ESTIMATING CONDITIONAL DISTRIBUTION FUNCTIONS USING DIMENSION REDUCTION

Peter Hall[1,2] Qiwei Yao[2]

ABSTRACT. We suggest methods for approximating the conditional distribution function of a random variable Y given a dependent random vector X, by the conditional distribution function of Y given $\theta^{\mathrm{T}}X$, where the unit vector θ is selected so that the approximation is optimal under a least-squares criterion. We show that θ may be estimated root-n consistently using local linear regression. Furthermore, estimation of the conditional distribution function of Y, given $\theta^{\mathrm{T}}X$, has the same first-order asymptotic properties that it would enjoy if θ were known. The proposed method is illustrated using both simulated and real-data examples, showing its effectiveness for both independent datasets and data from time-series. Its usefulness for prediction and forecasting is also demonstrated. An effective empirical technique for bandwidth selection is given.

KEY WORDS AND PHRASES. Conditional distribution, cross-validation, dimension reduction, forecasting, kernel methods, leave-one-out method, local linear regression, nonparametric regression, prediction, projection pursuit, root-n consistency, time-series analysis.

SHORT TITLE. Estimating conditional distributions.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[2] Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.

1. INTRODUCTION

Estimating a conditional distribution function is an important feature of many statistical problems, including, for example, regression analysis (see Yin and Cook 2002, and references therein) and nonlinear time-series forecasting (see, e.g., Chapter 10 of Fan and Yao 2003). In most of these applications we are interested in estimating the conditional distribution of a scalar random variable Y, given a random d-vector X. Even for small values of $d \geq 2$ a conventional nonparametric estimator can suffer poor accuracy, reflected in slow theoretical convergence rates. In this paper we suggest a solution to this difficulty, based on approximating the conditional distribution function of Y given X by that of Y given $\theta^{\mathrm{T}}X$, where θ is selected so that the approximation is optimal under an appropriate least-squares criterion.

We define the criterion as an accumulation of squared differences between the joint probabilities of (Y, X) and the expected conditional probabilities of Y given $\theta^{\mathrm{T}}X$, over a large class of subsets; see (2.2) and (2.4) in section 2 below. Our search for the global parameter θ is based on leave-one-out local linear regression estimators for conditional distribution functions. Under very mild assumptions the resulting estimator $\hat{\theta}$ is root-n consistent and asymptotically normally distributed. To achieve the root-n rate, the bandwidth used in local linear regression should be an order of magnitude smaller than a conventional bandwidth for estimating a univariate function by nonparametric regression. However, concise choice of the bandwidth is not a major problem, since any bandwidth within a broad range will produce an estimator of θ with the same first-order theoretical properties. Moreover, there exist simple and effective empirical ways of choosing the bandwidth.

Of course, our main purpose in computing $\hat{\theta}$ is so it can be used in a conditional distribution estimator. The root-n convergence rate achieved by our estimator is so fast that the estimator of the conditional distribution function of Y given $\hat{\theta}^{\mathrm{T}}X$ is first-order equivalent to its counterpart that would be used if the true value of θ were known. This high level of theoretical performance is borne out in numerical work, which shows that if sample size is at least moderately large then the error incurred by replacing the true value of θ by its estimator is particularly small.

There exists an extensive literature on nonparametric estimation of conditional distributions. It includes work of Bhattacharya and Gangopadhyay (1990), Sheather

and Marron (1990), Yu and Jones (1998) and Cai (2002) on conditional quantile regression; Rosenblatt (1969), Hyndman, Bashtannyk and Grunwald (1996), Fan, Yao and Tong (1996), Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002) on estimation of conditional density estimation; and Hall, Wolff and Yao (1999) on estimation of conditional distribution functions. Dimension reduction has been discussed extensively in the context of regression or density estimation; see, for example, work of Friedman and Stuetzle (1981), Friedman, Stuetzle and Schroeder (1984), Huber (1985), Friedman (1987), Jones and Sibson (1987), Li (1991) and Posse (1995). The case of conditional distribution estimation is unusual in that it combines aspects of both density and regression estimation as well as unique features of its own. The latter include new types of objective functions, which lead to unusual theoretical and numerical properties, as well as a broad new range of applications, encompassing prediction and forecasting.

The article is organised as follows. In section 2 we introduce our method for estimating θ. Asymptotic properties of estimators $\hat{\theta}$ and $\widehat{F}(\cdot \,|\, \theta^{\mathrm{T}} X)$ are presented in section 3. Numerical examples involving both simulated models and a real-data application are given in section 4. All theoretical arguments are relegated to section 5.

2. METHODOLOGY

2.1. Motivation. Assume we observe data (X_i, Y_i), for $1 \le i \le n$, from the distribution of (X, Y). Here, X is a d-vector and Y a scalar. Let Θ denote the set of d-variate unit vectors θ with first nonzero component nonnegative, write f for the density of X, and let $F_{Y|\theta^{\mathrm{T}}X}(\,\cdot\,|z)$ represent the distribution of Y conditional on $\theta^{\mathrm{T}} X = z$. Given subsets $\mathcal{A}$ and $\mathcal{B}$, of d-dimensional space and of the real line, respectively, define

$$\pi_\theta(\mathcal{A}, \mathcal{B}) = \int_{\mathcal{A}} F_{Y|\theta^{\mathrm{T}}X}\big(\mathcal{B} \big| \theta^{\mathrm{T}} x\big)\, f(x)\, dx\,, \quad \pi(\mathcal{A}, \mathcal{B}) = P(X \in \mathcal{A}, Y \in \mathcal{B})\,.$$

If, for some θ and all x, $F_{Y|\theta^{\mathrm{T}}X}(\,\cdot\,|\theta^{\mathrm{T}} x)$ is identical to the distribution of Y given that $X = x$, then for this θ, $\pi_\theta(\mathcal{A}, \mathcal{B}) = \pi(\mathcal{A}, \mathcal{B})$ for all $\mathcal{A}, \mathcal{B}$.

We can estimate $F_{Y|\theta^{\mathrm{T}}X}$ using nonparametric methods, permitting us to estimate $\pi_\theta(\mathcal{A}, \mathcal{B})$. And of course, we can estimate $\pi(\mathcal{A}, \mathcal{B})$ as the proportion of pairs (X_i, Y_i) that lie in $\mathcal{A} \times \mathcal{B}$. Hence, for each triple $(\theta, \mathcal{A}, \mathcal{B})$ we can estimate $\pi_\theta(\mathcal{A}, \mathcal{B})$ and $\pi(\mathcal{A}, \mathcal{B})$ under minimal conditions. (We shall denote estimators of π_θ and π

by $\hat{\pi}_\theta$ and $\hat{\pi}$, respectively.) Therefore we can check (or, more formally, test) the hypothesis that $F_{Y|\theta^{\mathrm{T}}X}(\,\cdot\,|\theta^{\mathrm{T}}x)$ is identical to the distribution of Y conditional on $X = x$, for all x, by examining the average value of $\{\hat{\pi}_\theta(\mathcal{A},\mathcal{B}) - \hat{\pi}(\mathcal{A},\mathcal{B})\}^2$ over a range of sets $\mathcal{A}$ and $\mathcal{B}$.

Although exact equality of π and π_θ is unlikely in practice, the difference-based criterion noted above can be used to empirically select θ such that, in a global sense, the distribution of Y given $\theta^{\mathrm{T}}X = \theta^{\mathrm{T}}x$ is a good approximation to the distribution of Y given that $X = x$. Indeed, the argument in the previous paragraph suggests that methodology of this type could be based on the difference measure,

$$S_1(\theta) = \int\int \left\{\hat{\pi}_\theta(\mathcal{A}_\alpha,\mathcal{B}_\beta) - \hat{\pi}(\mathcal{A}_\alpha,\mathcal{B}_\beta)\right\}^2 w(\alpha,\beta)\,d\alpha\,d\beta\,, \tag{2.1}$$

where w is a weight function and the integral is taken over a parameterisation (α,β) of $(\mathcal{A},\mathcal{B})$. We suggest taking the sets $\mathcal{A}$ to be d-variate spheres with differing centres, and the sets $\mathcal{B}$ to be intervals, for example semi-infinite intervals $(-\infty,\beta)$.

The sets $\mathcal{A}_\alpha$ should be such that the density $f_{\theta^{\mathrm{T}}X}$ of $\theta^{\mathrm{T}}X$ is bounded away from zero at all points $\theta^{\mathrm{T}}x$ with $\theta \in \Theta$ and $x \in \mathcal{A}_\alpha$. Otherwise, design sparseness problems can arise when nonparametrically estimating $F_{Y|\theta^{\mathrm{T}}X}$. Considerations of this type suggest taking the $\mathcal{A}_\alpha$'s to be d-variate spheres whose centres confine them to lie inside a larger, bounded region where f is bounded away from zero. Such restrictions are unnecessary when considering the sets $\mathcal{B}$, except that there is little point in giving emphasis to sets for which $P(Y \in \mathcal{B})$ is low.

For these reasons, when permitting $\mathcal{B}_\beta$ to be the interval $(-\infty,\beta)$ it is appropriate to take $w(\alpha,\beta)$ at (2.1) to be proportional to the density of Y at β, and to not depend on α. We shall achieve this end empirically, by replacing the double integral at (2.1) by a sum of integrals:

$$S(\theta) = \sum_{j=1}^{n} \int \left\{\hat{\pi}_\theta(\mathcal{A}_\alpha,\mathcal{B}_{Y_j}) - \hat{\pi}(\mathcal{A}_\alpha,\mathcal{B}_{Y_j})\right\}^2 d\alpha\,, \tag{2.2}$$

where $\mathcal{B}_\beta$ denotes the interval $(-\infty,\beta]$ and the integral is taken over an appropriate set of sphere centres and radii. Below we shall use the notation y instead of β.

2.2. Estimator of θ. With these principles in mind, let h be a bandwidth and K a kernel function, and define

$$T^{[k]}_{-i,-j}(\theta) = \frac{1}{(n-2)\,h} \sum_{i_1 :\, i_1 \neq i,j} K\left\{\frac{\theta^{\mathrm{T}}(X_i - X_{i_1})}{h}\right\} \left\{\frac{\theta^{\mathrm{T}}(X_i - X_{i_1})}{h}\right\}^k,$$

$$w_{i_1;-i,-j}(\theta) = K\Big\{\frac{\theta^{\mathrm{T}}(X_i - X_{i_1})}{h}\Big\}\Big\{T^{[2]}_{-i,-j}(\theta) - \frac{\theta^{\mathrm{T}}(X_i - X_{i_1})}{h}\,T^{[1]}_{-i,-j}(\theta)\Big\},$$

$$\widehat{F}_{-i,-j}\big(y \,\big|\, \theta^{\mathrm{T}}X_i\big) = \Big\{\sum_{i_1:\, i_1 \neq i,j} w_{i_1;-i,-j}(\theta)\, I(Y_{i_1} \le y)\Big\}\Big/\Big\{\sum_{i_1:\, i_1 \neq i,j} w_{i_1;-i,-j}(\theta)\Big\}. \tag{2.3}$$

Write simply $F(y|z)$ for $P(Y \le y|\theta^{\mathrm{T}}X = z)$, and let $\mathcal{A}$ be a subset of d-variate space. In this notation, $\widehat{F}_{-i,-j}(y \,|\, \theta^{\mathrm{T}}X_i)$ is a local linear estimator of $F(y \,|\, \theta^{\mathrm{T}}X_i)$, based on data pairs other than the ith and the jth; and

$$\frac{1}{n-1}\sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \widehat{F}_{-i,-j}\big(y \,\big|\, \theta^{\mathrm{T}}X_i\big)$$

is an estimator of $\pi_\theta(\mathcal{A},\mathcal{B})$ when $\mathcal{B} = (-\infty, y]$.

