

Smoothed quantile regression processes for binary response models.

Stanislav Volgushev

*Ruhr-Universität Bochum and University of Illinois at Urbana Champaign**

February 25, 2013

Abstract

In this paper, we consider binary response models with linear quantile restrictions. Considerably generalizing previous research on this topic, our analysis focuses on an infinite collection of quantile estimators. We derive a uniform linearization for the properly standardized empirical quantile *process* and discover some surprising differences with the setting of continuously observed responses. Moreover, we show that considering quantile processes provides an effective way of estimating binary choice probabilities without restrictive assumptions on the form of the link function, heteroskedasticity or the need for high dimensional non-parametric smoothing necessary for approaches available so far. A uniform linear representation and results on asymptotic normality are provided, and the connection to rearrangements is discussed.

1 Introduction

In various situations in daily life, individuals are faced with making a decision that can be described by a binary variable. Examples relevant to various fields of economics include the decision to participate in the labour market, to retire, to make a major purchase. From an econometric point of view, such decisions can be modelled by a binary response variable $Y = I\{Y^* \geq 0\}$ that depends on an unobserved continuous random variable Y^* which summarizes an individual's preferences. In the presence of covariates, say W , a natural question is: what can we infer about the distribution of the unobserved variable Y^* conditional on W from observations of i.i.d. replicates of (Y, W) . In a seminal paper, Manski (1975) assumed that $Y^* = W^T \beta + \varepsilon$ where the 'error' ε satisfies the conditional median restriction $P(\varepsilon \leq 0 | W = w) = 0.5$ and derived conditions on the distribution of (ε, W) that imply identifiability of the coefficient vector β up to scale. In later work, Manski (1985) extended those results to general quantile restrictions of the form $P(\varepsilon \leq 0 | W = w) = \tau$ for fixed $\tau \in (0, 1)$. A more detailed discussion of identification issues was provided in Manski (1988). Due to their importance in understanding binary decisions, binary choice models have ever since

*The idea of considering binary response quantile processes originated from discussions with Prof. Roger Koenker. I am thankful to him for the encouragement and many insightful discussions on this topic. The mistakes are of course my sole responsibility. This research was conducted while I was a visiting scholar at UIUC. I am very grateful to the Statistics and Economics departments for their hospitality. Financial support from the DFG (grant VO1799/1-1) is gratefully acknowledged.

aroused a lot of interest and many estimation procedures have been proposed [see Cosslett (1983), Horowitz (1992), Powell et al. (1989), Ichimura (1993), Klein and Spady (1993), Coppejans (2001), Kordas (2006) and Khan (2013) to name just a few].

A particularly challenging part of analysing binary response models lies in understanding the stochastic properties of corresponding estimation procedures. The asymptotic distribution of Manski's estimator was derived in Kim and Pollard (1990) under fairly general conditions, while a non-standard case was considered in Portnoy (1998). In particular, Kim and Pollard (1990) demonstrated that the convergence rate is $n^{-1/3}$ and that the limiting distribution is non-Gaussian. A different approach based on non-parametric smoothing that avoids some of the difficulties encountered by Manski's estimator was taken by Horowitz (1992). By smoothing the objective function, Horowitz (1992) obtained both better rates of convergence and a normal limiting distribution. However, note that the smoothness conditions on the underlying model are stronger than those of Kim and Pollard (1990).

The approaches of Manski and Horowitz have in common that only estimators for the coefficient vector β are provided. While those coefficients are of interest and can provide valuable structural information, their interpretation can be quite difficult since the scale of β is not identifiable from the observations. On the other hand, the 'binary choice probabilities' $p_w := P(Y = 1|W = w)$ provide a much simpler and more straightforward interpretation.

Most of the available methods for estimating binary choice probabilities are of two basic types. The first and more thoroughly studied approach is to assume a model of the form $Y_i^* = W_i^T \beta + \varepsilon_i$ where the ε_i are assumed to be either independent of X_i [see Cosslett (1983) and Coppejans (2001)], or admit a very special kind of heteroskedasticity [Klein and Spady (1993)]. Another popular approach has been to embed the problem into general estimation of single index models, see for example Powell et al. (1989) or Ichimura (1993). Here, it is again necessary to assume independence between ε and the covariate W .

While in the settings described above it is possible to obtain parametric rates of convergence for the coefficient vector β and also construct estimators for choice probabilities, in many cases the assumptions on the underlying model structure seem too restrictive.

An alternative approach allowing for general forms of heteroskedasticity was recently investigated by Khan (2013), who proved that under general smoothness conditions any binary response model with $Y_i^* = W_i^T \beta + \varepsilon_i$ is observationally equivalent to a Probit/Logit model with multiplicative heteroskedasticity, that is a model where $\varepsilon_i = \sigma_0(W_i)U_i$ with U_i independent of W_i and general scale function σ_0 . Khan (2013) also proposed to simultaneously estimate β and the function σ_0 by a semi-parametric sieve approach. The resulting model allows one to obtain an estimator of the binary choice probabilities. While this idea is extremely interesting, it effectively requires estimation of a d -dimensional function in a non-parametric fashion. For the purpose of estimating β , the function σ_0 can be viewed as nuisance parameter and its estimation does not have an impact on the rate at which β is estimable. However, the binary choice probabilities explicitly depend on σ_0 and can thus only be estimated at the corresponding d -dimensional non-parametric rate. In settings where d is moderately large this can be quite problematic.

In the classical setting where responses are observed completely, linear quantile regression models [see Koenker and Bassett (1978)] have proved useful in providing a model that can incorporate gen-

eral forms heteroskedasticity and at the same time avoid non-parametric smoothing. In particular, by looking at a collection of quantile coefficients indexed by the quantile level τ it is possible to obtain a broad picture of the conditional distribution of the response given the covariates. The aim of the present paper is to carry this approach into the setting of binary response models. In contrast to existing methods, we can on one hand allow for rather general forms of heteroskedasticity and at the same time estimate binary choice probabilities without the need of non-parametrically estimating a d -dimensional function.

The ideas explored here are closely related to the work of Kordas (2006). Yet, there are many important differences. First, in his theoretical investigations, Kordas (2006) considered only a finite collection of quantile levels. The present paper aims at considering the quantile *process*. Contrary to the classical setting, and also contrary to the results suggested by the analysis in Kordas (2006), we see that the asymptotic distribution is a white noise type process with limiting distributions corresponding to different quantile levels being independent. An intuitive explanation of this seemingly surprising fact along with rigorous theoretical results can be found in Section 2. We thus provide both a correction and considerable extension of the findings in Kordas (2006).

Further, our results on the quantile process pave the way to obtaining an estimator for the conditional probabilities p_w and derive its asymptotic representation. While a related idea was considered in Kordas (2006), no theoretical justification of its validity was provided. Moreover, we are able to considerably relax the identifiability assumptions that were implicitly made in there. Finally, we demonstrate that our ideas are closely related to the concept of rearrangement [see Dette et al. (2006) or Chernozhukov et al. (2010)] and provide new theoretical insights regarding certain properties of the rearrangement map that seem to be of independent interest.

The rest of the paper is organized as follows. In Section 2, we formally state the model and provide results on uniform consistency and a uniform linearization of the binary response quantile process. All results hold uniformly over an infinite collection of quantiles T . In Section 3, we show how the results from Section 2 can be used to obtain estimators of choice probabilities. We elaborate on the connection of this approach to rearrangements. Finally, a uniform asymptotic representation for a properly rescaled version of the proposed estimators is provided and their joint asymptotic distribution is briefly discussed. All proofs are deferred to an appendix.

2 Estimating the coefficients

Before we proceed to state our results, let us briefly recall some basic facts about identification in binary response models and provide some intuition for the estimators of Manski (1975) and Horowitz (1992). Assume that we have n i.i.d. replicates, say $(Y_i, W_i)_{i=1, \dots, n}$, drawn from the distribution $(Y = I\{Y^* \geq 0\}, W)$ with Y^* denoting the unobserved variable of interest and W denoting a vector of covariates. Further, denote by $q_\tau(w)$ the conditional quantile function of Y^* given $W = w$ and assume that for $\tau \in T \subset [0, 1]$ we have $q_\tau(w) = w^T \beta_\tau$ for some vectors β_τ . Observing that $I\{Y^* \geq 0\} = I\{aY^* \geq 0\}$ with $a > 0$ arbitrary directly shows that the scale of the vector β_τ can not be identified from (Y, W) . On the other hand, the vector β_τ is identified up to scale if for example $b \neq \beta_\tau$ implies that the distribution of Y conditional on $w^T \beta_\tau > 0$ differs from that conditional on $w^T b > 0$ on a sufficiently large set. More precisely, assume that the

function $u \mapsto F_{Y^*|W}(w^T \beta_\tau + u|w)$ is strictly increasing for u in a neighbourhood of zero and all $w \in \text{support}(W)$. In that case we have by the definition of the τ 'th quantile

$$P(Y = 1|W = w) \begin{cases} > 1 - \tau, & \text{if } w^T \beta_\tau > 0 \\ = 1 - \tau, & \text{if } w^T \beta_\tau = 0 \\ < 1 - \tau, & \text{if } w^T \beta_\tau < 0. \end{cases}$$

This already suggests that the expectation of $(Y - (1 - \tau))$ is positive for $W^T \beta_\tau > 0$ and negative for $W^T \beta_\tau < 0$. We thus expect that under appropriate conditions the function

$$S_\tau(\beta) := \mathbb{E}[(Y - (1 - \tau))I\{W^T \beta \geq 0\}]$$

should be maximal at $a\beta_\tau$ for any $a > 0$. Consider a vector $b \in \mathbb{R}^{d+1}$. Then

$$S_\tau(b) - S_\tau(\beta_\tau) = D_1(b, \tau) + D_2(b, \tau)$$

with

$$\begin{aligned} D_1(b, \tau) &:= \mathbb{E}[(Y - (1 - \tau))I\{W^T b \geq 0 > W^T \beta_\tau\}], \\ D_2(b, \tau) &:= -\mathbb{E}[(Y - (1 - \tau))I\{W^T b < 0 \leq W^T \beta_\tau\}]. \end{aligned}$$

Note that both quantities are non-positive, and at least one of them being strictly negative is sufficient for inferring $\beta_\tau/\|\beta_\tau\| \neq b/\|b\|$ from the observable data. An overview and more detailed discussion of related results is provided in Chapter 4 of Horowitz (2009).

