SEMIPARAMETRIC ESTIMATION OF SINGLE-INDEX CONDITIONAL DENSITIES FOR DEPENDENT DATA

By

Roy Rosemarin

Draft, Revised March 21, 2012

Abstract

Nonparametric methods of estimation of conditional density functions when the dimension of the explanatory variable is large are known to suffer from slow convergence rates due to the 'curse of dimensionality'. When estimating the conditional density of a random variable Y given random d-vector X, a significant reduction in dimensionality can be achieved by approximating the conditional density by that of a Y given $\theta^T X$, where the unit-vector θ is chosen to optimise the approximation under the Kullback-Leibler criterion. The proposed estimation procedure is based on standard kernel methods. Under strong-mixing conditions, we derive a general asymptotic representation for the orientation estimator, and as a result, the approximated conditional density is shown to enjoy the same first-order asymptotic properties as it would have if the optimal θ was known. The method is illustrated in a simulation study with nonlinear time series models.

1 Introduction

Conditional probability density functions play a key role in many statistical applications, including regression analysis (Yin and Cook 2002), forecasting (Hyndman 1995, Fan and Yao 2003), sensitivity to initial conditions in nonlinear stochastic dynamic systems (Yao and Tong 1994, Fan, Yao and Tong 1996), quantiles estimation (Engle and Manganelli 2004, Wu, Yu and Mitra 2008), and asset pricing (Aït-Sahalia 1999, Engle 2001), among many others. In this paper we consider estimation of the conditional density $f_{Y|X}(y|x)$ of Y given $X = x$, where Y is a random scalar X is a random d-vector.

If the conditional density is assumed to be normal, then the estimation of the predictive density reduces to estimation of the conditional mean and autocovariances. However, in real life probability densities are often charachterised by asymmetry, heavy-tails, and even multimodality. Moreover, even for a known parametric model, the conditional density may be difficult to derive analytically. In such cases, a nonparametric estimation of the conditional density can be useful. Nonetheless, even for small dimension of X, $d \geq 2$, nonparametric estimators are known to suffer from slow convergence rates and unstable performance in practice due to the 'curse of dimensionality' and the 'empty space phenomenon' (see Silverman 1986, section 4.5). For this reason, we suggest approximating the conditional density $f_{Y|X}(y|x)$ by $f_{Y|\theta^T X}\left(y|\theta^T x\right)$, the conditional density of Y given $\theta^T X = \theta^T x$, where the orientation θ is a scalar-valued d-vector that minimises the Kullback-Leibler (K-L) relative

entropy,

$$E \log f_{Y|X}(y|x) - E \log f_{Y|\theta^T X}\left(y|\theta^T x\right). \tag{1}$$

The approximated conditional density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is estimated nonparametrically by a kernel estimator. In doing so, our approach provides a low dimensional approximation of the conditional density which is optimal under the Kullback-Leibler criterion.

In the popular single-index regression model it is typically assumed that $Y = g\left(\theta^T X\right) + \varepsilon$, where g is some link function and ε is a noise term such that $E(\varepsilon|X) = 0$ (see Ichimura 1993). Our methodology differs from this regression model by aiming for the most informative projection $\theta^T X$ of X to explain the conditional density of Y given X, rather than just the conditional mean. However, that is not to say that the true conditional distribution of $Y|X$ is assumed to be the same as that of $Y|\theta^T X$. The method aims to provide the optimal single-index conditional density approximation possible for a general $f_{Y|X}(y|x)$.

The approach of using the K-L relative entropy for estimation of orientation has been utilised by Delecroix, Härdle and Hristache (2003) in single-index regression, Yin and Cook (2005) for dimension reduction subspace estimation, and by Fan et al (2009), who similar to us, dealt with conditional densities. Yin and Cook (2005) discuss several equivalent presentations of the K-L relative entropy and they show relations to inverse regression, maximum likelihood and other ideas from information theory. Our work extends the approaches taken by the above papers in two main

aspects.

First, by allowing the data to be stationary strong mixing, the suggested method is shown to be applicable for dependent data, and in particular to the estimation of predictive densities in high-dimensional time series. As an example, ARMA, GARCH and stochastic volatility processes were proved to be strong-mixing under some mild conditions (cf. Pham and Tran 1985, Carrasco and Chen 2002, Davis and Mikosch, 2009), and our method can be applied to these series when the assumption of Gaussianity is not applicable. For a general univariate strong mixing series $\{z_t\}_{t=1}^{n+d+k-1}$, let

$$y_t = Z_{t+d+k-1}, \quad x_t = (Z_{t+d-1}, ..., Z_t)^T, \quad t = 1, ..., n.$$

Then $f_{Y|\theta^T x}\left(y_t|\theta^T x_t\right)$ provides a k-steps ahead conditional density based on the d-lagged vector x_t, which allows generalising standard time series models to possibly nonlinear or nongaussian processes.

As a second contribution, we derive a general asymptotic representation for the difference between the orientation estimator $\widehat{\theta}$ and the unknown optimal orientation θ_0 that is equal to a sum of zero-mean asymptotic Gaussian components with $\sqrt{n}$-rate of convergence and two other, stochastic and deterministic, components. The representation holds for kernels of either order two or four, while the asymptotically dominant terms are determined by the order of kernels in use and the choice of kernel bandwidths.

Kernels of high-order benefit from reduced asymptotic bias in the estimation,

yet they take negative values and thus often produce negative density estimates. An investigation by Marron and Wand (1992) of higher order kernels for density estimation concluded that the practical gain from higher order kernels is often absent or insignificant for realistic sample sizes (see also Marron 1992 for graphical insight into the effectiveness high-order kernels). Our proposed procedure allows estimating θ_0 with high-order kernels, while then estimating the conditional density with non-negative second-order kernels.

We carry out a numerical study to compare the performances of the orientation estimators obtained with second and fourth order kernels. Our results indicate that despite having better asymptotic properties, orientation estimators obtained with fourth-order kernels perform poorly relative to those obtained with only second-order non-negative kernels. Our conclusion is that the 'failure' of high-order kernels to attain their theoretical benefit in realistic sample sizes carries through to the estimation of the orientation.

The outline for the rest of paper is as follows. Section 2 states the model's general setting and estimation methodology; Section 3 contains the assumptions and main theoretical results; and Section 4 presents a Monte-Carlo study with three simulated time series examples. The proofs of the main theorems are given in Appendix A, while some other technical lemmas are outlined in Appendix B.

2 Model and Estimation

Let $\{y_j, x_j\}_{j=1}^n$ be strictly stationary strong mixing observations with the same distribution as (Y, X), where Y is a random scalar and X is a random d-vector. Our aim is to estimate the conditional density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ of Y given a random d-vector $\theta^T X{=}\theta^T x$, where θ is a vector in $\mathbb{R}^d$ that minimises the K-L relative entropy (1). Since the first term of the K-L relative entropy does not depend on θ, minimising K-L relative entropy is equivalent to maximising the expected log-likelihood $E \log f_{Y|\theta^T X}\left(y|\theta^T x\right)$. Clearly, the orientation θ is identifiable only with regards to its direction, and we therefore consider unit-vectors that belong to the compact parameter space

$$\Theta = \left\{\theta \in \mathbb{R}^d : \theta^T\theta = 1,\ \theta_1 \geq c > 0\right\},$$

where θ_1 is the first element of the orientation and $c > 0$ is arbitrarily small. For example, if Y_t is the k-step ahead observation of a time series and X_t consists of d lagged values of the series, then the constraint that $\theta_1 \neq 0$ represents the belief that the k-step ahead observation depends on the most recent observed value.

In order to ensure the uniform convergence of our estimator, we need to restrict ourselves to a compact subset of the support of $Z = (Y, X)$ such that for any $\theta \in \Theta$ the probability density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is well defined and bounded away from 0. Denote such a subspace by $\mathbb{S}$, and let also $\mathbb{S}_X = \left\{x \in \mathbb{R}^d : \exists y \text{ s.t. } (y, x) \in \mathbb{S}\right\}$.

Let θ_0 be the maximiser of expected log-likelihood conditional on $Z \in \mathbb{S}$, that is,

$$\theta_0 = \arg\max_{\theta \in \Theta} E_{\mathbb{S}} \left(\log f_{Y|\theta^T X} \left(Y | \theta^T X \right) \right), \tag{2}$$

where $E_{\mathbb{S}}$ is the conditional expectation given $Z \in \mathbb{S}$. Note that the condition $Z \in \mathbb{S}$ should not have any significant effect on θ_0 if the subset $\mathbb{S}$ is large enough. For ease of presentation, we shall assume that all observations $\{y_j, x_j\}_{j=1}^n$ belong to $\mathbb{S}$.

To estimate θ_0 one can maximise a sample version of (2). Define the orientation estimator by $\widehat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$, where $\mathcal{L}(\theta)$ is the likelihood function

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}^{-i}_{Y|\theta^T X} \left(y_i | \theta^T x_i \right) \widehat{\rho}_i^{\theta}. \tag{3}$$

Here, $\widehat{\rho}_i^{\theta}$ is a trimming term, which is discussed below, and with probability 1 it is eventually equals to 1 for large enough n. The unknown conditional density is estimated by a nonparametric kernel estimator

$$\widehat{f}^{-i}_{Y|\theta^T X} \left(y_i | \theta^T x_i \right) = \frac{\widehat{f}^{-i}_{Y,\theta^T X} \left(y_i, \theta^T x_i \right)}{\widehat{f}^{-i}_{\theta^T X} \left(\theta^T x_i \right)}.$$

where $\widehat{f}_{Y,\theta^T X} \left(y, \theta^T x \right)$ and $\widehat{f}_{\theta^T X} \left(\theta^T x \right)$ denote the standard kernel probability density estimates, whereas the superscript '$-i$' indicates exclusion of the i'th observation from the calculation, that is,

$$\begin{aligned}
\widehat{f}^{-i}_{Y,\theta^T X} \left(y_i, \theta^T x_i \right) &= \left\{ (n-1) \, h_y h_x \right\}^{-1} \sum_{j \neq i} K \left(\frac{y_j - y_i}{h_y} \right) K \left(\frac{\theta^T \left(x_j - x_i \right)}{h_x} \right), \\
\widehat{f}^{-i}_{\theta^T X} \left(\theta^T x_i \right) &= \left\{ (n-1) \, h_x \right\}^{-1} \sum_{j \neq i} K \left(\frac{\theta^T \left(x_j - x_i \right)}{h_x} \right),
\end{aligned}$$

where $h_y, \; h_x$ are bandwidths and K is a fixed, bounded-support, kernel function.

