j) = m_j^{-1} \sum_{|i-j|>r} \tilde{w}_{1i}(X_j) Y_i^*, \tag{6.2}$$

where $m_j = \#\{i = 1, \ldots, n : \ |i - j| > r\}$ and r is a given integer satisfying $2r + 1 << n$. By leaving out more than one observation ($r > 0$), we attempt to drop from the sample all the data points that are close in 'time' to (X_j, Y_j^*). In other words, we omit the observations that are 'susceptible' to be highly correlated with (X_j, Y_j^*). The local modification of the CV method allows the adaptation to the concentration of the data, the variation of the noise level and the local behavior of the underlying regression function. Of course, the function (6.1) can be used only if G_x is known, which is not the case in real data analysis. However, when the censoring variable is independent of the covariate, $G_x(t) = P(C \leq t|X = x) = P(C \leq t)$ can be estimated by the

Kaplan-Meier estimator, and so (6.1) can still be used by pluging-in this estimator in (6.1). When C and X are correlated, an easy solution would be to select h_0 and h_1 by simultaneously minimizing

$$CV(x, h_0, h_1) = n_k^{-1} \sum_{j \in J_k} \left(\hat{m}_r^{\hat{G}}(X_j) - \hat{Y}_j^* \right)^2, \tag{6.3}$$

where $\hat{m}_r^{\hat{G}}(X_j)$ is like (6.2) but with $\hat{Y}_i^*$ instead of Y_i^*. We have checked this method via a simulation study, and the obtained results were globally unsatisfactory. For this reason we propose a modification of this approach. The idea behind our proposal is the following. We know that $\mathbb{E}(\zeta^1|X = x) - \mathbb{E}(\zeta^2|X = x) = \mathbb{E}(\zeta_\eta|X = x) = 0$, so for a good choice of h_0 and h_1, we should get a small value for both $|n^{-1} \sum_{i=1}^n \tilde{w}_{1i}(x)(\hat{\zeta}_i^1 - \hat{\zeta}_i^2)| = \Delta_{1n}(x, h_0, h_1)$ and $|n^{-1} \sum_{i=1}^n \tilde{w}_{1i}(x)\hat{\zeta}_{\eta i}| = \Delta_{2n}(x, h_0, h_1)$. This suggests to adjust (6.3) by including $\Delta_{1n}(x, h_0, h_1)$ and $\Delta_{2n}(x, h_0, h_1)$ in the calculation procedure. A simple way to do this is via the following calibrated CV criterion :

$$CCV(x, h_0, h_1) = \sqrt{CV(x, h_0, h_1)} + \Delta_{1n}^{(1-s)}(x, h_0, h_1)\Delta_{2n}^s(x, h_0, h_1), \tag{6.4}$$

with typically $s = 0$, $s = 1$ or $s = 1/2$. This is just one of many possible corrections that we have tested, the other ones like $\sqrt{CV} + \Delta_{1n} + \Delta_{2n}$, do not seem to work as good as (6.4). One can also plan to include α and η in this selection procedure, but doing so will make the computation somewhat complicated and may also increase the instability of the proposed CV function. Given our conclusions in the previous section we decide to run this procedure with $\alpha = 0$ for small proportions of censoring, say less than 50%, and with $\alpha = 0.5$ for large proportions of censoring. For the function η, we restrict our analysis to the case $\eta = \phi$. Due to the amount of calculations required by this CV procedure, we only run 500 Monte Carlo simulations with data of size $n = 350$ generated according to Model 1 with both $a_1 = 0.5$ and $a_1 = 1$. For each simulated data set, using $k = 0.25$ and $r = 2$, we select the pair $(h_0, h_1) \in \{0.2, 0.24, \ldots, 3\} \times \{0.2, 0.24, \ldots, 3\}$ that minimizes (6.4). Table 5 shows the mean of the squared error obtained over the 500 replications using $s = 0$ and $s = 1$ for low and high censoring respectively. Comparing these results with those of Table 1 and Table 4, we observe that globally this approach leads to reasonable results. In some cases, especially with small PC and small variance, the MSE that we obtain using our automatic bandwidth selection criterion is better than the coresponding MSE evaluated with the optimal fixed bandwidths. However, as censoring and/or variance increase, the results become worse.