A common assumption [see e.g. Chapter 4 in Horowitz (2009)] is that one component of β_τ is either constant or at least bounded away from zero. Without loss of generality, we assume that this holds for the first component of β_τ . In order to simplify the notation of what follows, write the covariate W in the form $W = (Z, X)$ with Z being the first component of W and X denoting the remaining components. Denote the supports of X, Z, W by $\mathcal{X}, \mathcal{Z}, \mathcal{W}$, respectively. Denote by $(Y_i, W_i)_{i=1}^n$ a sample of i.i.d. realizations of the random variable (Y, W) . Define the empirical counterpart of S_τ by

$$\tilde{S}_{n,\tau}(\beta) := \frac{1}{n} \sum_{i=1}^n (Y_i - (1 - \tau))I\{W_i^T \beta \geq 0\}$$

and consider a smoothed version

$$\hat{S}_{n,\tau}(\beta) := \frac{1}{n} \sum_{i=1}^n (Y_i - (1 - \tau))\mathcal{K}\left(\frac{W_i^T \beta}{h_n}\right)$$

with h_n denoting a bandwidth parameter and $\mathcal{K}(u) := \int_{-\infty}^u K(v)dv$ a smoothed version of the indicator function $I\{u \leq 0\}$. Following Horowitz (2009), define the estimator (\hat{s}, \hat{b}_τ) through

$$(\hat{s}_\tau, \hat{b}_\tau) = \operatorname{argmax}_{s=\pm 1, b \in \mathbb{R}^d} \hat{S}_{n,\tau}((s, b^T)^T)$$

Remark 2.1 The proofs of all subsequent results implicitly rely on the fact that we *know* which coefficient stays away from zero and that the covariate corresponding to this particular coefficient

has a 'nice' distribution conditional on all other coefficients [see assumptions (F1), (D2) etc.]. This is in line with the approach of Horowitz (1992) and Kordas (2006) and makes sense in many practical examples. Results similar to the ones presented below might continue to hold if we use Manski's normalization $\|\hat{\beta}\| = 1$ instead of setting the 'right' component to ± 1 . However, the asymptotic representation would be somewhat more complicated. For this reason, we leave this interesting question to future research.

In all of the subsequent developments we make the following basic assumption.

- (A) The coefficient $\beta_{\tau,1}$ satisfies $\inf_{\tau \in T} |\beta_{\tau,1}| > 0$ and the coefficient $\beta_{\tau,1}$ has the same sign on all of T . In what follows, denote this sign by s_0 .

Remark 2.2 Note that due to the scaling the estimator \hat{b} defined above is an estimator of the re-scaled quantity $\bar{b}_\tau := b_\tau / |\beta_{\tau,1}|$ where $b_\tau := (\beta_{\tau,2}, \dots, \beta_{\tau,d+1})^T$. When interpreting the estimator \hat{b}_τ , this must be taken into account. In particular, \hat{b}_τ can not be interpreted as classical quantile regression coefficient. This also explains the reason behind assumption (A).

In order to establish uniform consistency of the smoothed maximum score estimator, we need the following assumptions.

- (K1) The function \mathcal{K} is uniformly bounded and satisfies $\sup_{|v| \geq c} |\mathcal{K}(v) - I\{v \geq 0\}| \rightarrow 0$ as $c \rightarrow \infty$.
- (F1) The conditional distribution function of Z given X , say $F_{Z|X}$, is uniformly continuous uniformly over $x \in \mathcal{X}$, that is

$$\sup_{x \in \mathcal{X}} \sup_{v \in \mathbb{R}} \sup_{|u| \leq \delta} |F_{Z|X}(v|x) - F_{Z|X}(v+u|x)| \rightarrow 0 \quad \text{as } \delta \rightarrow 0$$

- (D1) For any fixed $\tau \in T$, $\bar{\beta}_\tau = (s_0, \bar{b}_\tau)$ is the unique minimizer of $S_\tau(\beta)$ on $\{-1, 1\} \times \mathbb{R}^d$ and additionally

$$d(\alpha) := \sup \left\{ \varepsilon \mid \inf_{\tau \in T} \inf_{\|\beta - \bar{\beta}_\tau\| \geq \varepsilon, |\beta_1| = 1} |S_\tau(\beta) - S_\tau(\bar{\beta}_\tau)| \geq \alpha \right\} \rightarrow 0 \quad \text{as } \alpha \rightarrow 0$$

In order to intuitively understand the meaning of condition (D1) above, note that conditions (K1) and (F1) imply that $\hat{S}_{n,\tau}(\beta) \rightarrow S_\tau(\beta)$ uniformly in τ, β . Condition (D1) essentially requires that the maximum of $S_\tau(\beta)$ is 'well separated' uniformly in τ , which allows to obtain uniform consistency of a sequence of maximizers of any function that uniformly converges to S_τ . Versions of this condition that are directly connected to densities and distributions of some of the regressors can for example be derived by considering a uniform version of Assumption 2 in Manski (1985) by arguments similar to the ones given in that paper, see also Assumptions 1-3 in Horowitz (1992).

Lemma 2.3 *Under assumptions (K1), (D1), (F1) let $h_n \rightarrow 0$. Then the estimator (\hat{s}, \hat{b}_τ) is weakly uniformly consistent, that is*

$$\sup_{\tau \in T} \|(\hat{s}_\tau, \hat{b}_\tau) - (s_0, \bar{b}_\tau)\|_\infty = o_P(1).$$

The next collection of assumptions is sufficient for deriving a uniform linearization [some kind of 'Bahadur-representation'] for \hat{b}_τ . Assume that there exists some $\eta > 0$ such that the following conditions hold.

(K2) The function \mathcal{K} is two times continuously differentiable and its second derivative is uniformly Hölder continuous of order $\gamma > 0$, that is it satisfies

$$\sup_{|x-y|\leq\delta} |\mathcal{K}''(x) - \mathcal{K}''(y)| \leq C_K |x - y|^\gamma.$$

Denote the derivative of \mathcal{K} by K . Assume that additionally, K, K' are uniformly bounded and we have $\int |v^2 K'(v)| dv < \infty$, $\int |K'(v)|^2 dv < \infty$ and additionally $\alpha_n := h_n^{-1} \int_{|v|>\eta} |K'(v)| dv = o(1)$.

(K3) $\int |v^k K(v)| < \infty$ for some $k \geq 2$ and $\int v^j K(v) dv = 0$ for $1 \leq j < k$ and additionally $\int_{|v|>\eta} |K(v)| dv = o(h_n^k)$ as well as $\sup_{|a|>x} |K(a)| = o(x^{-1})$ as $x \rightarrow \infty$.

(B) The bandwidth h_n satisfies $h_n = o(1)$ and additionally $(nh_n^3)^{-1/2}(\log n)^2 = o(1)$.

(D2) The distribution of X has bounded support \mathcal{X} . For almost every $x \in \mathcal{X}$, the covariate Z has a conditional density $f_{Z|X}(\cdot|x)$.

(D3) For any vector b with $\|b - \bar{b}_\tau\| \leq \eta$ the two functions $u \mapsto f_{Z|X}(s_0(-x^T \bar{b}_\tau + u)|x)$ and $u \mapsto F_{Y^*|X,Z}(0|x, s_0(-x^T \bar{b}_\tau + u))$ are two times continuously differentiable at every $u \in U_\eta(0)$ for almost every $x \in \mathcal{X}$ and the first and second derivatives are uniformly bounded [uniformly over $x \in \mathcal{X}, \tau \in T$].

(D4) The function $u \mapsto f_{Z|X}(s_0(-x^T \bar{b}_\tau + u)|x)$ is $k-1$ times continuously differentiable for every $x \in \mathcal{X}$ at every u with $|u| \leq \eta$. All derivatives are uniformly bounded and uniformly continuous uniformly in x, τ . The function $u \mapsto F_{Y^*|X,Z}(0|x, S - 0(-x^T \bar{b}_\tau + u))$ is k times continuously differentiable at every u with $|u| \leq \eta$ at almost every $x \in \mathcal{X}$ and all derivatives are uniformly bounded and uniformly continuous uniformly in x, τ .