For simplicity, when constructing the statistic $S(\theta)$ at (2.2) we shall take all the spheres $\mathcal{A}_\alpha$ to have the same radius, r say. (Nevertheless, all our results have analogues when we use an integral average over a range of radii.) Therefore, the index $\alpha \in \mathbb{R}^d$ will denote the centre of $\mathcal{A}_\alpha$. We shall average over all spheres $\mathcal{A}_\alpha$ that lie entirely within a given fixed set $\mathcal{R}$. Let $\mathcal{R}_r = \{\alpha \in \mathbb{R}^d : \mathcal{A}_\alpha \subseteq \mathcal{R}\}$ be the set of sphere centres.

Write $\widehat{F}_{-j}(\mathcal{A}, y)$ for the proportion of the $n-1$ values of (X_i, Y_i), for $i \neq j$, that satisfy $(X_i, Y_i) \in \mathcal{A} \times (-\infty, y]$. Put

$$S(\theta,\mathcal{A}) = \sum_{j=1}^{n}\Big\{\widehat{F}_{-j}(\mathcal{A}, Y_j) - \frac{1}{n-1}\sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \widehat{F}_{-i,-j}\big(Y_j \,\big|\, \theta^{\mathrm{T}}X_i\big)\Big\}^2,$$

$$S(\theta) = \int_{\mathcal{R}_r} S(\theta,\mathcal{A}_\alpha)\, d\alpha\,. \tag{2.4}$$

The latter represents a particular form of $S(\theta)$ at (2.2), in which leave-one-out methods are used to compute estimators.

We choose $\hat\theta$ to minimise $S(\theta)$ over $\theta \in \Theta$. Thus, $\hat\theta$ may be viewed as an estimator of θ_0, the minimiser (over $\theta \in \Theta$) of

$$S_0(\theta) = \int_{\mathcal{R}_r} d\alpha \int \big\{F(\mathcal{A}_\alpha, y) - G_\theta(\mathcal{A}_\alpha, y)\big\}^2 f(y)\, dy\,, \tag{2.5}$$

where $F(\mathcal{A}, y) = P\{(X,Y) \in \mathcal{A} \times (-\infty, y]\}$ and

$$G_\theta(\mathcal{A}, y) = \int_{\mathcal{A}} F\big(y \,\big|\, \theta^{\mathrm{T}}x\big)\, f(x)\, dx\,. \tag{2.6}$$

A low-dimensional approximation to $F_{Y|X}(y \,|\, X = x)$ is therefore $\widetilde{F}_{\hat\theta}(y \,|\, \hat\theta^{\mathrm{T}} x)$, where $\widetilde{F}_\theta(y|z)$ is an estimator of $P(Y \le y \,|\, \theta^{\mathrm{T}} X = \theta^{\mathrm{T}} x)$. Denoting by $\widehat{F}$ a local linear version of $\widetilde{F}$, we define

$$\widehat{F}_\theta\big(y \,|\, \theta^{\mathrm{T}} x\big) = \Big\{ \sum_{i=1}^n w_i(x,\theta)\, I(Y_i \le y) \Big\} \Big/ \Big\{ \sum_{i=1}^n w_i(x,\theta) \Big\}, \tag{2.7}$$

where

$$w_i(x,\theta) = K\Big\{ \frac{\theta^{\mathrm{T}}(x - X_i)}{h} \Big\} \Big\{ T^{[2]}(x,\theta) - \frac{\theta^{\mathrm{T}}(x - X_i)}{h}\, T^{[1]}(x,\theta) \Big\},$$

$$T^{[k]}(x,\theta) = \frac{1}{nh} \sum_{i=1}^n K\Big\{ \frac{\theta^{\mathrm{T}}(x - X_i)}{h} \Big\} \Big\{ \frac{\theta^{\mathrm{T}}(x - X_i)}{h} \Big\}^k .$$

Our empirical, low-dimensional approximation to $F_{Y|X}(y \,|\, X = x)$ is taken to be $\widehat{F}_{\hat\theta}(y \,|\, \hat\theta^{\mathrm{T}} x)$, and is of course an estimator of $P(Y \le y \,|\, \theta_0^{\mathrm{T}} X = \theta_0^{\mathrm{T}} x)$.

It is important that the bandwidth, h, used to construct $\widehat{F}_\theta$ be different from the bandwidth employed in the definition of $S(\theta)$ at (2.4). As we shall show in section 3, optimal performance is achieved if the latter bandwidth is much smaller than the former.

3. THEORY

First we define the vector of derivatives, $\dot a$, of a function a of $\theta \in \Theta$. Let $\omega_1, \ldots, \omega_{d-1}$ be orthonormal vectors all perpendicular to θ, put $\omega_{i\delta} = (1 - \delta^2)^{1/2}\, \theta + \delta\, \omega_i$ for a scalar δ, and set

$$b_i = \lim_{\delta \to 0} \delta^{-1} \{ a(\omega_{i\delta}) - a(\theta) \},$$

assuming the limit exists and is finite. Then, $\dot a(\theta) \equiv \sum_{1 \le i \le d-1} b_i\, \omega_i$, a vector in the plane perpendicular to θ. Similarly we may define the matrix, $\ddot a$, of second derivatives of a.

Let (X, Y) have the distribution of a generic pair (X_i, Y_i). We shall assume that

$$\text{the distribution of } (X,Y) \text{ has a compactly supported density with four bounded derivatives.} \tag{3.1}$$

The bandwidth, h, will be permitted to vary within a range, effectively from $n^{-1/3}$ to $n^{-1/4}$; see (3.4) below. If we were confining attention to the lower end of this range then we could reduce the smoothness assumption in (3.1) from four bounded

derivatives to three derivatives plus a Hölder continuity condition. In this sense, the smoothness required by (3.1) is excessive.

Recall that $\mathcal{R}_r$ denotes the set of all centres, α, of the spheres $\mathcal{A}_\alpha$, each of which is of radius $r > 0$ and completely contained within $\mathcal{R}$. We shall suppose that:

$$\begin{aligned}&\mathcal{R} \text{ is an open, bounded set; the density of } X \text{ is bounded}\\ &\text{away from zero on } \mathcal{R}\text{; and the content of } \mathcal{R}_r \text{ is nonzero.}\end{aligned} \tag{3.2}$$

In particular this ensures that the density of the distribution of $\theta^{\mathrm{T}}X$ is bounded away from zero on the set of points $\theta^{\mathrm{T}}x$ with $x \in \mathcal{A} \subseteq \mathcal{R}$. Assumption (3.2) may therefore be viewed as the analogue of the condition, imposed in more standard problems of nonparametric regression, that the design density is bounded above zero.

Conditions (3.1) and (3.2) imply a range of smoothness properties of the marginal density $f_{\theta^{\mathrm{T}}X}$ and the conditional distribution $F(y|z) = P(Y \le y|\theta^{\mathrm{T}}X = z)$. For example, the k_1th derivative with respect to θ, of the k_2th derivative with respect to z, of either $f_{\theta^{\mathrm{T}}X}(z)$ or $F(y|z)$, is well defined and bounded in $k_1 + k_2 \le 4$, y, $\theta \in \Theta$ and $z = \theta^{\mathrm{T}}x$ for $x \in \mathcal{R}$.

Recall the definition of $G_\theta(\mathcal{A}, y)$ at (2.6), and let $\dot{G}_\theta(\mathcal{A}, y)$ and $\ddot{G}_\theta(\mathcal{A}, y)$ denote, respectively, the vector of first derivatives and the matrix of second derivatives of $G_\theta(\mathcal{A}, y)$ with respect to θ, with $(\mathcal{A}, y)$ held fixed. Note that $\theta_0 = \operatorname{argmin}_\theta S_0(\theta)$, where S_0 is defined at (2.5).

Put

$$\begin{aligned}M(\theta) = \int_{\mathcal{R}_r} d\alpha \int_{\mathcal{A}_\alpha} \Big[&\dot{G}_\theta(\mathcal{A}_\alpha, y)\, \dot{G}_\theta(\mathcal{A}_\alpha, y)^{\mathrm{T}}\\ &- \big\{F(\mathcal{A}_\alpha, y) - G_\theta(\mathcal{A}_\alpha, y)\big\}\, \ddot{G}_\theta(\mathcal{A}_\alpha, y)\Big]\, f(y)\, dy\,,\end{aligned}$$

a $d \times d$ matrix. The assumption:

$$\begin{aligned}&\theta = \theta_0 \text{ gives a unique global minimum of } S_0(\theta)\text{, and}\\ &\omega^{\mathrm{T}} M(\theta_0)\, \omega > 0 \text{ for each nonvanishing vector } \omega \perp \theta_0\end{aligned} \tag{3.3}$$

is equivalent to asking that $S_0(\theta) \to S_0(\theta_0)$ at exactly the rate $\|\theta - \theta_0\|^2$ as $\theta \to \theta_0$. Of the kernel K and bandwidth h we shall assume that

$$\begin{aligned}&K \text{ is nonnegative, symmetric and compactly supported, and}\\ &\text{has a bounded derivative; and, for some } \epsilon > 0,\ h = h(n)\\ &\text{satisfies } h = O(n^{-\epsilon-(1/4)}) \text{ and } n^{-(1/3)+\epsilon} = O(h) \text{ as } n \to \infty.\end{aligned} \tag{3.4}$$

The most important aspect of this assumption is that it implies h should lie between $n^{-1/3}$ and $n^{-1/4}$, and so should be an order of magnitude smaller than a conventional bandwidth for estimating a univariate function by nonparametric regression. A conventional bandwidth would be of size $n^{-1/5}$.