The trimming term $\widehat{\rho}_i^{\theta}$ that appears in (3) is introduced to stabilise the finite-sample performances of the algorithm. To appreciate the role of this term, observe that even if observation (y_i, x_i) belongs to $\mathbb{S}$ it may still be the case that the kernel estimates $\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)$, $\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)$ rely on very few neighbouring observations, or even none, and as a result these estimates may be close to zero and even non-positive when high-order kernels are used. Including $\log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)$ in the computation of the likelihood function in such cases may have a drastic adverse effect on the accuracy of the likelihood surface estimates, and it is therefore preferable to trim such terms. In this paper, we adopted the following simple trimming scheme, which works very well in practice (For alternative trimming schemes cf. Härdle and Stoker 1989, Ichimura 1993, Delecroix, Hristache and Patilea 2006, Ichimura and Todd 2006 and Xia Härdle and Linton 2012). For a given observation (y_i, x_i) and $\theta \in \Theta$, let

$$I^i_{n,\theta} = \begin{cases} \mathbf{1}, & \text{if } \min\left\{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right), \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)\right\} > a_0 n^{-c}, \\ 0, & \text{otherwise,} \end{cases}$$

for some small constants $a_0, c > 0$. As $I^i_{n,\theta}$ depends on θ it needs to be normalised to account for the actual number of observations considered in the computation of $\mathcal{L}(\theta)$, and hence we take

$$\widehat{\rho}_i^{\theta} = I^i_{n,\theta} \Big/ \frac{1}{n}\sum_{i=1}^n I^i_{n,\theta}. \tag{4}$$

We show in appendix B that if c is sufficiently small, then $\widehat{\rho}_i^{\theta}$ eventually equals to 1 for any large enough n with probability 1. Therefore, $\widehat{\rho}_i^{\theta}$ has no asymptotic effect

on the method performance.

It is common in single-index regression models, that the kernel's bandwidths for orientation estimation are required to undersmooth the nonparametric estimator of the link function (cf. Hall 1989, p. 583). Our theory indicates that a similar property arises in single-index conditional density estimation. It is therefore the reason that a second stage of estimation is utilised when now $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is estimated with the already estimated orientation $\widehat{\theta}$ and with optimal bandwidths H_y and H_x.

Let the conditional density estimator be obtained with non-negative symmetric kernels $\widetilde{K}$ with all observations and bandwidths H_y and H_x in place of h_y and h_x,

$$\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) = \frac{\frac{1}{nH_yH_x}\sum_{j=1}^{n}\widetilde{K}\left(\frac{y_j-y}{H_y}\right)\widetilde{K}\left(\frac{\widehat{\theta}^T(x_j-x)}{H_x}\right)}{\frac{1}{nH_x}\sum_{j=1}^{n}\widetilde{K}\left(\frac{\widehat{\theta}^T(x_j-x)}{H_x}\right)}.$$

The following section presents the asymptotic properties of $\widehat{\theta}$ and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$.

3 Asymptotic Results

We introduce some new notations that will be used throughout the section and in the proofs. For a function $g(\theta)$ that possibly also depends on y and x, let $\nabla g(\theta)$ and $\nabla^2 g(\theta)$ be the vector and matrix of partial derivatives of $g(\theta)$ with respect to θ, i.e.,

$$\left\{\nabla g(\theta)\right\}_k = \frac{\partial g(\theta)}{\partial \theta_k} \quad \text{and} \quad \left\{\nabla^2 g(\theta)\right\}_{k,l} = \frac{\partial^2 g(\theta)}{\partial \theta_k \partial \theta_l}, \quad k,l \in \{1,...,d\}.$$

Denote now $Z = (X, Y)$ and

$$\begin{aligned}\Psi(\theta) &= E_{\mathbb{S}}\left[\nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right) \nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)^T\right], \\ \Omega(\theta) &= E_{\mathbb{S}}\left[-\nabla^2 \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)\right].\end{aligned}$$

where $E_{\mathbb{S}}$ is the conditional expectation given $Z \in \mathbb{S}$. For some small $\delta > 0$, define also the set $\mathbb{S}_\delta$ distant no further than $\delta > 0$ from some $\left(y, \theta^T x\right)$ such that $(y, x) \in \mathbb{S}$ and $\theta \in \Theta$.

The following assumptions are required to obtain the asymptotic results for the orientation estimator $\widehat{\theta}$.

(A1) $(Y, X) \in \mathbb{R}^{d+1}$ is strong mixing with mixing coefficients that satisfy $\alpha_t \leq At^{-\eta_1}$ with $0 < A < \infty$ and $\eta_1 > 3$, and $X \in \mathbb{R}^d$ is strictly stationary and strong mixing with $\alpha_t \leq Bt^{-\eta_2}$ with $0 < B < \infty$ and $\eta_2 > 2$.

(A2) $K(\cdot)$ is a symmetric, compactly supported, boundedly differentiable kernel.

(A3) The bandwidths satisfy $h_y, h_x = o(1)$ and $\frac{\ln n}{n^{\phi_1} h_y h_x} = o(1)$ with $\phi_1 = (\eta_1 - 3)/(\eta_1 + 1)$ and $\frac{\ln n}{n^{\phi_2} h_x} = o(1)$ with $\phi_2 = (\eta_2 - 2)/(\eta_2 + 2)$.

(A4) For all $\theta \in \Theta$, $\left(Y, \theta^T X\right)$ has probability density $f_{Y,\theta^T X}(y, t)$ with respect to Lebesgue measure on $\mathbb{S}_\delta$ and $\inf_{(y,t)\in\mathbb{S}_\delta} f_{Y,\theta^T X}(y, t) > 0$. $f_{Y,\theta^T X}(y, t)$ and $E\left(X|Y = y, \theta^T X = t\right)$ and $E\left(XX^T|Y = y, \theta^T X = t\right)$ are twice continuously differentiable with respect to $(y, t) \in \mathbb{S}_\delta$. Moreover, there is some j^* such that for all $j > j^*$ and $\left(Y_1, \theta^T X_1\right), \left(Y_j, \theta^T X_j\right) \in \mathbb{S}_\delta$ the joint probability density of $\left(Y_1, \theta^T X_1, Y_j, \theta^T X_j\right)$ is bounded.

(A5) For the trimming operator, we require that $a_0, c > 0$ and $n^c\left(h_y^2 + h_x^2\right) =$

$o(1)$ and $n^{1-2c}h_y h_x \to \infty$.

(A6) For all $\theta \in \Theta$, $E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\right)$ is finite and it has a unique global maximum θ_0 that lies in the interior of Θ.

We further require that $K(\cdot)$ is either a second-order or a fourth-order kernel function, such that

$$\int u^j K(u)\, du = 0 \quad \text{for} \quad j = 1, ..., p-1, \quad \text{and} \quad \int u^p K(u)\, du \neq 0,$$

for $p = 2$ or $p = 4$. We then make the following assumptions.

(A7) $K(\cdot)$ is p'th-order kernel with $p = 2$ or $p = 4$, and it is three times boundedly differentiable.

(A8) The bandwidths $h_y,\ h_x$ satisfy $n^{2-\delta}h_y h_x^5 \to \infty$ for some $\delta > 0$.

(A9) $f_{Y,\theta^T X}(y,t)$ and $E\left(X|Y=y, \theta^T X = t\right)$ and $E\left(XX^T|Y=y, \theta^T X = t\right)$ are $(2+p)$-times continuously differentiable with respect to $(y,t) \in \mathbb{S}_\delta$.

(A10) $w\Omega(\theta_0)w^T > 0$ for any non-zero d-vector $w \perp \theta_0$.

Conditions (A1)-(A6) are needed for uniform consistency of the log-likelihood function on $\Theta \times \mathbb{S}$, and therefore for consistency of $\widehat{\theta}$. In particular, condition (A1) allows the data to come from a strong mixing process. For many common time series processes (e.g. ARMA, GARCH), the mixing coefficients decay exponentially and η_1 and η_2 can be taken as $+\infty$. Condition (A2) requires that $K(\cdot)$ is symmetric and therefore it is of second-order at the least. Condition (A3) on the bandwidths is needed to obtain uniform convergence of the kernel density estimators. In condition (A4), the bound on the joint probability density of $\left(\theta^T X_1, Y_1, \theta^T X_j, Y_j\right)$ may not hold

for $j \leq j^*$, which allows components of X_1 and X_j to overlap for some small j's, as in the case where X_t consists of multiple lags of Y_t. Condition (A5) for the trimming operator terms is derived from Lemma 8 in the appendix. (A6) is an identifiability requirement for θ_0. Conditions (A7)-(A9) are stronger versions of (A2)-(A4) and are needed for the derivation of the rate of consistency of $\widehat{\theta}$. Condition (A8) discusses rate of decay for the bandwidths. It implies together with condition (A3) that the exponent for the mixing coefficients in (A1) cannot decay too slowly. For example, if both bandwidths h_y, h_x are taken to be proportional to $n^{-\gamma}$, $\gamma > 0$, then by (A3) and (A8) γ must satisfy

$$0 < \gamma < \min\left\{1/3, \frac{\eta_1 - 3}{2(\eta_1 + 1)}, \frac{\eta_2 - 2}{\eta_2 + 2}\right\}.$$

Finally, condition (A10) is a standard requirement (see Hall and Yao 2005).

We now turn to state the main theorems of the paper, proved in appendix A. The following theorem shows the consistency of $\widehat{\theta}$.

