On Average Derivative Quantile Regression

Probal Chaudhuri[1], Indian Statistical Institute, Calcutta,

Kjell Doksum[2], University of California, Berkeley, and

Alexander Samarov[3], University of Massachusetts and MIT.

For fixed $\alpha \in (0,1)$, the quantile regression function gives the αth quantile $\theta_\alpha(\mathbf{x})$ in the conditional distribution of a response variable Y given the value $\mathbf{X} = \mathbf{x}$ of a vector of covariates. It can be used to measure the effect of covariates not only in the center of a population, but also in the upper and lower tails. A functional that summarizes key features of the quantile specific relationship between $\mathbf{X}$ and Y is the vector $\boldsymbol{\beta}_\alpha$ of weighted expected values of the vector of partial derivatives of the quantile function $\theta_\alpha(\mathbf{x})$. In a nonparametric setting, $\boldsymbol{\beta}_\alpha$ can be regarded as a vector of quantile specific nonparametric regression coefficients. In survival analysis models (e.g. Cox's proportional hazard model, proportional odds rate model, accelerated failure time model) and in monotone transformation models used in regression analysis, $\boldsymbol{\beta}_\alpha$ gives the direction of the parameter vector in the parametric part of the model. $\boldsymbol{\beta}_\alpha$ can also be used to estimate the direction of the parameter vector in semiparametric single index models popular in econometrics. We show that, under suitable regularity conditions, the estimate of $\boldsymbol{\beta}_\alpha$ obtained by using the locally polynomial quantile estimate of Chaudhuri (1991 *Annals of Statistics*), is $n^{1/2}$-consistent and asymptotically normal with asymptotic variance equal to the variance of the influence function of the functional $\boldsymbol{\beta}_\alpha$. We discuss how the estimate of $\boldsymbol{\beta}_\alpha$ can be used for model diagnostics and in the construction of a link function estimate in general single index models.

AMS 1991 *subject classifications.* Primary 62J02; secondary 62G99.
1. Partially supported by a grant from Indian Statistical Institute.
2. Partially supported by NSF Grant DMS-93-07403.
3. Partially supported by NSF Grant DMS-93–06245 and by a grant from IFSRC at MIT.
Key words and phrases. Average derivative estimate, transformation model, projection pursuit model, index model, survival analysis, heteroscedasticity, reduction of dimensionality, quantile specific regression coefficients.

1. Introduction. The quantile regression function is defined as the αth quantile $\theta_\alpha(\mathbf{x})$ in the conditional distribution $F_{Y|\mathbf{X}}(y|\mathbf{x})$ of a response variable Y given the value $\mathbf{X} = \mathbf{x}$ of a d-vector of covariates: for fixed α, $0 < \alpha < 1$, $\theta_\alpha(\mathbf{x}) = inf\{y : F_{Y|\mathbf{X}}(y|\mathbf{x}) \geq \alpha\}$. It has the advantage, over the commonly used mean regression, that by considering different α, it can be used to measure the effect of covariates not only in the center of a population, but also in the upper and lower tails. For instance, the effect of a covariate can be very different for high and low income groups. Thus, in the latest presidential election, the Democrats produced data showing that between 1980 and 1992, there was an increase in the number of people in the high salary category as well as the number of people in the low salary category. This phenomena could be demonstrated by computing the $\alpha = .90$ quantile regression function $\theta_{.90}(x)$ of salary Y as a function of the covariate $x =$ time and comparing it with the $\alpha = .10$ quantile regression function $\theta_{.10}(x)$. An increasing $\theta_{.90}(x)$ and a decreasing $\theta_{.10}(x)$ would correspond to the Democrats' hypothesis that "the rich got richer and the poor got poorer" during the Republican administration. The US Government yearly conducts a sample survey of about 60,000 households [the yearly Current Population Survey (CPS)] from which estimates of various quantiles can be obtained. Rose (1992) reported data for 1979 and 1989, and there the 10th percentile and the 90th percentile of the family income indeed show opposite trends over time. Recently Buchinsky (1994) have reported an extensive study of changes in US wage structure during 1963–1987 using linear parametric quantile regression. Similarly, in survival analysis, it is of interest to study the effect of a covariate on high risk individuals as well the effect on median and low risk individuals. Thus one can be interested in the quantiles $\theta_{.1}(\mathbf{x})$, $\theta_{.5}(\mathbf{x})$ and $\theta_{.9}(\mathbf{x})$ of the survival time Y given a vector $\mathbf{x}$ of covariates. Quantile regression is also useful in marketing studies as the influence of a covariate may be very different on individuals who belong to high, median and low consumption groups. Hendricks and Koenker (1992) studied variations in electricity consumption over time using some nonparametric quantile regression techniques.

1.1. Nonparametric quantile regression coefficients. Statistical literature frequently focuses on the estimation of the mean conditional response $\mu(\mathbf{x}) = E(Y|\mathbf{x})$. In linear statistical inference, the partial derivatives $\partial\mu(\mathbf{x})/\partial x_i$, where $\mathbf{x} = (x_1, \cdots, x_d)$, are assumed to be constant and are called regression coefficients. They are of primary interest since they measure how much the mean response is changed as the ith covariate is perturbed while other covariates are held fixed. However, this does not reveal dependence on the covariates in the lower and upper tails of the response distribution [see e.g. Efron (1991) for a detail discussion of this latter issue]. The quantile dependent regression coefficient curves can be defined as

$$\theta'_{\alpha i}(\mathbf{x}) = \partial\theta_\alpha(\mathbf{x})/\partial x_i, \quad i = 1, \ldots, d$$

which measure how much the αth response quantile is changed as the ith covariate is perturbed while the other covariates are held fixed. We consider the nonparametric setting where the gradient vector $\nabla\theta_\alpha(\mathbf{x}) = (\theta'_{\alpha 1}(\mathbf{x}), \ldots, \theta'_{\alpha d}(\mathbf{x}))$ is estimated using some appropriate smoothing technique, and we will focus on the average gradient vector

$$\boldsymbol{\beta}_\alpha = (\beta_{\alpha 1}, \ldots, \beta_{\alpha d}) = E(\nabla\theta_\alpha(\mathbf{X})).$$

The vector $\boldsymbol{\beta}_\alpha$, which gives a concise summary of quantile specific regression effects, will be called the vector of (*nonparametric*) *quantile regression coefficients*. Note that $\beta_{\alpha i}$ gives the average change in the quantile of the response as the ith covariate is perturbed while the other covariates are held fixed. Note also that in the linear model $Y = \sum_{j=1}^{d} \gamma_j X_j + \epsilon$, the vector $\boldsymbol{\beta}_\alpha$ coincides with the vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_d)$ of regression coefficients.

We next consider two examples which illustrate quantile specific regression effects when the covariate is real valued.

EXAMPLE 1.1. ¿From Bailar (1991), we get Table 1 which gives the first, middle and third quartiles of statistics professor salaries for the academic year 1991-92. Departments of Biostatistics and Colleges of Education were excluded. The explanatory variable x is the number of years in the rank of full professor. ¿From Table 1 and Figure 1, we see somewhat different trends over time in the three quartiles. Note that there is nonlinearity and some heteroscedasticity in this data set. Table 2 illustrates the quantile regression coefficient curves for $\alpha = .25,\ .5,\ .75$, and gives the estimated nonparametric quantile regression coefficients

$$(\hat{\beta}_{.25}, \hat{\beta}_{.5}, \hat{\beta}_{.75}) = (0.31, 0.67, 1.01),$$

computed as a weighted average of $\hat{\theta}'_\alpha(x)$ using the weights $\hat{p}(x)$, where the $\hat{p}(x)$ are the relative frequencies of data points in the bins indicated in the top rows. Again, these coefficients reveal a big difference in the effects of the covariate on the three quantiles.

Table 1. Quartiles of salaries (in thousands of dollars) of Statistics Professors 1991–1992. x is the number of years as Full Professor. n_x is the sample size. $\sum n_x = 469$.

x	2	5	8	11	14	17	20	23	25^+
$\hat{\theta}_{.25}(x)$	50.1	51.5	56.7	54.5	55.5	56.0	60.5	60.6	54.8
$\hat{\theta}_{.50}(x)$	54.0	62.2	63.8	61.5	62.8	69.0	70.9	66.9	62.2
$\hat{\theta}_{.75}(x)$	61.9	71.4	71.8	72.4	75.7	77.7	76.9	80.6	83.4
n_x	79	69	48	65	63	52	30	27	36

(Figure 1 around here)

Table 2. Quartile specific rates of change in salaries of Statistics Professors as seniority increases. $\hat{p}(x)$ is the proportion of people in the indicated category.

x	2-5	5-8	8-11	11-14	14-17	17-20	20-23	23-25^{+}	$\hat{\beta}_\alpha$
$\hat{\theta}'_{.25}(x)$	0.47	1.73	-0.73	0.33	0.17	1.5	0.03	-1.93	0.31
$\hat{\theta}'_{.50}(x)$	2.73	0.53	-0.77	0.43	2.07	0.63	-1.33	-1.57	0.67
$\hat{\theta}'_{.75}(x)$	3.17	0.13	0.20	1.1	0.67	-0.27	1.23	0.93	1.01
$\hat{p}(x)$	.18	.14	.14	.16	.14	.10	.07	.08	

EXAMPLE 1.2. We next consider a model where the quantile regression coefficient vector reveals interesting aspects of the relationship between $\mathbf{X}$ and Y in the tails of the response distribution as well as the center. Consider the heteroscedastic model

$$Y = \mu(\mathbf{X}) + \tau[\mu(\mathbf{X})]^{\lambda}\epsilon$$

where ϵ and $\mathbf{X}$ are independent, ϵ has continuous distribution function F_ϵ, the mean of ϵ is zero, and τ and λ are real parameters. The log normal and gamma regression models are of this form with $\lambda = 1$ and $\mu(\mathbf{x}) = \sum_{j=1}^{d} x_j\gamma_j$, while the Poisson regression model is of this form with $\lambda = \frac{1}{2}$, [cf. Carroll and Ruppert (1988), p.12]. Let e_α be an αth quantile of F_ϵ, then

$$\begin{aligned}
\theta_\alpha(\mathbf{x}) &= \mu(\mathbf{x}) + \tau[\mu(\mathbf{x})]^{\lambda} e_\alpha \\
\nabla\theta_\alpha(\mathbf{x}) &= \nabla\mu(\mathbf{x}) + \tau\lambda[\mu(\mathbf{x})]^{\lambda-1}\nabla\mu(\mathbf{x}) e_\alpha \\
\boldsymbol{\beta}_\alpha &= E(\nabla\mu(\mathbf{X})) + \tau\lambda E\{[\mu(\mathbf{X})]^{\lambda-1}\nabla\mu(\mathbf{X})\} e_\alpha
\end{aligned}$$

When $\lambda = 0$, the quantile regression coefficient vector $\boldsymbol{\beta}_\alpha$ is, for any fixed α, equivalent to the average derivative functional of Härdle and Stoker (1989). Note that this model gives dramatically different $\theta_\alpha(\mathbf{x})$, $\nabla\theta_\alpha(\mathbf{x})$ and $\boldsymbol{\beta}_\alpha$ for different α. For instance, if $F_\epsilon = \Phi$, the $N(0,1)$ distribution, $d = \tau = \lambda = 1$, and $\mu(x) = \gamma_1 + \gamma_2 x$, we have $\beta_\alpha = [1 + \Phi^{-1}(\alpha)]\gamma_2$. Thus the quantile regression coefficients turn out to be

$$\beta_{.1} = -0.282\gamma_2, \qquad \beta_{.5} = \gamma_2, \qquad \beta_{.9} = 2.282\gamma_2.$$

This model, with $\gamma_2 > 0$, nicely captures the "the rich get richer and the poor get poorer" hypothesis.

1.2. Survival analysis and transformation models. Many models in statistics, in particular in survival analysis, can be written in the form of a transformation model

$$h(Y) = \sum_{j=1}^{d} X_j\gamma_j + \epsilon, \tag{1.1}$$

where Y is survival time, $\mathbf{X} = (X_1, \cdots, X_d)$ is a vector of covariates, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_d)$ is a vector of regression coefficients, ϵ is a residual independent of $\mathbf{X}$, and h is an increasing function specific to the model being considered. For instance, Cox's proportional hazard model is of this form with $h(y) = \log\{-\log[1-F_0(y)]\}$, and there the distribution F_ϵ of ϵ is equal to the extreme value distribution $1-\exp\{-\exp\{t\}\}$. Here F_0 is an unknown continuous distribution function referred to as the baseline distribution: it is the distribution of Y when the γ_i's are all zero. Dabrowska and Doksum (1987) considered the estimation of $\theta_\alpha(\mathbf{x})$ in this model. Similarly, the proportional odds rate model is of the form (1.1) with $h(y) = \log[F_0(y)/\{1-F_0(y)\}]$ and $F_\epsilon =$ the logistic distribution $1/[1+\exp\{-t\}]$. See Doksum and Gasko (1990) for the details and history of these two and similar models. A third important survival analysis model of the form (1.1) is the accelerated failure time model where $h(y) = \log y$ and F_ϵ is unknown. In the above three models, the first two have unknown h and known F_ϵ, while the third has known h and unknown F_ϵ. Other models of the form (1.1) have parametric h and F_ϵ. For instance, Box and Cox (1964) and Bickel and Doksum (1981) have h equal to a power transformation and let F_ϵ depend on a scale parameter. Box and Cox consider normal F_ϵ while Bickel and Doksum consider robustness questions for more general F_ϵ.