Let $\phi_{\theta^{\mathrm{T}}X|\mathcal{A}}$ denote the density of $\theta^{\mathrm{T}}X$ conditional on $X \in \mathcal{A}$, and define

$$\psi(\mathcal{A}, x_1, y_1, y, \theta) = \bigl\{I(y_1 \le y) - F\bigl(y \,\big|\, \theta^{\mathrm{T}}x_1\bigr)\bigr\} \times \biggl\{I(x_1 \in \mathcal{A}) - \frac{\phi_{\theta^{\mathrm{T}}X|\mathcal{A}}(\theta^{\mathrm{T}}x_1)\, P(X \in \mathcal{A})}{f_{\theta^{\mathrm{T}}X}(\theta^{\mathrm{T}}x_1)}\biggr\}. \tag{3.5}$$

(The ratio in this definition is guaranteed well-defined, since $P(X \in \mathcal{A})\, \phi_{\theta^{\mathrm{T}}X|\mathcal{A}} \le f_{\theta^{\mathrm{T}}X}$.) Let V denote the Gaussian d-vector with zero mean and covariance matrix equal to that of

$$V = \int_{\mathcal{R}_r} d\alpha \biggl[\int \psi(\mathcal{A}_\alpha, X, Y, y, \theta)\, \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y)\, f(y)\, dy + \{F(\mathcal{A}_\alpha, Y) - G_{\theta_0}(\mathcal{A}_\alpha, Y)\}\, \dot{G}_{\theta_0}(\mathcal{A}_\alpha, Y) \biggr] d\alpha\,.$$

Let $\|\cdot\|$ denote the Euclidean metric in d-variate space, and recall that $\hat{\theta}$ is defined to be the global minimiser of $S(\theta)$, at (2.4).

Theorem 3.1. *Assume conditions* (3.1)–(3.4). *Then* $\hat{\theta} \to \theta_0$ *with probability 1, and* $n^{1/2}\, M(\theta_0)\, (\hat{\theta} - \theta_0)$ *converges in distribution to* V *as* $n \to \infty$.

To appreciate the implications of this result, let $\hat{\theta}^{\perp}$ denote the projection of $\hat{\theta}$ into the plane $\Pi^{\perp}$ that is perpendicular to θ_0. (Equivalently, $\hat{\theta}^{\perp}$ is the projection of $\hat{\theta} - \theta_0$ into $\Pi^{\perp}$.) The first part of Theorem 3.1 implies that $\|\hat{\theta} - \theta_0\| \to 0$ with probability 1, from which it follows (since $\hat{\theta}$ and θ_0 are both unit vectors) that

$$\hat{\theta} - \theta_0 = \hat{\theta}^{\perp} + o\bigl(\|\hat{\theta} - \theta_0\|\bigr) \tag{3.6}$$

with probability 1. That is, in first-order asymptotic terms, $\hat{\theta} - \theta_0$ is completely describable through the projection of this vector into the plane perpendicular to θ_0.

Note that, by definition of differentiation with respect to θ, the vector $\dot{G}_\theta$ is perpendicular to θ. It therefore follows from the definition of V that, with probability 1, V lies completely in $\Pi^{\perp}$. Observe too that, in view of (3.3), there is a generalised inverse of $M_0 = M(\theta_0)$ (call it M_0^-) that is well-defined in $\Pi^{\perp}$. It has the

property that $M_0 M_0^- v = M_0^- M_0 v = v$ for all $v \in \Pi^\perp$. These results, Theorem 3.1 and (3.6) imply that $n^{1/2}\,(\hat\theta - \theta)$ converges in distribution to $M_0^- V$.

Of course, our main purpose in computing $\hat\theta$ is so it can be used in a conditional distribution estimator, such as $\widehat F_\theta$ introduced at (2.7). Our next result shows that the root-n consistency achieved by the estimator $\hat\theta$ makes that quantity so accurate that, from the viewpoint of first-order performance, the estimator $\widehat F_{\hat\theta}(y \,|\, \hat\theta^{\mathrm T} x)$ is equivalent to its counterpart which would be employed if the value of θ_0 were known. To interpret such a result, note that when using a bandwidth of size $n^{-1/5}$ the local-linear estimator $\widehat F_{\theta_0}(y \,|\, \theta_0^{\mathrm T} x)$ converges to its limit at rate $n^{-2/5}$, and in fact the exact convergence rates of the bias and standard deviation of $\widehat F_{\theta_0}(y \,|\, \theta_0^{\mathrm T} x)$ are both $n^{-2/5}$. We shall show that the difference between $\widehat F_{\hat\theta}(y \,|\, \hat\theta^{\mathrm T} x)$ and $\widehat F_{\theta_0}(y \,|\, \theta_0^{\mathrm T} x)$ is of strictly smaller order than $n^{-2/5}$.

Assume that when constructing $\widehat F_\theta$ we use a bandwidth which is of conventional size, giving optimal convergence properties for a local linear estimator:

> the bandwidth h_1 used to construct $\widehat F_\theta$ has the property that $n^{1/5}\,h_1$ is bounded away from zero and infinity as $n \to \infty$; and the kernel is non-negative, symmetric, compactly supported and has a bounded derivative. (3.7)

We shall reduce the stringency of (3.1), assuming instead that

> the distribution of (X, Y) has a compactly supported density with two continuous derivatives. (3.8)

As the following theorem shows, we do not need the full force of the result that $\hat\theta - \theta_0 = O_p(n^{-1/2})$; the convergence rate $o_p(n^{-2/5})$ suffices.

Theorem 3.2. *Assume* (3.2), (3.7), (3.8), *that* $x \in \mathcal{R}$, *and that* $\hat\theta - \theta_0 = o_p(n^{-2/5})$ *as* $n \to \infty$. *Then for each* y,

$$\widehat F_{\hat\theta}\big(y \,\big|\, \hat\theta^{\mathrm T} x\big) = \widehat F_{\theta_0}\big(y \,\big|\, \theta_0^{\mathrm T} x\big) + o_p(n^{-2/5})\,.$$

It follows from Remark 4 of Hall, Wolff and Yao (1999) that the estimator $\widehat F_{\theta_0}(y \,|\, \theta_0^{\mathrm T} x)$ is asymptotically normally distributed with convergence rate $n^{-2/5}$. By Theorem 3.2 above, $\widehat F_{\hat\theta}(y \,|\, \hat\theta^{\mathrm T} x)$ and $\widehat F_{\theta_0}(y \,|\, \theta_0^{\mathrm T} x)$ have the same asymptotic distribution.

Analogues of Theorems 3.1 and 3.2 may be derived for dependent data, in particular for sequences of pairs $\{(X_i, Y_i), -\infty < i < \infty\}$ that satisfy sufficiently strong mixing conditions.

4. NUMERICAL PROPERTIES

4.1. Preliminaries. We approximate the integral at (2.4) by a series,

$$S(\theta) = \frac{1}{B} \sum_{i=1}^{B} S(\theta, \mathcal{A}_i) \,, \tag{4.1}$$

where the $\mathcal{A}_i$'s are spheres of radius r contained within $\mathcal{R}$. Minimising $S(\theta)$ is computationally intensive but nevertheless straightforward. Indeed, since $S(\theta)$ is defined in terms of leave-one-out estimators then we may choose both the bandwidth h and the unit vector θ to minimise $S(\theta)$, as follows. First select $\tilde{\theta}(h)$ to minimise $S(\theta)$ for given h, then choose $\hat{h}$ to minimise $S\{\tilde{\theta}(h)\}$, and finally take $\hat{\theta} = \tilde{\theta}(\hat{h})$. We may apply the method suggested at (2.7), or one of the approaches discussed by Hall, Wolff and Yao (1999), to estimate $F(\cdot|\theta^{\mathrm{T}}x)$.

In the numerical examples below, we first standardised the dataset $\{X_i\}$ so that its sample covariance matrix was the identity, and then searched for θ (with h fixed) using either the golden section method (when $d = 2$) or the downhill simplex method (when $d \geq 3$); see sections 10.1 and 10.4 of Press, Teukolsky, Vetterling and Flannery (1992). Using the Epanechnikov kernel, the bandwidth was sought among values $h_i = 0.15 \times 1.13^{i-1}$ for $i = 1, \ldots, 15$.

4.2. Simulation studies. We illustrate the proposed method through two simulated models, one with independent observations and the other using time-series data.

Example 1. Here we consider the model

$$Y_i = \theta_1 \, X_{i1} + \theta_2 \, X_{i2} + \theta_3 \, X_{i3} + \theta_4 \, X_{i4} + \epsilon_i \,,$$

where $\theta^{\mathrm{T}} \equiv (\theta_1, \ldots, \theta_4) = (1, 2, 0, 3)/\sqrt{14}$ and the X_{ij}'s and ϵ_i's are totally independent standard Normal random variables. Thus, the conditional distribution of Y given $X \equiv (X_1, \ldots, X_4)^{\mathrm{T}}$ is Normal $\mathrm{N}(\theta^{\mathrm{T}}X, 1)$. We used $n = 200$ or 400, took sphere radii to be $r = 1$, and let sphere centres be points (x_1, x_2, x_3, x_4), where each x_j ranged over either 5 or 7 grid points between -1.5 and 1.5, with spacing 0.75 or 0.5 respectively, resulting in $B = 625$ or $B = 2401$. Each setting was replicated 100 times.

Figure 1(a) presents the boxplot of the inner product $\theta^{\mathrm{T}}\hat{\theta}$. Since both θ and $\hat{\theta}$ are unit vectors then $\theta^{\mathrm{T}}\hat{\theta} = 1$ if and only if $\theta = \hat{\theta}$. We see from Figure 1(a) that the estimates of θ become steadily more accurate as sample size increases.

Moreover, the algorithm is largely insensitive to the value of B; the estimates of θ with $B = 625$ and $B = 2401$ are virtually identical. Figure 1(b) displays boxplots of bandwidth. As expected, empirical bandwidth is a decreasing function of sample size.

We also calculated values of the local linear estimator defined at (2.7), using $\theta = \hat{\theta}$. Here we used the Gaussian kernel with $h = \hat{h}$, the latter estimated via cross-validation using the Epanechnikov kernel. Figure 1(c) gives average absolute errors, computed using a regular grid (with adjacent points distant 0.05 apart) in the (X, Y)-plane. For the sake of comparison we also report the errors for estimators based on the true θ. Clearly, accuracy increases with sample size, and estimators based on $\hat{\theta}$ are less accurate than those based on the true θ. However, the deficit due to errors in estimating θ is not great when $n = 200$, and is negligible when $n = 400$. Choice of radius r is not critical either; results with $r = 0.5$ and 1.5 are similar to those for $r = 1$, and therefore are not reported here.

Example 2. Next we consider an AR(2) model,

$$Y_t = \theta_1 \, Y_{t-1} + \theta_2 \, Y_{t-2} + \epsilon_t \, ,$$

where $\theta \equiv (\theta_1, \theta_2)^{\mathrm{T}} = (0.6, -0.8)^{\mathrm{T}}$ and the ϵ_t's are independent and standard Normal. Let $X_t = (Y_{t-1}, Y_{t-2})^{\mathrm{T}}$. The conditional distribution of Y given $\theta^{\mathrm{T}} X$, which is of course Normal $\mathrm{N}(\theta^{\mathrm{T}} X, 1)$, was estimated. We used $n = 100$ or 200, let all sphere radii equal 1, and chose sphere centres to be points (x_1, x_2) where each x_j took values over either 9 or 17 grid points between -4 and 4 with spacing either 1 or 0.5, respectively, resulting in $B = 81$ or $B = 289$. Each setting was replicated 100 times. The results are depicted in Figure 2. We may deduce from Figure 2(b) that when $n = 100$, cross-validation tends to choose bandwidths larger than the default upper bound. However, in spite of this, the estimates of θ are still very accurate; see Figure 2(a). Consequently, the difference between the average absolute errors based on $\hat{\theta}$, and those using the true θ, is negligible; see Figure 2(c).