Theorem 1 *Let (A1)-(A6) hold. Then as* $n \to \infty$

$$\widehat{\theta} \to_p \theta_0.$$

As an implication of Theorem 1 and the fact that both θ_0 and $\widehat{\theta}$ are unit-vectors, it follows from a simple geometric argument that the difference $\widehat{\theta} - \theta_0$ can be approximated up to first-order asymptotics by $\widehat{\theta}^{\perp}$, the projection of $\widehat{\theta}$ into the plane orthogonal to θ_0, i.e.,

$$\widehat{\theta} - \theta_0 = \widehat{\theta}^{\perp} + o_p\left(\left\|\widehat{\theta} - \theta_0\right\|\right).$$

Since $f_{Y|\theta^T X}\left(Y|\theta^T X\right)$ depends only on the direction of θ, then for any vector $\theta \in \mathbb{R}^d$ we get that both vector $\nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)$ and the column (row) space spanned by matrix $\nabla^2 \log f_{Y|\theta^T X}\left(y|\theta^T x\right)$ are perpendicular to θ. Indeed, this can also be seen directly from Lemma 4 in appendix B. Note, however, that by (A10) there is a generalised inverse of $\Omega(\theta_0)$, denoted $\Omega(\theta_0)^-$, that is well defined in the perpendicular space to θ_0. Let now $V(\theta_0) = \Omega(\theta_0)^- \Psi(\theta_0) \Omega(\theta_0)^-$. The next theorem gives a general second-order asymptotic representation for $\widehat{\theta} - \theta_0$.

Theorem 2 *Let (A1)-(A10) hold. Then*

$$\widehat{\theta} - \theta_0 = n^{-1/2} V(\theta_0)^{1/2} Z + O_p\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p),$$

where Z is asymptotically normal $N(0, I)$ random d-vector and $\delta > 0$ arbitrarily small.

It is clear from this theorem that for $\widehat{\theta}$ to be $\sqrt{n}$-consistent estimator of θ_0, one needs

$$\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} \leq n^{-1/2} \quad \text{and} \quad h_y^p + h_x^p \leq n^{-1/2}. \tag{5}$$

However, it is easy to see that both conditions cannot be satisfied if $p = 2$, and hence the $\sqrt{n}$-convergence rate is not achieved in that case (cf. Remark 2 of Fan et al 2009), although the convergence rate can still become arbitrarily close to $\sqrt{n}$. By increasing the order of the kernel to $p = 4$, the condition (5) can be fulfilled under $h_y, h_x \leq n^{-1/8}$ and $h_y h_x^3 \geq n^{\delta - 1}$, and if the two last inequalities are strict, then the Theorem implies asymptotic normality of the estimate.

The asymptotic expression given by Theorem 2 at the limit $\delta \to 0$ suggests that the optimal bandwidths h_y and h_x have both the asymptotic rate $n^{-1/(p+2)}$, where p is the kernel's order. Taking $p = 2$, for example, we have that the optimal bandwidths are of asymptotic order $n^{-1/4}$. This optimal rate reflects undersmoothing of the kernel estimator, which is a typical requirement in many single-index models.

Under appropriate choice of bandwidths, $\widehat{\theta}$ can converge fast enough to θ_0 so that $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$ estimates $f_{Y|\theta_0^T X}\left(y|\theta_0^T x\right)$ with the same first-order asymptotic properties as if θ_0 was known. The Theorem below formalises this idea.

Theorem 3 *Let (A1)-(A10) hold and* $H_y H_x / h_y h_x^3 = o\left(n^{1-\delta}\right)$ *for some* $\delta > 0$ *and* $H_y H_x \left(h_y^{2p} + h_x^{2p}\right) = o\left(n^{-1}\right)$. *In addition let* $\widetilde{K}$ *be a symmetric, compactly supported, boundedly differentiable kernel, and* $H_y, H_x = O\left(n^{-1/6}\right)$ *and* $\frac{\ln n}{n H_y H_x} = o(1)$. *Then for any* $\delta > 0$,

$$\sup_{(y,x)\in\mathbb{S}} \left| \widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) - f_{Y|\theta_0^T X}\left(y|\theta_0^T x\right) \right| = O_p\left(\left(\frac{\ln n}{n H_y H_x} \right)^{1/2} \right).$$

4 Implementation and Simulations

In this section, we discuss implementation of the proposed method and we examine its finite-sample properties over few simulated time series models.

In all of the simulations we used the three-time differentiable and IMSE optimal kernels with support $(-1, 1)$, derived by Müller (1984) and specified below. The second-order Müller's kernel, also known as the Triweight kernel, is given for $u \in$

$(-1,1)$ by

$$K(u) = 35/32 \cdot \left(1 - 3u^2 + 3u^4 - u^6\right), \tag{6}$$

and the fourth-order Müller's kernel is given for $u \in (-1,1)$ by

$$K(u) = 315/512 \cdot \left(3 - 20u^2 + 42u^4 - 36u^6 + 11u^8\right).$$

For the estimation of the conditional density $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$ we use only the non-negative Triweight kernel.

In order to facilitate the implementation, we standardised $x_j = (x_{j1}, ..., x_{jd})$ by setting $x_j \leftarrow S_x^{-1}(x_j - \overline{x})$ and we standardised y_j by setting $y_j \leftarrow (y_j - \overline{y})/s_y$, where $\overline{x}$ and $\overline{y}$ are the vector and scalar sample means of $\{x_j\}_{j=1}^n$ and $\{y_j\}_{j=1}^n$, and S_x^2 and s_y^2 are the $d \times d$-matrix and the scalar sample variances. Once the two-stage estimation procedure was complete, the estimates of the orientation and the conditional density were transformed back to the original coordinates by setting $\widehat{\theta} \leftarrow S_x^{-1}\widehat{\theta} / \left\| S_x^{-1}\widehat{\theta} \right\|$ and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) \leftarrow \widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)/s_y$.

We now provide a brief discussion on the topic of bandwidths selection. Typical bandwidths selection methods proposed in the literature of single-index models usually suffer from heavy computational burden (cf. Xia, Tong, Li 1999, Härdle, Hall, and Ichimura 1993, Hall and Yao 2005). Such burden may be particularly noticeable in models like ours, where the estimation requires solving a numerical multivariate optimisation problem. In practice, however, various prior numerical studies that we carried out with different selection rules for h_y and h_x demonstrated that the orientation estimator is very robust to the choice of bandwidths as long as the band-

widths are not too small. Motivated by the single-index regression algorithm of Xia, Härdle, Linton (2012), we propose the following iterative procedure that successfully reconciles effective bandwidth selection with fast and robust numerical optimisation.

Step 0. Let $\widehat{\theta}^0 \in \Theta$ be any initial guess for θ_0, for example $\widehat{\theta}^0 = (1, 0, ..., 0)$. Set also a finite sequences of decreasing bandwidths $h_y^\tau = h_x^\tau = a^\tau n^{-1/(p+2)}$, $\tau = 1, ..., T$, where p is the kernel-order and $\{a^\tau\} > 0$ is a decreasing sequence such that the first bandwidths notably oversmooth the unconditional density and the last one is chosen, e.g., by Scott's (1992) normal reference rule. In our simulations, we used $\left(a^1, a^2, ..., a^T\right) = (9, 8, ..., 3)$, which yields good results. Set the iteration number $\tau = 1$.

Step 1. Apply a multivariate variant of the Newton-Raphson method with starting point $\widehat{\theta}^{\tau-1}$ to find a maximum log-likelihood estimate $\widehat{\theta}^\tau$ numerically based on bandwidths h_y^τ and h_x^τ (e.g. use the Broyden-Fletcher-Goldfarb-Shanno BFGS method).

Step 2. Stop the procedure and use the estimate $\widehat{\theta} = \widehat{\theta}^\tau$ either if $\tau = T$ or if a certain convergence criterion is met, i.e. if $\left(\widehat{\theta}^\tau\right)^T \widehat{\theta}^{\tau-1} > 1 - \varepsilon$ for some small $\varepsilon > 0$. Otherwise, set $\tau \leftarrow \tau + 1$ and $h_y^\tau = h_x^\tau = a^\tau n^{-1/(p+2)}$, and return to Step 1.

Note that since $h_y^1 = h_x^1 = 9n^{-1/(p+2)}$ are chosen to oversmooth the conditional density in the first iteration of estimation, the corresponding likelihood surface is thus oversmoothed as well, and the optimisation algorithm is insensitive to the choice of $\widehat{\theta}^0$. On the other hand, if we simply use one step of maximization with

only $h_y = h_x = 3n^{-1/(p+2)}$, then the algorithm is very likely to converge to some local maximum, depending on the starting point $\widehat{\theta}^0$ provided.

For the second stage estimator of the conditional density, $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$, we used Scott's (1992) normal reference rule for bandwidth selection, which suggests using bandwidths given by $H_y = H_x = an^{-1/6}$, where for Triweight kernel (6) $a \approx 3$.

In all of our simulations, we used all observations in both stages of the estimation, but we set $\widehat{\rho}_i^{\theta}$ to trim down only observations whose density estimates were non-positive.

The performances of the proposed methods are demonstrated in the following three examples of simulated time series models.

Example 1. As a first example, we consider the linear AR(4) model

$$y_t = 0.5 \cdot \sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j} + 0.5 \cdot \varepsilon_t,$$

where $\theta_0^T \equiv (\theta_{0,1}, .., \theta_{0,4}) = (3, 2, 0, -1) / \sqrt{14}$ and ε_t are i.i.d. $N(0, 1)$.

Example 2. In the next example we consider the nonlinear AR(4) model

$$y_t = g\left(\sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j}\right) + 0.5 \cdot \varepsilon_t,$$

where $g(u) = \exp\left(\left(0.4 - 2u^2\right) u\right)$, $\theta_0^T \equiv (\theta_{0,1}, .., \theta_{0,4}) = (1, 2, -1, 0) / \sqrt{6}$, and the ε_t are as in Example 1.

Example 3. Finally we would like to examine how the method works where the optimal projection $\theta_0^T X$ is related to higher moments of X. For the third example,

TABLE 1: Mean and Standard error (in brackets) of the inner product $\widehat{\theta}^T\theta_0$.