We consider model (1.1) with both h and F_ϵ unknown, and assume that h is continuous and strictly monotone and F_ϵ is continuous. Since h is unknown, $\boldsymbol{\gamma}$ is only identifiable up to a multiplicative constant; in other words, only the direction of $\boldsymbol{\gamma}$ is identifiable. We drop the assumption that $\mathbf{X}$ and ϵ are independent and add instead a weaker assumption that the conditional quantile e_α of ϵ given $\mathbf{X} = \mathbf{x}$ does not depend on $\mathbf{x}$. Then, using the notation $g = h^{-1}$,

$$\theta_\alpha(\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\gamma} + e_\alpha) \quad \text{and} \quad \boldsymbol{\beta}_\alpha = c_\alpha\boldsymbol{\gamma}, \quad \text{where} \quad c_\alpha = E\left[g'(\mathbf{X}^T\boldsymbol{\gamma} + e_\alpha)\right] .$$

It follows that $\boldsymbol{\beta}_\alpha$ has the same direction as $\boldsymbol{\gamma}$, and we may without loss of generality estimate $\boldsymbol{\beta}_\alpha$. Note further that $(\beta_{\alpha i}/\beta_{\alpha j}) = (\gamma_i/\gamma_j)$ so that $\beta_{\alpha i}$ and $\beta_{\alpha j}$ give the relative importance of the covariates X_i and X_j.) One implication of this is that the coefficients in the Cox model can be given an interpretation similar to the usual intuitive idea of what regression coefficients are: the Cox regression coefficients give the average change in a quantile (e.g. median) survival time as the ith covariate is perturbed while the others are held fixed. The quantile regression vector $\boldsymbol{\beta}_\alpha$ is a unifying concept that represents the coefficient vectors in the standard linear model, the Cox model, the proportional odds rate model, the accelerated failure time model, etc.

REMARK 1.1. Let $\boldsymbol{\delta}_\alpha = \boldsymbol{\beta}_\alpha/|\boldsymbol{\beta}_\alpha|$, where $|\cdot|$ is the Euclidean norm. In model (1.1), $\boldsymbol{\delta}_\alpha = \boldsymbol{\delta}$ does not depend on α as long as ϵ and $\mathbf{X}$ are independent, and $\boldsymbol{\delta}$ represents the direction of $\boldsymbol{\gamma}$ so that estimates of $\boldsymbol{\delta}_\alpha$ obtained at grid points $\alpha_1, \cdots, \alpha_k$ can be combined into an estimate of $\boldsymbol{\delta}$ by computing their weighted average. Conversely, if $\boldsymbol{\delta}_{\alpha_1} \neq \boldsymbol{\delta}_{\alpha_2}$ for two different values of α, then the model (1.1) with $\mathbf{X}$ independent of ϵ does not hold, which suggests that the conditional quantile approach can also be used for model diagnostics (see Section 3).

REMARK 1.2. We obtain an estimating equation for $g = h^{-1}$ by introducing $Z = \sum_{j=1}^d X_j\beta_{\alpha j}$ and noting that, if we let $\xi_\alpha(Z)$ denote the αth quantile in the conditional distribution of Y given Z, then g can be expressed as

$$g(Z) = \xi_\alpha\left(c_\alpha(Z - e_\alpha)\right) ,$$

and we can estimate the "shape" of g and h using an estimate of the αth quantile function $\xi_\alpha(Z)$ (note that g is identifiable up to a location and scale transformation of its argument).

1.3. Reduction of dimensionality and single index models. Nonparametric estimation of the gradient vector $\nabla\theta_\alpha(\mathbf{x})$ is subject to the "curse of dimensionality" in the sense that accurate pointwise estimation is difficult with the sample sizes usually available in practice because of the sparsity of the data in subsets of R^d even for moderately large values of d. An important semiparametric regression class of models is projection pursuit regression, which has been used by a number of authors [e.g. Friedman and Tukey (1974), Huber (1985)] while analyzing high dimensional data in an attempt to cope with the "curse of dimensionality". The one term projection pursuit model, which gives the first step in projection pursuit regression, has the form

$$Y = g(\boldsymbol{\gamma}^T\mathbf{X}) + \epsilon, \tag{1.2}$$

where $\boldsymbol{\gamma}$ is a d-dimensional parameter vector (the projection vector), ϵ denotes random error, and g is a smooth real valued function of a real variable. Stigler (1986, pp. 283-290) pointed out that Francis Galton used a projection pursuit type analysis while computing "mid-parents' heights" in course of his analysis of the data on the heights of a group of parents and their adult children in the late 19th century. Note that when (1.2) holds, we must have $\theta_\alpha(\mathbf{x}) = g(\boldsymbol{\gamma}^T\mathbf{x}) + e_\alpha(\mathbf{x})$, where $e_\alpha(\mathbf{x})$ is the αth quantile in the conditional distribution of ϵ given $\mathbf{X} = \mathbf{x}$. Therefore, if $e_\alpha(\mathbf{x})$ is a constant free from $\mathbf{x}$ for some $0 < \alpha < 1$, the gradient vector $\nabla\theta_\alpha(\mathbf{x})$ will be equal to a scalar multiple of $\boldsymbol{\gamma}$ for all $\mathbf{x}$. Consequently, an estimate of $\boldsymbol{\beta}_\alpha$ gives an estimate of the projection direction $|\boldsymbol{\gamma}|^{-1}\boldsymbol{\gamma}$. Note that when the smooth function g is completely unspecified, only the direction of $\boldsymbol{\gamma}$ (and not its magnitude) is identifiable as in the transformation model (1.1).

In recent econometric literature, there is a considerable interest in the so called single index model [see, e.g., Han (1987), Powell, et al. (1989), Newey and Ruud (1991), Sherman (1993)] defined by

$$Y = \phi(\boldsymbol{\gamma}^T\mathbf{X}, \epsilon), \tag{1.3}$$

where ϵ is a random error independent of $\mathbf{X}$, and ϕ, which is a real valued function of two real variables, is typically assumed to be monotonic in both of its arguments. Duan and Li (1991) considered a very similar model in their regression analysis under link violations. They did not assume any monotonicity condition on the unknown link function ϕ. Their sliced inverse regression approach for estimating the direction of $\boldsymbol{\gamma}$ is applicable under the assumption of elliptic symmetry on the distribution of the regressor and the independence between $\mathbf{X}$ and ϵ. Härdle and Stoker (1989) and Samarov (1993) investigated procedures for estimating the direction of $\boldsymbol{\gamma}$ in (1.2) and (1.3), using estimates of the gradient of the conditional mean of Y given $\mathbf{X} = \mathbf{x}$. Their approach requires neither the elliptic symmetry of the regressors nor the monotonicity of ϕ. However, the use of the conditional mean of the response makes the procedure non-robust, and it does not allow for the estimation of the function ϕ in (1.3) (see Section 3 on the estimation of ϕ).

It is important to note that most of these earlier approaches require independence between the errors ϵ and the regressor $\mathbf{X}$, thus imposing a strong homoscedasticity condition.

The approach of this paper allows one to weaken this assumption and only requires that, for some $0 < \alpha < 1$, the αth conditional quantile $e_\alpha(\mathbf{x})$ is a constant free from $\mathbf{x}$, which is some kind of a centering assumption for the distribution of the error ϵ. It was considered, e.g., by Manski (1988) in the context of binary response models, who called this assumption quantile independence. Typically one would center the conditional distribution of the response at $\theta_{.5}(\mathbf{x})$, and in that case $e_{.5}(\mathbf{x})$ is assumed to be a constant free from $\mathbf{x}$, which can be taken as zero without loss of generality. This centering device allows one to work under possible dependence between the covariate $\mathbf{X}$ and the error ϵ.

Note that model (1.1) is a special case of model (1.3), and model (1.2) is not a special case of model (1.3) unless g is assumed to be monotonic. We will drop the assumption of monotonicity of ϕ with respect to its first argument and assume that ϕ is strictly increasing in its second argument. Note that this will cover (i) the regression model with product error $Y = e\psi(\boldsymbol{\gamma}^T\mathbf{X})$, where ψ is smooth and positive, (ii) the heteroscedastic one-term projection pursuit model $Y = g(\boldsymbol{\gamma}^T\mathbf{X}) + e\psi(\boldsymbol{\gamma}^T\mathbf{X})$, where g is smooth and ψ is smooth and positive, and (iii) the heteroscedastic one-term projection pursuit model with transformation

$$h(Y) = \psi_1(\boldsymbol{\gamma}^T\mathbf{X}) + e\psi_2(\boldsymbol{\gamma}^T\mathbf{X}),$$

where ψ_1 is smooth, h is smooth and monotonic, and ψ_2 is smooth and positive.

In model (1.3) with ϕ monotonic only in its second argument,

$$\theta_\alpha(\mathbf{x}) = \phi\{\boldsymbol{\gamma}^T\mathbf{x}, e_\alpha(\mathbf{x})\},$$

and if there exists $0 < \alpha < 1$ such that $e_\alpha(\mathbf{x})$ is a constant free from $\mathbf{x}$, $\nabla\theta_\alpha(\mathbf{x})$ will again be a scalar multiple of $\boldsymbol{\gamma}$ for all $\mathbf{x}$. Hence, an estimate for $\boldsymbol{\beta}_\alpha$ can be used to estimate the direction of $\boldsymbol{\gamma}$ in this case too.

The rest of the paper is organized as follows. In the next section we consider nonparametric estimation of the average gradient functional $\boldsymbol{\beta}_\alpha$. We report some results from a numerical study to illustrate the implementation of the methodology and discuss large sample statistical properties of the estimate of $\boldsymbol{\beta}_\alpha$ in detail. A discussion of efficiency, diagnostic applications, and estimation of the link function in model (1.3) are given in Section 3 while Section 4 contains the proofs.

2. Estimation and main results. Let $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ be n independent random vectors distributed as $(\mathbf{X}, Y)$, $\mathbf{X} \in R^d, Y \in R^1$. For fixed $0 < \alpha < 1$, let $\theta(\mathbf{x})$ be the conditional αth quantile of Y given $\mathbf{X} = \mathbf{x}$ and let $f(\mathbf{x})$ denote the density of $\mathbf{X}$. We want to estimate

$$\boldsymbol{\beta} = \int \{\nabla\theta(\mathbf{x})\}\, w(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x}\,, \tag{2.1}$$

where the dependence of $\theta(\mathbf{x})$ and $\boldsymbol{\beta}$ on α is suppressed as long as it does not cause an ambiguity, and $w(\mathbf{x})$ is a smooth weight function with a compact support within the interior of the support of $f(\mathbf{x})$.

The weight function is introduced to obtain functionals and estimates that are not overly influenced by outlying $\mathbf{x}$ values (high leverage points). It allows our functional $\boldsymbol{\beta}$ to focus on quantile dependent regression effects without being unduly influenced by the tail behaviour of $f(\mathbf{x})$. It also reduces boundary effects that occur in nonparametric smoothing. The weight function does not alter the fact that $\boldsymbol{\beta}$ has the same direction as $\boldsymbol{\gamma}$ in the single index model with an unknown monotonic link. In a more general nonparametric setting, we would recommend using a smooth weight function which equals one except in the extreme tails of the $\mathbf{X}$ distribution.

We will consider two estimators of $\boldsymbol{\beta}$. The first one is the direct plug-in estimator

$$\hat{\boldsymbol{\beta}}_1 = n^{-1} \sum \left\{ \nabla \hat{\theta}(\mathbf{X}_i) \right\} \; w(\mathbf{X}_i), \tag{2.2}$$

where $\nabla\hat{\theta}(\mathbf{X}_i)$ is a nonparametric estimator of the gradient of the conditional quantile $\theta(\mathbf{x})$ at $\mathbf{x} = \mathbf{X}_i$. The second estimator is based on the observation that, under the above assumptions on the weight function $w(\mathbf{x})$, integration by parts gives:

$$\boldsymbol{\beta} = -\int \theta(\mathbf{x}) \nabla \{w(\mathbf{x}) f(\mathbf{x})\} d\mathbf{x},$$

the sample version of which gives:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= -\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(\mathbf{X}_i) \frac{\nabla w(\mathbf{X}_i) \hat{f}(\mathbf{X}_i) + w(\mathbf{X}_i) \nabla \hat{f}(\mathbf{X}_i)}{\hat{f}(\mathbf{X}_i)} \\ &= -\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(\mathbf{X}_i) \{ \nabla w(\mathbf{X}_i) + w(\mathbf{X}_i) \hat{\ell}(\mathbf{X}_i) \}, \end{aligned} \tag{2.3}$$

where $\hat{\ell}(\mathbf{X}_i) = \nabla \hat{f}(\mathbf{X}_i) / \hat{f}(\mathbf{X}_i)$, and $\hat{f}$ and $\nabla \hat{f}$ are some nonparametric estimators of the density and its gradient. We will use here leave-one-out kernel estimators

$$\hat{f}(\mathbf{X}_i) = \frac{1}{(n-1) h_n^d} \sum_{j \neq i} W(\frac{\mathbf{X}_j - \mathbf{X}_i}{h_n}), \tag{2.4}$$

and

$$\nabla \hat{f}(\mathbf{X}_i) = \frac{1}{(n-1) h_n^{d+1}} \sum_{j \neq i} W^{(1)}(\frac{\mathbf{X}_j - \mathbf{X}_i}{h_n}), \tag{2.5}$$

where $W : R^d \to R^1$ and $W^{(1)} : R^d \to R^d$ are multivariate kernels for the density and its gradient, respectively, and h_n is a (scalar) bandwidth such that $h_n \to 0$ as $n \to \infty$. The bandwidth in $\nabla \hat{f}$ does not have to be same as that in $\hat{f}$ (cf. Lemma 4.3).