(D5) The map $\tau \mapsto \beta_\tau$ is uniformly on T Hölder continuous of order $\gamma > 0$, that is $\sup_{\tau, \tau' \in T, |\tau - \tau'| \leq \delta} \|\beta_\tau - \beta_{\tau'}\| \leq C\delta^\gamma$ for some universal constant C , some $\gamma > 0$ and all $\delta \leq \delta_0$ with $\delta_0 > 0$.

(D6) For any $\tau_1 \neq \tau_2$ with $\tau_1, \tau_2 \in T$ there exists $\varepsilon > 0$ such that $P(\|W^T(\beta_{\tau_1} - \beta_{\tau_2})\| < \varepsilon) = 0$.

(Q) We have $\inf_{\tau \in T} |\lambda_{\max}(Q(s_0, \bar{b}_\tau, \tau))| > 0$ where $\lambda_{\max}(A)$ denotes the largest eigenvalue of the matrix A and we defined

$$Q(s, b, \tau) := \int \frac{\partial}{\partial u} \left((\tau - F_{Y^*|X,Z}(0|x, s(-x^T b + u))) f_{Z|X}(s(-x^T b + u)|x) \right) \Big|_{u=0} x x^T dP_X(x).$$

The conditions on the kernel function K are standard in the binary response setting and were for example considered in Horowitz (2009) and Kordas (2006). Assumptions (D2)-(D4) and (Q) are uniform versions of the conditions in Horowitz (1992) and are needed to obtain results holding uniformly in an infinite collection of quantiles. Condition (D5) is needed to obtain a rate in the

uniform representation below. Condition (D6) implies asymptotic independence of the limiting variables corresponding to different quantile levels. Essentially, it states that quantile curves corresponding to different quantile levels should be 'uniformly separated' which is reasonable in most applications. In particular, (D6) follows if the conditional density $f_{Y^*|W}(y|w)$ of Y^* given W is uniformly bounded away from zero for all (y, w) with $F_{Y^*|W}(y|w) \in T$ and w in the support of W .

Remark 2.4 Some straightforward calculations show that under assumption (D3) and the boundedness of the support of Z the matrix $Q(s_0, \bar{b}_\tau, \tau)$ in condition (Q) is the second derivative of the function $b \mapsto S_\tau((s_0, b^T)^T)$ evaluated in \bar{b}_τ . Since S is assumed to be maximal in this point, the matrix $Q(s_0, \bar{b}_\tau, \tau)$ is negative definite, and thus we need to bound its largest eigenvalue away from zero in order to obtain a uniform version of the non-singularity of Q .

Theorem 2.5 *Under assumptions (A), (B), (D1)-(D5), (F1), (K1)-(K3), we have*

$$Q(s_0, \bar{b}_\tau, \tau)(\hat{b}_\tau - \bar{b}_\tau) = -\tilde{T}_n(s_0, \bar{b}_\tau, \tau) + R_n(\tau) \quad (2.1)$$

where

$$\begin{aligned} \tilde{T}_n(s, b, \tau) &:= \frac{\partial \tilde{S}_n(s, b, \tau)}{\partial b} = \frac{1}{nh_n} \sum_{i=1}^n (Y_i - (1 - \tau)) X_i K\left(\frac{X_i^T b + s Z_i}{h_n}\right), \\ \sup_{\tau \in T} \|R_n(\tau)\| &= O_P(\kappa_n) := O_P\left((h_n^k + (nh_n)^{-1/2} \log n)((nh_n^3)^{-1/2} \log n + h_n + \alpha_n)\right). \end{aligned}$$

In particular, $\kappa_n = o(h_n^k + (nh_n)^{-1/2})$ and thus negligible compared to $\tilde{T}_n(s_0, \bar{b}_\tau, \tau)$.

Now assume that additionally condition (D6) holds. Then, for any finite collection $\tau_1, \dots, \tau_k \in T$ we obtain

$$\left(\sqrt{nh_n}(\hat{b}_{\tau_j} - \bar{b}_{\tau_j} - T_n(s_0, \bar{b}_{\tau_j}, \tau_j))\right)_{j=1, \dots, k} \xrightarrow{\mathcal{D}} \left((Q_{\tau_j}(s_0, \bar{b}_{\tau_j}, \tau_j))^{-1} M_{\tau_j}\right)_{j=1, \dots, k} \quad (2.2)$$

where

$$\begin{aligned} T_n(s, b, \tau) &:= \frac{h_n^k}{k!} \int v^k K(v) dv \int g_k(s, b, x) x dP_X(x), \\ g_j(s, x, b) &:= \frac{\partial^j}{\partial u^j} \left((\tau - F_{Y^*|X, Z}(0|x, s(-x^T b + u))) f_{Z|X}(s(-x^T b + u)|x) \right) \Big|_{u=0}, \end{aligned}$$

M_{τ_i}, M_{τ_j} are independent for $j \neq i$ and

$$M_\tau \sim \mathcal{N}(0, \Sigma_\tau).$$

where

$$\Sigma_\tau := \tau(1 - \tau) \int K^2(u) du \int x x^T f_{Z|X}(-s_0 x^T \bar{b}_\tau | x) dP_X(x).$$

Compared to the results available in the literature [e.g. in Kordas (2006) and Horowitz (2009)], the preceding theorem provides two important new insights. To the best of our knowledge, it is the first time that the estimator is simultaneously considered at an infinite collection of quantiles. Equally importantly, it demonstrates that the joint asymptotic distribution of several quantiles differs substantially from what both intuition and results in Kordas (2006) seem to suggest.

Remark 2.6 In contrast to the 'classical' case, the properly normalized quantile process at different quantile levels converges to independent random variables. An intuitive explanation for this surprising fact can be obtained from the asymptotic linearization in (2.1). For simplicity, assume that the kernel K has compact support, say $[-1, 1]$. Then all observations that have a non-zero contribution to $\tilde{T}_n(s_0, \bar{b}_\tau, \tau)$ will need to satisfy $|W_i^T \beta_\tau| \leq h_n |\beta_{\tau,1}|$. In particular, letting $h_n \rightarrow 0$ implies that asymptotically for different values of τ disjoint sets of observations will be driving the distribution of T_n . Similar phenomena can be observed in other settings that include non-parametric smoothing, a classical example being density estimation. Note that regarding this particular point the paper of Kordas (2006) contained a mistake. More precisely, Kordas (2006) claimed that the asymptotic distributions corresponding to different quantiles have a non-trivial covariance which is not the case.

In particular, the above findings imply that there can be no weak convergence of the normalized process $\left(\sqrt{nh_n}(\hat{b}_\tau - \bar{b}_\tau - T_n(s_0, \bar{b}_\tau, \tau))\right)_{\tau \in T}$ in a reasonable functional sense since the candidate 'limiting process' has a 'white noise' structure and is not tight. This will present an additional challenge for the analysis of estimators for binary choice probabilities constructed in the following section.

3 Estimating conditional probabilities

Partially due to the lack of complete identification, the coefficients estimated in the preceding section might be hard to interpret. A more tractable quantity is given by the conditional probability $p_w := P(Y = 1|W = w)$. One possible way to estimate this probability would be local averaging. However, due to the curse of dimensionality, this becomes impractical if the length of W exceeds 2 or 3. An alternative is to assume that the linear model $q_\tau(w) = w^T \beta_\tau$ holds for all $\tau \in T \subset [0, 1]$. By definition of $Y = I\{Y^* \geq 0\}$, the existence of $\tau_w \in T$ with $w^T \beta_{\tau_w} = 0$ implies that $p_w = 1 - \tau_w$. On the other hand, the quantile function of Y^* is given by $w^T \beta_\tau$ and thus $P(Y^* \leq 0|W = w) = w^T \beta_{\tau_w}$. By definition of the quantile function and the assumptions on Y^* , $\tau < \tau_w \Leftrightarrow w^T \beta_\tau < w^T \beta_{\tau_w}$. This implies the equality $\tau_w = \int_0^1 I\{w^T \beta_\tau \leq 0\} d\tau$. In particular, we have for any $(a, b) \subset T$ with $a < \tau_w < b$

$$p_w = \int_0^1 I\{w^T \beta_\tau \geq 0\} d\tau = 1 - b + \int_a^b I\{w^T \beta_\tau \geq 0\} d\tau = 1 - b + \int_a^b I\{w^T \bar{\beta}_\tau \geq 0\} d\tau$$

This suggests to estimate p_w by replacing β_τ in the above representation with the estimator $\hat{\beta}_\tau$ from the preceding section after choosing (a, b) in some sensible manner. The fact that $\hat{\beta}$ is an estimator of the re-scaled version $\bar{\beta}_\tau$ is not important here since multiplication by a positive number does not affect the inequality $w^T \beta_\tau \geq 0$. From here on, define

$$\hat{p}_w(a, b) := 1 - b + \int_a^b I\{w^T \hat{\beta}_\tau \geq 0\} d\tau. \quad (3.1)$$

This also indicates that in order to estimate p_w we do not need the linear model $q_\tau(w) = w^T \beta_\tau$ to hold globally and also do not require that β_τ can be estimated for all $\tau \in T$. In fact, the validity

of the linear model $q_\tau(w) = w^T \beta_\tau$ for τ in a neighbourhood of τ_w and estimability of β_τ on this region is sufficient for the asymptotic developments provided below.