	b_1	0.95		0	
	b_0	0	-0.27	0	-0.238
$a_1 = 0.5$					
a_0	-0.25	0.0051	0.0047	0.0498	0.0512
	0	0.0049	0.0047	0.0490	0.0504
	0.25	0.0046	0.0053	0.0480	0.0499
$a_1 = 1$					
a_0	-0.25	0.0863	0.0419	0.0475	0.0745
	0	0.0472	0.0242	0.0491	0.0795
	0.25	0.0265	0.0152	0.0573	0.0733

Table 5: *MSE under Model 1, obtained using CCV to select h_0 and h_1.*

7 Appendix

In this appendix, we establish a uniform consistency rate for a kernel type estimator of a conditional functional. The results that are shown here are of general interest and can be used for many other estimation problems associated with strong mixing conditions.

Let (X_i, Y_i) be a strictly stationary α-mixing process, having the same distribution as the random vector (X, Y), with mixing coefficient $\alpha(i) \leq \phi i^{-\nu}$ $(i \to \infty)$ for some $\nu > 2$ and $\phi > 0$, joint density $f(x, y)$, marginal DF F_0 and marginal density f_0 for X. The support of X is denoted by J and is supposed to be an interval in $\mathbb{R}$ that may be infinite. We require that $\{X_i\}$ satisfies :

Assumption (A).

(A1) $0 < m_1 \leq f_0(x) \leq M_1 < \infty$, for all $x \in J$.

(A2) $f_{0j}(u, v) \leq M_* < \infty$ for each $j \geq j_*$ and $u, v \in J :\ |u - v| \leq \varepsilon$ for some $\varepsilon > 0$ and $j_* \geq 1$, where f_{0j} denotes the joint density function of (X_1, X_{j+1}).

(A3) $f_0(.)$ is uniformly locally Lipschitz (ulL) on J; i.e., for some $\delta > 0$ and $M < \infty$,

$$\sup_{x,x' \in J, |x-x'| \leq \delta} |f_0(x) - f_0(x')| \leq M |x - x'| .$$

Let $I \subset \mathbb{R}$ and denote by $\{\beta_t,\ t \in I\}$ a family of real-valued measurable functions and let $r_t(x)$ be the conditional expectation of $\beta_t(Y)$ given that $X = x$. We denote by r_{tn} the NW estimator of $r_t(x)$, that is

$$r_{tn}(x) = \frac{n^{-1} \sum_{i=1}^{n} \beta_t(Y_i) K_h(x - X_i)}{n^{-1} \sum_{i=1}^{n} K_h(x - X_i)} \equiv \frac{d_{tn}(x)}{f_n(x)}, \qquad (7.1)$$

where $K_h(.) = 1/hK(./h)$, K is a symmetric density that has a bounded support with a bounded first derivative and $0 < h \equiv h_n \to 0$.
We will show, in Theorem 7.3 below, that if assumption (A) holds, and if $\{\beta_t,\ t \in I\}$ and h_n satisfy certain regularity conditions, then

$$\sup_{t\in I, x\in J} |r_{tn}(x) - r_t(x)| = O_p\left(\sqrt{\frac{\log n}{nh}} + h\right)$$

This is the main result of this Appendix. It is a generalization of the result of Härdle et al. (1988) to the dependent case. Our proofs follow the same methodology. We therefore omit certain derivations and refer to their paper for more details.

In a first step we develop a general result for a class of functions $\{\gamma_t,\ t \in I\}$, that satisfies the following assumptions. Later on we will take β_t equal to a linear combination of these γ_t-functions.

Assumption (B).

(B1) $\sup_{t\in I, x\in J} \int \gamma_t^2(y) f(x,y)dy = M_0^* < \infty$.

(B2) $0 \le \gamma_t(y) \le \gamma_{t'}(y), \quad t < t' \in I,\ y \in \mathbb{R}$.