While various nonparametric estimators of conditional quantiles could be used in (2.2) and (2.3), including kernel, nearest neighbor, and spline estimators [see, e.g., Truong (1989), Bhattacharya and Gangopadhyay (1990), Dabrowska (1992), Koenker, et al (1992, 1994)], we will consider here the locally polynomial estimators [cf. Chaudhuri (1991a,b)]. The reason is that in order to develop asymptotic results for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, we need to consider local polynomials in d variables with arbitrary degrees, and Chaudhuri's results provide Bahadur-type expansions of estimators of $\theta(\mathbf{x})$ as well as of estimators of $\nabla\theta(\mathbf{x})$ which can be readily adapted for our purposes.

Consider a positive real sequence $\delta_n \to 0$, which will be chosen more explicitly later. Let $C_n(\mathbf{X}_i)$ be a cube in R^d centered at $\mathbf{X}_i$ with side legth $2\delta_n$, and let $S_n(\mathbf{X}_i)$ be the index set defined by

$$S_n(\mathbf{X}_i) = \{j : 1 \le j \le n, j \ne i, \mathbf{X}_j \in C_n(\mathbf{X}_i)\}, \text{ and } N_n(\mathbf{X}_i) = \#(S_n(\mathbf{X}_i)).$$

For $\mathbf{u} = (u_1, \ldots, u_d)$, a d-dimensional vector of nonnegative integers, set $[\mathbf{u}] = u_1 + \ldots + u_d$. Let A be the set of all d-dimensional vectors $\mathbf{u}$ with nonnegative integer components such that $[\mathbf{u}] \le k$ for some integer $k \ge 0$. Let $s(A) = \#(A)$ and let $\mathbf{c} = (c_{\mathbf{u}})_{\mathbf{u} \in A}$ be a vector of dimension $s(A)$. Also, given $\mathbf{X}_1, \mathbf{X}_2 \in R^d$, define $P_n(\mathbf{c}, \mathbf{X}_1, \mathbf{X}_2)$ to be the polynomial $\sum_{\mathbf{u} \in A} c_{\mathbf{u}}[(\mathbf{X}_1 - \mathbf{X}_2)/\delta_n]^{\mathbf{u}}$ (here, if $\mathbf{z} \in R^d$ and $\mathbf{u} \in A$, we set $\mathbf{z}^{\mathbf{u}} = \prod_{i=1}^d z_i^{u_i}$ with the convention that $0^0 = 1$). Let $\hat{\mathbf{c}}_n(\mathbf{X}_i)$ be a minimizer with respect to $\mathbf{c}$ of

$$\sum_{j \in S_n(\mathbf{X}_i)} \rho_\alpha\{Y_j - P_n(\mathbf{c}, \mathbf{X}_i, \mathbf{X}_j)\}, \tag{2.6}$$

where $\rho_\alpha(s) = |s| + (2\alpha - 1)s$. Since $0 < \alpha < 1$, $\rho_\alpha(s)$ tends to ∞ as $|s| \to \infty$, and so the above minimization problem always has a solution [see Chaudhuri (1991a, b) for more on the uniqueness and other properties of the solution of this minimization problem]. We now set $\hat{\theta}(\mathbf{X}_i) = \hat{c}_{n,0}(\mathbf{X}_i)$ and $\nabla\hat{\theta}(\mathbf{X}_i) = \hat{\mathbf{c}}_{n,1}(\mathbf{X}_i)/\delta_n$, where $\hat{c}_{n,0}(\mathbf{X}_i)$ and $\hat{\mathbf{c}}_{n,1}(\mathbf{X}_i)$ are the components of the minimizing vector of coefficients $\hat{\mathbf{c}}_n(\mathbf{X}_i)$ corresponding to the zero and first degree coefficients, respectively.

Note that (2.6) defines a leave-one-out estimator, i.e. $\hat{\mathbf{c}}_n(\mathbf{X}_i)$ does not involve Y_i. This simplifies the use of the conditioning argument at various places in the proofs in Section 4. It may be pointed out however that even if $\hat{\mathbf{c}}_n(\mathbf{X}_i)$ is allowed to involve all the data points including the ith one, the asymptotic behavior of the resulting estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ remains same. As a matter of fact, the leave-one-out and the non-leave-one-out versions of the estimates of $\boldsymbol{\beta}$ are asymptotically first order equivalent in the sense that their difference converges to zero at a rate faster than $n^{-1/2}$.

2.1. Some numerical results. We consider "Boston housing data" that has been analyzed by several statisticians in the recent past [see e.g. Doksum and Samarov (1995) for a recent analysis of the data and other related references]. There are $n = 506$ observations in the data set and the response variable (Y) is the median price of a house in a given area. We focus on three important covariates that are RM = average number of rooms per house in the area, $LSTAT$ = the percentage of population having lower economic status in the area and DIS = weighted distance to five Boston employment centers from houses of the area. One note-worthy feature of the data is that the Y-values larger or equal to \$50,000 have been recorded as \$50,000 (the data was collected in early 70's). Such a truncation in the upper tail of the response variable makes quantile regression, which is not influenced very much by extreme values of the response, a very appropriate methodology.

We computed normalized nonparametric quantile regression coefficients $\hat{\boldsymbol{\delta}}_\alpha = \hat{\boldsymbol{\beta}}_\alpha|\hat{\boldsymbol{\beta}}_\alpha|^{-1}$ using locally quadratic quantile regression. All covariates were standardized so that each of them has zero mean and unit variance. For weighted averaging, we used the weight function defined as : $w(z_1, z_2, z_3) = w_0(z_1)w_0(z_2)w_0(z_3)$, where $w_0(z) = 1$ if $|z| \le 2.4$, $w_0(z) = [1 - \{(z + 2.4)/0.2\}^2]^2$ if $-2.6 \le z \le -2.4$, $w_0(z) = [1 - \{(z - 2.4)/0.2\}^2]^2$ if

$2.4 \leq z \leq 2.6$ and $w_0(z) = 0$ for all other values of z. We considered estimation of $\boldsymbol{\beta}_\alpha$ with varying choices of the bandwidth δ_n in order to get a feeling for the effect of bandwidth selection on the resulting estimates. $\hat{\boldsymbol{\delta}}_\alpha$ was observed to be fairly stable with respect to different choices of the bandwidth δ_n as we tried 1.0, 1.2, and 1.4 as values for δ_n. Table 3 summarizes the results for $\delta_n = 1.2$. The local quadratic fit requires the local fitting of ten parameters. For three points near the boundary in $\mathbf{x}$ space with positive $w(\mathbf{x})$, there were not enough data points in the δ_n neighborhood to do a local quadratic fit. For these three points we doubled δ_n (see, e.g., Rice (1984) for a similar approach to the boundary problem).

Table 3. Normalized nonparametric quantile regression coefficients for "Boston housing data".

α	0.10	0.25	0.50	0.75	0.90
RM	0.438	0.443	0.533	0.553	0.505
$LSTAT$	-0.676	-0.848	-0.844	-0.814	-0.812
DIS	0.593	0.291	0.066	-0.178	-0.292

The following conclusions are immediate from the figures in Table 3. Firstly, $LSTAT$ appears to be the most important covariate for all percentile levels by comparing the absolute values of the normalized coefficients. This observation is in conformity with the findings reported in Doksum and Samarov (1995). Secondly, covariates do seem to have different effects on different percentiles of the conditional distribution of the response. In particular, the sign of the coefficient of DIS changes from positive to negative as we move from lower percentiles to upper ones.

2.2. Asymptotic behavior of the estimators. In this section we give results on the asymptotic behaviour of the estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. We find that by assuming certain smoothness conditions on $f(\mathbf{x})$ and $\theta(\mathbf{x})$ and by using local polynomials of sufficiently high degree, we can establish the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}), j = 1, 2$, in a nonparametric setting. Moreover, we show that $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ have the same influence function and this influence function equals the influence function of the functional $\boldsymbol{\beta}$, which indicates that, with additional regularity conditions, asymptotic nonparametric efficiency can be achieved. We also investigate how much efficiency $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ loose in parametric models by comparing them with the Koenker and Basset (1978) quantile regression estimator in a linear model, and find that the efficiency loss is small.

In what follows, the asymptotic relations such as $\mathbf{a} = O(1), o(1), O_p(1)$, or $o_p(1)$, applied to a vector $\mathbf{a}$, will be understood componentwise. We will also use notation $r_n(\mathbf{X}) = O_{L_2}(a_n)$ and $r_n(\mathbf{X}) = o_{L_2}(a_n)$, with a real sequence a_n, meaning that, as $n \to \infty$, $E(r_n(\mathbf{X})/a_n)^2$ is bounded and converges to zero, respectively.

Let V be an open convex set in R^d. We will say that a function $m : R^d \to R^1$ has the order of smoothness p on V with $p = l + \gamma$, where $l \geq 0$ is an integer and $0 < \gamma \leq 1$, and will write $m \in H_p(V)$, if (i) partial derivatives $D^{\mathbf{u}} m(\mathbf{x}) := \partial^{[\mathbf{u}]} m(\mathbf{x}) / \partial x_1^{u_1} \ldots \partial x_d^{u_d}$ exist and are continuous for all $\mathbf{x} \in V$ and $[\mathbf{u}] \leq l$. (ii) there exists a constant $C > 0$ such that

$$|D^{\mathbf{u}} m(\mathbf{x}_1) - D^{\mathbf{u}} m(\mathbf{x}_2)| \leq C|\mathbf{x}_1 - \mathbf{x}_2|^\gamma \text{ for all } \mathbf{x}_1, \mathbf{x}_2 \in V \text{ and } [\mathbf{u}] = l.$$

The orders of smoothness p_j, $j = 1, \ldots, 4$, in Conditions 1 through 4 below will be specified later.

Condition 1 : The marginal density $f(\mathbf{x})$ of $\mathbf{X}$ is positive on V and $f \in H_{p_1}(V)$.

Condition 2 : The weight function w is supported on a compact set with nonempty interior, $supp(w) \subset V$ and $w \in H_{p_2}(V)$.

Condition 3: The conditional density $f_{\epsilon|\mathbf{X}}(e|\mathbf{x})$ of $\epsilon = Y - \theta(\mathbf{X})$ given $\mathbf{X} = \mathbf{x}$, considered as a function of $\mathbf{x}$, belongs to $H_{p_3}(V)$ for all e in a neighborhood of zero (zero being the αth quantile of the conditional distribution). Further, the conditional density is positive for $e = 0$ for all values of $\mathbf{x} \in V$, and its first partial derivative w.r.t. e exists continuously for values of e in a neighborhood of zero for all $\mathbf{x} \in V$.

Condition 4 : The conditional αth quantile function $\theta(\mathbf{x})$ of Y given $\mathbf{X} = \mathbf{x}$ has the order of smoothness p_4 , i.e. $\theta(\mathbf{x}) \in H_{p_4}(V)$.

Condition 4 implies that for every $\mathbf{x} \in V$, $k = [p_4]$, and all sufficiently large n, $\theta(\mathbf{x} + \mathbf{t}\delta_n)$ can be approximated by the k-order Taylor polynomial

$$\theta_n^*(\mathbf{x} + \mathbf{t}\delta_n, \mathbf{x}) = \sum_{\mathbf{u} \in A} c_{n,\mathbf{u}}(\mathbf{x}) \mathbf{t}^{\mathbf{u}}, \tag{2.7}$$

with the coefficients $\mathbf{c}_{n,\mathbf{u}}(\mathbf{x}) = (\mathbf{u}!)^{-1} D^{\mathbf{u}} \theta(\mathbf{x}) \delta_n^{[\mathbf{u}]}$, where $\mathbf{u}! = u_1! \ldots u_d!$, and the remainder $r(\mathbf{t}\delta_n, \mathbf{x}) = \theta(\mathbf{x} + \mathbf{t}\delta_n) - \theta_n^*(\mathbf{x} + \mathbf{t}\delta_n, \mathbf{x})$ satisfies the inequality

$$|r(\mathbf{t}\delta_n, \mathbf{x})| \leq C(|\mathbf{t}|\delta_n)^{p_4}, \tag{2.8}$$

uniformly over $|\mathbf{t}| \leq 1$ and $\mathbf{x} \in V$.