4.3. Real-data example: Canadian lynx data. Here we illustrate our method with $\{Y_t, 1 \le t \le 114\}$, using the classic Canadian lynx dataset. The variables Y_t were taken to be the natural logarithms of lynx numbers, recorded annually in the years 1821–1934. For $d = 2, 3, 4$, let $X_t = (Y_{t-1}, \ldots, Y_{t-d})^{\mathrm{T}}$. We estimated the value of $\theta \equiv (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$ such that conditional distribution function of Y_t, given $\theta^{\mathrm{T}} X_t$, was the best approximation for the conditional distribution of Y_t given X_t, in the sense

that $S(\theta)$, defined at (4.1), was minimised. Sphere centres were taken to be the points X_t (so that $B = n$), and each sphere radius was $r = 1.5$.

Once $\hat{\theta}$ was obtained we constructed the adjusted Nadaraya-Watson estimator $\widehat{F}(\cdot|z)$ (see Hall, Wolff and Yao, 1999) of the conditional distribution of Y_t given $\hat{\theta}^{\mathrm{T}} X_t = z$. The resulting quantile prediction interval is $[\widehat{F}^{-1}(\frac{1}{2}\alpha|z), \widehat{F}^{-1}(1-\frac{1}{2}\alpha|z)]$, for $\alpha \in (0,1)$. To check on performance, we used the data for 1821–1924 to estimate $\hat{\theta}$ and $\widehat{F}(\cdot|z)$, and employed the last 10 data points to validate the predicted values. Results with $\alpha = 0.1$ are reported in Table 1. For $d = 2$, $\hat{\theta} = (0.856, -0.516)^{\mathrm{T}}$. Note that Moran (1953) fitted an AR(2) model to the same data and obtained autoregressive coefficients $(1.41, -0.77)$; the latter vector is proportional to $(0.877, -0.479)$.

For $d = 3$ and 4 the quantities $\hat{\theta}^{\mathrm{T}} X_t$ are contrasts between the first lagged value and the other lagged values, and are largely determined by the first two lagged values. All the prediction intervals in Table 1 contain the subsequent 10 data points. The average interval lengths, for $d = 2$, 3 and 4, are respectively 2.05, 1.96 and 1.74. Alternative prediction intervals, based on an estimated conditional distribution given a single lagged value, were reported by Hall, Wolff and Yao (1999) and had average length 2.80. By combining the information in several lagged values into a linear combination, our method provides more accurate predictions without using multivariate smoothing techniques, which are susceptible to the "curse of dimensionality."

5. THEORETICAL ARGUMENTS

5.1. Proof of Theorem 3.1. We shall derive only the second, weak convergence part of the theorem, since the first is relatively straightforward to establish. Our derivation of the second part has two main features: first, showing that $\|\hat{\theta} - \theta_0\| = O_p(n^{\epsilon-(1/4)})$ for each $\epsilon > 0$, and secondly, deriving the claimed weak convergence result. These programs will be carried out in Steps (i) and (ii), respectively. In the first we shall state five lemmas, each of which has two parts and where the second part is needed only for Step (ii). Outline proofs of the lemmas will be given in Step (iii). In each step we shall make implicit use of the argument given immediately below the statement of Theorem 3.1, and in particular of the fact that $\|\theta - \theta_0\| \sim \|\theta^{\perp}\|$ as $\theta \to \theta_0$.

Step (i): Proof that $\|\hat{\theta} - \theta_0\| = O_p(n^{\epsilon-(1/4)})$ *for each* $\epsilon > 0$. We begin by expanding both the numerator and the denominator in the definition of $\widehat{F}_{-i,-j}(Y_j \,|\, \theta^{\mathrm{T}} X_i)$. Put

$K_k(u) = u^k\,K(u)$ and let $W_{ij}^{[1]} \equiv 1$ and $W_{ij}^{[2]} = I(Y_i \le Y_j)$. Define

$$S_{-i,-j}^{[k,\ell]}(x,\theta) = \frac{1}{(n-2)\,h} \sum_{i_1:\, i_1 \neq i,j} K_k\left\{\frac{\theta^{\mathrm{T}}(x - X_{i_1})}{h}\right\} W_{i_1 j}^{[\ell]}\,,$$

$$w_{-i,-j}(x,y,\theta) = K\left\{\frac{\theta^{\mathrm{T}}(x-y)}{h}\right\}\left\{S_{-i,-j}^{[2,1]}(x,\theta) - \frac{\theta^{\mathrm{T}}(x-y)}{h}\,S_{-i,-j}^{[1,1]}(x,\theta)\right\}.$$

For $\ell = 1$ and 2, let

$$\begin{aligned} U_{ij}^{[\ell]}(x,\theta) &= \frac{1}{(n-2)\,h} \sum_{i_1:\, i_1 \neq i,j} w_{-i,-j}(x, X_{i_1}, \theta)\, W_{i_1 j}^{[\ell]} \\ &= S_{-i,-j}^{[2,1]}(x,\theta)\, S_{-i,-j}^{[0,\ell]}(x,\theta) - S_{-i,-j}^{[1,1]}(x,\theta)\, S_{-i,-j}^{[1,\ell]}(x,\theta)\,. \end{aligned} \tag{5.1}$$

Set $s_j^{[k,\ell]} = E(S_{-i,-j}^{[k,\ell]} | Y_j)$ and $\Delta S_{ij}^{[k,\ell]} = S_{-i,-j}^{[k,\ell]} - s_j^{[k,\ell]}$. (The subscript j on $s_j^{[k,\ell]}$ may be dropped when $\ell = 1$, since there the dependence on Y_j is degenerate.) Let $\mathcal{S}_r$ denote the set of all points that are contained in a sphere of radius r which is in turn completely contained within $\mathcal{R}$. We shall shortly prove the following result.

Lemma 5.1. *Under the conditions of Theorem 3.1, and for each $\epsilon > 0$,*

$$\left|\Delta S_{ij}^{[k,\ell]}(x,\theta)\right| = O_p\left\{(nh)^{-1/2}\, n^{\epsilon}\right\}, \tag{5.2}$$

uniformly in $x \in \mathcal{S}_r$, $\theta \in \Theta$, $1 \le i, j \le n$ with $i \neq j$ and $X_i \in \mathcal{S}_r$, values of Y_j, and $k = 0, 1, 2$ and $\ell = 1, 2$. Furthermore, for each $\epsilon > 0$,

$$\left|\Delta S_{ij}^{[k,\ell]}(x,\theta) - \Delta S_{ij}^{[k,\ell]}(x,\theta_0)\right| = O_p\left\{\|\theta - \theta_0\|\,(nh^3)^{-1/2}\, n^{\epsilon} + (nh)^{\epsilon-1}\right\}, \tag{5.3}$$

uniformly in the same sense.

Recall that $f_{\theta^{\mathrm{T}}X}$ denotes the density of $\theta^{\mathrm{T}}X$. Let $F(\cdot|z) = F_{Y|\theta^{\mathrm{T}}X}(\cdot|z)$ be the distribution function of Y given that $\theta^{\mathrm{T}}X = z$. For $k \ge 0$, define $\kappa_k = \int u^k\,K(u)\,du$. Without loss of generality, $\kappa_0 = 1$. Note that $\kappa_1 = 0$. Put

$$g_j^{[1]}(x,\theta) = g^{[1]}(x,\theta) = \kappa_2\, f_{\theta^{\mathrm{T}}X}(\theta^{\mathrm{T}}x)^2\,, \quad g_j^{[2]}(x,\theta) = \kappa_2\, f_{\theta^{\mathrm{T}}X}(\theta^{\mathrm{T}}x)^2\, F(Y_j|\theta^{\mathrm{T}}x)\,.$$

Lemma 5.2. *Under the conditions of Theorem 3.1,*

$$s_j^{[0,1]}(x,\theta) = f_{\theta^{\mathrm{T}}X}\big(\theta^{\mathrm{T}}x\big) + O\big(h^2\big)\,, \tag{5.4}$$

$$s_j^{[0,2]}(x,\theta) = f_{\theta^{\mathrm{T}}X}\big(\theta^{\mathrm{T}}x\big)\, F\big(Y_j\big|\theta^{\mathrm{T}}x\big) + O\big(h^2\big)\,, \tag{5.5}$$

$$s_j^{[1,\ell]}(x,\theta) = O(h)\,, \quad s_j^{[2,1]}(x,\theta) = \kappa_2\, f_{\theta^{\mathrm{T}}X}\big(\theta^{\mathrm{T}}x\big) + O\big(h^2\big)\,, \tag{5.6}$$

$$s_j^{[2,1]}(x,\theta)\, s_j^{[0,\ell]}(x,\theta) - s_j^{[1,1]}(x,\theta)\, s_j^{[1,\ell]}(x,\theta) = g_j^{[\ell]}(x,\theta) + O\big(h^2\big)\,, \tag{5.7}$$

uniformly in $x \in \mathcal{S}_r$, $\theta \in \Theta$ *and* $1 \le i, j \le n$. *Furthermore, if any one of the relations* (5.4)–(5.7) *is written in the form "*$a_j(x,\theta) = b(x, Y_j, \theta) + O(h^\nu)$*," then*

$$a_j(x,\theta) - a_j(x,\theta_0) = b(x, Y_j, \theta) - b(x, Y_j, \theta_0) + O\big(\|\theta - \theta_0\|\, h^\nu\big) + O\big(h^{\nu+2}\big)\,, \quad (5.8)$$

uniformly in x, θ *and* j.

Although the remainder terms represented by $O(h^\nu)$, for $\nu = 1$ or 2, in (5.4)–(5.8) may be stochastic, since they depend on the random variable Y_j, we do not denote them by $O_p(h^\nu)$ since the claimed bound is by a constant multiple of h^ν, not depending on j or Y_j.