Remark 3.1 Assume that the estimator $\hat{\beta}_\tau$ is uniformly consistent and that additionally for any $\delta > 0$ there exists $\varepsilon_\delta > 0$ such that $\inf_{|u - \tau_w| > \delta} |w^T(\hat{\beta}_u - \bar{\beta}_{\tau_w})| \geq \varepsilon_\delta$. Then uniform consistency of $\hat{\beta}_\tau$ will directly yield that, with probability tending to one, $\hat{p}_w(a, b) = \hat{p}_w(a', b')$ as long as $a < \tau_w < b$ and $a' < \tau_w < b'$. This suggests that from an asymptotic point of view the choice of a, b in the estimator $\hat{p}_w(a, b)$ is not very critical. On the other hand, τ_w is unknown in practice. Thus choosing a, b as small and large as the data allow, respectively, seems to be a sensible practical approach. At the same time, identifiability at infinity is not needed to obtain an estimator of probabilities for points that are bounded away from the boundary of the covariate space.

Remark 3.2 The definition of \hat{p}_w is closely connected to the concept of rearrangement [see Hardy et al. (1988)]. More precisely, recall that the monotone rearrangement Φ of a function $g : [0, 1] \rightarrow \mathbb{R}$ is defined as

$$\Phi_g(u) = \Psi_g^-(u), \quad \Psi_g(v) = \int_0^1 I\{g(u) \leq v\} du$$

where Ψ_g^- denotes the generalized inverse of the function $u \mapsto \Psi_g(u)$ and the first step of the rearrangement, Ψ_g , is the *distribution function of g with respect to Lebesgue measure*. Thus we can interpret the integral $\int_0^1 I\{w^T \beta_\tau \leq 0\} d\tau$ in the definition of τ_w as the distribution function of the map $\tau \mapsto w^T \beta_\tau$. Previously, a smoothed version of the first step of the rearrangement was used by Dette and Volgushev (2008) to invert a non-increasing estimator of an increasing function in the setting of quantile regression. On the other hand, it is not obvious if the function $\tau \mapsto w^T \bar{\beta}_\tau$ is increasing since $\bar{\beta}_\tau$ is a re-scaled version of the quantile coefficient β_τ . However, as we already pointed out in Remark 3.1, the function $\tau \mapsto w^T \bar{\beta}_\tau$ will still have a unique zero. As we shall argue next, the first step of the rearrangement map can provide a way to estimate this zero point in a sensible way.

The properties of the rearrangement viewed as mapping between function spaces were considered in Dette et al. (2006) for estimating a monotone function and Chernozhukov et al. (2010) for monotone crossing quantile curves. In particular, the last-named authors derived a kind of compact differentiability of the rearrangement mapping at functions that are not necessarily increasing. However, those results can not be directly applied here since Dette et al. (2006) and Dette and Volgushev (2008) applied smoothing while the compact differentiability result of Chernozhukov et al. (2010) requires a process based functional central limit theorem. Due to the asymptotic independence of the limiting distributions in Theorem 2.5, such a result is impossible in our setting. Still, a general analysis of the rearrangement map is possible and will be presented next. The crucial insight is that the process $\tau \mapsto (nh_n)^{1/2}(\hat{\beta}_\tau - \bar{\beta}_\tau)$ is still sufficiently smooth on domains of size decreasing at the rate h_n while its convergence to the limit takes place at a faster rate.

We begin by stating a general result that allows to derive a uniform linearization of the map Ψ defined above. In situations where a functional central limit result does not hold (this will often be the case in the situation of estimators build from local windows), this result seems to be of

independent interest. In particular, it can be used to derive a uniform Bahadur representation for the estimator \hat{p}_w in the present setting.

Theorem 3.3 *Consider a collection of functions $g_q : [0, 1] \rightarrow \mathbb{R}$ indexed by a general set Q and assume that for all $q \in Q$ there exists $u_{0,q} \in (0, 1)$ with $g_q(u_{0,q}) = 0$. Additionally, assume that each g_q is continuously differentiable in a neighbourhood $U_\delta(u_{0,q}) \subset (0, 1) \forall q \in Q$ and that its derivative is uniformly Hölder continuous of order γ with constant C_H both not depending on q , and that for any $\varepsilon > 0$ we have $\inf_q \inf_{|u-u_{0,q}|>\varepsilon} |g_q(u) - g_q(u_{0,q})| > 0$ and $\inf_q g'_q(u_{0,q}) > 0$. Denote by $g_{n,q} : [0, 1] \rightarrow \mathbb{R}$ a collection of estimators for g_q . Assume that*

$$\sup_{q \in Q} \sup_{|u-u_{0,q}| \leq \varepsilon_n} |g_{n,q}(u_{0,q}) - g_{n,q}(u) - (g_q(u_{0,q}) - g_q(u))| = O_P(\xi_n(\varepsilon_n)), \quad (3.2)$$

that

$$\sup_q \sup_{u \in [0,1]} |g_{n,q}(u) - g_q(u)| = o_P(1) \quad (3.3)$$

and that

$$\sup_{|u-u_{0,q}| \leq \delta} |g_{n,q}(u) - g_q(u)| = O_P(R_n). \quad (3.4)$$

If for all $c_n = o(C_n)$ with some given $C_n \rightarrow \infty$ we have $\xi_n(R_n c_n) = o(R_n)$ it follows that for any collection of points $a_q, b_q \in (0, 1)$ with $a_q + \varepsilon \leq u_q \leq b_q - \varepsilon$ with $\varepsilon > 0$ fixed we have with probability tending to one

$$\Psi_{g_{n,q}}(0) = a_q + \int_{a_q}^{b_q} I\{g_{n,q}(u) \leq 0\} du \quad \forall q \in Q \quad (3.5)$$

and

$$\sup_{q \in Q} \left| \Psi_{g_q}(0) - \Psi_{g_{n,q}}(0) \right| = O_P\left(R_n^{1+\gamma} + \xi_n(R_n+)\right). \quad (3.6)$$

where $f(x+) := \lim_{\varepsilon \searrow 0} f(x + \varepsilon)$.

We now state the additional assumptions that are needed to derive the limiting distribution of \hat{p}_w . Assume that for some $\delta > 0$ the conditions of Theorem 2.5 hold on the set $T^\delta := [t_L - \delta, t_U + \delta]$ with $T := [t_L, t_U] \subset (\delta, 1 - \delta)$. For this T , we will need the following conditions.

(T1) Define the set $W_{T^\delta} := \{w \in \mathcal{W} \mid \exists \tau_w \in T^\delta : w^T \beta_{\tau_w} = 0\}$. Assume that for every $w \in W_{T^\delta}$ there exists a unique $\tau_w \in T$ with $w^T \beta_{\tau_w} = 0$. Assume that the function $\tau \mapsto \bar{\beta}_\tau$ is continuously differentiable on T^δ , that its derivative, say Δ_τ , is uniformly Hölder continuous of order $\gamma > 0$ and that $L_T := \inf_{\tau \in T^\delta} |w^T \Delta_{\tau_w}| > 0$.

(T2) The function $\tau \mapsto Q(s_0, \bar{b}_\tau, \tau)$ is Hölder continuous of order $\gamma > 0$ uniformly on T^δ .

(K4) We have $\sup_{|x| \geq c} |K'(x)| \leq c^{-1/2-\varepsilon}$ for some $\varepsilon > 0$ and all $c \geq c_0$.

The above conditions ensure that the collection of estimators $g_{n,w}(\tau) := w^T \hat{\beta}_\tau$ satisfies the conditions of Theorem 3.3 [see Lemma 4.4 for condition (3.2)]. An application of this result thus directly yields the following result.