Condition 5 : Let $k' \geq 1$ be an integer. a) The kernel $W : R^d \to R^1$ is a bounded continuous function with bounded variation on its support, which is contained in the unit cube $[-1, 1]^d$. Further, $W(\mathbf{t}) = W(-\mathbf{t})$, $\int W(\mathbf{t}) d\mathbf{t} = 1$, and

$$\int W(\mathbf{t}) \mathbf{t}^{\mathbf{u}} d\mathbf{t} = 0 \text{ for } [\mathbf{u}] \leq k'.$$

b) The components $W_\nu^{(1)}(\mathbf{t})$, $\nu = 1, \ldots, d$, of the kernel $W^{(1)} : R^d \to R^d$ are bounded continuous functions with bounded variation on their support (contained in $[-1, 1]^d$), $W_\nu^{(1)}(\mathbf{t}) = -W_\nu^{(1)}(-\mathbf{t})$, and

$$\int W_\nu^{(1)}(\mathbf{t}) \mathbf{t}^{\mathbf{u}} d\mathbf{t} = -\delta_{1[\mathbf{u}]} \delta_{1 u_\nu}$$

for $[\mathbf{u}] \leq k'$, where δ_{ab} is the Kroneker delta.

THEOREM. *Let γ be a real number in $(0, 1]$. For the "plug in" estimator $\hat{\boldsymbol{\beta}}_1$, assume that conditions 1, 2, and 3 hold with $p_1 = p_2 = p_3 = 1 + \gamma$, condition 4 holds with*

$p_4 > 3+3d/2$, that the order of the polynomial in (2.6) is $k = [p_4]$, and that the "bandwidth" δ_n in the definition (2.6) of the conditional quantile estimator is such that

$$\delta_n \asymp n^{-\kappa} \text{ with } \frac{1}{2(p_4 - 1)} < \kappa < \frac{1}{4+3d}. \tag{2.9}$$

For the "by parts" estimator $\hat{\boldsymbol{\beta}}_2$, assume that the conditions 1, 2, and 4 hold with $p_1 = p_2 = p_4 = p > 3 + 2d$ and condition 3 holds with $p_3 = \gamma$, and condition 5 holds with $k' = [p]$. Let q be a real number such that $3d/2 < q \le p$ and suppose that the order of the polynomial in (2.6) is $k = [q]$. Assume also that

$$\delta_n \asymp n^{-\kappa} \text{ with } \frac{1}{2q} < \kappa < \frac{1}{3d}, \tag{2.10}$$

and the bandwidth h_n of the kernel estimators (2.4), (2.5) is chosen such that

$$h_n \asymp n^{-\tau} \text{ with } \frac{1}{2(p-1)} \le \tau \le \frac{1}{4(d+1)}. \tag{2.11}$$

Then for $j = 1, 2$, as $n \to \infty$,

$$\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta} = \tag{2.12}$$

$$\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\nabla\theta(\mathbf{X}_i) - (\alpha - 1\{\epsilon_i \le 0\})\frac{\nabla w(\mathbf{X}_i) + w(\mathbf{X}_i)\ell(\mathbf{X}_i)}{f_{Y|\mathbf{X}}\{\theta(\mathbf{X}_i)|\mathbf{X}_i)\}} - \boldsymbol{\beta} + o_p(n^{-1/2}),$$

where $\epsilon_i = Y_i - \theta(\mathbf{X}_i)$, $\ell(\mathbf{X}) = \nabla f(\mathbf{X})/f(\mathbf{X})$, and $1\{\cdot\}$ is the indicator function.

REMARK 2.1. Note that the nonparametric estimates of the quantile surface $\theta_\alpha(\mathbf{x})$ and its derivative $\nabla\theta_\alpha(\mathbf{x})$ converge at a rate slower than $n^{-1/2}$. Their rates of convergence are quite slow when the number of covariates (i.e. the dimension of $\mathbf{X}$) is large. We obtain $n^{-1/2}$ rate of convergence for the estimate of the vector of quantile regression coefficients $\boldsymbol{\beta}$ even in a non-parametric setting. The "weighted averaging" of the derivative estimates leads to a concise summary of the quantile specific relationship between the response Y and the covariate $\mathbf{X}$ and enables us to escape the "curse of dimensionality" that occurs in non-parametric function estimation at least asymptotically. To achieve this, we need to assume in Condition 4 that the degree of smoothness p_4 of $\theta(\mathbf{x})$ grows with the dimensionality d, as required by Lemmas 4.1 and 4.3.

REMARK 2.2. Note that even though both estimators $\hat{\boldsymbol{\beta}}_j$, $j = 1, 2$, have the same asymptotic expansion, the first one needs less smoothness of the marginal density $f(\mathbf{x})$ and the weight function $w(\mathbf{x})$ in conditions 1 and 2, respectively. Also, the second one requires nonparametric estimation of $f(\mathbf{x})$ and its derivative. We hope to make a comparison of the finite sample performance of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ in terms of their mean square error in a separate paper.

3. Discussion.

Efficiency considerations. The theorem in Section 2 shows that the estimators $\hat{\boldsymbol{\beta}}_j$, $j = 1, 2$, are, using the terminology of Bickel, et al. (1994), asymptotically linear with the influence function

$$IF_\alpha(\mathbf{X}, Y) = w(\mathbf{X})\nabla\theta(\mathbf{X}) - (\alpha - 1\{\epsilon \le 0\})\frac{\nabla w(\mathbf{X}) + w(\mathbf{X})\ell(\mathbf{X})}{f_{Y|\mathbf{X}}(\theta(\mathbf{X})|\mathbf{X})} - \boldsymbol{\beta}, \tag{3.1}$$

and hence are asymptotically normal with covariance matrix $Var(IF_\alpha(\mathbf{X}, Y))$. A straightforward computation shows that $IF_\alpha(\mathbf{x}, y)$ is, in fact, the efficient influence function, i.e. it coincides with the influence function of the functional $\boldsymbol{\beta}$, so that Proposition 3.3.1 of Bickel, et al. (1994) implies that, under additional regularity conditions [such regularity conditions have been discussed in Newey and Stoker (1993)] guaranteeing pathwise differentiability of the functional $\boldsymbol{\beta}$, the estimators $\hat{\boldsymbol{\beta}}_j$, $j = 1, 2$, are asymptotically efficient in the class of regular estimators.

Note that the asymptotic efficiency of nonparametric estimators $\hat{\boldsymbol{\beta}}_j$ of the functional $\boldsymbol{\beta}$ does not imply their efficiency as estimates of the coefficients $\boldsymbol{\gamma}$ in the semiparametric models (1.1)-(1.3), cf. Klaassen (1992), Horowitz (1993), Klein and Spady (1993), Bickel and Ritov (1994). Example 3.1 below demonstrates that the loss in efficiency of our nonparametric estimates, when applied to some parametric models, may not be very large. Even though the estimators $\hat{\boldsymbol{\beta}}_j$ will not typically be fully efficient in specific parametric versions of models (1.1)-(1.3), the fact that they are $\sqrt{n}$ consistent means that they can serve as initial estimators for various "one-step" and other "improved" estimators in those models, see Klaasen (1992), Bickel, et al. (1994).

EXAMPLE 3.1. Consider the transformation model (1.1), where $\mathbf{X}$ and ϵ are independent, h is increasing and differentiable, and $\mathbf{X}$ is multivariate normal $N(\boldsymbol{\mu}, \Sigma)$. In this case $\nabla\theta_\alpha(\mathbf{x}) = \boldsymbol{\gamma}\{h'(\theta_\alpha(\mathbf{x}))\}^{-1}$, $\ell(\mathbf{x}) = -\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$ and $f_{Y|\mathbf{x}}(\theta_\alpha(\mathbf{x})|\mathbf{x}) = f_\epsilon(e_\alpha)h'(\theta_\alpha(\mathbf{x}))$, where e_α is the αth quantile of ϵ. We have from (2.12) that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_1$ (and $\hat{\boldsymbol{\beta}}_2$) is

$$\frac{\alpha(1-\alpha)}{nf_\epsilon^2(e_\alpha)} E\left\{\frac{-w(\mathbf{X})\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) + \nabla w(\mathbf{X})}{h'(\theta_\alpha(\mathbf{X}))}\right\}\left\{\frac{-w(\mathbf{X})\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) + \nabla w(\mathbf{X})}{h'(\theta_\alpha(\mathbf{X}))}\right\}^T$$

$$+ \ \boldsymbol{\gamma}\boldsymbol{\gamma}^T n^{-1} Var\left\{\frac{w(\mathbf{X})}{h'(\theta_\alpha(\mathbf{X}))}\right\}.$$

If we take $w(\mathbf{x})$ equal to one except in the extreme tails of the density of $\mathbf{X}$, then, to a very close approximation, this asymptotic variance-covariance matrix is equal to

$$\{\alpha(1-\alpha)/nf_\epsilon^2(e_\alpha)\}E[\{h'(\theta_\alpha(\mathbf{X}))\}^{-2}\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}] + \boldsymbol{\gamma}\boldsymbol{\gamma}^T n^{-1} Var\Big(\frac{1}{h'(\theta_\alpha(\mathbf{X}))}\Big).$$

In the case when $h(y) = y$, we have $\boldsymbol{\beta}_\alpha = \boldsymbol{\gamma}$ and this expression reduces to

$$\frac{\alpha(1-\alpha)}{nf_\epsilon^2(e_\alpha)}\Sigma^{-1},$$

which we recognize as the asymptotic variance-covariance matrix of the quantile regression estimate of the coefficient vector in the linear model, see Koenker and Basset (1978). This means that our estimator, which is constructed without knowing h, is nearly as efficient in this case as the Koenker-Basset estimator which uses the linearity of $h(y)$. We also note that for this model and the same weight function $w(\mathbf{x})$, the asymptotic variance-covariance matrix of the Härdle-Stoker estimator $\hat{\boldsymbol{\beta}}_{HS}$ of $E\{w(\mathbf{X})\nabla\mu(\mathbf{X})\} = \boldsymbol{\gamma}$ [recall that $\mu(\mathbf{X}) = E(Y|\mathbf{X})$] is equal to $\sigma_\epsilon^2 n^{-1}\Sigma^{-1}$. Therefore, the asymptotic efficiency of our estimator of $\boldsymbol{\gamma}$ relative to the Härdle-Stoker estimator is

$$\frac{\sigma_\epsilon^2 f_\epsilon^2(e_\alpha)}{\alpha(1-\alpha)},$$

which is equal to the relative asymptotic efficiency of the sample α-quantile estimator vs. the sample mean, which may be greater or less than one depending on α and the distribution of ϵ.

The choice of bandwidth. Note that the choices (2.9) and (2.10) of the bandwidth δ_n "undersmooth" compared to the optimal nonparametric function estimation bandwidth $\delta_n \asymp n^{-(2p+d)^{-1}}$ [cf. Chaudhuri (1991a,b)]. The "undersmoothing" is needed to make the bias of the estimators of the order $o(n^{-1/2})$; the variance attains the order $1/n$ because of the averaging over different $\mathbf{X}_i$'s. As long as the bandwidth δ_n satisfies conditions (2.9) or (2.10), the choice of bandwidth only has a second order effect on the mean squared error (MSE) of $\hat{\beta}_j$, $j = 1, 2$. In the case of average derivative estimation of $\boldsymbol{\gamma}$ in model (1.2), Härdle, Hart, Marron and Tsybakov (1992) and Härdle and Tsybakov (1993) have used the second order term in the MSE to obtain an expression for the asymptotically optimal bandwidth. Note that in their approach also, undersmoothing is needed to obtain the desired asymptotic results. Recently, Härdle, Hall and Ichimura (1993) have investigated simultaneous estimation of the optimal bandwidth and the vector $\boldsymbol{\gamma}$ in model (1.2).

Estimating the "link" functions in semiparametric models. Assume now that in the semiparametric models (1.1)-(1.3), for a given $0 < \alpha < 1$, the conditional α-quantile of ϵ given $\mathbf{X} = \mathbf{x}$ is constant in $\mathbf{x}$, i.e. $e_\alpha(\mathbf{x}) = e_\alpha$. Set $Z = \boldsymbol{\gamma}^T\mathbf{X}$, and denote by $\xi_\alpha(z)$ the conditional αth quantile of Y given $Z = z$. Then we have $\xi_\alpha(z) = h^{-1}(z + e_\alpha)$ in model (1.1), $\xi_\alpha(z) = g(z) + e_\alpha$ in model (1.2), and $\xi_\alpha(z) = \phi(z, e_\alpha)$ in model (1.3). So, after getting an estimate of the direction of $\boldsymbol{\gamma}$, one can project the observed $\mathbf{X}$'s on that estimated direction and then use those real valued projections to construct non-parametric estimates of h, g and ϕ in model (1.1), (1.2) and (1.3) respectively (keeping in mind the identifiability constraints in each of these models). This can be viewed as dimensionality reduction before constructing nonparametric estimates of the functional parameters in the models (1.1), (1.2) and (1.3). Under suitable regularity conditions, it is easy to construct an estimate $\hat{\xi}_\alpha(z)$ of $\xi_\alpha(z)$ that will converge at the rate $O_p(n^{-2/5})$, which is the usual rate for nonparametric pointwise estimation of a function of a single real variable. Properties of some nonparametric estimates of the conditional quantile function $\xi_\alpha(z)$ constructed following the above strategy will be investigated in detail in a separate paper. Note, however, that such estimates of $\xi_\alpha(z)$ are not necessarily monotonic and one needs to establish asymptotic results for isotonic versions of the estimates. Nonparametric estimates of an unknown monotone transformation in regression models similar to (1.1) can be found in Doksum (1987), Cuzick (1988), Horowitz (1993) and Ye and Duan (1994).