It follows from (5.1) that

$$\begin{aligned} U_{ij}^{[\ell]} = {} & s_j^{[2,1]}\, s_j^{[0,\ell]} - s_j^{[1,1]}\, s_j^{[1,\ell]} + \Delta S_{ij}^{[2,1]}\, s_j^{[0,\ell]} + s_j^{[2,1]}\, \Delta S_{ij}^{[0,\ell]} \\ & - \Delta S_{ij}^{[1,1]}\, s_j^{[1,\ell]} - s_j^{[1,1]}\, \Delta S_{ij}^{[1,\ell]} + \Delta S_{ij}^{[2,1]}\, \Delta S_{ij}^{[0,\ell]} - \Delta S_{ij}^{[1,1]}\, \Delta S_{ij}^{[1,\ell]}\,. \quad (5.9) \end{aligned}$$

Combining (5.2), (5.7) and (5.9), defining $\eta = (nh)^{-1}\, n^\epsilon + h^2$, and suppressing the argument (x,θ), we deduce that

$$U_{ij}^{[\ell]} = g_j^{[\ell]} + \Delta S_{ij}^{[2,1]}\, s_j^{[0,\ell]} + s_j^{[2,1]}\, \Delta S_{ij}^{[0,\ell]} - \Delta S_{ij}^{[1,1]}\, s_j^{[1,\ell]} - s_j^{[1,1]}\, \Delta S_{ij}^{[1,\ell]} + O_p(\eta)\,, \quad (5.10)$$

uniformly in $x \in \mathcal{S}_r$, $\theta \in \Theta$ and $1 \le i, j \le n$. Since, by (5.2) and (5.6), $|\Delta S_{ij}^{[1,\ell]}| = O_p\{(nh)^{-1/2}\, n^\epsilon\}$ and $|s_j^{[1,\ell]}| = O(h)$; and since $\frac{1}{2}\,(nh)^{-1/2}\, n^\epsilon\, h \le (nh)^{-1}\, n^{2\epsilon} + h^2$; then (5.10) reduces to:

$$U_{ij}^{[\ell]} = g_j^{[\ell]} + \Delta S_{ij}^{[2,1]}\, s_j^{[0,\ell]} + s_j^{[2,1]}\, \Delta S_{ij}^{[0,\ell]} + O_p(\eta)\,. \quad (5.11)$$

Note too that there exists $C > 0$ such that, for all sufficiently large n,

$$\inf_{x \in \mathcal{S}_r} \inf_{\theta \in \Theta} g^{[1]}(x,\theta) \ge C\,. \quad (5.12)$$

Combining (5.11) and (5.12), and noting that

$$\widehat{F}_{-i,-j}(Y_j \,|\, \theta^{\mathrm{T}} X_i) = U_{ij}^{[1]}(X_i,\theta)/U_{ij}^{[2]}(X_i,\theta) \quad \text{and} \quad g_j^{[2]}(x,\theta)/g^{[1]}(x,\theta) = F(Y_j|\theta^{\mathrm{T}} x)\,,$$

we deduce that, uniformly in $\theta \in \Theta$ and $1 \le i, j \le n$ with $i \ne j$ and $X_i \in \mathcal{S}_r$,

$$\begin{aligned} \widehat{F}_{-i,-j}\big(Y_j \,\big|\, \theta^{\mathrm{T}} X_i\big) - F(Y_j|\theta^{\mathrm{T}} X_i) & \\ = {} & \Big(\Delta S_{ij}^{[2,1]}\, s_j^{[0,2]} + s_j^{[2,1]}\, \Delta S_{ij}^{[0,2]}\Big)\big(g^{[1]}\big)^{-1} \\ & - \Big(\Delta S_{ij}^{[2,1]}\, s_j^{[0,1]} + s_j^{[2,1]}\, \Delta S_{ij}^{[0,1]}\Big)\, \frac{g_j^{[2]}}{(g^{[1]})^2} + O_p(\eta)\,, \quad (5.13) \end{aligned}$$

where arguments on the right-hand side have been suppressed.

Using (5.4)–(5.6) we may show that, up to terms that go into the remainder $O_p(\eta)$, the terms in $\Delta S_{ij}^{[2,1]}$ on the right-hand side of (5.13) cancel. Therefore, (5.13) may be rewritten as:

$$\widehat{F}_{-i,-j}\big(Y_j \,\big|\, \theta^{\mathrm{T}} X_i\big) - F(Y_j|\theta^{\mathrm{T}} X_i) = \frac{\Delta S_{ij}^{[0,2]}(X_i,\theta)}{f_{\theta^{\mathrm{T}} X}\big(\theta^{\mathrm{T}} X_i\big)} - \frac{\Delta S_{ij}^{[0,1]}(X_i,\theta)\, F(Y_j|\theta^{\mathrm{T}} X_i)}{f_{\theta^{\mathrm{T}} X}\big(\theta^{\mathrm{T}} X_i\big)} + O_p(\eta)\,,$$

uniformly in the same sense as (5.13). Therefore, summing over i, and defining $n_1 = n - 1$, we have:

$$\frac{1}{n_1} \sum_{i\,:\, i\neq j,\, X_i \in \mathcal{A}} \widehat{F}_{-i,-j}\big(Y_j|\theta^{\mathrm{T}} X_i\big) = \frac{1}{n_1} \sum_{i\,:\, i\neq j,\, X_i \in \mathcal{A}} \bigg\{ F(Y_j|\theta^{\mathrm{T}} X_i) + \frac{\Delta S_{ij}^{[0,2]}(X_i,\theta)}{f_{\theta^{\mathrm{T}} X}\big(\theta^{\mathrm{T}} X_i\big)} - \frac{\Delta S_{ij}^{[0,1]}(X_i,\theta)\, F(Y_j|\theta^{\mathrm{T}} X_i)}{f_{\theta^{\mathrm{T}} X}\big(\theta^{\mathrm{T}} X_i\big)} \bigg\} + O_p(\eta), \quad (5.14)$$

uniformly in j.

Define $W_i^{[1]}(y) \equiv 1$ and $W_i^{[2]}(y) = I(Y_i \le y)$. For a given sphere $\mathcal{A}$ of radius r, let

$$A^{[1]}(u,y,\theta) = \frac{1}{f_{\theta^{\mathrm{T}} X}(u)}\,, \quad A^{[2]}(u,y,\theta) = \frac{F(y|u)}{f_{\theta^{\mathrm{T}} X}(u)}\,,$$

$$\Delta_{1j}^{[\ell,m]}(\mathcal{A},y,\theta) = \frac{1}{n} \sum_{i_1\,:\, i_1\neq j} \bigg[\frac{1}{h} \int K\bigg\{ \frac{\theta^{\mathrm{T}}(x - X_{i_1})}{h} \bigg\}\, W_{i_1}^{[\ell]}(y) \times A^{[m]}\big(\theta^{\mathrm{T}} x, y, \theta\big)\, I(x \in \mathcal{A})\, f(x)\, dx - \mu_1^{[\ell,m]}(\mathcal{A},y,\theta) \bigg]\,,$$

where

$$\begin{aligned} &\mu_1^{[\ell,m]}(\mathcal{A},y,\theta) \\ &\quad = h^{-1} \int \bigg[K\bigg\{ \frac{\theta^{\mathrm{T}}(x - X_{i_1})}{h} \bigg\}\, W_{i_1}^{[\ell]}(y)\, A^{[m]}\big(\theta^{\mathrm{T}} x, y, \theta\big)\, I(x \in \mathcal{A})\, f(x) \bigg]\, dx \\ &\quad = h^{-1}\, P(X \in \mathcal{A}) \int E\bigg\{ K\bigg(\frac{u - \theta^{\mathrm{T}} X_{i_1}}{h} \bigg)\, W_{i_1}^{[\ell]}(y) \bigg\}\, A^{[m]}(u,y,\theta)\, \phi_{\theta^{\mathrm{T}} X|\mathcal{A}}(u)\, du \\ &\quad = \mu_2^{[\ell,m]}(\mathcal{A},y,\theta) + o(1) \end{aligned}$$

and

$$\mu_2^{[\ell,m]}(\mathcal{A},y,\theta) = P(X \in \mathcal{A})\, E\Big\{ W_i^{[\ell]}(y)\, A^{[m]}(\theta^{\mathrm{T}} X_i, y, \theta)\, \phi_{\theta^{\mathrm{T}} X|\mathcal{A}}\big(\theta^{\mathrm{T}} X_i\big) \Big\}\,.$$

Put

$$Q_{1j}(\mathcal{A},\theta) = \frac{1}{n_1} \sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \Delta S_{ij}^{[0,2]}(X_i,\theta)\,\frac{1}{f_{\theta^{\mathrm{T}}X}\bigl(\theta^{\mathrm{T}}X_i\bigr)} - \Delta_{1j}^{[2,1]}(\mathcal{A},Y_j,\theta)\,,$$

$$Q_{2j}(\mathcal{A},\theta) = \frac{1}{n_1} \sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \Delta S_{ij}^{[0,1]}(X_i,\theta)\,\frac{F(Y_j|\theta^{\mathrm{T}}X_i)}{f_{\theta^{\mathrm{T}}X}\bigl(\theta^{\mathrm{T}}X_i\bigr)} - \Delta_{1j}^{[1,2]}(\mathcal{A},Y_j,\theta)\,.$$

We shall shortly prove the following two lemmas.

Lemma 5.3. *Under the conditions of Theorem 3.1, and for each $\epsilon > 0$,*

$$|Q_{1j}(\mathcal{A},\theta)| + |Q_{2j}(\mathcal{A},\theta)| = O_p\bigl\{(nh)^{-1}\,n^{\epsilon}\bigr\}\,, \tag{5.15}$$

uniformly in choices $\mathcal{A}\subseteq\mathcal{R}$ of the sphere, $\theta\in\Theta$, and $1\le j\le n$. Moreover, for each $\epsilon>0$,

$$\begin{aligned}\bigl|Q_{1j}(\mathcal{A},\theta) - Q_{1j}(\mathcal{A},\theta_0)\bigr| + \bigl|Q_{2j}(\mathcal{A},\theta) - Q_{2j}(\mathcal{A},\theta_0)\bigr| \\ = O_p\bigl\{\|\theta-\theta_0\|\,(nh^{3/2})^{-1}\,n^{\epsilon} + (nh)^{\epsilon-2}\bigr\}\,,\end{aligned} \tag{5.16}$$

uniformly in $\mathcal{A}$, θ and j.

Define

$$\begin{aligned}\Delta_{2j}^{[\ell,m]}(\mathcal{A},y,\theta) = \frac{1}{n_1} \sum_{i_1\,:\,i_1\neq j} \Bigl\{ W_{i_1}^{[\ell]}(y)\,A^{[m]}\bigl(\theta^{\mathrm{T}}X_{i_1},y,\theta\bigr) \\ \times \phi_{\theta^{\mathrm{T}}X|\mathcal{A}}\bigl(\theta^{\mathrm{T}}X_i\bigr)\,P(X\in\mathcal{A}) - \mu_2^{[\ell,m]}(y,\theta)\Bigr\}\,.\end{aligned}$$

Lemma 5.4. *Under the conditions of Theorem 3.1, and for some $t>0$,*

$$\Delta_{1j}^{[\ell,m]}(\mathcal{A},y,\theta) - \Delta_{2j}^{[\ell,m]}(\mathcal{A},y,\theta) = O_p\bigl(n^{-(1/2)-t}\bigr)\,, \tag{5.17}$$

uniformly in choices $\mathcal{A}\subseteq\mathcal{R}$ of the sphere, y, $\theta\in\Theta$, $1\le j\le n$, and $[\ell,m]=[1,2]$ and $[2,1]$. Furthermore, for some $t>0$,

$$\begin{aligned}\Bigl|\bigl\{\Delta_{1j}^{[\ell,m]}(\mathcal{A},y,\theta) - \Delta_{2j}^{[\ell,m]}(\mathcal{A},y,\theta)\bigr\} - \bigl\{\Delta_{1j}^{[\ell,m]}(\mathcal{A},y,\theta_0) - \Delta_{2j}^{[\ell,m]}(\mathcal{A},y,\theta_0)\bigr\}\Bigr| \\ = O_p\bigl(\|\theta-\theta_0\|\,n^{-t-(1/2)} + n^{-t-1}\bigr)\,,\end{aligned} \tag{5.18}$$

uniformly in $\mathcal{A}$, y, θ, j, and $[\ell,m]=[1,2],[2,1]$.