(B3) $D_t(.) := \int \gamma_t(y) f(., y)dy$ is ulL on J, uniformly in $t \in I$; i.e., for some $\delta > 0$ and $M < \infty$,

$$\sup_{t\in I} \sup_{x,x'\in J, |x-x'|\le\delta} |D_t(x) - D_t(x')| \le M\,|x - x'|\,.$$

(B4) $\mathbb{E}\gamma_t(Y)$ is a continuous function of t in I.

(B5) The limit functions $\gamma_{t_*} = \lim_{t\to t_*} \gamma_t$ and $\gamma_{t^*} = \lim_{t\to t^*} \gamma_t$ exist and are finite a.s. (w.r.t. the DF of Y), where $t_* = \inf I$ and $t^* = \sup I$.

(B6) $\|\gamma_{t^*}(Y)\|_\lambda = M_\lambda < \infty$, for some $2(\nu - 1)/(\nu - 2) < \lambda \le \infty$, where $\|\cdot\|_\lambda$ is the L_λ-norm.

(B7) $\exists\varepsilon > 0$ and $j_* \ge 1$ such that for each $j \ge j_*$,
$\sup_{u,v\in J, |u-v|\le\varepsilon} \int \gamma_{t^*}(u')\gamma_{t^*}(v') f_j(u, v, u', v')du'dv' \le M_* < \infty$,
where f_j denotes the joint density function of $(X_1, X_{j+1}, Y_1, Y_{j+1})$.

Remark 7.1
Assumptions (B1)-(B6) correspond, respectively, to assumptions (B2), (A3)-(A6) and (A7), with $\lambda > 2$, in Härdle et al. (1988). In the case $\lambda = \infty$, i.e., γ_t is bounded, assumption (B7) reduces to the assumption (A2) given above. In the case that $\gamma_t \equiv 1$, i.e., $D_t(x) = f_0(x)$, assumption (B3) reduces to the assumption (A3) given above. Assumptions (A1) and (B1) imply that

$$\sup_{t\in I, x\in J} \int \gamma_t^2(y) f(y|x)dy \le M_0^*/m_1 \equiv M_0, \tag{7.2}$$

where $f(y|x)$ is the conditional density of Y given X.

Let $D_{tn}(x) = c_n^{-1}\left[G_{tn}(x + c_n') - G_{tn}(x - c_n'')\right]$, c_n' and c_n'' are positive sequences tending to 0, $c_n = c_n' + c_n''$ and $G_{tn}(x) = n^{-1}\sum_{i=1}^{n} \gamma_t(Y_i) I(X_i \le x)$.

Theorem 7.1 *Assume (A1) and (B). Let c_n satisfy (i) $0 < c_n \to 0$ and (ii) $n^{(4\nu-2)\lambda^{-1}-(2\nu-7)} c_n^{2\lambda^{-1}-(2\nu+7)} (\log n)^{-2(2\nu+1)\lambda^{-1}+(2\nu-3)} \to 0$. Then,*

$$\sup_{t\in I, x\in J} |D_{tn}(x) - D_t(x)| = O_p(\Delta_n^{-1/2} + c_n),$$

with $\Delta_n = nc_n/\log n$.
Further, in the case $\lambda = \infty$, for each $\mathcal{B} > 0$ there exists a constant $\mathcal{C}_{\mathcal{B}} > 0$ such that, for n sufficiently large,

$$P\left(\sup_{t\in I, x\in J} |D_{tn}(x) - D_t(x)| \ge \mathcal{B}\Delta_n^{-1/2} + Mc_n\right) \le \mathcal{C}_{\mathcal{B}}\left[n^{-2\nu+7} c_n^{-2\nu-7} (\log n)^{2\nu-3}\right]^{1/4}.$$

Remark 7.2
- Condition (ii) implies that (iii) $c_n\,(n/\log n)^{1-2/\lambda} \to \infty$, which itself implies that (iv) $\Delta_n \to \infty$ and (v) $c_n^{-1} < (n/\log n)^{1-2/\lambda}$ $(n \to \infty)$.
- If instead of (B3) we assume: **(B3')** $\dfrac{\partial D_t(x)}{\partial x}$ *exists and is ulL on J uniformly in* $t \in I$, then by taking $c_n' = c_n''$, we get, instead of Theorem 7.1,

$$\sup_{t\in I, x\in J} |D_{tn}(x) - D_t(x)| = O_p(\Delta_n^{-1/2} + c_n^2).$$

- In the case that $\gamma_t \equiv 1$, assumption (B3') reduces to the assumption **(A3')** $f_0'(x)$ *exists and is ulL on J.*