Model diagnostics. The nonparametric estimates of the average derivatives of conditional quantiles (or quantile regression coefficients) $\boldsymbol{\beta}_\alpha$ lead to some useful model diagnostic techniques [cf. related works on heteroscedasticity by Hendricks and Koenker (1992) and Koenker, et al. (1992)]. Note first that if conditions in Section 2 hold for several conditional quantiles $\boldsymbol{\theta}_{\alpha_1}(\mathbf{x}), \ldots, \boldsymbol{\theta}_{\alpha_k}(\mathbf{x})$, where $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_k < 1$, Theorem 2.1 implies that our estimates of $\boldsymbol{\beta}_{\alpha_1}, \ldots, \boldsymbol{\beta}_{\alpha_k}$ are jointly asymptotically normal. Using the asymptotic normal distribution of estimators of $\boldsymbol{\beta}_{\alpha_1}, \ldots, \boldsymbol{\beta}_{\alpha_k}$, we can construct asymptotic tests of their equality when $d = dim(X) = 1$, and thereby test homoscedasticity in such situations as mentioned in example 2 in Section 1.

In the models (1.1)-(1.3) in the presence of strong homoscedasticity, i.e. when ϵ and $\mathbf{X}$ are independent, $\nabla\boldsymbol{\theta}_\alpha(\mathbf{x})$ will be proportional to the parameter vector $\boldsymbol{\gamma}$ for all α and $\mathbf{x}$, and hence the estimated directions of $\nabla\boldsymbol{\theta}_\alpha(\mathbf{x})$'s for different values of α and $\mathbf{x}$ should be closely aligned, and so should be the estimates of quantile regression coefficients $\boldsymbol{\beta}_\alpha$ for different α's. Using again the joint asymptotic normality of the estimates of $\boldsymbol{\beta}_{\alpha_j}$ for $j = 1, \ldots, k$, we can construct asymptotic tests of homoscedasticity for the models (1.1)-(1.3) by testing the hypothesis of identical directions of $\boldsymbol{\beta}_{\alpha_j}$'s.

Further diagnostic information can be obtained by using nonparametric estimates of the $d \times d$ matrix functional

$$\Gamma_\alpha = Ew(\mathbf{X})\{\nabla\boldsymbol{\theta}_\alpha(\mathbf{X})\}\{\nabla\boldsymbol{\theta}_\alpha(\mathbf{X})\}^T, \tag{3.2}$$

which can be estimated in a way essentially similar to $\boldsymbol{\beta}_\alpha$ (asymptotic properties of the estimates of (3.2) will be considered in a separate paper). In particular, the validity of the single index models (1.3) can be tested by testing that the rank of Γ_α is one. More generally, Γ_α can be used to identify the linear subspace spanned by the vectors $\boldsymbol{\gamma}_j$, $j = 1, \ldots, k$ in the general dimensionality reduction (or multiple index) model $Y = G(\mathbf{x}^T\boldsymbol{\gamma}_1, \ldots, \mathbf{x}^T\boldsymbol{\gamma}_k, \epsilon)$ of Li (1991). Just note that, provided the function G is monotonic in ϵ and the αth conditional quantile of ϵ given $\mathbf{X}$ is free from $\mathbf{X}$, this subspace coincides with the subspace of those eigenvectors of Γ_α which have nonzero eigenvalues [cf. Samarov (1993)].

Further work. A number of important issues remains to be addressed: (i) The finite sample size performance of the estimators has to be investigated using Monte Carlo methods. This would include an investigation of bandwidth selection rules for the smoothers used in $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ as well as a comparison of the mean squared errors of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. (ii) Statistical properties of the estimates of the link function in models (1.1), (1.2), and (1.3) remain to be more fully investigated. In particular, the estimates of $\xi_\alpha(z)$ mentioned earlier in this section which converge at the rate $O_p(n^{-2/5})$ are not necessarily monotone. We need to establish asymptotic results for the isotonic versions of our estimators. (iii) While Example 3.1 suggests that the loss in efficiency of our nonparametric estimators, when applied to some parametric models, may be not very large, it is of interest to find out how close the asymptotic variance of $\hat{\boldsymbol{\beta}}_\alpha$ is to the asymptotic efficiency bounds in the semiparametric models (1.1), (1.2), and (1.3). (iv) In our examples of transformation model, we included some important models used in survival analysis. We are currently working on extending the results of this paper to censored data, see Dabrowska (1992).

4. Proofs. We will first prove three lemmas. The first lemma is an extension of the Bahadur type representation for the local polynomial conditional quantile estimators and their derivatives given in Theorem 3.3 in Chaudhuri (1991a), which is uniform in the conditioning variables and does not assume the independence between $\mathbf{X}$ and the residual $\epsilon = Y - \boldsymbol{\theta}(\mathbf{X})$.

Denote by $\mathbf{c}_n(\mathbf{x}) = (c_{n,\mathbf{u}}(\mathbf{x}))_{\mathbf{u}\in A}$ the $s(A)$- vector of Taylor coefficients in (2.7) and let $I(w) = \{i : \mathbf{X}_i \in \mathrm{supp}(w),\ i = 1, 2, \ldots, n\}$.

LEMMA 4.1. *Assume that the density of $\mathbf{X}$ is positive and continuous on V and the weight function w has a compact support in V. Then, under the conditions 3 with $p_3 = \gamma$, $\gamma > 0$, and condition 4 with $p_4 > 0$ and $k = [p_4]$, we have*

$$\hat{\mathbf{c}}_n(\mathbf{X}_i) - \mathbf{c}_n(\mathbf{X}_i) = \{N_n(\mathbf{X}_i)G_n(\mathbf{X}_i)\}^{-1} \tag{4.1}$$

$$\sum_{j=1, j\neq i}^{n} \mathbf{b}(\delta_n, \mathbf{X}_j - \mathbf{X}_i)(\alpha - 1\{Y_j \le \boldsymbol{\theta}_n^*(\mathbf{X}_j, \mathbf{X}_i)\})1\{|\mathbf{X}_j - \mathbf{X}_i| \le \delta_n\} + R_n(\mathbf{X}_i),$$

where the $s(A)$-vector

$$\mathbf{b}(\delta_n, \mathbf{X}_j - \mathbf{X}_i) = \{\delta_n^{-[\mathbf{u}]}(\mathbf{X}_j - \mathbf{X}_i)^{\mathbf{u}},\ [\mathbf{u}] \le k\} \tag{4.2}$$

has "naturally" ordered components, G_n is the $s(A) \times s(A)$ matrix

$$G_n(\mathbf{X}_i) = \{q_n^{\mathbf{u},\mathbf{v}} = \frac{\int_{[-1,1]^d} \mathbf{t}^{\mathbf{u}}\mathbf{t}^{\mathbf{v}} f_{\epsilon|\mathbf{X}}(0|\mathbf{X}_i + \delta_n\mathbf{t}) f(\mathbf{X}_i + \delta_n\mathbf{t})d\mathbf{t}}{\int_{[-1,1]^d} f(\mathbf{X}_i + \delta_n\mathbf{t})d\mathbf{t}},\ [\mathbf{u}] \le k, [\mathbf{v}] \le k\}, \tag{4.3}$$

$\boldsymbol{\theta}_n^*(\mathbf{x}_1, \mathbf{x}_2)$ is defined in (2.7), and the remainder term $R_n(\mathbf{X}_i)$ satisfies

$$\max_{i\in I(w)} |R_n(\mathbf{X}_i)| \; = \; O\left(n^{-3(1-\kappa d)/4}[\log n]^{3/4}\right) \text{ almost surely as } n \to \infty\ ,$$

provided that $\delta_n \asymp n^{-\kappa}$ with $1/(2p_4 + d) < \kappa < 1/d$.

REMARK 4.1. Under the conditions of Lemma 4.1, we have

$$(i) \max_{i\in I(w)} \delta_n^{-1}|R_n(\mathbf{X}_i)| \; = \; o(n^{-1/2}) \text{ almost surely as } n \to \infty\ ,$$

provided that $\delta_n \asymp n^{-\kappa}$ with

$$1/(2p_4 + d) < \kappa < 1/(4 + 3d) \tag{4.4}$$

and

$$(ii) \max_{i\in I(w)} |R_n(\mathbf{X}_i)| \; = \; o(n^{-1/2}) \text{ almost surely as } n \to \infty\ ,$$

provided that $\delta_n \asymp n^{-\kappa}$ with

$$1/(2p_4 + d) < \kappa < 1/3d. \tag{4.5}$$

The item (i) will be used for the "plug in" estimator $\hat{\boldsymbol{\beta}}_1$ and (ii) for the "by parts" estimator $\hat{\boldsymbol{\beta}}_2$.

Proof of Lemma 4.1 : The proof, which is based on modifications and extensions of the corresponding proofs in Chaudhuri (1991a, b), will be presented in steps. We will provide only the main ideas and skip technical details, which are fairly routine in view of the proofs already documented in Chaudhuri (1991a, b).

Step 1 : Let $\delta_n \asymp n^{-\kappa}$, where $0 < \kappa < (1/d)$, and for a pair of positive constants $c_1 < c_2$ define the event E_n as

$$E_n = \left\{ \; c_1 n^{1-\kappa d} \leq N_n(\mathbf{X}_i) \leq c_2 n^{1-\kappa d} \text{ for all } \mathbf{X}_i \in supp(w) \; \right\}$$

Then in view of the conditions assumed on the marginal density of $\mathbf{X}$ and the weight function w, it follows by a straight forward modification of the arguments used in the proofs of Theorem 3.1 in Chaudhuri (1991a) and Theorem 3.1 in Chaudhuri (1991b) that it is possible to choose the constants c_1 and c_2 so that

$$\Pr\left(\liminf E_n\right) \;=\; 1.$$

In fact, $\Pr(\limsup E_n^c)$ converges to zero at an exponential rate.

Step 2 : For a constant $K_1 > 0$, let F_n be the event defined as

$$F_n \;=\; \left\{ \; |\hat{\mathbf{c}}_n(\mathbf{X}_i) - \mathbf{c}_n(\mathbf{X}_i)| \leq K_1 n^{-(1-\kappa d)/2} (\log n)^{1/2} \;\text{ for all } \mathbf{X}_i \in supp(w) \; \right\}$$

and $\kappa > 1/(2p_4 + d)$. Once again, in view of the conditions assumed on the conditional density of the error ϵ given $\mathbf{X}$, simple modifications of the arguments used in the proofs of Theorem 3.2 in Chaudhuri (1991a) and Theorems 3.2 and 3.3 in Chaudhuri (1991b) yield the following. There exists a choice of K_1 such that

$$\Pr\left(\; \liminf F_n \;\right) \;=\; 1\;.$$

In fact, here also $\Pr(\; \limsup F_n^c \;)$ converges to zero at an exponential rate. Observe that Fact 6.5 in Chaudhuri (1991a) and Fact 5.2 in Chaudhuri (1991b), which play very crucial role, were stated in a set up in which the error ϵ and the regressor $\mathbf{X}$ are independent. However, as long as the conditional distribution of ϵ given $\mathbf{X}$ satisfies Condition 3 the main implication of those facts remain unaltered and they can be restated to serve our purpose.

Step 3 : Finally, some routine modifications and extensions of the argument used in the proof of Theorem 3.3 in Chaudhuri (1991a) exploiting Bernstein's inequality and Theorems 3.1 and 3.3 in Koenker and Bassett (1978) [see Facts 6.3 and 6.4 in Chaudhuri (1991a)] yield the following

$$\max_{i \in I(w)} |R_n(\mathbf{X}_i)| \;=\; O\left(n^{-3(1-\kappa d)/4} [\log n]^{3/4}\right) \;\text{ almost surely as } n \;\longrightarrow \infty\;,$$

provided that $\kappa > 1/(2p_4 + d)$. This completes the proof of Lemma 4.1.

Before stating Lemma 4.2 and its proof, we need to introduce some notations. Let Q be the $s(A) \times s(A)$ matrix with a typical entry $q_{\mathbf{u},\mathbf{v}} = \int_{[-1,1]^d} \mathbf{t}^{\mathbf{u}} \mathbf{t}^{\mathbf{v}} d\mathbf{t}$, where $\mathbf{u}, \mathbf{v} \in A$, cf. Chaudhuri (1991a,b). Denote by e_i, $1 \le i \le d$, the i-th column of the $d \times d$ identity matrix. Denote by Q_i^* the $s(A)\times s(A)$ matrix with a typical entry $q^*_{\mathbf{u},\mathbf{v},i} = \int_{[-1,1]^d} \mathbf{t}^{\mathbf{u}+\mathbf{v}+e_i} d\mathbf{t}$, $\mathbf{u},\mathbf{v} \in A$.