Combining (5.15) and (5.17) we deduce that

$$\frac{1}{n_1} \sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \Delta S_{ij}^{[0,2]}(X_i,\theta)\,\frac{1}{f_{\theta^{\mathrm{T}}X}\bigl(\theta^{\mathrm{T}}X_i\bigr)} - \Delta_{2j}^{[2,1]}(\mathcal{A},Y_j,\theta) = o_p\bigl(n^{-1/2}\bigr)\,,$$

$$\frac{1}{n_1} \sum_{i\,:\,i\neq j,\,X_i\in\mathcal{A}} \Delta S_{ij}^{[0,1]}(X_i,\theta)\,\frac{F(Y_j|\theta^{\mathrm{T}}X_i)}{f_{\theta^{\mathrm{T}}X}\bigl(\theta^{\mathrm{T}}X_i\bigr)} - \Delta_{2j}^{[1,2]}(\mathcal{A},Y_j,\theta) = o_p\bigl(n^{-1/2}\bigr)\,,$$

each relation holding uniformly in $\theta \in \Theta$ and $1 \le j \le n$. Combining these results with (5.14) we find that

$$\begin{aligned}
&\frac{1}{n_1} \sum_{i\,:\,i \ne j,\, X_i \in \mathcal{A}} \widehat{F}_{-i,-j}\big(Y_j \,\big|\, \theta^{\mathrm{T}} X_i\big) \\
&\quad = \frac{1}{n_1} \sum_{i\,:\,i \ne j,\, X_i \in \mathcal{A}} F(Y_j | \theta^{\mathrm{T}} X_i) + \Delta_{2j}^{[2,1]}(\mathcal{A}, Y_j, \theta) - \Delta_{2j}^{[1,2]}(\mathcal{A}, Y_j, \theta) + o_p\big(n^{-1/2}\big)\,,
\end{aligned}$$

uniformly in spheres $\mathcal{A}$ of radius r centred in $\mathcal{R}_r$, $\theta \in \Theta$ and $1 \le j \le n$. Therefore, recalling the definition (3.5) of ψ,

$$\begin{aligned}
&\widehat{F}_{-j}(\mathcal{A}, Y_j) - \frac{1}{n_1} \sum_{i\,:\,i \ne j,\, X_i \in \mathcal{A}} \widehat{F}_{-i,-j}\big(Y_j \,\big|\, \theta^{\mathrm{T}} X_i\big) \\
&\qquad = F(\mathcal{A}, Y_j) - G_\theta(\mathcal{A}, Y_j) + \xi_{nj}(\mathcal{A}, Y_j, \theta) + o_p\big(n^{-1/2}\big)\,, \qquad (5.19)
\end{aligned}$$

uniformly in the same sense, where $F(\mathcal{A}, y) = P(X \in \mathcal{A}, Y \le y)$ and

$$\xi_{nj}(\mathcal{A}, y, \theta) = \frac{1}{n_1} \sum_{i\,:\,i \ne j} \Big[\psi(\mathcal{A}, X_i, Y_i, y, \theta) - E\{\psi(\mathcal{A}, X, Y, y, \theta)\}\Big]\,.$$

Lemma 5.5. *Under the conditions of Theorem 3.1,*

$$|\xi_{nj}(\mathcal{A}_\alpha, y, \theta)| = O_p\big(n^{\epsilon - (1/2)}\big)\,, \qquad (5.20)$$

uniformly in $1 \le j \le n$, $\alpha \in \mathcal{R}_r$, y *and* $\theta \in \Theta$. *Furthermore, for each* $\epsilon > 0$,

$$\big|\xi_{nj}(\mathcal{A}, y, \theta) - \xi_{nj}(\mathcal{A}, y, \theta_0)\big| = O_p\big(\|\theta - \theta_0\|\, n^{\epsilon - (1/2)} + n^{\epsilon - 1}\big)\,, \qquad (5.21)$$

uniformly in y *and* $\theta \in \Theta$.

We may deduce from (5.19), (5.20) and the definition of $S(\theta)$ at (2.4) that for each $\epsilon > 0$,

$$n^{-1}\, S(\theta) = \int d\widehat{F}(y) \int_{\mathcal{R}_r} \big\{F(\mathcal{A}_\alpha, y) - G_\theta(\mathcal{A}_\alpha, y) + O_p\big(n^{\epsilon - (1/2)}\big)\big\}^2\, d\alpha\,, \qquad (5.22)$$

where $\widehat{F}$ denotes the empirical distribution function of the sample $Y_1, \ldots, Y_n$, and the remainder $O_p(n^{\epsilon-(1/2)})$ is of the stated order uniformly in α, y and θ. Put

$$\begin{aligned}
&D_\theta(\alpha, y) = F(\mathcal{A}, y) - G_\theta(\mathcal{A}_\alpha, y)\,, \quad \widehat{S}(\theta) = \int d\widehat{F}(y) \int_{\mathcal{R}_r} D_\theta(\alpha, y)^2\, d\alpha\,, \\
&H_0(\alpha, y) = D_{\theta_0}(\alpha, y)\,, \quad H_1(\alpha, y) = -2\, D_{\theta_0}(\alpha, y)\, \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y)\,, \\
&H_2(\alpha, y) = \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y)\, \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y)^{\mathrm{T}} - D_{\theta_0}(\alpha, y)\, \ddot{G}_{\theta_0}(\mathcal{A}_\alpha, y)\,.
\end{aligned}$$

Note that by (5.22),

$$\{n^{-1}S(\theta)\}^{1/2} - \widehat{S}(\theta)^{1/2} = O_p\big(n^{\epsilon-(1/2)}\big)\,, \tag{5.23}$$

uniformly in θ; and that by Taylor expansion,

$$D_\theta(\alpha,y)^2 = D_{\theta_0}(\alpha,y)^2 + (\theta-\theta_0)^{\mathrm{T}} H_1(\alpha,y) + (\theta-\theta_0)^{\mathrm{T}} H_2(\alpha,y)\,(\theta-\theta_0) + o\big(\|\theta-\theta_0\|^2\big)\,, \tag{5.24}$$

where the remainder is of the stated size uniformly in $\alpha \in \mathcal{R}_r$ and in y, as $\theta \to \theta_0$. Define $\widehat{S}_j = \int d\widehat{F}(y) \int_{\mathcal{R}_r} H_j(\alpha,y)\,d\alpha$. It follows from (5.24) that

$$\widehat{S}(\theta) = \widehat{S}_0 + (\theta-\theta_0)^{\mathrm{T}}\widehat{S}_1 + (\theta-\theta_0)^{\mathrm{T}}\widehat{S}_2\,(\theta-\theta_0) + o_p\big(\|\theta-\theta_0\|^2\big)\,, \tag{5.25}$$

uniformly in n, as $\theta \to \theta_0$.

Since θ_0 gives a turning point of $S_0(\theta)$ then $\int dF(y) \int_{\mathcal{R}_r} H_1(\alpha,y)\,d\alpha = 0$, i.e. $E(\widehat{S}_1) = 0$. Hence it may be proved that $\widehat{S}_1 = O_p(n^{-1/2})$. Similarly, $\widehat{S}_2 = M_0 + o_p(1)$, where $M_0 = M(\theta_0)$ and M is as defined in section 3. Therefore, by (5.23) and (5.25),

$$S(\theta) = \widehat{S}_0 + (\theta-\theta_0)^{\mathrm{T}} M_0\,(\theta-\theta_0) + o_p\big(\|\theta-\theta_0\|^2\big) + O_p\big(n^{\epsilon-(1/2)}\big)\,, \tag{5.26}$$

uniformly in n, as $\theta \to \theta_0$. Result (5.26), and the fact that $\hat\theta \to \theta_0$ in probability (in fact, almost surely) by the first half of Theorem 3.1, imply that $\|\hat\theta - \theta_0\| = O_p(n^{\epsilon-(1/4)})$ for each $\epsilon > 0$, concluding Step (i) of the proof.

Step (ii): Completion of proof of Theorem 3.1. First we prove that

$$\|\hat\theta - \theta_0\| = O_p\big(n^{\epsilon-(1/2)}\big) \quad \text{for each} \quad \epsilon > 0\,. \tag{5.27}$$

Observe that if one carries through the argument leading to (5.22), bounding remainder terms as before but this time using the second part of each lemma rather than the first, one obtains the following refinement of (5.22):

$$n^{-1}\,S(\theta) = \int d\widehat{F}(y) \int_{\mathcal{R}_r} \Big\{F(\mathcal{A}_\alpha,y) - G_\theta(\mathcal{A}_\alpha,y) + O_p\langle n^{-1/2}\rangle + O_p\big(\textstyle\sum \|\theta-\theta_0\|^u\, n^{\epsilon-v}\big)\Big\}^2 d\alpha\,, \tag{5.28}$$

where $O_p\langle n^{-1/2}\rangle$ denotes a term that is of the stated order uniformly in α and y *but does not depend on* θ, and $O_p(\sum \|\theta-\theta_0\|^u\, n^{\epsilon-v})$ denotes terms which, uniformly

in α, y and θ, are of order no more than that of the sum of $\|\theta - \theta_0\|^{u_i}\, n^{\epsilon_i - v_i}$ over a bounded number of indices i, where each $\epsilon_i > 0$ and may be taken arbitrarily small, each $u_i, v_i \geq 0$, and $\frac{1}{2}u_i + v_i \geq 1$ for each i. Some of these terms arise through products of two quantities in the remainders treated by the lemmas, whereas others come from a single quantity.