Theorem 3.4 *Assume that for some $\delta > 0$ the conditions of Theorem 2.5 hold on the set $T^\delta := [t_L - \delta, t_U + \delta]$ with $T := [t_L, t_U] \subset (\delta, 1 - \delta)$ and let conditions (T1), (T2), (K4) hold. Assume that for each $w \in W_T$ we have $(a_w, b_w) \subset T^\delta$. Then for any $w \in W_T := \{w \in \mathcal{W} \mid \exists \tau_w \in T : w^T \beta_{\tau_w} = 0\}$*

$$\hat{p}_w(a_w, b_w) - p_w = -w^T (\hat{\beta}_{\tau_w} - \bar{\beta}_{\tau_w}) |w^T \Delta_{\tau_w}|^{-1} + R_n^{(2)}(w)$$

with

$$\sup_{w \in W_T} |R_n^{(2)}(w)| = O_P \left(\frac{1}{(nh_n)^{1/2}} \left(\frac{(\log n)^{1+\gamma_\beta}}{(nh_n)^{\gamma_\beta/2}} + \frac{(\log n)^2}{(nh_n^3)^{1/2}} \right) \right) + o_P(h_n^k) + O_P(\kappa_n)$$

where κ_n was defined in Theorem 2.5. In particular, the remainder is negligible compared to $-w^T (\hat{\beta}_{\tau_w} - \bar{\beta}_{\tau_w}) |w^T \Delta_{\tau_w}|^{-1}$. Moreover, for any finite collection $w_1, \dots, w_k \in W_T$ with $w_j^T = (z_j, x_j)$ we obtain

$$\begin{aligned} & \left(\sqrt{nh_n} \left(\hat{p}_{w_j} - p_{w_j} + |w^T \Delta_{\tau_w}|^{-1} x_j^T T_n(s_0, \bar{b}_{\tau_{w_j}}, \tau_{w_j}) \right) \right)_{j=1, \dots, k} \\ & \xrightarrow{\mathcal{D}} \left((w^T \Delta_{\tau_{w_j}})^{-1} x_j^T (Q_{\tau_{w_j}}(s_0, \bar{b}_{\tau_{w_j}}, \tau_{w_j}))^{-1} M_{\tau_{w_j}} \right)_{j=1, \dots, k} \end{aligned}$$

where T_n, M_τ is as defined in Theorem 2.5.

From the results derived above, we see that the convergence rate of the estimators for binary choice probabilities corresponds to the rate typically encountered if one-dimensional smoothing is performed. Compared to the results of Khan (2013) whose rates correspond to d -dimensional smoothing, this can be a very substantial improvement. While our assumptions are of course more restrictive than those of Khan (2013), the form of allowed heteroskedasticity is somewhat more general than the simple multiplicative heteroskedasticity or even homoskedasticity assumed in previous work. While we of course do not suggest to completely replace the methodologies developed in the literature, we feel that our approach can be considered as a good compromise between flexibility of the underlying model and convergence rates. It thus provides a valuable supplement and extension of available procedures.

4 Proofs

Proof of Lemma 2.3 By Lemma 2.6.15 and Lemma 2.6.18 in van der Vaart and Wellner (1996), the classes of functions $\{(y, w) \mapsto yI\{w^T \beta \geq 0\} | \beta \in \mathbb{R}^{d+1}\}$ and $\{w \mapsto I\{w^T \beta \geq 0\} | \beta \in \mathbb{R}^{d+1}\}$ are VC-subgraph classes of functions. Together with Theorem 2.6.7 and Theorem 2.4.3 in the same reference this implies

$$\sup_{\tau \in [0,1], \beta \in \mathbb{R}^{d+1}} |S_\tau(\beta) - \tilde{S}_{n,\tau}(\beta)| = o(1) \quad a.s. \quad (4.1)$$

Next, observe that almost surely

$$|\hat{S}_{n,\tau}(\beta) - \tilde{S}_{n,\tau}(\beta)| \leq \sup_{|v| \geq c} |\mathcal{K}(v/h_n) - I\{v \geq 0\}| + (1 + \sup_v |\mathcal{K}(v)|) \frac{1}{n} \sum_{i=1}^n I\{|\beta^T W_i| \leq ch_n\}.$$

Moreover,

$$I\{|\beta^T W_i| \leq ch_n\} = I\{\beta^T W_i \leq ch_n\} - I\{\beta^T W_i > -ch_n\},$$

and the classes of functions $\{w \mapsto I\{w^T \beta \leq c\} | \beta \in \mathbb{R}^{d+1}, c \in \mathbb{R}\}$, $\{w \mapsto I\{w^T \beta > c\} | \beta \in \mathbb{R}^{d+1}, c \in \mathbb{R}\}$ are VC-subgraph by Lemma 2.6.15 and Lemma 2.6.18(viii) in van der Vaart and Wellner (1996). In combination with Theorem 2.6.7 and Theorem 2.4.3 from the same reference this implies

$$\sup_{\beta \in \mathbb{R}^{d+1}} \sup_{c \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I\{|\beta^T W_i| \leq ch_n\} - P(|\beta^T W_i| \leq ch_n) \right| \rightarrow 0 \quad a.s.$$

Setting $c = c_n = h_n^{-1/2}$ in the bound for $|\hat{S}_{n,\tau}(\beta) - \tilde{S}_{n,\tau}(\beta)|$ we see that the first term, which is independent of β , converges to zero by assumption (K1). Moreover, by assumption (F1) we have for $\beta = (1, b^T)^T$

$$\begin{aligned} \sup_{b \in \mathbb{R}^d} P(Z_i + b^T X_i \leq c_n h_n) &= \sup_{b \in \mathbb{R}^d} \int F_{Z|X}(-b^T x + h_n^{1/2}|x) - F_{Z|X}(-b^T x - h_n^{1/2}|x) dP_X(x) \\ &= o(1) \end{aligned}$$

almost surely. A similar results holds for $\beta = (-1, b^T)^T$. Combining all the results so far we thus see that

$$s_n := \sup_{b \in \mathbb{R}^d, s = \pm 1} \left| \hat{S}_{n,\tau}((s, b^T)^T) - S_\tau((s, b^T)^T) \right| = o(1) \quad a.s.$$

Finally, observe that almost surely

$$\sup_{\tau \in T} \|\hat{\beta}_{\tau,n} - \bar{\beta}_\tau\| \leq 2d(s_n) = o(1)$$

where $d(s)$ was defined in condition (D1). To see that this is the case, observe that $\hat{\beta}_{\tau,n}$ maximizes $\tilde{S}_{n,\tau}(\beta)$, and thus we have a.s. for every $\tau \in T$

$$0 \leq \tilde{S}_{n,\tau}(\hat{\beta}_\tau) - \tilde{S}_{n,\tau}(\bar{\beta}_\tau) \leq S_\tau(\hat{\beta}_\tau) - S_\tau(\bar{\beta}_\tau) + 2s_n,$$

which implies $\sup_{\tau} |S_\tau(\hat{\beta}_\tau) - S_\tau(\bar{\beta}_\tau)| \leq 2s_n$ since $S_\tau(\bar{\beta}_\tau) \geq S_\tau(\hat{\beta}_\tau)$ for all $\tau \in T$. \square

Proof of Theorem 2.5 Define the quantities

$$\tilde{S}_n(s, b, \tau) := \frac{1}{n} \sum_{i=1}^n (Y_i - (1 - \tau)) \mathcal{K} \left(\frac{X_i^T b + s Z_i}{h_n} \right). \quad (4.2)$$

First, by uniform consistency of $\hat{\beta}_\tau$ and given the fact that $s_0 \in \{-1, 1\}$ we see that with probability tending to one $\hat{s}_\tau = s_0$ for all $\tau \in T$. Thus defining

$$\hat{b}_\tau := \operatorname{argmax}_{b \in \mathbb{R}^d} \hat{S}_n(s_0, b, \tau)$$

we see that $P(\hat{b}_\tau = \tilde{b}_\tau \forall \tau \in T) \rightarrow 1$. Moreover, uniform consistency of \hat{b}_τ implies that with probability tending to one it will satisfy $\sup_{\tau \in T} \|\hat{b}_\tau - \bar{b}_\tau\| \leq \eta$, and continuous differentiability of $b \mapsto \tilde{S}_n(s, b, \tau)$ for b with $\|b - \bar{b}_\tau\| \leq \eta$ implies that

$$\tilde{T}_n(s_0, \hat{b}_\tau, \tau) = 0 \quad \forall \tau \in T.$$

A Taylor expansion now yields that with probability tending to one

$$\tilde{T}_n(s_0, \bar{b}_\tau, \tau) + \tilde{Q}_n(s_0, b_\tau^*, \tau)(\hat{b}_\tau - \bar{b}_\tau) = 0 \quad \forall \tau \in T$$

where

$$\tilde{Q}_n(s, b, \tau) := \frac{\partial \tilde{T}_n(s, b, \tau)}{\partial b} = \frac{1}{nh_n^2} \sum_{i=1}^n (Y_i - (1 - \tau)) X_i X_i^T K' \left(\frac{X_i^T b + s Z_i}{h_n} \right) \quad (4.3)$$

and $b_\tau^* = \xi_n(\tau) \bar{b}_\tau + (1 - \xi_n(\tau)) \hat{b}_\tau$ for some $\xi_n(\tau) \in [0, 1]$. Rearranging we obtain

$$Q(s_0, \bar{b}_\tau, \tau)(\hat{b}_\tau - \bar{b}_\tau) = -\tilde{T}_n(s_0, b_\tau, \tau) + (\tilde{Q}_n(s, b_\tau^*, \tau) - Q(s, \bar{b}_\tau, \tau))(\hat{b}_\tau - \bar{b}_\tau). \quad (4.4)$$

Since $\|\hat{b}_\tau - \bar{b}_\tau\|_\infty = o_P(1)$ uniformly over $\tau \in T$ and since the same holds for $\tilde{T}_n(s_0, \bar{b}_\tau, \tau)$ [see Lemma 4.2], there exists a $\gamma_n \rightarrow 0$ such that

$$\sup_{\tau \in T} \left\| -\tilde{T}_n(s_0, \bar{b}_\tau, \tau) + (\tilde{Q}_n(s, b_\tau^*, \tau) - Q(s, \bar{b}_\tau, \tau))(\hat{b}_\tau - \bar{b}_\tau) \right\| = o_P(\gamma_n).$$

By the conditions on $Q(s_0, \bar{b}_\tau, \tau)$ this implies $\sup_{\tau} \|\hat{b}_\tau - \bar{b}_\tau\| = o_P(\gamma_n)$. By Lemma 4.1 this in turn implies

$$\sup_{\tau \in T} \left\| \tilde{Q}_n(s, b_\tau^*, \tau) - Q(s, \bar{b}_\tau, \tau) \right\| = O_P((nh_n^3)^{-1/2} \log n) + O(h_n) + o_P(\gamma_n).$$

Plugging this into (4.4) and repeating this argument [note that every application yields an improvement of the bound until $\gamma_n \sim \sup_{\tau} \|\tilde{T}_n(s_0, \bar{b}_\tau, \tau)\|$] yields the assertion (2.1).