$k.order$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
Example 1				
$p = 2$	0.9241 (0.0776)	0.9630 (0.0312)	0.9758 (0.0258)	0.9864 (0.0182)
$p = 4$	0.8958 (0.1026)	0.8962 (0.0949)	0.8953 (0.0799)	0.9113 (0.0529)
Example 2				
$p = 2$	0.8867 (0.2193)	0.9632 (0.1325)	0.9809 (0.1043)	0.9936 (0.0654)
$p = 4$	0.7114 (0.2617)	0.7203 (0.2969)	0.7122 (0.2970)	0.7439 (0.3154)
Example 3				
$p = 2$	0.6412 (0.2864)	0.7374 (0.2636)	0.8703 (0.1689)	0.9301 (0.0961)
$p = 4$	0.6858 (0.2757)	0.8131 (0.2201)	0.8914 (0.1534)	0.9195 (0.0975)

we consider the nonlinear ARCH(4) model

$$y_t = g\left(\sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j}\right) \cdot \varepsilon_t,$$

where $g(u) = 0.5\sqrt{1+u^2}$. Here, $\theta_{0,j} = \exp(-j) / \sqrt{\sum_{k=1}^{4} \exp(-2k)}$, $j = 1, ..., 4$, and the ε_t are as in the previous examples.

All the three models can easily be verified to be geometrically ergodic by either Theorem 3.1 or Theorem 3.2 of An and Huang (1996), and hence they are strictly stationary and strong mixing with exponential decaying rates (see Fan and Yao 2003, p. 70). In all examples, our goal was to estimate the optimal orientation θ_0 and the single-index predictive density $f_{Y|\theta^T x}\left(y_t|\theta^T x_t\right)$ of y_t given the lagged observations

$x_t = (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. For each model 200 replications were generated with sample sizes $n = 100, 200, 400$ and 800, and we implemented the method to produce the corresponding estimates $\widehat{\theta}$ and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$.

Table 1 presents the average and standard error (over 200 replications) of the inner products $\left|\widehat{\theta}^T \theta_0\right|$ obtained for the three models with different sample sizes. Note that since $\widehat{\theta}$ and θ_0 are unit vectors, $\left|\widehat{\theta}^T \theta_0\right|$ is just $|\cos\alpha|$, where α is the angle between $\widehat{\theta}$ and θ_0. Therefore the closer $\left|\widehat{\theta}^T \theta_0\right|$ is to 1, the more accurate the estimate $\widehat{\theta}$.

As a general conclusion from Table 1, we can see that the orientation estimates become more accurate as the sample size increases, although the rate of improvement is not as fast as suggested by the theoretical asymptotic results. Two exceptions appear for the fourth order kernel in Examples 1 and 2, where the average accuracy of the estimates did not improve between $n = 200$ and $n = 400$. For both Examples 1 and 2, the fourth-order kernel yields consistently much inferior estimates with higher standard errors in comparison to the second-order kernel. We therefore attribute these two exceptions to sample fluctuations.

Comparing between the accuracy of the orientation estimates across the three different models, one can see that the method seems to be less accurate for the nonlinear models, and in particular for the nonlinear ARCH model with relatively small sample sizes ($n = 100$ or 200). However, when the number of observations is increased to 800, the average of the inner product $\left|\widehat{\theta}^T \theta_0\right|$ is consistently higher than

0.9 for all of the models with second-order kernels, and two out of the three models with fourth-order kernels.

The generally better performances of the second-order kernels compared with the fourth-order kernels in terms of the accuracy of the corresponding orientation estimates are particularly striking in Examples 1 and 2. In Example 3, on the other hand, the fourth-order kernel yields some more accurate estimates for θ_0 with sample sizes $n = 100,\ 200$ or 400. However, when the sample size is increased to $n = 800$ the accuracy of the second-order kernels 'catches up' with that of the fourth-order kernels. An extensive investigation performed by Marron and Wand (1992) of the effectiveness of high-order kernels in nonparametric density estimation provides an explanation for this discrepancy between theory and practice as it shows that it may take extremely large sample sizes (with a typical order of magnitude of few thousands and up to hundreds of thousands) for the asymptotic dominant effect to begin to be realised, and for the high-order kernels to produce more accurate estimates. In particular, Marron and Wand (1992) conclude that high-order kernels are not recommended in practice for kernels density estimation with realistic sample sizes.

In order to assess the accuracy of the conditional density estimator, $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$, we used the sample Root Mean Square Percentage Error (RMSPE),

$$RMSPE = \sum_{i=1}^{n}\left[\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y_i|\widehat{\theta}^T x_i\right) - f_{Y|X}\left(y_i|x_i\right)\right]^2 \Big/ \sum_{i=1}^{n} f_{Y|X}\left(y_i|x_i\right)^2,$$

where $f_{Y|X}\left(y_i|x_i\right)$ is the real conditional density of the model. The average and

TABLE 2: Mean and standard error (in brackets) of the sample RMSPE.

$k.order$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
Example 1				
$p = 2$	0.0460 (0.0201)	0.0327 (0.0117)	0.0245 (0.0097)	0.0167 (0.0055)
$p = 4$	0.0497 (0.0213)	0.0415 (0.0156)	0.0363 (0.0130)	0.0289 (0.0102)
Example 2				
$p = 2$	0.0756 (0.0333)	0.0511 (0.0205)	0.0370 (0.0167)	0.0272 (0.0086)
$p = 4$	0.1050 (0.0355)	0.0939 (0.0376)	0.0866 (0.0413)	0.0722 (0.0453)
Example 3				
$p = 2$	0.0712 (0.0271)	0.0500 (0.0172)	0.0374 (0.0148)	0.0276 (0.0100)
$p = 4$	0.0626 (0.0241)	0.0455 (0.0165)	0.0347 (0.0135)	0.0256 (0.0093)

standard error (over 200 replications) of the sample RMSPE are given in Table 2.

Here, we see that the estimation error given by the sample RMSPE consistently decreases as the sample size increases for all the simulation settings. Observe that although the average accuracy of the orientation estimates did not improve in Examples 1 and 2 between $n = 200$ and $n = 400$, the approximated conditional density obtained at the second stage is more accurate on average for the larger sample size $n = 400$. Finally, as a consequence of the orientation estimation performances, we see that in Examples 1 and 2 the conditional density estimates obtained by using second-order kernels (at the first-stage of the estimation) outperforms the ones

obtained with fourth-order kernels. In Example 3, however, the estimates corresponding to fourth-order kernels are slightly more accurate on average.

5 Appendix A - Proofs of the Theorems

Proof of Theorem 1. By (A6) it is sufficient to prove that

$$\sup_{\theta\in\Theta}\left|\left(\mathcal{L}(\theta)-E_{\mathbb{S}}\left(\log f_{Y|\theta^TX}\left(y|\theta^Tx\right)\right)\right)\right|=o_p(1). \tag{7}$$

By Lemma 8 we can ignore the trimming-terms, i.e. set $\widehat{\rho}_i^{\theta}\equiv 1$, in the sense of almost sure consistency. Since $f_{Y,\theta^TX}\left(y,\theta^Tx\right)$, $f_{\theta^TX}\left(\theta^Tx\right)$ are bounded from below by $\varepsilon>0$ on $\Theta\times\mathbb{S}$ and $\Theta\times\mathbb{S}_X$, by Lemma 5 and the continuous mapping theorem we get with $z_j=(y_j,x_j)$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\log\widehat{f}^{-i}_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right)-\frac{1}{n}\sum_{i=1}^{n}\log f_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right)\right|\\
\leq\ &\max_{1\leq i\leq n}\sup_{\theta\in\Theta}\left|\log\widehat{f}^{-i}_{Y,\theta^TX}\left(y_i,\theta^Tx_i\right)-\log f_{Y,\theta^TX}\left(y_i,\theta^Tx_i\right)\right|\\
&+\max_{1\leq i\leq n}\sup_{\theta\in\Theta}\left|\log\widehat{f}^{-i}_{\theta^TX}\left(\theta^Tx_i\right)-\log f_{\theta^TX}\left(\theta^Tx_i\right)\right|\\
\leq\ &\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\log\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)-\log f_{Y,\theta^TX}\left(y,\theta^Tx\right)\right|\\
&+\sup_{\theta\in\Theta,x\in\mathbb{S}_X}\left|\log\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-\log f_{\theta^TX}\left(\theta^Tx\right)\right|+o(1)\\
=\ &o_p(1).
\end{aligned} \tag{8}$$

Next, the series $\log f_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right)$ is itself strong mixing with $\alpha_t=O\left(t^{-\eta_1}\right)$ (see, for instance, White 1984). By the ergodic theory we get for any $\theta\in\Theta$

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\log f_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^TX}\left(y|\theta^Tx\right)\right)\right)\right|\to 0\quad a.s. \tag{9}$$

By smoothness condition (A4) we have that for any $\varepsilon > 0$ there exists a positive constant $\delta > 0$ such that for any $(\theta_1, y, x) \in \Theta \times \mathbb{S}$ and $\theta \in U_\delta(\theta_1)$, a δ-ball with centre at θ_1,

$$\left|\log f_{Y|\theta^T X}\left(y|\theta^T x\right) - \log f_{Y|\theta^T X}\left(y|\theta_1^T x\right)\right| < \varepsilon.$$

As a result, we have

$$\begin{aligned} & \sup_{\theta \in U_\delta(\theta_1)} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right)\right| \\ = \; & 2\varepsilon + \left| \left(\frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta_1^T x_i\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta_1^T x\right)\right)\right)\right|. \end{aligned} \tag{10}$$

Note also that since is Θ compact, it is possible to construct a finite open covering of Θ by δ-balls $U_\delta(\theta_k)$, $k = 1, ..., K$. Thus, using (10) we have that for any $\varepsilon > 0$

$$\begin{aligned} & P\left(\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right| > 4\varepsilon \right) \\ \leq \; & P\left(\sup_{\theta \in \Theta} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) - \frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)\right)\right| > \varepsilon \right) \\ & + K \max_{k=1,...,K} P\left(\sup_{\theta \in U_\delta(\theta_k)} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right)\right| > 3\varepsilon \right) \\ \leq \; & P\left(\sup_{\theta \in \Theta} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) - \frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)\right)\right| > \varepsilon \right) \\ & + K \max_{k=1,...,K} P\left(\left| \left(\frac{1}{n} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta_k^T x_i\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta_k^T x\right)\right)\right)\right| > \varepsilon \right) \end{aligned}$$

Results (8), (9) imply that the last expression approaches zero as $n \to \infty$, i.e., (7) is proved. ■

Proof of Theorem 2. Since θ_0 lies in the interior of Θ and $\widehat{\theta}$ converges in probability to θ_0, then $\nabla \mathcal{L}\left(\widehat{\theta}\right) = o_p(1)$. By an application of the mean value theorem it is sufficient to prove the following assertions.