Examples include terms of order

$$\bigl\{\|\theta - \theta_0\|\,(nh^3)^{-1/2}\,n^{\epsilon} + (nh)^{\epsilon-1}\bigr\}^2 = O\bigl\{\|\theta - \theta_0\|^2\,(nh^3)^{-2}\,n^{2\epsilon} + (nh)^{2\epsilon-2}\bigr\}\,,$$

arising through squaring the remainder at (5.3). Here, bearing in mind the range of values of h allowed by (3.4), see that we may take $(u, v) = (2, 0)$ in the first instance and $(u, v) = (0, \frac{4}{3})$ in the second. Another order that arises is $\|\theta - \theta_0\|\,h^2$, coming from the first of the remainders at (5.8) with $\nu = 2$. Here, again noting (3.4), we see that $(u, v) = (1, \frac{1}{2})$. A further example is represented by the order

$$\|\theta - \theta_0\|\,(nh^{3/2})^{-1}\,n^{\epsilon} + (nh)^{\epsilon-2}\,,$$

representing the right-hand side of (5.16) and where, once more by (3.4), $(u, v) = (1, \frac{1}{2})$ in the case of the first term, and $(0, \frac{4}{3})$ for the second. The value $(u, v) = (1, \frac{1}{2})$ is also appropriate when representing the orders of the first terms on the right-hand sides (5.18) and (5.21), and $(u, v) = (0, 1)$ is adequate for the second terms in each of those formulae.

Substituting for $F(\mathcal{A}, y) - G_\theta(\mathcal{A}_\alpha, y) = D_\theta(\mathcal{A}_\alpha, y)$ in (5.28), using (5.24); and expanding the square on the right-hand side of (5.28); we deduce instead of (5.26) that

$$S(\theta) = T + (\theta - \theta_0)^{\mathrm{T}} M_0\,(\theta - \theta_0) + o_p\bigl(\|\theta - \theta_0\|^2\bigr) + O_p\bigl(\textstyle\sum \|\theta - \theta_0\|^u\, n^{\epsilon - v}\bigr)\,, \quad (5.29)$$

where T denotes terms that do not depend on θ. (We have placed the product of $O_p\langle n^{-1/2}\rangle$ and terms in $\theta - \theta_0$ into the $O_p(\ldots)$ remainder at (5.29), since it admits the required bound with $(u, v) = (1, \frac{1}{2})$.)

Assume we have already proved that $\|\hat\theta - \theta_0\| = O_p(n^{-c})$ for some $c > 0$; we know from Step (i) that $c = \frac{1}{4} - \epsilon$, for any $\epsilon > 0$, is appropriate. It now follows from (5.29) that for each $\epsilon > 0$, $\|\hat\theta - \theta_0\|^2$ is of order equal to the minimum, over the bounded number of pairs (u, v) that satisfy $\frac{1}{2}u + v \geq 1$, of $n^{\epsilon - cu - v}$. That is, $\|\hat\theta - \theta_0\| = O_p(n^{\epsilon - (cu+v)/2})$ for all $\epsilon > 0$, where, in a minor abuse of notation, here

and below we write (u, v) for the value that gives the minimum. Repeating this argument with c replaced by $\epsilon - \frac{1}{2}(cu + v)$, for arbitrarily small $\epsilon > 0$; noting that the limiting value obtained by infinitely iterating the transformation that takes c to $\frac{1}{2}(cu + v)$, is $L(u, v) = v/2(1 - \frac{1}{2}u)$ if $u < 2$ (and infinity otherwise); and observing that the property $\frac{1}{2}u + v \geq 1$ is equivalent to $L(u, v) \geq \frac{1}{2}$; we deduce that (5.27) holds.

Finally, we complete the proof of the theorem. The argument leading to the remainder $O_p(\ldots)$ at (5.28) was conservative, in that each time the $n^{\epsilon - v}$ term there arose from a quantity of the form $n^{c_1} h^{c_2}$ for some $c_2 > 0$, we replaced h by its crude upper bound, $n^{-1/4}$ (see (3.4)); and each time $c_2 < 0$ we replaced h by its crude lower bound, $n^{-1/3}$. A more refined argument would use the property that $n^{\zeta_1 - (1/3)} < h < n^{\zeta_2 - (1/4)}$ for all sufficiently large n, where $\zeta_1, \zeta_2 > 0$ are fixed and in particular do not depend on the constants $\epsilon > 0$ in the $O_p(\ldots)$ remainder at (5.28), or in the $O_p(n^{\epsilon - (1/2)})$ term at (5.27). Since the ϵ's may be chosen arbitrarily small, whereas $\zeta_1, \zeta_2 > 0$ are fixed, then we may reduce the $O_p(\ldots)$ term at (5.28) to $O_p(\sum \|\theta - \theta_0\|^u \, n^{\epsilon - v - \zeta})$, for all $\epsilon > 0$ and some fixed $\zeta > 0$, excepting for the contribution to that remainder where the $n^{\epsilon - v}$ factor does not arise from a term $n^{c_1} h^{c_2}$ for some $c_2 \neq 0$.

There is only one contribution of this type, and it comes from the product of $\xi_{nj}(\mathcal{A}, Y_j, \theta)$ with $F(\mathcal{A}, Y_j) - G_\theta(\mathcal{A}, Y_j)$ at (5.19). Taylor-expanding the latter about $\theta = \theta_0$ we deduce the following refined version of (5.29):

$$\begin{aligned} S(\theta) = T + (\theta - \theta_0)^{\mathrm{T}} M_0 \, (\theta - \theta_0) - 2 \, (\theta - \theta_0)^{\mathrm{T}} (V_1 + V_2) \\ + o_p\big(\|\theta - \theta_0\|^2\big) + O_p\big(\textstyle\sum \|\theta - \theta_0\|^u \, n^{\epsilon - v - \zeta}\big) \,, \end{aligned} \tag{5.30}$$

where T does not depend on θ (although it is different from T at (5.29)), $\zeta > 0$ is fixed, $\epsilon > 0$ is arbitrarily small,

$$\begin{aligned} V_1 &= \int d\widehat{F}(y) \int_{\mathcal{R}_r} \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y) \, \xi_n(\mathcal{A}_\alpha, y, \theta_0) \, d\alpha \,, \\ V_2 &= \int d\widehat{F}(y) \int_{\mathcal{R}_r} D_{\theta_0}(\alpha, y) \, \dot{G}_{\theta_0}(\mathcal{A}_\alpha, y) \, d\alpha \,, \\ \xi_n(\mathcal{A}, y, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \Big[\psi(\mathcal{A}, X_i, Y_i, y, \theta) - E\{\psi(\mathcal{A}, X, Y, y, \theta)\} \Big] \,, \end{aligned}$$

and ψ is as at (3.5). (In deriving (5.30) we have placed an additional term into the $O_p(\ldots)$ remainder there, coming from the difference between ξ_n with θ and its

value with $\theta = \theta_0$. For the added term, $u = \frac{3}{2}$ and $v = \frac{1}{2}$. Since, without loss of generality, $\|\theta - \theta_0\| \le n^{-1/3}$, then we may alternatively write this remainder term as $\|\theta - \theta_0\|^u n^{\epsilon - v - \zeta}$, where $(u, v) = (1, \frac{1}{2})$, $0 < \zeta \le \frac{1}{6}$ and $\epsilon > 0$ is arbitrary.)

By (5.27), $\|\hat\theta - \theta_0\| = O_p(n^{\epsilon - (1/2)})$ for each $\epsilon > 0$. Therefore, on taking $\theta = \hat\theta$ the $O_p(\ldots)$ remainder at (5.30) may be written as $O_p(\sum n^{\epsilon - (u/2) - v - \zeta})$ for each $\epsilon > 0$. Since $\frac{1}{2}u + v \ge 1$ for each pair (u, v) contributing to the series, then the $O_p(\ldots)$ remainder in fact equals $O_p(\sum n^{\epsilon - 1 - \zeta}) = o_p(n^{-1})$, on choosing ϵ sufficiently small. Theorem 3.1 follows from this result, from (5.30) with $\theta = \hat\theta$, and from the fact that $n^{1/2}(V_1 + V_2)$ converges in distribution to V, the latter defined a little before the statement of the theorem.

Step (iii): Proofs of lemmas. Derivation of Lemma 5.2 is straightforward. Formulae (5.4)–(5.7) there are derived by standard Taylor expansion, in the bandwidth h, familiar for computing biases of kernel estimators. Formula (5.8) follows in the same manner, using Taylor expansion in both θ and h.

The idea behind the proofs of each part of each of the other four lemmas is as follows. Write the result as "$B = O(b)$," uniformly over a range of parameter values λ, which we shall represent as subscripts when it is necessary to refer to them. Let Λ denote the set of all values of λ. Derive a uniform upper bound for the $2m$th moment of B/b, for each integer $m \ge 1$, and in particular show that, for any given $c > 0$, $E(B/b)^{2m} = O(n^{-c})$ uniformly in λ, provided $m = m(c)$ is sufficiently large. It follows from this property, and Markov's inequality, that for any $c_1, c_2 > 0$, any $\zeta > 0$, and any subset Λ_n of Λ that has at most $O(n^{c_1})$ elements,

$$P\Big(\sup_{\lambda \in \Lambda_n} |B_\lambda / b_\lambda| > \zeta\Big) \le (\#\Lambda_n) \sup_{\lambda \in \Lambda_n} P\big(|B_\lambda / b_\lambda| > \zeta\big) = O\big(n^{-c_2}\big)\,. \tag{5.31}$$

Using the smoothness properties assumed of both the sampled distribution and K, it is straightforward to prove that if $q > 0$ is given then for $c = c(q) > 0$ sufficiently large we may choose a set $\Lambda_n \subseteq \Lambda$ that has at most n^c elements and for which

$$\sup_{\lambda_1 \in \Lambda} \inf_{\lambda_2 \in \Lambda_n} \big|(B_{\lambda_1}/b_{\lambda_1}) - (B_{\lambda_2}/b_{\lambda_2})\big| = O\big(n^{-q}\big)\,. \tag{5.32}$$

(Here the $O(n^{-q})$ bound is deterministically of that order, for any $n \ge 2$, and so does not have to be written as $O_p(n^{-q})$.) Any $q > 0$ will serve our present purpose; take $q = 1$ for definiteness.

Together results (5.31) and (5.32) imply that

$$P\Big(\sup_{\lambda\in\Lambda}|B_\lambda/b_\lambda|>\zeta\Big)\to 0\,,$$

for all $\zeta>0$. This is sufficient to prove the part of the lemma we are addressing.

Therefore, proofs of the lemmas reduce to deriving appropriate upper bounds to $2m$th moments of the left-hand sides. The bounds should not exceed constant multiples (depending only on m) of $2m$th moments of the right-hand sides. For the sake of brevity we shall give the arguments only in the cases of (5.3) and (5.16), showing in each that the expected value of the $2m$th power of the left-hand side is, for any given $c>0$ and all sufficiently large m, bounded above by n^{-c} times the $2m$th power of the right-hand side.