Note that matrices Q and Q_i^* can be written as

$$Q = \int_{[-1,1]^d} \mathbf{b}(1,\mathbf{t})\mathbf{b}(1,\mathbf{t})^T d\mathbf{t}. \tag{4.6}$$

and

$$Q_i^* = \int_{[-1,1]^d} t_i \mathbf{b}(1,\mathbf{t})\mathbf{b}(1,\mathbf{t})^T d\mathbf{t}, \tag{4.7}$$

respectively.

Set $p_n(\mathbf{X}) = \delta_n^d \int_{[-1,1]^d} f(\mathbf{X} + \delta_n \mathbf{t}) d\mathbf{t}$. Note that under condition 1 with $p_1 = 1+\gamma$, $\gamma > 0$, we have for $\mathbf{X} \in V$

$$\frac{p_n(\mathbf{X})}{(2\delta_n)^d} = f(\mathbf{X}) + O_{L_2}(\delta_n^{1+\gamma}). \tag{4.8}$$

We will denote by $f^{(i)}_{\epsilon,\mathbf{X}}(0,\mathbf{x})$ the first order partial derivative of $f_{\epsilon,\mathbf{X}}(0,\mathbf{x})$ w.r.t. the ith co-ordinate of $\mathbf{x}$.

LEMMA 4.2.
a) $max_{1\le i\le n}|N_n(\mathbf{X}_i) - np_n(\mathbf{X}_i)| = O((n\log n)^{1/2})$ *almost surely as* $n \longrightarrow \infty$.
b) If Conditions 1 and 3 hold with $p_1 = p_3 = 1+\gamma$, $\gamma > 0$, *we have for* $\mathbf{X} \in V$ *the following expansion*

$$\delta_n^d \{p_n(\mathbf{X}) G_n(\mathbf{X})\}^{-1} = \{f_{\epsilon,\mathbf{X}}(0,\mathbf{X})\}^{-1} Q^{-1} -$$

$$\delta_n \{f_{\epsilon,\mathbf{X}}(0,\mathbf{X})\}^{-2} \sum_{i=1}^d Q^{-1} Q_i^* Q^{-1} f^{(i)}_{\epsilon,\mathbf{X}}(0,\mathbf{X}) + r_n(\mathbf{X}), \tag{4.9}$$

where $r_n(\mathbf{X}) = O_{L_2}(\delta_n^{1+\gamma})$, *with* $O_{L_2}(\cdot)$ *interpreted here componentwise.*

Proof : Part a) follows immediately from Bernstein's inequality, since $np_n(\mathbf{X}_i) = E(N_n(\mathbf{X}_i)|\mathbf{X}_i)$.

To prove part b), note that the numerator of a typical entry of $G_n(\mathbf{X})$ (see (4.3)) has an expansion of the form

$$\int_{[-1,1]^d} \mathbf{t}^{\mathbf{u}} \mathbf{t}^{\mathbf{v}} f_{\epsilon,\mathbf{X}}(0, \mathbf{X} + \delta_n \mathbf{t}) d\mathbf{t} \;=\; f_{\epsilon,\mathbf{X}}(0,\mathbf{X}) q_{\mathbf{u},\mathbf{v}} \;+\; \delta_n \sum_{i=1}^d q^*_{\mathbf{u},\mathbf{v},i} f^{(i)}_{\epsilon,\mathbf{X}}(0,\mathbf{X}) \;+\; O_{L_2}(\delta_n^{1+\gamma}),$$

and use von Neumann expansion for the inverse matrix, see, e.g., Stuart and Sun (1990).

The next lemma will be used only for the "by parts" estimator $\hat{\boldsymbol{\beta}}_2$.

LEMMA 4.3. *a) Assume that the density* $f(\mathbf{x})$ *of* $\mathbf{X}$ *is positive and continuous on* V *and the weight function* w *has a compact support in* V. *Then, under condition 3 with* $p_3 = 1+\gamma$, $\gamma > 0$, *condition 4 with* $p_4 > d/2$, *and*

$$\delta_n \asymp n^{-\kappa}, \text{ with } 1/(2p_4 + d) < \kappa < 1/(2d), \tag{4.10}$$

we have

$$max_{i\in I(w)}|\hat{\boldsymbol\theta}_n(\mathbf{X}_i)-\boldsymbol\theta(\mathbf{X}_i)| = o_p(n^{-1/4}).$$

b) Under condition 1 with $p_1 > d$ *and condition (5) with* $k=[p_1]$, *and*

$$h_n \asymp n^{-\tau},\ \frac{1}{4p_1} < \tau < \frac{1}{4d}, \tag{4.11}$$

we have for the density estimator (2.4)

$$max_{i\in I(w)}|\hat f(\mathbf{X}_i) - f(\mathbf{X}_i)| = o_p(n^{-1/4}).$$

c) Under condition 1 with $p_1 > d+2$ *and condition 5 with* $k=[p_1]$, *and*

$$h_n \asymp n^{-\tau},\ \frac{1}{4(p_1-1)} < \tau < \frac{1}{4(d+1)}, \tag{4.12}$$

we have for the estimator (2.5)

$$max_{i\in I(w)}|\nabla\hat f(\mathbf{X}_i) - \nabla f(\mathbf{X}_i)| = o_p(n^{-1/4}).$$

Proof: Claim a) follows from Step 2 in the proof of Lemma 4.1.
Next, it follows from theorem 3.1.12, claim (i), in Prakasa Rao (1983) that

$$\sup_{\mathbf{x}\in V}|\hat f(\mathbf{x}) - E\hat f(\mathbf{x})| = O(\frac{(\log\log\ n)^{1/2}}{h_n^d n^{1/2}}),\ \text{almost surely as } n\to\infty.$$

Combining this result with $\sup_{\mathbf{x}\in V}|E\hat f(\mathbf{x}) - f(\mathbf{x})| = O(h_n^{p_1})$, which is obtained, under conditions 1 and 5, by applying the standard Taylor expansion argument [see, e.g., Lemma 1 in Samarov (1993)], and choosing h_n as in (4.11), we get the claim (b).

Applying the proof of claim (i) of theorem 3.1.12 from Prakasa Rao (1983) to the components of the vector $\nabla\hat f(\mathbf{x})$, we get

$$\sup_{\mathbf{x}\in V}|\nabla\hat f(\mathbf{x}) - E\nabla\hat f(\mathbf{x})| = O(\frac{(\log\log n)^{1/2}}{h_n^{(d+1)} n^{1/2}}), \text{almost surely as } n\to\infty. \tag{4.13}$$

Applying the argument of claim (b) to the components of $\nabla\hat f(\mathbf{x})$, we get $\sup_{\mathbf{x}\in V}|E\nabla\hat f(\mathbf{x}) - \nabla f(\mathbf{x})| = O(h_n^{p_1-1})$, which, together with (4.13) and (4.12), proves claim (c).

Proof of the Theorem for $\hat{\boldsymbol\beta}_1$. Setting $w_i = w(\mathbf{X}_i)$, we have

$$\hat{\boldsymbol\beta}_1 - \boldsymbol\beta = \frac{1}{n}\sum_{i=1}^n w_i(\nabla\hat{\boldsymbol\theta}(\mathbf{X}_i) - \nabla\boldsymbol\theta(\mathbf{X}_i)) + \frac{1}{n}\sum_{i=1}^n w_i\nabla\boldsymbol\theta(\mathbf{X}_i) - \boldsymbol\beta, \tag{4.14}$$

and to prove the asymptotic expansion (2.12) it is sufficient to obtain the corresponding scalar expansion for $\frac{1}{n}\sum_{i=1}^n w_i\mathbf{a}^T(\nabla\hat{\boldsymbol\theta}(\mathbf{X}_i) - \nabla\boldsymbol\theta(\mathbf{X}_i))$ with an arbitrary d-vector $\mathbf{a}$. For two positive constants $c_3 < c_4$ define the event $D_n(\mathbf{X}_i) = \{c_3 n\delta_n^d \le N_n(\mathbf{X}_i) \le c_4 n\delta_n^d\}$. Note that there exist appropriate choices for c_3 and c_4 such that $Pr\{\max_{1\leq i\leq n,\ \mathbf{X}_i\in supp(w)} 1\{D_n^c(\mathbf{X}_i)\} > 0\}$ converges to zero at an exponential rate in view of Bernstein's inequality, and we, therefore, have

$$\frac{1}{n}\sum_{i=1}^{n} w_i\mathbf{a}^T(\nabla\hat{\boldsymbol{\theta}}_n(\mathbf{X}_i) - \nabla\ \boldsymbol{\theta}(\mathbf{X}_i)) \tag{4.15}$$

$$= \frac{1}{n}\sum_{i=1}^{n} w_i\mathbf{a}^T(\nabla\hat{\boldsymbol{\theta}}_n(\mathbf{X}_i) - \nabla\ \boldsymbol{\theta}(\mathbf{X}_i))1\{D_n(\mathbf{X}_i)\} + o_p(n^{-1/2}).$$

Applying now Lemma 4.1 to the RHS of (4.15), we have

$$\frac{1}{n}\sum_{i=1}^{n} w_i\mathbf{a}^T(\nabla\hat{\boldsymbol{\theta}}_n(\mathbf{X}_i) - \nabla\ \boldsymbol{\theta}(\mathbf{X}_i))1\{D_n(\mathbf{X}_i)\} = \tag{4.16}$$

$$\frac{1}{n\delta_n}\sum_{i=1}^{n}\frac{w_i}{N_n(\mathbf{X}_i)}\mathbf{A}^T G_n^{-1}(\mathbf{X}_i)\sum_{j=1,j\neq i}^{n}\mathbf{b}(\delta_n,\mathbf{X}_j-\mathbf{X}_i)(\alpha - 1\{Y_j\leq\boldsymbol{\theta}_n^*(\mathbf{X}_j,\mathbf{X}_i)\})\times$$

$$1\{|\mathbf{X}_j-\mathbf{X}_i|\leq\delta_n\}1\{D_n(\mathbf{X}_i)\} + \frac{1}{n\delta_n}\sum_{i=1}^{n} w_i\mathbf{A}^T R_n(\mathbf{X}_i)1\{D_n(\mathbf{X}_i)\} + o_p(n^{-1/2}),$$

where the $s(A)$ vector $\mathbf{A}^T = (0,\mathbf{a}^T,0,\ldots,0)$ selects in the expansion (4.1) the terms corresponding to the first order partial derivatives of $\boldsymbol{\theta}$.

It follows from Remark 4.1 that, when δ_n is chosen as in (4.4),

$$\frac{1}{n\delta_n}\sum_{i=1}^{n} w_i\mathbf{A}^T R_n(\mathbf{X}_i)1\{D_n(\mathbf{X}_i)\} = o_p(n^{-1/2}). \tag{4.17}$$

We will next replace $\boldsymbol{\theta}_n^*(\mathbf{X}_j,\mathbf{X}_i)$ in the leading term in the RHS of (4.16) with $\boldsymbol{\theta}(\mathbf{X}_j)$, and will denote the resulting expression U_n^*. The error which results from this replacement is of the order $o_p(n^{-1/2})$ in view of the fact that the smallest eigenvalue of $G_n(\mathbf{x})$ is bounded away from zero uniformly over $\mathbf{x}\in supp(w)$ as $n\to\infty$, (2.8), and of the left inequality in (2.9).

Writing now U_n^* as

$$U_n^* = U_n + J_n,$$

where U_n is obtained from U_n^* by replacing $N_n(\mathbf{X}_i)$ with its conditional expectation $np_n(\mathbf{X}_i)$ and then dropping $1\{D_n(\mathbf{X}_i)\}$, we show that $J_n = o_p(n^{-1/2})$. Note first that, using part a) of Lemma 4.2 and (4.8), we have

$$max_{1\leq i\leq n} w_i\left\{\frac{1}{N_n(\mathbf{X}_i)} - \frac{1}{np_n(\mathbf{X}_i)}\right\}1\{D_n(\mathbf{X}_i)\} = O(n^{-3/2}\delta_n^{-2d}\sqrt{\log n}),$$

almost surely as $n\longrightarrow\infty$. Next, by Bernstein's inequality

$$max_{1\leq i\leq n}\sum_{j=1,j\neq i}^{n}\mathbf{b}(\delta_n,\mathbf{X}_j-\mathbf{X}_i)(\alpha - 1\{Y_j\leq\boldsymbol{\theta}(\mathbf{X}_j)\})1\{|\mathbf{X}_j-\mathbf{X}_i|\leq\delta_n\}1\{D_n(\mathbf{X}_i)\}$$

$$= O\left(\sqrt{n\delta_n^d \ \log n}\right),$$

almost surely as $n \longrightarrow \infty$. Since $|G_n^{-1}(\mathbf{x})|$ remains uniformly bounded for $\mathbf{x} \in supp(w)$ and $Pr\{\max_{1\le i\le n,\ \mathbf{X}_i\in supp(w)} 1\{D_n^c(\mathbf{X}_i)\} > 0\}$ goes to zero at an exponential rate, we obtain, using also (2.9), $J_n = o_p(n^{-1/2})$.