The proof of Theorem 7.1 can be split in two parts. In the first part one can check

$$\sup_{t\in I, x\in J} |\mathbb{E}D_{tn}(x) - D_t(x)| \le Mc_n, \tag{7.3a}$$

by using (B3), Taylor's Theorem and the fact that $\mathbb{E}G_{tn}(x) = \int_{-\infty}^{x} D_t(z)dz := G_t(x)$ and $G_t'(x) = D_t(x)$. The second step is to show that

$$\sup_{t\in I, x\in J} |D_{tn}(x) - \mathbb{E}D_{tn}(x)| = O_p(\Delta_n^{-1/2}). \tag{7.3b}$$

The crucial ingredient to prove this is Lemma 7.1 below.

Put $a_n = \Delta_n^{-1/2} c_n$, $Q_n = M_\lambda a_n^{-1/(\lambda-1)}$, $w_n = \left\lfloor 2Q_n\Delta_n^{1/2} + 1\right\rfloor$. For a given $t \in I$, $v \in [0,1]$, $r = -w_n, -w_n+1, \ldots, w_n$ and $j = 1, \ldots, n$, define

$$Z_{trj}(v) = \gamma_t(Y_j) I\left(\gamma_t(Y_j) \le Q_n\right)\left[I\left(F_0(X_j) \le \eta_{rv}\right) - I\left(F_0(X_j) \le v\right)\right],$$

with $\eta_{rv} = v + rM_1c_n/w_n$. Let

$$\tilde{Z}_{trj}(v) = Z_{trj}(v) - \mathbb{E}Z_{trj}(v) \text{ and } \xi_{trn}(v) = \left| n^{-1}\sum_{j=1}^{n} \tilde{Z}_{trj}(v)\right|.$$

When no confusion is possible, we will write Z_j, $\tilde{Z}_j$ and ξ_n instead of $Z_{trj}(v)$, $\tilde{Z}_{trj}(v)$ and $\xi_{trn}(v)$, respectively.

Lemma 7.1 *Assume (A1), (B1), (B7). Let c_n satisfy (i) and (iii) from Theorem 7.1 and Remark 7.2. For each $\mathcal{B} > 0$ there exists a constant $\mathcal{C}_{\mathcal{B}} > 0$ such that, for n sufficiently large,*

$$P\left(\xi_{trn}(v) \geq \mathcal{B}a_n\right) \leq \mathcal{C}_{\mathcal{B}} n a_n^{-\frac{\lambda}{2(\lambda-1)}} q_n^{\nu+1}, \tag{7.4}$$

with $q_n = \Delta_n^{-1/2} a_n^{-\frac{1}{\lambda-1}}$.

Proof. First note that $|\tilde{Z}_j| \leq 2Q_n$ and $\mathbb{E}\tilde{Z}_j = 0$. Using (7.2),

$$\mathbb{V}ar\tilde{Z}_j \leq \mathbb{E}Z_j^2 \leq M_0 M_1 c_n. \tag{7.5a}$$

Put $C_j = \mathbb{C}ov\left(\tilde{Z}_1, \tilde{Z}_{j+1}\right)$. From (7.5a) and using the Cauchy-Schwartz inequality, we obtain for $1 \leq j \leq n$,

$$|C_j| \leq M_0 M_1 c_n. \tag{7.5b}$$

Also, note that $|C_j| \leq \mathbb{E}|Z_1 Z_{j+1}| + (\mathbb{E}|Z_1|)^2$. For a positive r, we have by (B7),

$$\begin{aligned}\mathbb{E}|Z_1 Z_{j+1}| &\leq M_* \left[\int I\left(v < F_0(x) < \eta_{rv}\right) dx\right]^2 \\ &\leq M_*(M_1/m_1)^2 c_n^2 \quad \text{for } j \geq j_*,\end{aligned}$$

where, in the last inequality, we have used (A1) and applied the mean-value theorem. On the other hand, by (7.2),

$$\begin{aligned}\mathbb{E}|Z_1| &\leq \int \mathbb{E}\left(\gamma_t(Y)|X = x\right) I\left(v < F_0(x) < \eta_{rv}\right) f_0(x)dx \\ &\leq M_0^{1/2} P\left(v < F_0(X) < \eta_{rv}\right) \\ &\leq M_0^{1/2} M_1 c_n.\end{aligned}$$