To derive a bound for the $2m$th moment of the left-hand side of (5.3), observe that it equals a sum of n independent and identically distributed random variables with zero mean, which we denote by $\sum_i U_i$. By Rosenthal's inequality (see p. 23 of Hall and Heyde (1980)),

$$E\Big(\sum_{i=1}^n U_i\Big)^{2m}\le C(m)\left[\big\{n\,E\big(U_1^2\big)\big\}^m+n\,E\big(U_1^{2m}\big)\right], \tag{5.33}$$

where $C(m)>0$ depends only on m. Let $\mathcal{S}$ denote the support of K, and let $C_1,C_2,\ldots$ be positive constants. Then $E(U_1^2)$ is bounded above by

$$\begin{aligned}
&\frac{C_1}{(nh)^2}\int\left[K_k\left\{\frac{\theta^{\mathrm{T}}(x-v)}{h}\right\}-K_k\left\{\frac{\theta_0^{\mathrm{T}}(x-v)}{h}\right\}\right]^2 f(v)\,dv\\
&\le\frac{C_2}{(nh)^2}\int\left|\frac{(\theta-\theta_0)^{\mathrm{T}}(x-v)}{h}\right|^2 I\left\{\frac{\theta^{\mathrm{T}}(x-v)}{h}\in\mathcal{S}\quad\text{or}\quad\frac{\theta_0^{\mathrm{T}}(x-v)}{h}\in\mathcal{S}\right\}f(v)\,dv\\
&\le\frac{C_3}{(nh)^2}\left(\frac{\|\theta-\theta_0\|}{h}\right)^2\int\left[I\left\{\frac{\theta^{\mathrm{T}}(x-v)}{h}\in\mathcal{S}\right\}+I\left\{\frac{\theta_0^{\mathrm{T}}(x-v)}{h}\in\mathcal{S}\right\}\right]f(v)\,dv\\
&\le C_4\,\|\theta-\theta_0\|^2/n^2h^3\,.
\end{aligned}\tag{5.34}$$

More simply, $E(U_1^{2m})$ is bounded above by a constant multiple of $(nh)^{-2m}h$. Therefore, the right-hand side of (5.33) is dominated by a constant multiple of

$$\big(\|\theta-\theta_0\|^2/nh^3\big)^m+(nh)^{1-2m}\,.$$

For any $c,\epsilon>0$, and sufficiently large m, this is bounded above by n^{-c} multiplied by the $2m$th power of the right-hand side of (5.3), as had to be shown.

Moments of order $2m$ of the quantities on the left-hand side of (5.16) may be bounded in similar fashion. Note that $Q_{\ell j}(\mathcal{A}, \theta)$, and hence also $\Delta \equiv Q_{\ell j}(\mathcal{A}, \theta) - Q_{\ell j}(\mathcal{A}, \theta_0)$, is constructed to be a degenerate U-statistic. In particular,

$$\Delta = \sum_{i, i_1 \,:\, i \neq i_1} \sum U(X_i, Y_i; X_{i_1}, Y_{i_1}) \,,$$

where the deterministic function U has the property

$$E\{U(X, Y, x, y)\} = E\{U(x, y, X, Y)\} = 0 \,.$$

The $2m$th moment of $Q_{\ell j}$ may thus be shown to be bounded above by a constant multiple of

$$\Big[n^2 \, E\big\{U(X_i, Y_i; X_{i_1}, Y_{i_1})^2\big\} \Big]^m + n^2 \, h^{-2m-1} \, E\big\{U(X_i, Y_i; X_{i_1}, Y_{i_1})^{2m}\big\} \,.$$

An argument similar to that at (5.34) shows that $E\{U(X_i, Y_i; X_{i_1}, Y_{i_1})^2\}$ is bounded above by a constant multiple of

$$\big(n^2 h\big)^{-2} \cdot h \cdot \big(\|\theta - \theta_0\| / h\big)^2 = n^{-2} \left\{ \|\theta - \theta_0\| \left(n h^{3/2}\right)^{-1} \right\}^2 .$$

Likewise, $E\{U(X_i, Y_i; X_{i_1}, Y_{i_1})^{2m}\}$ is bounded above by a constant multiple of $(n^2 h)^{-2m} \, h$. Therefore, $(E Q_{\ell j}^{2m})^{1/2m}$ is bounded above by a constant multiple of $\|\theta - \theta_0\| \, (n h^{3/2})^{-1} + n^{1/m} \, (nh)^{-2}$. For any $c, \epsilon > 0$, and sufficiently large m, this is bounded above by n^{-c} multiplied by the $2m$th power of the right-hand side of (5.16), as had to be proved.

5.2. Proof of Theorem 3.2. The theorem follows from the following result. To state it, let Θ_n denote the set of all $\theta \in \Theta$ that satisfy $\|\theta - \theta_0\| \le \delta(n) \, n^{-2/5}$, where $\delta(n) \downarrow 0$ as $n \to \infty$.

Lemma 5.6. *Assume* (3.2), (3.7), (3.8), *and that* $x \in \mathcal{R}$. *Then for each* y,

$$\sup_{\theta \in \Theta} \left| \widehat{F}_\theta\big(y \,\big|\, \theta^{\mathrm{T}} x\big) - \widehat{F}_{\theta_0}\big(y \,\big|\, \theta_0^{\mathrm{T}} x\big) \right| = o_p(n^{-2/5}) \,.$$

Next we outline the proof. Treat $\widehat{F}_\theta$ as the ratio expressed at (2.7), although multiply top and bottom there by $(nh)^{-1}$ (here $(nh_1)^{-1}$, since we take the bandwidth to be h_1) in order to ensure that neither the numerator nor the denominator converges to zero or diverges to infinity. The numerator and denominator are now

each in the form $T_1T_2 - T_3T_4$, where each T_j is linear in functions of the data X_i and has a proper limit as n diverges. Additively decompose each T_j into its expected value (or mean), and the difference between it and its mean. Each mean is of course purely deterministic. In the remainder of this section we shall outline the technique, starting from this decomposition, for treating T_1 and T_2; a similar argument may be given in the case of T_3 or T_4.

The expected value of T_1 or T_2 may be written as its "$h_1 \to 0$ limit," plus a term that equals h_1^2 multiplied by a function of θ, plus a remainder that equals $o(h_1^2)$ uniformly in θ. The "$h_1 \to 0$ limit," evaluated at θ, equals the same quantity evaluated at θ_0 rather than at θ, plus a remainder of order $O\{\delta(n)\, n^{-2/5}\} = o(n^{-2/5})$, uniformly in $\theta \in \Theta_n$; and similarly, the coefficients of h_1^2 (for θ and θ_0, respectively) are identical, up to a term that converges to 0 uniformly in $\theta \in \Theta_n$ as $n \to \infty$. These arguments require only Taylor expansion, and prove that the mean of each of the T_j's equals its counterpart when θ is replaced by θ_0, plus terms that are of size $o(n^{-2/5})$ uniformly in $\theta \in \Theta_n$. Call this result (R); it applies to each of the terms $T_1, \ldots, T_4$.

It remains only to show that the same is true for the stochastic differences from the means. This is accomplished by parallelling the argument in Step (iii) of the proof of Theorem 3.1, as follows. Subtract from each difference its counterpart when $\theta = \theta_0$. This "difference of differences" is a sum of independent random variables with zero expected value. The $2m$th moment of the sum is bounded by a constant multiple of the mth power of its variance, the constant not depending on θ or n. Using the fact that K has a bounded derivative, we may show that the variance is bounded by a constant multiple of $(nh_1)^{-1} \{\|\theta - \theta_0\|/h_1\}^2 = O(n^{-6/5})$, uniformly in $\theta \in \Theta_n$. Therefore, by Markov's inequality based on $2m$th moments, the probability that any of the differences (for the T_j's) exceeds $\epsilon\, n^{-2/5}$, is bounded above by a constant multiple of $n^{-2m/5}$. If $C > 0$ is given, and we choose $m > 5C/2$, we may deduce from this result that the probability that at least one of the differences exceeds $\epsilon\, n^{-2/5}$ for at least one point θ in a subset $\mathcal{S}_n$ of $O(n^C)$ values of $\theta \in \Theta_n$, converges to zero as $n \to \infty$. Choosing C large, and using the Hölder continuity of K (which is a consequence of the boundedness of K'), we deduce that the same result applies if "at least one point θ in a subset of $\mathcal{S}_n$..." is replaced by "at least one point in Θ_n."

Therefore, the stochastic differences from the means (of the T_j's), for general $\theta \in \Theta_n$, differ from their respective values when $\theta = \theta_0$, by only $o_p(n^{-2/5})$ uniformly in $\theta \in \Theta_n$. The theorem follows from this property and result (R).

REFERENCES

BASHTANNYK, D.M. AND HYNDMAN, R.J. (2001). Bandwidth selection for kernel conditional density estimation. *Computat. Statist. Data Anal.* **36**, 279–298.

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

CAI, Z. (2002). Regression quantiles for time series. *Econometric Theory* **18**, 169–192.

FAN, J. AND YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

FAN, J., YAO, Q. AND TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

FRIEDMAN, J.H. AND STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817–823.

FRIEDMAN, J.H., STUETZLE, W. AND SCHROEDER, A. (1987). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79**, 599–608.

FRIEDMAN, J.H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82**, 249–266.

HALL, P. AND HEYDE, C.C. (1980). *Martingale Limit Theory and its Application*. Academic, New York.

HALL, P., WOLFF, R.C.L. AND YAO, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**, 154–163.

HUBER, P.J. (1985). Projection pursuit. (With discussion.) *Ann. Statist.* **13**, 435–525.

HYNDMAN, R.J., BASHTANNYK, D.M. AND GRUNWALD, G.K. (1996). Estimating and visualizing conditional densities. *J. Comp. Graph. Statist.* **5**, 315–336.

HYNDMAN, R.J. AND YAO, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametric Statist.*, to appear.

JONES, M.C. AND SIBSON, R. (1987). What is projection pursuit? (With discussion.) *J. Roy. Statist. Soc.* Ser. A **150**, 1–36.

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

MORAN, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle, I: structure and prediction. *Aust. J. Zool.* **1**, 163–173.

POSSE, C. (1995). Projection pursuit exploratory data analysis. *Comput. Statist. Data Anal.* **20**, 669–687.

PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. AND FLANNERY, B.P. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge.

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In: P. Krishnaiah, ed., *Multivariate Analysis II*. Academic, New York, pp. 25–31.

SHEATHER, S.J. AND MARRON, J.S. (1990). Kernel quantile estimators. *J. Amer. Statist. Assoc.* **85**, 410–416.

YIN, X. AND COOK, R.D. (2002). Dimension reduction for conditional kth moment in regression. *J. Royal Statist. Soc.* **B 64**, 159–175.

YU, K. AND JONES, M.C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228–237.

Caption for Figure 1: *Simulation results for Example 1.* Boxplots of (a) inner product $\theta^{\mathrm{T}}\hat{\theta}$, (b) bandwidths estimated by cross-validation, and (c) average absolute errors of estimated conditional distribution of Y given $\theta^{\mathrm{T}}X$ with either $\theta = \hat{\theta}$ (denoted by "E") or θ equal to its true value (denoted by "T").

Caption for Figure 2: *Simulation results for Example 2.* Panels show the same information as in Figure 1.

Caption for Table 1: *Prediction intervals for Canadian lynx data in 1925–1934, based on data observed in 1821–1924.*