(a) $\nabla\mathcal{L}(\theta_0) = n^{-1/2}\Psi(\theta_0)^{1/2} Z + O_p\left(n^{2-\delta}h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p)$ for some $\delta > 0$,

(b) $\overline{\theta} \to_p \theta_0$ implies $\nabla^2\mathcal{L}\left(\overline{\theta}\right) \to_p -\Omega(\theta_0)$.

Denote $\nabla L(\theta_0)$ and $\nabla^2 L(\theta)$ as the versions of $\nabla\mathcal{L}(\theta_0)$ and $\nabla^2\mathcal{L}(\theta)$ when conditional density estimates are replaced by the true conditional densities, that is,

$$\nabla^k L(\theta_0) = n^{-1}\sum_{i=1}^{n} \nabla^k \log f_{Y,\theta_0^T X}\left(y_i|\theta_0^T x_i\right),$$

for $k = 1, 2$. By the central limit theorem (CLT) for α-mixing processes (cf. Fan and Yao 2003, Theorem 2.21),

$$n^{1/2}\nabla L(\theta_0) \to_d N(0, \Psi(\theta_0)),$$

and with smoothness condition (A9), it follows from $\overline{\theta} \to_p \theta_0$ that

$$\nabla^2 L\left(\overline{\theta}\right) + \Omega(\theta_0) = o_p(1).$$

We therefore get that assertions (a) and (b) will be established if we show the following two assertions.

(a') $\nabla\mathcal{L}(\theta_0) - \nabla L(\theta_0) = O_p\left(n^{2-\delta}h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p)$ for some $\delta > 0$,

and

(b') $\sup_{\theta\in U_\delta(\theta_0)}\left|\nabla^2\mathcal{L}(\theta) - \nabla^2 L(\theta)\right| = o_p(1)$, where $U_\delta(\theta_0)$ is an arbitrarily small neighborhood of θ_0.

We can simplify the new assertions (a') and (b') somewhat further. Setting $\widehat{\rho}_i^\theta \equiv 1$ by Lemma 8, we have

$$\nabla\mathcal{L}(\theta_0) = n^{-1}\sum_{i=1}^{n}\left(\frac{\nabla\widehat{f}^{-i}_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)}{\widehat{f}^{-i}_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)} - \frac{\nabla\widehat{f}^{-i}_{\theta_0^T X}\left(\theta_0^T x_i\right)}{\widehat{f}^{-i}_{\theta_0^T X}\left(\theta_0^T x_i\right)}\right),$$

and

$$\nabla^2 \mathcal{L}(\theta) = n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla^2 \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)}{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)} - \frac{\nabla^2 \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)}{\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)} \right)$$
$$-n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right) \nabla \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)^T}{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)^2} - \frac{\nabla \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right) \nabla \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)^T}{\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)^2} \right).$$

Recalling that $n^{2-\delta} h_y h_x^5 = o(1)$, assertions (a') and (b') will follow if we prove the following six assertions. For some $\delta > 0$,

$$\text{(a')(i)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{\theta_0^T X}\left(\theta_0^T x_i\right)}{\widehat{f}^{-i}_{\theta_0^T X}\left(\theta_0^T x_i\right)} - \frac{\nabla f_{\theta_0^T X}\left(\theta_0^T x_i\right)}{f_{\theta_0^T X}\left(\theta_0^T x_i\right)} \right)$$
$$= O_p\left(n^{2-\delta} h_x^3\right)^{-1/2} + O(h_x^p),$$
$$\text{(a')(ii)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)}{\widehat{f}^{-i}_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)} - \frac{\nabla f_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)}{f_{Y,\theta_0^T X}\left(y_i, \theta_0^T x_i\right)} \right)$$
$$= O_p\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p).$$

In addition, uniformly on a small neighborhood of θ_0,

$$\text{(b')(i)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla^2 \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)}{\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)} - \frac{\nabla^2 f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)} \right)$$
$$= O_p\left(n^{2-\delta} h_x^5\right)^{-1/2} + O(h_x^p),$$
$$\text{(b')(ii)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla^2 \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)}{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)} - \frac{\nabla^2 f_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)}{f_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)} \right)$$
$$= O_p\left(n^{2-\delta} h_y h_x^5\right)^{-1/2} + O(h_y^p + h_x^p),$$
$$\text{(b')(iii)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right) \nabla \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)^T}{\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)^2} - \frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right) \nabla f_{\theta^T X}\left(\theta^T x_i\right)^T}{f_{\theta^T X}\left(\theta^T x_i\right)^2} \right)$$
$$= O_p\left(n^{2-\delta} h_x^3\right)^{-1/2} + O(h_x^p),$$
$$\text{(b')(iv)} \qquad n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right) \nabla \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)^T}{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right)^2} - \frac{\nabla f_{Y,\theta^T X}\left(y_i, \theta_0^T x_i\right) \nabla f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)^T}{f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)^2} \right)$$
$$= O_p\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p),$$

The proofs of (a')(i)-(ii) and (b')(i)-(iv) are long and tedious. However, they are all proved similarly with the theory for U-statistics given in Lemma 7. For the sake

of brevity, we shall focus here on proving (a')(i), while the rest of the assertions follow by the same line of arguments. The uniformity of arguments (b')(i)-(iv) is achieved from the regularity conditions on the kernel and the density functions. Note also that the proof of assertion (a')(i) involves handling some third-order U-statistic remainder terms that are similar to the terms in assertion (b')(iii)-(iv).

In the following, whenever confusion does not occur we denote $f_{\theta^T X} \equiv f_{\theta^T X}\left(\theta_0^T x_i\right)$ and $\widehat{f}^{-i}_{\theta^T X} \equiv \widehat{f}^{-i}_{Y,\theta^T X}\left(\theta_0^T x_i\right)$ for some $x_i \in \mathbb{S}_X$. We now have by the mean-value theorem with $\left|\overline{f}_{\theta^T x} - f_{\theta^T x}\right| < \left|\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T x}\right|$,

$$\frac{1}{\widehat{f}^{-i}_{\theta^T X}} - \frac{1}{f_{\theta_0^T X}} = -\frac{1}{\left(f_{\theta_0^T X}\right)^2}\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta_0^T X}\right) + \frac{2}{\left(\overline{f}_{\theta_0^T X}\right)^3}\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta_0^T X}\right)^2. \tag{11}$$

We then obtain

$$\begin{aligned}
\frac{\nabla \widehat{f}^{-i}_{\theta^T X}}{\widehat{f}^{-i}_{\theta^T X}} &= \frac{\nabla \widehat{f}^{-i}_{\theta^T X}}{f_{\theta_0^T X}} + \left[\frac{1}{\widehat{f}^{-i}_{\theta^T X}} - \frac{1}{f_{\theta_0^T X}}\right]\left[\nabla f_{Y,\theta_0^T X} + \left(\nabla \widehat{f}^{-i}_{\theta^T X} - \nabla f_{Y,\theta_0^T X}\right)\right] \\
&= \frac{\nabla f_{\theta_0^T X}}{f_{\theta_0^T X}} + \left(\frac{\nabla \widehat{f}^{-i}_{\theta^T X}}{f_{\theta_0^T X}} - \frac{\widehat{f}^{-i}_{\theta^T X} \nabla f_{Y,\theta_0^T X}}{\left(f_{\theta_0^T X}\right)^2}\right) - \frac{\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta_0^T X}\right)\left(\nabla \widehat{f}^{-i}_{\theta^T X} - \nabla f_{\theta_0^T X}\right)}{\left(f_{Y,\theta_0^T X}\right)^2} \\
&\quad + \frac{2\nabla f_{\theta_0^T X}\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta_0^T X}\right)^2}{\left(\overline{f}_{\theta_0^T X}\right)^3} + \frac{2\left(\nabla \widehat{f}^{-i}_{\theta^T X} - \nabla f_{\theta_0^T X}\right)\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta_0^T X}\right)^2}{\left(\overline{f}_{\theta_0^T X}\right)^3}.
\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla \widehat{f}^{-i}_{\theta^T X}}{\widehat{f}^{-i}_{\theta^T X}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\nabla f_{\theta_0^T X}}{f_{\theta_0^T X}} + U^{(A)}_{\theta_0} - U^{(B)}_{\theta_0} - R_1 + R_2 + R_3, \tag{12}$$

where

$$U^{(A)}_{\theta} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \zeta^{(A)}_{\theta}(x_i, x_j), \quad U^{(B)}_{\theta} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \zeta^{(B)}_{\theta}(x_i, x_j),$$

are second order $\mathbb{R}^d$-vector U-statistics with arguments

$$\begin{aligned}
\varsigma_\theta^{(A)}(x_i,x_j) &= \frac{1}{h_x^2}\frac{1}{f_{\theta^T X}\left(\theta^T x_i\right)}(x_j-x_i)K'\left(\frac{\theta^T(x_j-x_i)}{h_x}\right)-\frac{1}{h_x}\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f^2_{\theta^T X}\left(\theta^T x_i\right)}K\left(\frac{\theta^T(x_j-x_i)}{h_x}\right),\\
\varsigma_\theta^{(B)}(x_i,x_j) &= \frac{1}{h_x^3}\int\left(x_j-E\left(X|\theta^T X=t\right)\right)K\left(\frac{\theta^T x_i-t}{h_x}\right)K'\left(\frac{\theta^T x_j-t}{h_x}\right)f(t)^{-1}dt\\
&\quad -E\left(\nabla\log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X=\theta^T x_i\right)-E\left(\nabla\log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X=\theta^T x_j\right)\\
&\quad +E\left(\nabla\log f_{\theta^T X}\left(\theta^T X\right)\right).
\end{aligned}$$