For a proof of assertion (2.2) note that for $\kappa \neq \tau$ and any $\delta > 0$ we have as $n \rightarrow \infty$

$$\begin{aligned} & \mathbb{E} \left[\left\| \tilde{T}_n(s_0, b_\tau, \tau)^T \tilde{T}_n(s_0, b_\kappa, \kappa) \right\| \right] \\ & \leq \frac{1}{nh_n^2} \int \int \left| K \left(\frac{x^T b_\tau + s_0 z}{h_n} \right) K \left(\frac{x^T b_\kappa + s_0 z}{h_n} \right) \right| f_{Z|X}(z|x) dz dP_X(x). \\ & \leq \frac{1}{nh_n^2} \int \left(\sup_{|a| \geq \delta h_n^{-1}} |K(a)| + \|K\|_\infty I\{\|w^T \beta_\tau\| < |\beta_{\tau,1}|^{-1} \delta\} \right) \\ & \quad \times \left(\sup_{|a| \geq \delta h_n^{-1}} |K(a)| + \|K\|_\infty I\{\|w^T \beta_\kappa\| < |\beta_{\kappa,1}|^{-1} \delta\} \right) dP_W(w) \end{aligned}$$

The assumptions on K now imply that for any $\delta > 0$ we have $\sup_{|a| \geq \delta h_n^{-1}} |K(a)| = o(h_n^2)$. Thus it remains to consider the integral

$$\int \|K\|_\infty^2 I\{\|W^T \beta_\kappa\| < |\beta_{\kappa,1}|^{-1} \delta\} I\{\|W^T \beta_\tau\| < |\beta_{\tau,1}|^{-1} \delta\} dP_W(w).$$

By condition (D6), this integral equals zero if we choose δ such that $|\beta_{\kappa,1}|^{-1} \delta$ and $|\beta_{\tau,1}|^{-1} \delta$ are below the ε specified in that condition. Thus $\mathbb{E}\left[\left\|\tilde{T}_n(s_0, b_\tau, \tau)^T \tilde{T}_n(s_0, b_\kappa, \kappa)\right\|\right] = o(nh_n^{-1})$ and the covariance is negligible compared to the variance which is of order nh_n^{-1} . The rest of the proof follows by standard arguments and is omitted. \square

Lemma 4.1 *Under assumptions (K1), (K2), (B), (D2), (D3), (D5) we have [α_n was defined in assumption (K2)]*

$$\sup_{\tau \in T} \sup_{\|b - \bar{b}_\tau\| \leq \eta} \left\| \mathbb{E}[\tilde{Q}_n(s_0, b, \tau)] - Q(s_0, b, \tau) \right\|_\infty = O(h_n + \alpha_n), \quad (4.5)$$

$$\sup_{\tau \in T} \sup_{\|b - \bar{b}_\tau\| \leq \eta} \left\| \tilde{Q}_n(s_0, b, \tau) - \mathbb{E}[\tilde{Q}_n(s_0, b, \tau)] \right\|_\infty = O_P((nh_n^3)^{-1/2} \log n), \quad (4.6)$$

where $\tilde{Q}_n(s, b, \tau)$ was defined in (4.3). Moreover, we have for any $a_n \rightarrow 0$

$$\sup_{\tau \in T} \sup_{\|b - \bar{b}_\tau\| \leq a_n} \left\| Q(s_0, \bar{b}_\tau, \tau) - Q(s_0, b, \tau) \right\|_\infty \leq C a_n \quad (4.7)$$

for a_n small enough and some universal constant C .

Proof We begin by considering assertion (4.5). Observe that

$$\begin{aligned} \mathbb{E}[\tilde{Q}_n(s_0, b, \tau)] &= \frac{1}{h_n^2} \int \int (\tau - F_{Y^*|X,Z}(0|x, z)) f_{Z|X}(z|x) x x^T K' \left(\frac{x^T b + s_0 z}{h_n} \right) dz dP_X(x) \\ &= \frac{1}{h_n} \int \int (\tau - F_{Y^*|X,Z}(0|x, s_0(vh_n - x^T b))) f_{Z|X}(s_0(vh_n - x^T b)|x) x x^T K'(v) dv dP_X(x). \end{aligned}$$

The assertion now follows from a Taylor expansion, the assumptions on K , and standard arguments similar to those given in Horowitz (2009). For a proof of (4.6) note that for $i, j = 1, \dots, d$ we have

$$\sup_{\tau \in T} \sup_{\|b - \bar{b}_\tau\| \leq \eta} \left| \left(\tilde{Q}_n(s_0, b, \tau) - \mathbb{E}[\tilde{Q}_n(s_0, b, \tau)] \right)_{i,j} \right| \leq h_n^{-2} n^{-1/2} \sup_{f \in \mathcal{F}_n^{i,j}} |\mathbb{G}_n(f)|$$

where $\mathcal{F}_n^{i,j}$ denotes the n -dependent class of functions

$$\mathcal{F}_n^{i,j} := \left\{ f_{n,b,\tau}(x, y, z) = K' \left(\frac{x^T (\bar{b}_\tau + b) + s_0 z}{h_n} \right) (y - (1 - \tau)) x_i x_j \mid b \in \mathbb{R}^d, \|b\| \leq \eta, \tau \in T \right\}$$

and

$$\mathbb{G}_n(f) := n^{-1/2} \sum_{i=1}^n (f(X_i, Y_i, Z_i) - \mathbb{E}[f(X_i, Y_i, Z_i)]).$$

Now uniform Hölder continuity of K' , uniform Hölder continuity of $\tau \mapsto b_\tau$, and uniform boundedness of x implies that for every sufficiently small $\delta > 0$ we have for all $|\tau - \tau'| \leq \delta$, $\|b - b'\| \leq \delta$ for some $\gamma > 0$

$$\|f_{n,b,\tau}(x, y, z) - f_{n,b',\tau'}(x, y, z)\|_\infty \leq \frac{C}{h_n}(|\tau - \tau'|^\gamma + \|b - b'\|^\gamma)$$

with C denoting some constant independent of n, τ, τ', b, b' . This shows that for sufficiently small ε the $\|\cdot\|_\infty$ -bracketing number [see van der Vaart and Wellner (1996), Chapter 2] of the class $\mathcal{F}_n^{i,j}$ is bounded by

$$\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}_n^{i,j}, \|\cdot\|_\infty) \leq C_\gamma h_n^{-(d+1)/\gamma} \varepsilon^{-(d+1)/\gamma}.$$

Next, observe that for any $\tau \in T$, $\|b\| \leq \eta$

$$\begin{aligned} \mathbb{E}[f_{n,b,\tau}^2(X, Y, Z)] &\leq C \int \int_{\mathbb{R}} \left| K' \left(\frac{x^T(\bar{b}_\tau + b) + s_0 z}{h_n} \right) \right|^2 dz dP_X(x) = C \int \int_{\mathbb{R}} |K'(zh_n^{-1})|^2 dz dP_X(x) \\ &= h_n C \int_{\mathbb{R}} |K'(z)|^2 dz. \end{aligned}$$

Combining this with Lemma A.1 yields

$$\sup_{f \in \mathcal{F}_n^{i,j}} |\mathbb{G}_n(f)| = O_P(h_n^{1/2} \log n),$$

and thus the proof of (4.6) is complete. Finally, assertion (4.7) follows by the smoothness properties of $F_{Y^*|X,Z}$ and $f_{Z|X}$. Thus the proof is complete. \square

Lemma 4.2 *Under assumptions (K1)-(K3), (B), (D2), (D4), (D5) we have*

$$\sup_{\tau \in T} \left\| \mathbb{E}[\tilde{T}_n(s_0, \bar{b}_\tau, \tau)] - T_n(s_0, \bar{b}_\tau, \tau) \right\|_\infty = o(h_n^k), \quad (4.8)$$

$$\sup_{\tau \in T} \left\| \tilde{T}_n(s_0, \bar{b}_\tau, \tau) - \mathbb{E}[\tilde{T}_n(s_0, \bar{b}_\tau, \tau)] \right\|_\infty = O_P((nh_n)^{-1/2} \log n). \quad (4.9)$$