So we have shown that

$$|C_j| \leq M_1^2(M_* m_1^{-2} + M_0)c_n^2 \quad \text{for } j \geq j_*, \text{ and } r = 0, \ldots, w_n. \tag{7.5c}$$

The same inequality remains true for $r = -w_n, \ldots, 0$. Now, by Billingsley's inequality, see e.g. Corollary 1.1 in Bosq (1998),

$$|C_j| \leq 4\phi Q_n^2 j^{-\nu} \quad (j \to \infty). \tag{7.5d}$$

Let $0 < k_n \to \infty$. From (7.5) it follows that, for each $m > 1$ and for n sufficiently large,

$$\begin{aligned}\sigma_m^2 := \mathbb{V}ar\left(\sum_{j=1}^m \tilde{Z}_j\right) &= m\mathbb{V}ar\tilde{Z}_1 + 2m\sum_{j=1}^m (1 - j/m)C_j \\ &\leq mM_0M_1c_n + 2m\left(\sum_{j=1}^{j_*}|C_j| + \sum_{j=j_*+1}^{k_n}|C_j| + \sum_{j\geq k_n+1}|C_j|\right) \\ &\leq mM_0M_1c_n + 2m\left(j_*M_0M_1c_n + M_1^2(M_*m_1^{-2} + M_0)k_nc_n^2 + \frac{4\phi}{\nu-1}Q_n^2k_n^{1-\nu}\right),\end{aligned}$$

where, in the last inequality, we have used the fact that $\sum_{j=k_n+1}^{\infty} j^{-\nu} \leq k_n^{1-\nu}/(\nu-1)$. Taking $k_n = \lfloor c_n^{-1} \rfloor$ yields

$$\sigma_m^2 \leq m \left(M_0 M_1 + 2 j_* M_0 M_1 + 2 M_1^2 (M_* m_1^{-2} + M_0) + \frac{8\phi}{\nu - 1} Q_n^2 c_n^{\nu-2} \right) c_n.$$

Using (v) and the fact that $\nu > 2$ and $\lambda > 2(\nu-1)/(\nu-2)$, one can check that $Q_n^2 c_n^{\nu-2} \leq M_\lambda^2$ $(n \to \infty)$. This shows that, for n sufficiently large,

$$\sigma_m^2 \leq C_\lambda m c_n \quad \text{for } m > 1, \tag{7.6}$$

with $C_\lambda = M_0 M_1 + 2 j_* M_0 M_1 + 2 M_1^2 (M_* m_1^{-2} + M_0) + \frac{8\phi}{\nu-1} M_\lambda^2$. By applying Theorem 1.3 in Bosq (1998), we have that for each $\epsilon > 0$ and $0 < q \leq 1$,

$$P(\xi_{trn}(v) \geq \epsilon) \leq 4 \exp \left(- \frac{\epsilon^2 n}{32 q \sigma_{\lfloor q^{-1} \rfloor}^2 + 16 Q_n q^{-1} \epsilon} \right) + 11 \left(1 + \frac{8 Q_n}{\epsilon} \right)^{1/2} n q \alpha \left(\lfloor q^{-1} \rfloor \right). \tag{7.7}$$

Taking $\epsilon = \mathcal{B} a_n$ $(\mathcal{B} > 0)$ and $q_n = \Delta_n^{-1/2} a_n^{-\frac{1}{\lambda - 1}}$, using (7.6) after some development, (7.7) can be written as (7.4).□

Note that in Theorem 7.1, assumptions (B2) and (B4)-(B6) are required to show that the supremum over I and J in (7.3b) can be reduced to $\max_{t \in I_n} \max_{v \in \tilde{J}_n} \max_{|r| \leq w_n} \xi_{trn}(v)$, where I_n and $\tilde{J}_n$ are finite sets corresponding to some suitable partitions of I and $[0,1]$, respectively. For a detailed justification, see pages 1444-1447 in Härdle et al. (1988). Once this is done, some easy elaborations lead to the result in Theorem 7.1.