Note that $U_\theta^{(B)}$ was added to (12) simply to make R_1 a degenerate U-statistic. Now, R_1, R_2, R_3 are the high-order remainder terms,

$$\begin{aligned}
R_1 &= \frac{1}{n}\sum_{i=1}^n\frac{\left(\widehat{f}^{-i}_{\theta^T X}-f_{\theta_0^T X}\right)\left(\nabla\widehat{f}^{-i}_{\theta^T X}-\nabla f_{\theta_0^T X}\right)}{\left(f_{Y,\theta_0^T X}\right)^2}-U_{\theta_0}^{(B)},\\
R_2 &= \frac{1}{n}\sum_{i=1}^n\frac{2\nabla f_{\theta_0^T X}\left(\widehat{f}^{-i}_{\theta^T X}-f_{\theta_0^T X}\right)^2}{\left(\overline{f}_{\theta_0^T X}\right)^3},\\
R_3 &= \frac{1}{n}\sum_{i=1}^n\frac{2\left(\nabla\widehat{f}^{-i}_{\theta^T X}-\nabla f_{\theta_0^T X}\right)\left(\widehat{f}^{-i}_{\theta^T X}-f_{\theta_0^T X}\right)^2}{\left(\overline{f}_{\theta_0^T X}\right)^3}.
\end{aligned}$$

We now handle the terms in the expansion (12) and we prove that for $\delta>0$ an arbitrarily small constant

$$U_{\theta_0}^{(A)}, U_{\theta_0}^{(B)} = O_p\left(n^{2-\delta}h_x^3\right)^{-1/2}+O\left(h_x^p\right), \tag{13}$$

and

$$R_1, R_2, R_3 = o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right)+O\left(h_x^p\right). \tag{14}$$

The asymptotic bounds (13) are derived with Lemma 7. Consider first $U_{\theta_0}^{(A)}$. Write

$U_{\theta_0}^{(A)}$ as

$$\begin{aligned} U_{\theta}^{(A)} &= \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \frac{1}{2}\left(\varsigma_{\theta}^{(A)}(x_i,x_j)+\varsigma_{\theta}^{(A)}(x_j,x_i)\right) \\ &\equiv \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \phi_{\theta}^{(A)}(x_i,x_j). \end{aligned}$$

We also show now that $U_{\theta}^{(A)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term. Let

$$\begin{aligned} Z_{\theta}^{(1)}(x_i,x_j) &= \frac{1}{h_x} f_{\theta^T X}\left(\theta^T x_i\right)^{-1} K\left(\frac{\theta^T(x_j-x_i)}{h_x}\right), \\ Z_{\theta}^{(2)}(x_i,x_j) &= \frac{1}{h_x^2} f_{\theta^T X}\left(\theta^T x_i\right)^{-1}(x_j-x_i) K'\left(\frac{\theta^T(x_j-x_i)}{h_x}\right), \end{aligned} \tag{15}$$

so that

$$\varsigma_{\theta}^{(A)}(x_i,x_j) = Z_{\theta}^{(2)}(x_i,x_j) - \frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}^2\left(\theta^T x_i\right)} Z_{\theta}^{(1)}(x_i,x_j).$$

For a fixed $x_i \in \mathbb{S}_X$ we obtain with a change of variables, integration by parts and Taylor expansion and Lemma 4 that $E\left(\frac{\nabla f_{\theta^T X}\left(\theta^T X\right)}{f_{\theta^T X}\left(\theta^T X\right)} Z^{(1)}(x_i,X)\right)$ and $E\left(Z^{(2)}(x_i,X)\right)$ are both equal to

$$\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)} + O(h_x^p), \tag{16}$$

while for a fixed $x_j \in \mathbb{S}_X$, $E\left(\frac{\nabla f_{\theta^T X}\left(\theta^T x_j\right)}{f_{\theta^T X}\left(\theta^T x_j\right)} Z_{\theta}^{(1)}(X,x_j)\right)$ and $E\left(Z_{\theta}^{(2)}(X,x_j)\right)$ are equal to

$$\begin{aligned} & -\frac{d}{dt}\bigg|_{t=\theta^T x} E\left(X|\theta^T X=t\right) + O(h_x^p) \qquad (17) \\ = \; & E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X=\theta^T x\right) + O(h_x^p), \end{aligned}$$

uniformly on $\Theta\times\mathbb{S}_X$. By denoting $\eta_{\theta}^{(A)}(\cdot) \equiv E\left(\phi_{\theta}^{(A)}(X,\cdot)\right)$ and $\mu_{\theta}^{(A)} = E\left(\eta_{\theta}^{(A)}(X)\right)$, it follows from (16)-(17) that $E\left(\phi_{\theta}^{(A)}(X,x)\right) = O(h_x^p)$ and $\mu_{\theta}^{(A)} = O(h_x^p)$. Hence,

$$U_{\theta}^{(A)} = U_{\theta}^{*(A)} + O(h_x^p), \tag{18}$$

where $U_\theta^{*(A)}$ is the degenerate U-statistic,

$$U_\theta^{*(A)} = \frac{2}{n\left(n-1\right)} \sum_{1 \leq i < j \leq n} \phi_\theta^{*(A)}\left(x_i, x_j\right),$$

with elements

$$\phi_\theta^{*(A)}\left(x_i, x_j\right) = \phi_\theta^{(A)} - \eta_\theta^{(A)}\left(x_i\right) - \eta_\theta^{(A)}\left(x_j\right) + \mu_\theta^{(A)}.$$

Applications of Chebyshev's inequality and Lemma 7 then yield

$$\begin{aligned} U_\theta^{*(A)} &= O_p\left(\frac{2}{n\left(n-1\right)}[E(\sum_{1 \leq i < j \leq n} \phi_\theta^{*(A)}\left(x_i, x_j\right))^2]^{1/2}\right) \\ &= O_p\left(n^{-1}\left(M_\theta^{(A)}\right)^{1/(2+\delta)}\right), \end{aligned} \tag{19}$$

where

$$M_\theta^{(A)} = \max_{1 \leq i < j \leq T} \max\left\{E\left|\phi_\theta^{*(A)}\left(x_i, x_j\right)\right|^{2+\delta}, \int\left|\phi_\theta^{*(A)}\left(x_i, x_j\right)\right|^{2+\delta} dP\left(x_i\right) dP\left(x_j\right)\right\}.$$

Here, $P\left(X\right)$ denotes the probability measure of r.v. X and $0 < \delta < 1$. Since $\phi_\theta^{*(A)}\left(x_i, x_j\right) = \phi_\theta^{(A)}\left(x_i, x_j\right) + O\left(h_x^p\right)$, we get with the C_r inequality,

$$\begin{aligned} M_\theta^{(A)} &= \max_{1 \leq i < j \leq T} \max\left\{E\left|\phi_\theta^{(A)}\left(x_i, x_j\right)\right|^{2+\delta}, \int\left|\phi_\theta^{(A)}\left(x_i, x_j\right)\right|^{2+\delta} dP\left(x_i\right) dP\left(x_j\right)\right\} \\ &\quad + O\left(h_x^{(2+\delta)p}\right), \end{aligned}$$

A standard calculation leads to

$$M_\theta^{(A)} = O\left(h_x^{-2(2+\delta)+1}\right) = O\left(h_x^{-(3+2\delta)}\right). \tag{20}$$

Hence, we conclude with results (18)-(20) that $U_{\theta_0}^{(A)} = O_p\left(n^{2-\delta} h_x^3\right)^{-1/2} + O\left(h_x^p\right)$.

We now turn to deal with $U_{\theta_0}^{(B)}$. Note that the first term in the definition of $\varsigma_\theta^{(B)}(x_i, x_j)$ is

$$\varsigma_\theta^{(B,1)}(x_i, x_j) \equiv \frac{1}{h_x^3}\int \left(x_j - E\left(X|\theta^T X = t\right)\right) K\left(\frac{\theta^T x_i - t}{h_x}\right) K'\left(\frac{\theta^T x_j - t}{h_x}\right) f(t)^{-1}\, dt. \tag{21}$$

Using Fubini's theorem and applying the standard argument, it is easy to see that both $E\left(\varsigma_\theta^{(B,1)}(X, x)\right)$ and $E\left(\varsigma_\theta^{(B,1)}(x, X)\right)$ are equal to

$$\begin{aligned} & -\left.\frac{d}{dt}\right|_{t=\theta^T x} E\left(X|\theta^T X = t\right) + O(h_x^p) \\ = \;& E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X = \theta^T x\right) + O(h_x^p), \end{aligned} \tag{22}$$

for fixed $x \in \mathbb{S}_X$. Thus, we have again that $U_\theta^{(B)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term. Applying Lemma 7 to the U-statistic $U_{\theta_0}^{(B)}$ in a similar way to $U_{\theta_0}^{(A)}$, it is possible to show now that $U_{\theta_0}^{(B)} = O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right)$. Thus, (13) is proved.