Observe now that U_n is a U-statistic with the kernel dependent on n:

$$U_n = \sum_{1\le i<j\le n} \xi_n(\mathbf{Z}_i,\mathbf{Z}_j), \text{ with } \mathbf{Z}_i = (\mathbf{X}_i, Y_i),\ \xi_n(\mathbf{Z}_i,\mathbf{Z}_j) = \eta_n(\mathbf{Z}_i,\mathbf{Z}_j) + \eta_n(\mathbf{Z}_j,\mathbf{Z}_i),$$

$$\eta_n(\mathbf{z}_1,\mathbf{z}_2) = \frac{1}{n\delta_n} w(\mathbf{x}_1)\mathbf{A}^T\{np_n(\mathbf{x}_1)G_n(\mathbf{x}_1)\}^{-1}\times$$
$$\mathbf{b}(\delta_n, \mathbf{x}_2 - \mathbf{x}_1)(\alpha - 1\{y_2 \le \boldsymbol{\theta}(\mathbf{x}_2)\})1\{|\mathbf{x}_2 - \mathbf{x}_1| \le \delta_n\},$$

where $\mathbf{z}_k = (\mathbf{x}_k, y_k)$, $k = 1, 2$.

To analyze U_n, we note first that the standard conditioning argument gives $E\xi_n(\mathbf{Z}_i,\mathbf{Z}_j) = E\eta_n(\mathbf{Z}_i,\mathbf{Z}_j) = 0$. The usual Hoeffding decomposition of U_n, [see, e.g., Serfling (1980)], now gives

$$E(U_n - P_n)^2 = \frac{n(n-1)}{2}\Big(E\xi_n^2(\mathbf{Z}_1,\mathbf{Z}_2) - 2Eg_n^2(\mathbf{Z})\Big) \le \frac{n(n-1)}{2}E\xi_n^2(\mathbf{Z}_1,\mathbf{Z}_2), \tag{4.18}$$

where P_n is the projection of U_n:

$$P_n = (n-1)\sum_{i=1}^n g_n(\mathbf{Z}_i), \tag{4.19}$$

and

$$g_n(\mathbf{z}) = E\xi_n(\mathbf{z},\mathbf{Z}) = E\eta_n(\mathbf{Z},\mathbf{z}). \tag{4.20}$$

We evaluate first $E\xi_n^2(\mathbf{Z}_1,\mathbf{Z}_2)$ in (4.18). Conditioning on $(\mathbf{X}_1,\mathbf{X}_2)$, we have:

$$E\xi_n^2(\mathbf{Z}_1,\mathbf{Z}_2) \le 4E\eta_n^2(\mathbf{Z}_1,\mathbf{Z}_2) = \frac{4}{n^2\delta_n^2}Ew^2(\mathbf{X}_1)(\alpha - 1\{Y_2 \le \boldsymbol{\theta}(\mathbf{X}_2)\})^2\times \tag{4.21}$$

$$(\mathbf{A}^T\{np_n(\mathbf{X}_1)G_n(\mathbf{X}_1)\}^{-1}\mathbf{b}(\delta_n, \mathbf{X}_2 - \mathbf{X}_1))^2 1\{|\mathbf{X}_2 - \mathbf{X}_1| \le \delta_n\} =$$
$$\frac{4\alpha(1-\alpha)}{n^2\delta_n^2}Ew^2(\mathbf{X}_1)(\mathbf{A}^T\{np_n(\mathbf{X}_1)G_n(\mathbf{X}_1)\}^{-1}\mathbf{b}(\delta_n, \mathbf{X}_2 - \mathbf{X}_1))^2 1\{|\mathbf{X}_2 - \mathbf{X}_1| \le \delta_n\}.$$

Applying (4.8), the fact that the smallest eigenvalue $G_n(\mathbf{x})$ is bounded away from zero, as $n \to \infty$, uniformly over $\mathbf{x} \in supp(w)$, and that each component of $\mathbf{b}(\delta_n, \mathbf{X}_2 - \mathbf{X}_1)1\{|\mathbf{X}_2 - \mathbf{X}_1| \le \delta_n\}$ is bounded by 1, we get $E\xi_n^2(\mathbf{Z}_1,\mathbf{Z}_2) = O(\frac{1}{n^4\delta_n^{d+2}})$, which together with (4.18) implies

$$E(U_n - P_n)^2 = O(\frac{1}{n^2\delta_n^{d+2}}),$$

and, hence, under (2.9),

$$U_n = P_n + o_p(\frac{1}{n^{1/2}}). \tag{4.22}$$

To complete the proof, we need to extract from the projection P_n in (4.19) the part which is free from n, i.e. to show that

$$V_n = Var\Big((n-1)\sum_{i=1}^n g_n(\mathbf{Z}_i) - \frac{(n-1)}{n^2}\sum_{i=1}^n(\alpha - 1\{Y_i \le \boldsymbol{\theta}(\mathbf{X}_i)\})\mathbf{a}^T M(\mathbf{X}_i)\Big) = o(n^{-1}), \tag{4.23}$$

where

$$M(\mathbf{x}) = -\nabla(\frac{w(\mathbf{x})f(\mathbf{x})}{f_{\epsilon,\mathbf{X}}(0,\mathbf{x})}) - \nabla f_{\epsilon,\mathbf{X}}(0,\mathbf{x})\frac{w(\mathbf{x})f(\mathbf{x})}{f^2_{\epsilon,\mathbf{X}}(0,\mathbf{x})} = -\frac{\nabla(w(\mathbf{x})f(\mathbf{x}))}{f_{\epsilon,\mathbf{X}}(0,\mathbf{x})}. \tag{4.24}$$

We have

$$V_n = \frac{n(n-1)^2}{n^4}Var\Big((\alpha - 1\{\epsilon_1 \le 0\})\times \tag{4.25}$$

$$[\frac{1}{\delta_n}\int w(\mathbf{x})\mathbf{A}^T\{p_n(\mathbf{x})G_n(\mathbf{x})\}^{-1}\mathbf{b}(\delta_n, \mathbf{X}_1 - \mathbf{x})1\{|\mathbf{X}_1 - \mathbf{x}| \le \delta_n\}f(\mathbf{x})d\mathbf{x} - \mathbf{a}^T M(\mathbf{X}_1)]\Big)$$

$$= \frac{(n-1)^2\alpha(1-\alpha)}{n^3}\times$$

$$E(\frac{1}{\delta_n}\int w(\mathbf{x})\mathbf{A}^T\{p_n(\mathbf{x})G_n(\mathbf{x})\}^{-1}\mathbf{b}(\delta_n, \mathbf{X}_1 - \mathbf{x})1\{|\mathbf{X}_1 - \mathbf{x}| \le \delta_n\}f(\mathbf{x})d\mathbf{x} - \mathbf{a}^T M(\mathbf{X}_1))^2.$$

Using now Lemma 4.2 and making a change of variables $\mathbf{x} = \mathbf{X}_1 - \mathbf{t}\delta_n$ in the integral in (4.25), we get

$$\frac{1}{\delta_n}\int w(\mathbf{x})\mathbf{A}^T\{p_n(\mathbf{x})G_n(\mathbf{x})\}^{-1}\mathbf{b}(\delta_n, \mathbf{X}_1 - \mathbf{x})1\{|\mathbf{X}_1 - \mathbf{x}| \le \delta_n\}f(\mathbf{x})d\mathbf{x}$$

$$= \frac{1}{\delta_n}\int_{[-1,1]^d} w(\mathbf{X}_1 - \mathbf{t}\delta_n)K(\mathbf{t})\frac{f(\mathbf{X}_1 - \mathbf{t}\delta_n)}{f_{\epsilon,\mathbf{X}}(0, \mathbf{X}_1 - \mathbf{t}\delta_n)}d\mathbf{t} \tag{4.26}$$

$$-\int_{[-1,1]^d}\frac{f(\mathbf{x}_1 - \mathbf{t}\delta_n)w(\mathbf{X}_1 - \mathbf{t}\delta_n)}{f^2_{\epsilon,\mathbf{X}}(0, \mathbf{X}_1 - \mathbf{t}\delta_n)}\sum_{i=1}^d L_i(\mathbf{t})f^{(i)}_{\epsilon,\mathbf{X}}(0, \mathbf{X}_1 - \mathbf{t}\delta_n)d\mathbf{t} + O_{L_2}(\delta_n^\gamma),$$

where $K(\mathbf{t}) = \mathbf{A}^T Q^{-1}\mathbf{b}(1,\mathbf{t})$ and $L_i(\mathbf{t}) = \mathbf{A}^T Q^{-1}Q_i^* Q^{-1}\mathbf{b}(1,\mathbf{t})$. Note that (4.6) implies that

$$\int_{[-1,1]^d} K(\mathbf{t})\mathbf{b}^T(1,\mathbf{t})d\mathbf{t} = \mathbf{A}^T, \tag{4.27}$$

i.e. $K(\mathbf{t})$ is a multivariate kernel of the order k for the first derivative. Similarly, (4.6) and (4.7) imply that

$$\int_{[-1,1]^d} L_i(\mathbf{t})\mathbf{b}^T(1,\mathbf{t})d\mathbf{t} = \mathbf{A}^T Q^{-1}Q_i^* = (a_i, 0, \ldots, 0, \mathbf{c}_k), i = 1, \ldots, d, \tag{4.28}$$

where a_i is the i-th component of the d-vector $\mathbf{a}$ and $\mathbf{c}_k$ is a vector filling in the components corresponding to the components of $\mathbf{b}^T(1,\mathbf{t})$ with the powers $\mathbf{t}^{\mathbf{u}}$ with $[\mathbf{u}]=k$. (4.28) means that the functions $L_i(\mathbf{t}),\ i=1,\dots,d$, are multivariate kernels of the order $k-1$.

We next expand the multiplier of $K(\mathbf{t})$ in (4.26) into the first order Taylor expansion and the multipliers of $L_i(\mathbf{t})$ into their zero order Taylor expansions, note that the remainders, in both cases, are of the order $O_{L_2}(\delta_n^\gamma)$, and apply (4.27) and (4.28): (4.23) then follows from (4.25) and (4.26).

Combining now all of the above results, we obtain the needed expansion for $\frac{1}{n}\sum_{i=1}^n w_i\mathbf{a}^T(\nabla\hat{\boldsymbol{\theta}}(\mathbf{X}_i)-\nabla\boldsymbol{\theta}(\mathbf{X}_i))$ with an arbitrary d-vector $\mathbf{a}$:

$$\frac{1}{n}\sum_{i=1}^n w_i\mathbf{a}^T(\nabla\hat{\boldsymbol{\theta}}(\mathbf{X}_i)-\nabla\boldsymbol{\theta}(\mathbf{X}_i))=\frac{1}{n}\sum_{i=1}^n(\alpha-1\{\epsilon_i\le 0\})\mathbf{a}^T M(\mathbf{X}_i))+o_p(n^{-1/2}),$$

which completes the proof of the Theorem for $\hat{\boldsymbol{\beta}}_1$.

Proof of the Theorem for $\hat{\boldsymbol{\beta}}_2$ is similar to that for $\hat{\boldsymbol{\beta}}_1$, and we will only indicate the differences. We have

$$\hat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}=-\frac{1}{n}\sum_{i=1}^n(\hat{\boldsymbol{\theta}}(\mathbf{X}_i)-\boldsymbol{\theta}(\mathbf{X}_i))(\nabla w_i+w_i l(\mathbf{X}_i))-\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\theta}}(\mathbf{X}_i)w_i(\hat{l}(\mathbf{X}_i)-l(\mathbf{X}_i))-$$
$$\frac{1}{n}\sum_{i=1}^n\boldsymbol{\theta}(\mathbf{X}_i)(\nabla w_i+w_i l(\mathbf{X}_i))+E\boldsymbol{\theta}(\mathbf{X})\frac{\nabla(w(\mathbf{X})f(\mathbf{X}))}{f(\mathbf{X})}, \tag{4.29}$$

and we need to obtain expansions for the first two sums in (4.29), which we will denote by I_1 and I_2, respectively. To obtain the expansion for I_1 we repeat the arguments given in the proof of Theorem 1 with the following modifications:

(i) w_i is replaced with the vector $\nabla w_i+w_i l(\mathbf{X}_i)$,

(ii) In (4.16) the factor $1/\delta_n$ is dropped and the vector $\mathbf{A}^T$ is replaced with the $s(A)$-vector $\mathbf{A}^T=(1,0,\dots,0)$ and the d-vector $\mathbf{a}^T$ becomes $(1,0,\dots,0)$.

(iii) Lemma 4.1 is applied with $k=[q]$ instead of $k=[p_4]$ and, accordingly, q replaces p_4 in (4.5).

(iv) The kernel $K(\mathbf{t})$ becomes here a k-order kernel for the function itself and not for its derivative, and only the first term of the expansion in part (b) of Lemma 4.2 is used in (4.26), so that the kernels L_i do not appear at all.

(v) The function $M(\mathbf{x})$ in (4.24) here becomes $M(\mathbf{x})=(\nabla w(\mathbf{x})+w(\mathbf{x})l(\mathbf{x}))/f_{\epsilon|\mathbf{X}}(0|\mathbf{x})$.