Proof The proof of (4.9) follows by arguments very similar to those used to establish (4.6) and is therefore omitted. For the proof of (4.8), note that

$$\begin{aligned} \mathbb{E}[\tilde{T}_n(s_0, \bar{b}_\tau, \tau)] &= \frac{1}{h_n} \int \int x(\tau - F_{Y^*|X,Z}(0|x, z)) K \left(\frac{s_0 z + x^T \bar{b}_\tau}{h_n} \right) f_{Z|X}(z|x) dz dP_X(x) \\ &= |s_0| \int \int x \left(\tau - F_{Y^*|X,Z}(0|x, s_0(vh_n - x^T \bar{b}_\tau)) \right) f_{Z|X}(s_0(vh_n - x^T \bar{b}_\tau)|x) K(v) dv dP_X(x). \\ &= \int \int_{|vh_n| > \eta} x \left(\tau - F_{Y^*|X,Z}(0|x, s_0(vh_n - x^T \bar{b}_\tau)) \right) f_{Z|X}(s_0(vh_n - x^T \bar{b}_\tau)|x) K(v) dv dP_X(x) \\ &\quad + \int \int_{|vh_n| \leq \eta} x \left(\tau - F_{Y^*|X,Z}(0|x, s_0(vh_n - x^T \bar{b}_\tau)) \right) f_{Z|X}(s_0(vh_n - x^T \bar{b}_\tau)|x) K(v) dv dP_X(x). \end{aligned}$$

The order of the first integral is $o(h_n^k)$ by the assumptions on K . The assertion now follows by a Taylor expansion of the function

$$u \mapsto \left(\tau - F_{Y^*|X,Z}(0|x, s_0(u - x^T \bar{b}_\tau)) \right) f_{Z|X}(s_0(u - x^T \bar{b}_\tau)|x),$$

which holds for $|u| \leq \eta$ the assumptions on K and standard arguments. \square

Proof of Theorem 3.3 The statement (3.5) is a direct consequence of the condition on the collection of functions g_q and the uniform consistency in (3.3).

The main technical ingredient for the remaining proof are the bounds provided in Lemma 4.3. Consider an arbitrary sequence $c_n \rightarrow \infty$ with $\xi_n(R_n c_n) = o(R_n)$ and set $r_n := R_n c_n$. Then, with probability tending to one we have for all $q \in Q$

$$\Psi_{g_q}(0) - \Psi_{g_{q,n}}(0) = \int_{u_{0,q}-r_n}^{u_{0,q}+r_n} I\{g_q(u) \leq 0\} - I\{g_{n,q}(u) \leq 0\} du$$

and for all $p > 0$ there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $\sup_q |g_{n,q}(u_{0,q})| + \frac{2(\xi_n(r_n) + C_H r_n^{1+\gamma})}{|g'_q(u_{0,q})|} \leq r_n$ with probability at least $1 - p$ [this follows from]. Applying the bound in Lemma 4.3 we thus see that for each $q \in Q$, as soon as $r_n < \delta$ and n large enough we have with probability at least $1 - p$

$$\begin{aligned} & \sup_q \left| \int_{u_{0,q}-r_n}^{u_{0,q}+r_n} I\{g_q(u) \leq 0\} - I\{g_{n,q}(u) \leq 0\} du - \frac{g_{n,q}(u_{0,q})}{|g'_q(u_{0,q})|} \right| \\ & \leq \sup_q \frac{2\xi_n(r_n) + 4C_H r_n^{1+\gamma}}{|g'_q(u_{0,q})|}. \end{aligned}$$

Since p was arbitrarily small, the claim of the Theorem follows with r_n instead of R_n . Since c_n can converge to infinity arbitrarily slowly, the claim also holds with R_n instead of r_n . This completes the proof. \square

Lemma 4.3 Consider functions $g, h : [0, 1] \rightarrow \mathbb{R}$ and assume that for some $u_0 \in (0, 1)$ we have $g(u_0) = 0$. Additionally, assume that g is continuously differentiable in a neighbourhood $U_\delta(u_0) \subset (0, 1)$ and that its derivative is uniformly Hölder continuous of order γ with constant C_H . Define $\xi(\varepsilon) := \sup_{|u-u_0| \leq \varepsilon} |h(u_0) - h(u) - (g(u_0) - g(u))|$. Then for any $\varepsilon < \delta$ such that $|h(u_0)| + \frac{2(\xi(\varepsilon) + C_H \varepsilon^{1+\gamma})}{|g'(u_0)|} \leq \varepsilon$ we have for $[u_0 - \varepsilon, u_0 + \varepsilon] \subset [a, b]$

$$\left| \int_a^b I\{g(u) \leq 0\} du - \int_a^b I\{h(u) \leq 0\} du - |g'(u_0)|^{-1} h(u_0) \right| \leq \frac{2\xi(\varepsilon) + 4C_H \varepsilon^{1+\gamma}}{|g'(u_0)|}.$$

Proof. Rewrite

$$I\{h(u) \leq 0\} = I\{h(u_0) - (g(u_0) - g(u)) - (h(u_0) - h(u) - (g(u_0) - g(u)))\}$$

and observe that by the properties of g we have

$$\sup_{|u-u_0| \leq \varepsilon} |g(u) - g(u_0) - g'(u_0)(u - u_0)| \leq C_H \varepsilon^{1+\gamma}.$$

Thus the indicators $I\{h(u_0) + g'(u_0)(u - u_0) \leq 0, |u - u_0| \leq \varepsilon\}$ and $I\{h(u_0) - (g(u_0) - g(u)) - (h(u_0) - h(u) - (g(u_0) - g(u))), |u - u_0| \leq \varepsilon\}$ can only take different values on an interval with length at most $2(\xi(\varepsilon) + C_H \varepsilon^{1+\gamma})/|g'(u_0)|$. Thus we see that

$$\left| \int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{h(u) \leq 0\} du - \int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{h(u_0) + g'(u_0)(u - u_0) \leq 0\} du \right| \leq \frac{2(\xi(\varepsilon) + C_H \varepsilon^{1+\gamma})}{|g'(u_0)|}.$$

Recalling that $g(u_0) = 0$, similar arguments yield the bound

$$\left| \int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{g(u) \leq 0\} du - \int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{g'(u_0)(u - u_0) \leq 0\} du \right| \leq \frac{2C_H \varepsilon^{1+\gamma}}{|g'(u_0)|}.$$

Finally, a simple computation shows that under the assumption $|h(u_0)| + \frac{2(\xi(\varepsilon) + C_H \varepsilon^{1+\gamma})}{|g'(u_0)|} \leq \varepsilon$ we have

$$\int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{g'(u_0)(u - u_0) \leq 0\} du - \int_{u_0-\varepsilon}^{u_0+\varepsilon} I\{h(u_0) + g'(u_0)(u - u_0) \leq 0\} du = \frac{h(u_0)}{|g'(u_0)|}.$$

Thus the proof is complete. \square

Lemma 4.4 *Under the assumptions of Theorem 3.4 we have for any $r_n = o(h_n)$*

$$\sup_{\tau_w \in T^\delta} \sup_{|\tau - \tau_w| \leq r_n} \|\hat{\beta}_{\tau_w} - \hat{\beta}_\tau - (\bar{\beta}_{\tau_w} - \bar{\beta}_\tau)\|_\infty = O_P(n^{-1/2} h_n^{-3/2} r_n \log n) + o(h_n^k) + O_P(\kappa_n)$$

where κ_n was defined in Theorem 2.5.

Proof From Theorem 2.5 we know that $\hat{s}_\tau = s_0$ for all $\tau \in T^\delta$ with probability tending to one, and thus it suffices to find a bound for $\|\hat{b}_\tau - \bar{b}_\tau - (\hat{b}_{\tau_w} - \bar{b}_{\tau_w})\|_\infty$. Here we have for any $\tau, \tau_w \in T^\delta$

$$\begin{aligned} & \hat{b}_\tau - \bar{b}_\tau - (\hat{b}_{\tau_w} - \bar{b}_{\tau_w}) \\ &= -Q(s_0, \bar{b}_\tau, \tau)^{-1} \tilde{T}_n(s_0, \bar{b}_\tau, \tau) + Q(s_0, \bar{b}_{\tau_w}, \tau_w)^{-1} \tilde{T}_n(s_0, \bar{b}_{\tau_w}, \tau_w) + R_n(\tau) + R_n(\tau_w) \\ &= Q(s_0, \bar{b}_{\tau_w}, \tau_w)^{-1} \left(\tilde{T}_n(s_0, \bar{b}_{\tau_w}, \tau_w) - \tilde{T}_n(s_0, \bar{b}_\tau, \tau) \right) \\ & \quad + \left(Q(s_0, \bar{b}_{\tau_w}, \tau_w)^{-1} - Q(s_0, \bar{b}_\tau, \tau)^{-1} \right) \tilde{T}_n(s_0, \bar{b}_\tau, \tau) + R_n(\tau) + R_n(\tau_w). \end{aligned}$$