We continue to prove the asymptotic bounds in probability for the remainder terms R_1, R_2 and R_3. We start with

$$R_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta_0^T X}\right)\left(\nabla \widehat{f}_{\theta^T X}^{-i} - \nabla f_{\theta_0^T X}\right)}{\left(f_{Y,\theta_0^T X}\right)^2} - U_{\theta_0}^{(B)}.$$

Put

$$\rho_{1,\theta}(x_i, x_j, x_k) \equiv Z_\theta^{(1)}(x_i, x_j) Z_\theta^{(2)}(x_i, x_k) - Z_\theta^{(1)}(x_i, x_j)\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)} - Z_\theta^{(2)}(x_i, x_k) + \frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)},$$

where $Z^{(1)}(\cdot,\cdot)$ and $Z^{(2)}(\cdot,\cdot)$ are defined in (15). Note that R_1 can be redefined as

a sum of third and second order $\mathbb{R}^d$-vector U-statistics in the following way

$$\begin{aligned} R_1 &= \frac{1}{n(n-1)^2}\sum_i\sum_{j\neq i}\sum_{k\neq i}\rho_\theta^{(1)}(x_i,x_j,x_k) - U_{\theta_0}^{(B)} \\ &= \frac{n-2}{n-1}\cdot\frac{1}{n(n-1)(n-2)}\sum_{1\leq i\neq j\neq k\leq n}\left(\rho_{1,\theta}(x_i,x_j,x_k) - \varsigma_\theta^{(B)}(x_j,x_k)\right) \\ &\quad + \frac{1}{n-1}\cdot\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\rho_{1,\theta}(x_i,x_j,x_j) \\ &\equiv \frac{n-2}{n-1}\cdot U^{(1,A)} + \frac{1}{n-1}\cdot U^{(1,B)}. \end{aligned} \tag{23}$$

We have for fixed $x_j, x_k \in \mathbb{S}_X$,

$$E\left(Z_\theta^{(1)}(X,x_j)\,Z_\theta^{(2)}(X,x_k)\right) = \varsigma_\theta^{(B,1)}(x_j,x_k), \tag{24}$$

with $\varsigma_\theta^{(B,1)}(x_i,x_j)$ as in (21), and it is easy to verify with results (16), (17), (22) and (24) that $U^{(1,A)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term, in the sense that for any fixed $x_i, x_j, x_k \in \mathbb{S}_X$,

$$E\left(\rho_{1,\theta}(X,x_j,x_k) - \varsigma_\theta^{(B)}(x_j,x_k)\right) = O(h_x^p), \quad E\left(\rho_{1,\theta}(x_i,X,x_k) - \varsigma_\theta^{(B)}(X,x_k)\right) = O(h_x^p),$$

$$\text{and} \quad E\left(\rho_\theta^{(1)}(x_i,x_j,X) - \varsigma_\theta^{(B)}(x_j,X)\right) = O(h_x^p).$$

As an Applications of Lemma 7 we now obtain

$$U^{(1,A)} = O_p\left(n^{3-\delta}h_x^5\right)^{-1/2} + O\left(h_x^p\right). \tag{25}$$

For $U^{(1,B)}$, it is enough to note that by Lemma 5 and the continuous mapping theorem, we have $\rho(x_i,x_j,x_k) = O_p\left(\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right) + O\left(h_x^p\right)$ and hence

$$U^{(1,B)} = O_p\left(n^{3-\delta}h_x^3\right)^{-1/2} + O\left(n^{-1}h_x^p\right). \tag{26}$$

Thus, it is clear from (23), (25) and (26) that

$$R_1 = o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right) + O\left(h_x^p\right).$$

Next, we show a stochastic bound for

$$|R_2| = \left|2\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla f_{\theta_0^T X}\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta_0^T X}\right)^2}{\left(\overline{f}_{\theta_0^T X}\right)^3}\right| \leq 2\sup_{x\in\mathbb{S}_X}\left|\frac{\nabla f_{\theta_0^T X} f_{\theta_0^T X}^2}{\left(\overline{f}_{\theta_0^T X}\right)^3}\right| \cdot \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}_{\theta^T X}^{-i}}{f_{\theta_0^T X}} - 1\right)^2. \tag{27}$$

Note as that as the first term in the RHS is bounded, it is enough to bound the second term in probability. Let now

$$\rho_{2,\theta}\left(x_i, x_j, x_k\right) = Z^{(1)}\left(x_i, x_j\right) Z^{(1)}\left(x_i, x_k\right) - Z^{(1)}\left(x_i, x_j\right) - Z^{(1)}\left(x_i, x_k\right) + 1,$$

where $Z^{(1)}(\cdot,\cdot)$ is defined in (15), and

$$\varsigma_\theta^{(C)}\left(x_i, x_j\right) = \frac{1}{h_x^2}\int K\left(\frac{\theta^T x_j - t}{h_x}\right) K\left(\frac{\theta^T x_i - t}{h_x}\right) f\left(t\right)^{-1} dt - 1.$$

We have for large enough n,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}_{\theta^T X}^{-i}}{f_{\theta_0^T X}} - 1\right)^2 \\
= \; & \frac{1}{n\left(n-1\right)^2}\sum_{i}\sum_{j\neq i}\sum_{k\neq i}\rho_{2,\theta}\left(x_i, x_j, x_k\right) \\
= \; & \frac{n-2}{n-1}\cdot\frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{1\leq i\neq j\neq k\leq n}\left(\rho_{2,\theta}\left(x_i, x_j, x_k\right) - \varsigma_\theta^{(C)}\left(x_j, x_k\right)\right) \\
& + \frac{1}{n-1}\cdot\frac{1}{n\left(n-1\right)}\sum_{1\leq i\neq j\leq n}\rho_{2,\theta}\left(x_i, x_j, x_j\right) + \frac{1}{n\left(n-1\right)}\sum_{1\leq i\neq j\leq n}\varsigma_\theta^{(C)}\left(x_i, x_j\right) \\
\equiv \; & \frac{n-2}{n-1}\cdot U^{(2,A)} + \frac{1}{n-1}\cdot U^{(2,B)} + U^{(2,C)}.
\end{aligned}$$

Here, $U^{(2,A)}$ is a third order $\mathbb{R}^d$-vector U-statistic, and $U^{(2,B)}$ and $U^{(2,C)}$ are second order $\mathbb{R}^d$-vector U-statistics. Following a similar treatment as above, these three

U-statistics are shown to be degenerate up to a $O(h_x^p)$ term, and by Lemma 7 we have

$$U^{(2,A)} = O_p\left(n^{3-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right), \quad U^{(2,B)} = O_p\left(n^{3-\delta}h_x\right)^{-1/2} + O\left(h_x^p\right),$$

$$\text{and} \quad U^{(2,C)} = O_p\left(n^{2-\delta}h_x\right)^{-1/2} + O\left(h_x^p\right).$$

The last arguments imply that

$$R_2 = o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right) + O\left(h_x^p\right).$$

Finally, bounding R_3 is trivial with the continuous mapping theorem. We have therefore established (14). Retracing through results (12)-(14), we have completed the proof of assertion (a')(i). ■

Proof of Theorem 3. By the mean-value theorem with mean value $\overline{\theta}$ and Theorems 2 and 5,

$$\begin{aligned} & \sup_{(y,x)\in\mathbb{S}} \left| \widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) - \widetilde{f}_{Y|\theta_0^T X}\left(y|\theta_0^T x\right) \right| \\ \leq & \left\|\widehat{\theta} - \theta_0\right\| \left\| \sup_{(y,x)\in\mathbb{S}} \nabla f_{Y|\overline{\theta}^T X}\left(y|\overline{\theta}^T x\right) + o_p(1) \right\| = o_p\left(\left(\frac{\ln n}{nH_yH_x}\right)^{1/2}\right). \quad \blacksquare \end{aligned}$$

6 Appendix B - Technical Lemmas

This section gives some useful technical results that are needed in the proofs of the main theorems.

Recall that for a function $g(\theta)$ that depends on $\theta \in \Theta$ and possibly also on other variables we denote $\nabla g(\theta)$ and $\nabla^2 g(\theta)$ as the vector and matrix of partial

derivatives of $g(\theta)$ with respect to θ. As a convention, we also use $\nabla^0 g(\theta) = g(\theta)$.

The following Lemma gives the forms of the partial derivatives of $f_{Y,\theta^T X}\left(y, \theta^T x\right)$ and $f_{\theta^T X}\left(\theta^T x\right)$ with respect to θ. One has to remember that θ affects the value of the probability densities $f_{Y,\theta^T X}\left(y, \theta^T x\right)$ and $f_{\theta^T X}\left(\theta^T x\right)$ not only through the variable $\theta^T x$, but it also defines the density functions $f_{Y,\theta^T X}(\cdot,\cdot)$ and $f_{\theta^T X}(\cdot)$ themselves.

Lemma 4 *Let* $E\left(X|Y=y, \theta^T X=t\right), E\left(XX^T|Y=y, \theta^T X=t\right), E\left(X|\theta^T X=t\right),$ $E\left(XX^T|\theta^T X=t\right)$ *and* $f_{Y,\theta^T X}(y,t)$ *and* $f_{\theta^T X}(t)$ *exist and they are twice differentiable with respect to* $y, t \in \mathbb{R}$. *Then*

$$\begin{aligned}
\nabla f_{Y,\theta^T X}\left(y, \theta^T x\right) &= \left.\frac{d}{dt}\right|_{t=\theta^T x} \left\{E\left(x - X|Y=y, \theta^T X=t\right) f_{Y,\theta^T X}(y,t)\right\}, \\
\nabla^2 f_{Y,\theta^T X}\left(y, \theta^T x\right) &= \left.\frac{d^2}{dt^2}\right|_{t=\theta^T x} \left\{E\left((x-X)(x-X)^T|Y=y, \theta^T X=t\right) f_{Y,\theta^T X}(y,t)\right\},
\end{aligned}$$

and similarly,

$$\begin{aligned}
\nabla f_{\theta^T X}\left(\theta^T x\right) &= \left.\frac{d}{dt}\right|_{t=\theta^T x} \left\{E\left(x - X|\theta^T X=t\right) f_{\theta^T X}(t)\right\}, \\
\nabla^2 f_{\theta^T X}\left(\theta^T x\right) &= \left.\frac{d^2}{dt^2}\right|_{t=\theta^T x} \left\{E\left((x-X)(x-X)^T|\theta^T X=t\right) f_{\theta^T X}(t)\right\}.
\end{aligned}$$

Proof. We prove here only the last two identities of the Lemma as the first two follow similarly. Assume $\theta_d \neq 0$ since otherwise we may reduce the dimension to $d-1$. Let $f_X(\xi_1, ..., \xi_d)$ be the probability density of X at $(\xi_1, ..., \xi_d)$. We now have

$$f_{\theta^T X}(t) = \theta_d^{-1} \int f_X\left(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)\right) d\xi_1 ... d\xi_{d-1},$$

where θ_d^{-1} is the determinant of the Jacobian matrix. Thus, for $t = \theta^T x$,

$$f_{\theta^T X}\left(\theta^T x\right) = \theta_d^{-1} \int f_X(\xi_1, ..., \xi_{d-1}, x_d + \theta_d^{-1} \sum_{j=1}^{d-1} \theta_j \left(x_j - \xi_j\right)) d\xi_1 ... d\xi_{d-1}.$$