Theorem 7.1 has many applications. Here we will restrict ourselves to the case where β_t may be written as

$$\beta_t(y) = \sum_{i=1}^{i_0} q_i \gamma_{ti}(y), \quad y \in \mathbb{R}, t \in I, \tag{7.8}$$

with fixed and finite i_0,$q_1, \ldots, q_{i_0}$ and with $\{\gamma_{ti},\ t \in I, 1 \leq i \leq i_0\}$ satisfying assumptions (B1)-(B7), with common $\lambda = \infty$ in (B6).

Theorem 7.2 *Let $d_t(x) = \int \beta_t(y) f(x,y) dy$ with $\{\beta_t,\ t \in I\}$ having representation (7.8). Let $d_{tn}(x)$ be defined by (7.1). Assume (A1) and (A2). If $n^{-2\nu+7} (\log n)^{2\nu-3}$ $h_n^{-3(2\nu+7)} \to 0$, then*

$$\sup_{t \in I, x \in J} |d_{tn}(x) - d_t(x)| = O_p(\Delta_n^{-1/2} + h_n),$$

with $\Delta_n = n h_n / \log n$.

This theorem follows from Theorem 7.1 by using similar techniques as in the proof of Theorem 2.3 in Härdle et al. (1988). The key idea here is to first consider only a discrete kernel, say K_n, for which Theorem 7.1 can directly be applied. The second step is to extend the result to the smooth kernel K by uniformly controling the remaining term $|d_{tn}(x, K_n) - d_{tn}(x, K)|$. To do so the bandwidth h_n needs to be chosen carefully. This explains in part the 3 that appears in $h_n^{-3(2\nu+7)}$ above.

In the case where $\beta_t \equiv 1$, d_t and d_{tn} become f_0 and f_n, respectively, and we have :

Corollary 7.1 *Let $f_n(x)$ be defined by (7.1). Assume (A). If $n^{-2\nu+7}(\log n)^{2\nu-3}$ $h_n^{-3(2\nu+7)} \to 0$, then*

$$\sup_{x\in J} |f_n(x) - f_0(x)| = O_p(\Delta_n^{-1/2} + h_n).$$

As a direct consequence of Corollary 7.1 and Theorem 7.2 we obtain our main theorem given below.

Theorem 7.3 *Let $r_t(x) = \mathbb{E}(\beta_t(Y)|X = x)$ with $\{\beta_t,\ t \in I\}$ having representation (7.8). Let $r_{tn}(x)$ be defined by (7.1). Assume (A). If $n^{-2\nu+7}(\log n)^{2\nu-3}$ $h_n^{-3(2\nu+7)} \to 0$, then*

$$\sup_{t\in I, x\in J} |r_{tn}(x) - r_t(x)| = O_p(\Delta_n^{-1/2} + h_n).$$

From Remark 7.2 it is clear that in the results given above, h_n may be replaced by h_n^2 if we substitute (A3) and (B3) by (A3') and (B3') respectively. As an application of this theorem, let us take $\beta_t(y) = I(y \le t)$. In this case $r_t(x)$ becomes the CDF of Y given X, i.e. $F(t|x)$, and $r_{tn}(x)$ becomes the NW estimator of $F(t|x)$ that we shall denote by $F_n(t|x)$.

Theorem 7.4 *Suppose that the marginal distribution function of Y is continuous and that $n^{-2\nu+7}(\log n)^{2\nu-3}h_n^{-3(2\nu+7)} \to 0$. Assume (A1) and (A2).*
(i) If (A3) holds, and $F(t|.)$ is ulL on J uniformly in $t \in \mathbb{R}$, then

$$\sup_{t\in\mathbb{R}, x\in J} |F_n(t|x) - F(t|x)| = O_p(\Delta_n^{-1/2} + h_n).$$

(ii) If (A3') holds, $\partial F(t|x)/\partial x := \dot{F}(t|x)$ exists and $\dot{F}(t|.)$ is ulL on J uniformly in $t \in \mathbb{R}$, then

$$\sup_{t\in\mathbb{R}, x\in J} |F_n(t|x) - F(t|x)| = O_p(\Delta_n^{-1/2} + h_n^2).$$

References

Beran, R. (1981). *Nonparametric regression with randomly censored survival data.* Technical report, Dept. Statist., Univ. California, Berkeley.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes.* Springer, New York.