Combining condition (T2) with the results from Theorem 2.5 and Lemma 4.2 we see that the term in the last line is of order $O_P(((nh_n)^{-1/2} \log n + h_n^k)^{1+\gamma} + \kappa_n)$. For the first term, note that

$$\sup_{\tau_w \in T^\delta} \sup_{|\tau - \tau_w| \leq r_n} \left\| \tilde{T}_n(s_0, \bar{b}_{\tau_w}, \tau_w) - \tilde{T}_n(s_0, \bar{b}_\tau, \tau) \right\|_\infty \leq D_{1,n} + D_{2,n}$$

where

$$\begin{aligned} D_{1,n} &:= \sup_{j=1, \dots, d} \sup_{\|f-g\|_2 \leq h_n^{-1/2} r_n, f, g \in \mathcal{F}_{n,j}} h_n^{-1} n^{-1/2} |\mathbb{G}_n(f) - \mathbb{G}_n(g)|, \\ D_{2,n} &:= \sup_{\tau_w \in T^\delta} \sup_{|\tau - \tau_w| \leq r_n} \left\| \mathbb{E}[\tilde{T}_n(s_0, \bar{b}_{\tau_w}, \tau_w) - \tilde{T}_n(s_0, \bar{b}_\tau, \tau)] \right\|_\infty, \\ \mathbb{G}_n(f) &:= \frac{1}{n} \sum_{i=1}^n (f(X_i, Y_i, Z_i) - \mathbb{E}[f(X_i, Y_i, Z_i)]), \end{aligned}$$

and the classes of functions $\mathcal{F}_{n,j}$ are given by

$$\mathcal{F}_{n,j} = \left\{ (z, x, y) \mapsto K \left(\frac{x^T (\bar{b}_\tau + b) + s_0 z}{h_n} \right) x_j (y - (1 - \tau)) \mid b \in \mathbb{R}^d, \|b\| \leq \eta, \tau \in T^\delta \right\}.$$

In order to see that the representation for $D_{1,n}$ is true, note that

$$\begin{aligned} & \tilde{T}_n(s_0, \bar{b}_{\tau_w}, \tau_w) - \tilde{T}_n(s_0, \bar{b}_\tau, \tau) \\ &= \frac{1}{nh_n} \sum_{i=1}^n \left(K\left(\frac{X_i^T \bar{b}_\tau + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau)) - K\left(\frac{X_i^T \bar{b}_{\tau_w} + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau_w)) \right) X_i \end{aligned}$$

In particular we have for n large enough

$$\begin{aligned} & \left| K\left(\frac{X_i^T \bar{b}_\tau + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau)) - K\left(\frac{X_i^T \bar{b}_{\tau_w} + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau_w)) \right| \\ & \leq \sup_{|v| \leq 1} \left| K'\left(v + \frac{X_i^T \bar{b}_\tau + s_0 Z_i}{h_n}\right) \right| \frac{|X_i^T \bar{b}_\tau - X_i^T \bar{b}_{\tau_w}|}{h_n} + \sup_v |K(v)| |\tau - \tau_w| \end{aligned}$$

This shows that for sufficiently small ε the $\|\cdot\|_\infty$ -bracketing number [see van der Vaart and Wellner (1996), Chapter 2] of the class $\mathcal{F}_{n,j}$ is bounded by

$$\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}_{n,j}, \|\cdot\|_\infty) \leq Ch_n^{-(d+1)/\gamma} \varepsilon^{-(d+1)/\gamma}.$$

Moreover, the above bound implies that for $|\tau - \tau_w| \leq r_n$ we have for n large enough

$$\begin{aligned} & \sup_w \sup_{|\tau - \tau_w| \leq r_n} \left\| K\left(\frac{X_i^T \bar{b}_\tau + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau)) - K\left(\frac{X_i^T \bar{b}_{\tau_w} + s_0 Z_i}{h_n}\right) (Y_i - (1 - \tau_w)) \right\|_2 \\ & \leq Ch_n^{-1/2} r_n. \end{aligned}$$

Applying Lemma A.1 to the classes of functions $\{f - g | f, g \in \mathcal{F}_{n,j}, \|f - g\| \leq Ch_n^{-1/2} r_n\}$ thus shows that $D_{n,1} = O_P(n^{-1/2} h_n^{-3/2} r_n \log n)$.

Next, consider $D_{n,2}$. By the results in Lemma 4.2 we have

$$\begin{aligned} D_{n,2} &= \sup_t \left| \frac{\partial^k}{\partial u^k} \left(\left((\tau - F_{Y^*|X,Z}(0|x, s_0(-x^T \bar{b}_\tau + u))) f_{Z|X}(s_0(-x^T \bar{b}_\tau + u)|x) \right. \right. \right. \\ & \quad \left. \left. \left. - (\tau_w - F_{Y^*|X,Z}(0|x, s_0(-x^T \bar{b}_{\tau_w} + u))) f_{Z|X}(s_0(-x^T \bar{b}_{\tau_w} + u)|x) \right) \right) \Big|_{u=0} \right| + o(h_n^k). \end{aligned}$$

From condition (D4) we see that the left-hand side in the above expression is of order $o(h_n^k)$. Thus the proof is complete. \square

A Technical details

Lemma A.1 *Assume that the classes of measurable functions \mathcal{F}_n consist of uniformly bounded functions (by a constant not depending on n). If additionally*

$$N_{[\cdot]}(\mathcal{F}_n, \varepsilon, L^2(P)) \leq Cn^a \varepsilon^{-a}$$

for every $\varepsilon \leq \delta_n$, some $a > 0$ and constants C, b not depending on n , then we have for any $\delta_n \sim n^{-b}$ with $b < 1/2$

$$\sqrt{n} \sup_{f \in \mathcal{F}_n, \|f\|_{P,2} \leq \delta_n} \left(\int f dP_n - \int f dP \right) = O_P\left(\delta_n (|\log \delta_n| + \log n)\right).$$

Proof. Start by observing that the uniform boundedness of elements of \mathcal{F}_n by D implies that $F \equiv D$ is a measurable envelope function with L_2 -norm D . Note that for η_n sufficiently small

$$\begin{aligned} a(\eta_n) &:= \eta_n D / \sqrt{1 + \log N_{[]}(\eta_n D, \mathcal{F}_n, L_2(P))} \geq D \eta_n / \sqrt{1 + \log C + b \log n - a \log(D \eta_n)} \\ &\geq D \tilde{C} \eta_n / \sqrt{|\log \eta_n| + \log n} \end{aligned}$$

for some finite constant \tilde{C} depending only on a, b, C, D . Thus the bound in Theorem 2.14.2 in van der Vaart and Wellner (1996) yields for δ_n sufficiently small

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}_n} \int f d\alpha_n \right]^* &\leq D J_{[]}(\delta_n, \mathcal{F}_n, L_2(P)) + \sqrt{n} \int F(u) I\{F(u) > \sqrt{n} a(\delta_n)\} P(du) \\ &\leq DC_1 \int_0^{\delta_n} |\log \varepsilon| + \log n d\varepsilon + D \sqrt{n} I \left\{ D > \frac{D \tilde{C} \sqrt{n} \delta_n}{|\log \delta_n| + \log n} \right\} \\ &\leq DC_2 \delta_n (|\log \delta_n| + \log n) + D \sqrt{n} I \left\{ 1 > \frac{\tilde{C} \sqrt{n} \delta_n}{|\log \delta_n| + \log n} \right\} \end{aligned}$$

where $\alpha_n := \sqrt{n}(P_n - P)$, P_n denotes the empirical measure, and C_1, C_2 are some finite constants. Here, the second inequality follows by a straightforward calculation and the first inequality is due to the fact that for δ_n sufficiently small by definition

$$J_{[]}(\delta_n, \mathcal{F}_n, L_2(P)) = \int_0^{\delta_n} \sqrt{1 + \log N_{[]}(\varepsilon D, \mathcal{F}_n, L_2(P))} d\varepsilon \leq C_1 \int_0^{\delta_n} |\log \varepsilon| + \log n d\varepsilon.$$

Now under the assumption on δ_n , the indicator in the last line will be zero for n large enough and thus the proof is complete. \square

References

- Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.
- Coppejans, M. (2001). Estimation of the binary response model using a mixture of distributions estimator (mod). *Journal of Econometrics*, 102(2):231–269.
- Cosslett, S. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica: Journal of the Econometric Society*, pages 765–782.
- Dette, H., Neumeyer, N., and Pilz, K. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, 12(3):469–490.
- Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):609–627.
- Hardy, G., Littlewood, J., and Polya, G. (1988). *Inequalities*. Cambridge University Press.
- Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica: Journal of the Econometric Society*, pages 505–531.

- Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer Series in Statistics. Springer, New York.
- Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120.
- Khan, S. (2013). Distribution free estimation of heteroskedastic binary response models using probit/logit criterion functions. *Journal of Econometrics*, 172(1):168 – 182.
- Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, pages 191–219.
- Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421.
- Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.
- Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407.
- Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.
- Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.
- Manski, C. (1988). Identification of binary response models. *Journal of the American Statistical Association*, 83(403):729–738.
- Portnoy, S. (1998). Convergence rates for maximal score estimators in binary response regressions. *Asymptotic Methods in Probability and Statistics*, Editor: B. Szyszkowicz, Elsevier, Amsterdam, pages 775–783.
- Powell, J., Stock, J., and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York.