With these modifications, we obtain, using (2.10),

$$I_1=-\frac{1}{n}\sum_{i=1}^n(\alpha-1\{\epsilon_i\le 0\})\frac{\nabla w(\mathbf{X}_i)+w(\mathbf{X}_i)l(\mathbf{X}_i)}{f_{\epsilon|\mathbf{X}}(0|\mathbf{X}_i)}+o_p(n^{-1/2}). \tag{4.30}$$

For I_2 we have, using lemma 4.3 and the assumption (2.11),

$$I_2=-\frac{1}{n}\sum_{i=1}^n\boldsymbol{\theta}(\mathbf{X}_i)w_i(\nabla\hat{f}(\mathbf{X}_i)-l(\mathbf{X}_i)\hat{f}(\mathbf{X}_i))/f(\mathbf{X}_i)-\frac{1}{n}\sum_{i=1}^n(\hat{\boldsymbol{\theta}}_n(\mathbf{X}_i)-\boldsymbol{\theta}(\mathbf{X}_i))w_i(\hat{l}(\mathbf{X}_i)-l(\mathbf{X}_i))-$$

$$(4.31) \qquad \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}(\mathbf{X}_i)w_i(\nabla\hat{f}(\mathbf{X}_i)-l(\mathbf{X}_i)\hat{f}(\mathbf{X}_i))(f(\mathbf{X}_i)-\hat{f}(\mathbf{X}_i))/(\hat{f}(\mathbf{X}_i)f(\mathbf{X}_i)) =$$

$$-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}(\mathbf{X}_i)w_i(\nabla\hat{f}(\mathbf{X}_i)-l(\mathbf{X}_i)\hat{f}(\mathbf{X}_i))/f(\mathbf{X}_i)+o_p(n^{-1/2}).$$

Plugging into (4.31) the expressions (2.4) and (2.5) for $\hat{f}(\mathbf{X}_i)$ and $\nabla\hat{f}(\mathbf{X}_i)$, we see that the leading term in the RHS of (4.31) is a U-statistic with the kernel $\tilde{U}_n$ dependent on n:

$$\tilde{U}_n = \sum_{1\le i<j\le n}\tilde{\eta}_n(\mathbf{X}_i,\mathbf{X}_j)+\tilde{\eta}_n(\mathbf{X}_j,\mathbf{X}_i),$$

where

$$\tilde{\eta}_n(\mathbf{X}_i,\mathbf{X}_j) = -\frac{\boldsymbol{\theta}(\mathbf{X}_i)w_i}{n(n-1)h_n^d f(\mathbf{X}_i)}\Big(\frac{1}{h_n}W^{(1)}\Big(\frac{\mathbf{X}_j-\mathbf{X}_i}{h_n}\Big)-l(\mathbf{X}_i)W\Big(\frac{\mathbf{X}_j-\mathbf{X}_i}{h_n}\Big)\Big).$$

The mean of the kernel $\mu_n = E\tilde{\eta}_n(\mathbf{X}_1,\mathbf{X}_2) = E\tilde{\eta}_n(\mathbf{X}_2,\mathbf{X}_1)$ is

$$(4.32) \qquad \mu_n = EE(\tilde{\eta}_n(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{X}_1) =$$

$$E\frac{\boldsymbol{\theta}(\mathbf{X}_1)w(\mathbf{X}_1)}{n(n-1)h_n^d f(\mathbf{X}_1)}E\Big(\Big(\frac{1}{h_n}W^{(1)}\Big(\frac{\mathbf{X}_2-\mathbf{X}_1}{h_n}\Big)-l(\mathbf{X}_1)W\Big(\frac{\mathbf{X}_2-\mathbf{X}_1}{h_n}\Big)\Big)\Big|\mathbf{X}_1\Big) =$$

$$E\frac{\boldsymbol{\theta}(\mathbf{X}_1)w(\mathbf{X}_1)}{n(n-1)f(\mathbf{X}_1)}\int_{[-1,1]^d}\Big(\frac{1}{h_n}W^{(1)}(\mathbf{t})-l(\mathbf{X}_1)W(\mathbf{t})\Big)f(\mathbf{X}_1+\mathbf{t}h_n)d\mathbf{t}.$$

Using now the usual Taylor expansion argument [see, e.g., Lemma 1 in Samarov (1993)] and conditions (1) and (5), we obtain

$$(4.33) \qquad \mu_n = O\Big(\frac{h_n^{p-1}}{n(n-1)}\Big).$$

The projection of $\tilde{U}_n$ is

$$(4.34) \qquad \tilde{P}_n = (n-1)\sum_{i=1}^{n}(\tilde{g}_n(\mathbf{X}_i)-2\mu_n),$$

with $\tilde{g}_n(\mathbf{x}) = E\tilde{\eta}_n(\mathbf{x},\mathbf{X})+E\tilde{\eta}_n(\mathbf{X},\mathbf{x})$. Repeating the argument given in (4.32), (4.33), we get for the first term $E\tilde{\eta}_n(\mathbf{x},\mathbf{X}) = O(h_n^{p-1}/(n(n-1)))$ uniformly over $\mathbf{x}\in supp(w)$, while its second term is

$$E\tilde{\eta}_n(\mathbf{X},\mathbf{x}) = -\frac{1}{n(n-1)h_n^d}E\Big(\frac{\boldsymbol{\theta}(\mathbf{X})w(\mathbf{X})}{f(\mathbf{X})}\Big(\frac{1}{h_n}W^{(1)}\Big(\frac{\mathbf{x}-\mathbf{X}}{h_n}\Big)-l(\mathbf{X})W\Big(\frac{\mathbf{x}-\mathbf{X}}{h_n}\Big)\Big)\Big).$$

Relying here again on the same Taylor expansion and higher order kernel argument as in (4.32), (4.33), and using conditions (1), (2), (4), and (5), we obtain from (4.34), using again (4.33),

$$\tilde{P}_n = \frac{1}{n}\sum_{i=1}^{n}\Big(\nabla(w(\mathbf{X}_i)\boldsymbol{\theta}(\mathbf{X}_i)) + w(\mathbf{X}_i)\boldsymbol{\theta}(\mathbf{X}_i)l(\mathbf{X}_i)\Big) + O_p(h_n^{p-1}). \tag{4.35}$$

As in case of (4.18)-(4.21), we have

$$E(\tilde{U}_n - \tilde{P}_n)^2 \le 2n(n-1)E\tilde{\eta}_n^2(\mathbf{X}_1,\mathbf{X}_2) + (\binom{n}{2}\mu_n)^2 =$$

$$E(\frac{\boldsymbol{\theta}(\mathbf{X}_1)w(\mathbf{X}_1)}{f(\mathbf{X}_1)})^2\frac{1}{n(n-1)h_n^{2d}}\int(\frac{1}{h_n}W^{(1)}(\frac{\mathbf{x}-\mathbf{X}_1}{h_n}) - l(\mathbf{X}_1)W(\frac{\mathbf{x}-\mathbf{X}_1}{h_n}))^2 f(\mathbf{x})d\mathbf{x} + O(h_n^{2(p-1)}) \le$$

$$\frac{1}{n(n-1)h_n^d}E(\frac{\boldsymbol{\theta}(\mathbf{X}_1)w(\mathbf{X}_1)}{f(\mathbf{X}_1)})^2\int_{[-1,1]^d}(\frac{1}{h_n}W^{(1)}(\mathbf{t}) - l(\mathbf{X}_1)W(t))^2 f(\mathbf{X}_1 + \mathbf{t}h_n)d\mathbf{t} + O(h_n^{2(p-1)})$$

$$= O(\frac{1}{n^2 h_n^{d+2}}) + O(h_n^{2(p-1)}). \tag{4.36}$$

Choosing now h_n as in (2.11), we obtain, combining (4.31) through (4.36),

$$I_2 = \frac{1}{n}\sum_{i=1}^{n}\nabla(w(\mathbf{X}_i)\boldsymbol{\theta}(\mathbf{X}_i)) + w(\mathbf{X}_i)\boldsymbol{\theta}(\mathbf{X}_i)l(\mathbf{X}_i) + o_p(n^{-1/2}),$$

which together with (4.29) and (4.30) completes the proof of Theorem for $\hat{\boldsymbol{\beta}}_2$.

Acknowledgments. The authors thank Hank Ibser for programming and computing the quantile regression coefficients. We thank an Associate Editor and two referees for their useful comments and suggestions.

References.

Bailar, B.A. (1991). Salary survey of U.S. colleges and universities offering degrees in statistics. *Amstat News* No. 182, page 3.

Bhattacharya, P.K. and Gangopadhyay, A. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist.* **18** 1400-1415.

Bickel, P.J. and Doksum, K.A. (1981). An analysis of transformation revisited. *J. Amer. Statist. Assoc.* **76** 296-311.

Bickel, P.J., Klaassen, C., Ritov, Y. and Wellner, J. (1994). *Efficient and Adaptive Estimation for Semiparametric Models.* John Hopkins Univ. Press, Baltimore.

Bickel, P.J. and Ritov, Y. (1994). Local asymptotic normality of ranks and covariates in transformation models. *Le Cam Volume.*

Box, G.E.P, and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211-252.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963–1987 : application of quantile regression. *Econometrica.* **62** 405-458.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

Chaudhuri, P. (1991a). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.* **19** 760-777.

Chaudhuri, P. (1991b). Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Multivariate Aanalysis* **39** 246-269.

Cuzick, J. (1988). Rank regression. *Ann. Statist.* **16** 1369-1389.

Dabrowska, D. (1992). Nonparametric quantile regression with censored data. *Sankhyā, Ser. A*, **54** 252-259.

Dabrowska, D. and Doksum, K. (1987). Estimates and confidence intervals for median and mean life in the proportional hazard model. *Biometrika* **74** 799-807.

Doksum, K. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15** 325-345.

Doksum, K. and Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review* **58** 243-252.

Doksum, K. and Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression. *The Ann. Statist.* **23** xxx-xxx.

Duan, N. and Li, K-C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19** 505-530.

Efron, B. (1991). Regression percentiles using asymmetric square error loss. *Statistica Sinica* **1** 93-125.

Friedman, J. and Tukey, J. (1974). A projection pursuit algorithm for exploratory data anlysis. *IEEE Transactions on Computers* **C-23** 881-889.

Han, A. (1987). A non-parametric analysis of transformations. *J. of Econometrics* **35** 191-209.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157-178.

Härdle, W., Hart, J., Marron, J. S. and Tsybakov, A. B. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87** 218-226.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986-995.

Härdle, W. and Tsybakov, A.B. (1993). How sensitive are average derivatives? *J. of Econometrics* **58**, 31-48.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline model for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.* **87** 58-68.

Horowitz, J.L. (1993). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Tech Report., Dept. of Statistics, Univ. of Iowa.

Huber, P. (1985). Projection pursuit. *Ann. of Statist.* **13** 435-525.

Klaassen, C. (1992). Efficient estimation in the Clayton-Cuzick model for survival data. Tech Report., Univ. of Amsterdam.

Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387-421.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica* **46** 33-50.

Koenker, R., Portnoy, S., and Ng, P. (1992). Nonparametric estimation of conditional quantile function. Proc. of the conference on L_1-Statistical Analysis and Related Methods (Y. Dodge, ed.), 217-229, Elsevier.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673-680.

Li, K-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316-342.

Manski, C. (1988). Identification of binary response models. *J. Amer. Stat. Assoc.* **83** 729-738.

Newey, W.K. and Ruud, P.A. (1994). Density weighted linear least squares. Working Paper No. 228. Dept. of Econ., U. C. Berkeley.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* **61** 1199-1223.

Powell, J., Stock, J., and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403-1430.

Prakasa Rao, B.L.S. (1983). *Nonparametric Function Estimation.* Academic Press.

Rice, J. (1984). Boundary modification for kernel regression. *Comm. in Statist.* 1 **13** 893-900.

Rose, S.J. (1992). *Social Stratification in the United States.* The New Press, New York.

Samarov, A. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Stat. Assoc.* **88** 836-849.

Serfling, R. (1980). *Approximation theorems of mathematical statistics.* John Wiley and Sons, New York.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61** 123-137.

Stigler, S.M. (1986). The History of Statistics: *The Measurement of Uncertainty before* 1900. Belknap Press, Cambridge, MA.

Stuart, G. and Sun, J. (1990). *Matrix Perturbation Theory.* Academic Press, New York.

Truong, Y. (1989). Asymptotic properties of kernel estimates based on local medians. *Ann. Statsit.* 17 606-617.

Ye, J. and Duan, N. (1994). Nonparametric $n^{-\frac{1}{2}}$ consistent estimation for the general transformation models. Tech Report No. 138, Grad. School of Bus., Univ. of Chicago.

Division of Theoretical Statistics & Mathematics
Indian Statistical Institute
203 B. T. Road
Calcutta 700035, India

Department of Statistics, 367 Evans Hall
University of California, Berkeley
Berkeley, CA 94720-3860

Sloan School of Management, E40-131
Massachusetts Institute of Technology,
Cambridge, MA 02139