Bougerol, P. and N. Picard (1992). Strict stationarity of generalized autoregressive processes. *Ann. Probab. 20*, 1714–1730.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics: A Survey of Recent Results*, pp. 165–192. Birkhuser, Boston (eds. E. Eberlein and M. S. Taqqu).

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

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

Doksum, K. A. and B. S. Yandell (1983). Properties of regression estimates based on censored survival data. In : *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pp. 140–156.

Doukhan, P. (1994). *Mixing: Properties and Examples.* Lecture Notes in Statistics. Springer, New York.

Eberlein, E. and M. S. Taqqu (Eds.) (1986). *Dependence in Probability and Statistics*, Volume 11 of *Progress in Probability and Statistics.* Birkhäuser Boston Inc. A survey of recent results, Papers from the conference held at the Mathematical Research Institute Oberwolfach, Oberwolfach, April 1985.

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

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*, Volume 66 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series.* Springer, New York.

González-Manteiga, W. and C. Cadarso-Suarez (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *J. Nonpar. Statist. 4*, 65–78.

Györfi, L., W. Härdle, P. Sarda, and P. Vieu (1989). *Nonparametric Curve Estimation from Time Series*, Volume 60 of *Lecture Notes in Statistics.* Springer, New York.

Hall, P., S. N. Lahiri, and Y. K. Truong (1995). On bandwidth choice for density estimation with dependent data. *Ann. Statist. 23*, 2241–2263.

Härdle, W., P. Janssen, and R. Serfling (1988). Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist. 16*, 1428–1449.

Härdle, W. and P. Vieu (1992). Kernel regression smoothing of time series. *J. Time Ser. Anal. 13*, 209–232.

Hoel, P., S. Port, and C. Stone (1971). *Introduction to Probability Theory.* Boston, MA: Houghton Mifflin.

Koul, H., V. Susarla, and J. Van Ryzin (1981). Regression analysis with randomly right-censored data. *Ann. Statist. 9*, 1276–1288.

Lai, T. L., Z. Ying, and Z. K. Zheng (1995). Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *J. Multiv. Anal. 52*, 259–279.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika 74*, 301–309.

Masry, E. (1996). Multivariate local polynomial regression estimation for time series: Uniform strong consistency and rates. *J. Time Ser. Anal. 17*, 571–599.

Masry, E. and J. Fan (1997). Local polynomial estimation of regression functions for mixing processes. *Scand. J. Statist. 24*, 165–179.

Masry, E. and D. Tjøstheim (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory 11*, 258–289.

Pham, T. D. and L. T. Tran (1985). Some mixing properties of time series models. *Stoch. Process. Appl. 19*, 297–303.

Quintela del Río, A. and J. M. Vilar Fernández (1992). A local cross-validation algorithm for dependent data. *Test 1*, 123–153.

Singh, R. S. and X. Lu (1999). Nonparametric synthetic data regression estimation for censored survival data. *J. Nonpar. Statist. 11*, 13–31.

Srinivasan, C. and M. Zhou (1994). Linear regression with censoring. *J. Multiv. Anal. 49*, 179–201.

Van Keilegom, I. and N. Veraverbeke (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math. 49*, 467–491.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. Roy. Statist. Soc.-Ser. B 60*, 797–811.

Zheng, Z. (1984). *Regression Analysis with Censored Data.* Ph.D. dissertation, Columbia University.

Zheng, Z. (1987). A class of estimators for the parameters in linear regression with censored data. *Acta Mathematicae Applicatae Sinica 3*, 231–241.

Zhou, M. (1992). Asymptotic normality of the "synthetic data" regression estimator for censored survival data. *Ann. Statist. 20*, 1002–1021.