Note also that for $k, l \in \{1, 2, ..., d-1\}$ we have

$$\begin{aligned}
E\left(X_k|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-1} \int \xi_k f_X(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1 ... d\xi_{d-1}, \\
E\left(X_d|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-2} \int (t - \sum_{j=1}^{d-1} \theta_j \xi_j) f_X(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1 ... d\xi_{d-1}, \\
E\left(X_k X_l|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-1} \int \xi_k \xi_l f_X(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1 ... d\xi_{d-1}, \\
E\left(X_k X_d|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-2} \int \xi_k (t - \sum_{j=1}^{d-1} \theta_j \xi_j) f_X(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1 ... d\xi_{d-1}, \\
E\left(X_d^2|\theta^T X = t\right) f_\theta(t) &= \theta_d^{-3} \int (t - \sum_{j=1}^{d-1} \theta_j \xi_j)^2 f_X(\xi_1, ..., \xi_{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1 ... d\xi_{d-1}.
\end{aligned}$$

Using the above expressions one can use direct differentiation to verify the last two identities of the Lemma. ■

The proofs of Theorems 1 and 2 rely heavily on the uniform consistency of the kernel density estimators' derivatives with respect to θ. The next two Lemmas are direct modifications of the results of Hansen (2008), but unlike Hansen's (2008) theory, they concern with partial derivatives of the kernel estimates with respect to θ, rather than with derivatives with respect to the density variables themselves.

Lemma 5 *Let (A1)-(A4) hold. Then*

$$\begin{aligned}
\sup_{\theta \in \Theta, z \in \mathbb{S}} \left| \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - f_{Y,\theta^T X}\left(y, \theta^T x\right) \right| &= O_p\left(\left(\frac{\ln n}{n h_y h_x} \right)^{1/2} + h_y^2 + h_x^2 \right), \\
\sup_{\theta \in \Theta, x \in \mathbb{S}_X} \left| \widehat{f}_{\theta^T X}\left(\theta^T x\right) - f_{\theta^T X}\left(\theta^T x\right) \right| &= O_p\left(\left(\frac{\ln n}{n h_x} \right)^{1/2} + h_x^2 \right)
\end{aligned}$$

If, in addition, also (A7) and (A9) hold. Then for $k = 0, 1, 2$,

$$\begin{aligned}
\sup_{\theta\in\Theta, z\in\mathbb{S}} \left|\nabla^k \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - \nabla^k f_{Y,\theta^T X}\left(y, \theta^T x\right)\right| &= O_p\left(\left(\frac{\ln n}{n h_y h_x^{1+2k}}\right)^{1/2} + h_y^p + h_x^p\right), \\
\sup_{\theta\in\Theta, x\in\mathbb{S}_X} \left|\nabla^k \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \nabla^k f_{\theta^T X}\left(\theta^T x\right)\right| &= O_p\left(\left(\frac{\ln n}{n h_x^{1+2k}}\right)^{1/2} + h_x^p\right).
\end{aligned}$$

Proof of Lemma 5. We prove here only that

$$\sup_{\theta\in\Theta, x\in\mathbb{S}_X} \left|\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \nabla f_{\theta^T X}\left(\theta^T x\right)\right| = O_p\left(\left(\frac{\ln n}{n h_x^3}\right)^{1/2} + h_x^p\right).$$

The proofs for the rest of the arguments are very similar. By Lemma 6, it is sufficient to prove that $\sup_{\Theta\times\mathbb{S}_X} \left|E\left(\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right) - \nabla f_{\theta^T X}\left(\theta^T x\right)\right| = O(h_x^p)$. A change of variables, integration by parts, and a Taylor expansion around $h_x = 0$ yield with (A7) and (A9) that uniformly in $x \in \mathbb{S}_X$,

$$\begin{aligned}
& E\left(\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right) \\
= \; & \frac{1}{h_x^2} \int \left(x - E\left(X|\theta^T X = t\right)\right) K'\left(\frac{\theta^T x - t}{h_x}\right) f_{\theta^T X}(t)\, dt \\
= \; & \frac{1}{h_x} \int \left(x - E\left(X|\theta^T X = \theta^T x - h_x u\right)\right) K'(u)\, f_{\theta^T X}\left(\theta^T x - h_x u\right) du \\
= \; & \int \left.\frac{d}{dt}\right|_{t=\theta^T x - h_x u} \left[\left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t)\right] K(u)\, du \\
= \; & \int \left\{\sum_{j=1}^{p-1} \left[\left.\frac{d^{1+j}}{dt^{1+j}}\right|_{t=\theta^T x} \left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t) (-h_x u)^j\right] + O\left(h_x^p\right)\right\} K(u)\, du \\
= \; & \left.\frac{d}{dt}\right|_{t=\theta^T x} \left[\left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t)\right] + O(h_x^p).
\end{aligned}$$

By Lemma 4, the last expression is just $\nabla f_{\theta^T X}\left(\theta^T x\right) + O(h^p)$. ■

Lemma 6 *Let (A1)-(A4) hold. Then*

$$\begin{aligned}\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)-E\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_yh_x}\right)^{1/2}\right),\\ \sup_{\theta\in\Theta,x\in\mathbb{S}_X}\left|\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_x}\right)^{1/2}\right),\end{aligned}$$

If, in addition, also (A7) holds. Then for $k=0,1,2$,

$$\begin{aligned}\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\nabla^k\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)-E\nabla^k\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_yh_x^{1+2k}}\right)^{1/2}\right),\\ \sup_{\theta\in\Theta,x\in\mathbb{S}_X}\left|\nabla^k\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla^k\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_x^{1+2k}}\right)^{1/2}\right).\end{aligned}$$

Proof. We prove here only that

$$\sup_{\Theta\times\mathbb{S}_X}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|=O_p\left(\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right).$$

The proofs for the rest of the arguments in the Theorem are very similar. Let $\overline{\theta}_1\in\Theta$, $\overline{x}_1\in\mathbb{S}_X$ and define

$$A_1=\left\{\theta,x:\left\|\theta-\overline{\theta}_1\right\|\leq\left(\frac{h_x\ln n}{n}\right)^{1/2},\|x-\overline{x}_1\|\leq\left(\frac{h_x\ln n}{n}\right)^{1/2}\right\}. \tag{28}$$

Since $\Theta\times\mathbb{S}_X$ is compact, then it can be covered by $J(n)=O\left(\frac{n}{h_x\ln n}\right)$ such subspaces $A_1,...,A_J$ around centres $\left\{\left(\overline{\theta}_k,\overline{x}_k\right)\right\}_{j=1}^J$. Since

$$\begin{aligned}&P\left(\sup_{\Theta\times\mathbb{S}_X}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)\\ \leq\ & J(n)\max_{j=1,...,J}P\left(\sup_{(\theta,x)\in A_j}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right),\end{aligned}$$

it is therefore suffice to prove that for any $\left(\overline{\theta}_1,\overline{x}_1\right)\in\Theta\times\mathbb{S}_X$ and A_1 as in (28), the following holds

$$P\left(\sup_{(\theta,x)\in A_1}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)=o\left(\frac{h_x\ln n}{n}\right), \tag{29}$$

where the constant in the $o(\cdot)$ term is independent of $\left(\overline{\theta}_1, \overline{x}_1\right)$ and n.

Define the functions $\widetilde{K}_j,\ j = 1, 2, 3$, on $T = \left\{t \in \mathbb{R} : t = \frac{\theta^T x}{h_x} \text{ for some } \theta \in \Theta \text{ and } \overline{x}_1 - x \in \mathbb{S}_X\right\}$ by

$$\widetilde{K}_1(t) = \sup_{H(t)} \left\{ \|x\| \left| K'' \left(\frac{\theta^T x}{h_x} \right) x \right| \right\}, \qquad \widetilde{K}_2(t) = \sup_{H(t)} \left\{ \|\theta\| \left| K'' \left(\frac{\theta^T x}{h_x} \right) x \right| \right\},$$

and

$$\widetilde{K}_3(t) = \sup_{H(t)} \left| K' \left(\frac{\theta^T x}{h_x} \right) \right|.$$

where all the sups are taken over $\theta \in \Theta$ and $x \in \mathbb{S}_X$ such that $\frac{\theta^T X}{h_x}$ is not too far from t in the sense that

$$H(t) \equiv \left\{ (\theta, x) : \|\theta - \theta_*\| \leq \left(\frac{h_x \ln n}{n} \right)^{1/2}, \|x - x_*\| \leq \left(\frac{h_x \ln n}{n} \right)^{1/2} \text{ and } \frac{\theta_*^T x_*}{h_x} = t \right\}$$

Note that $\widetilde{K}_j,\ j = 1, 2, 3$, are well-defined and finite for any $t \in T$ by assumption (A7) and compactness of Θ and $\mathbb{S}_X$. Let x_i denote the i'th X-observation, and for any $(\theta, x) \in A_1$ we have with mean-values θ_*, x_* such that $\left\|\overline{\theta}_1 - \theta_*\right\| \leq \left\|\overline{\theta}_1 - \theta\right\| \leq$

$\left(\frac{h_x \ln n}{n}\right)^{1/2}$ and $\|\overline{x}_1 - x_*\| \leq \|\overline{x}_1 - x\| \leq \left(\frac{h_x \ln n}{n}\right)^{1/2}$ that

$$
\begin{aligned}
& \left| \nabla K\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) - \nabla K\left(\frac{\theta^T(x - x_i)}{h_x}\right) \right| \\
\leq\ & \frac{1}{h_x}\left| \left[K'\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) - K'\left(\frac{\theta^T(x - x_i)}{h_x}\right) \right](\overline{x}_1 - x_i) \right| \\
& + \frac{1}{h_x}\left| K'\left(\frac{\theta^T(x - x_i)}{h_x}\right)(\overline{x}_1 - x) \right| \\
\leq\ & \frac{1}{h_x^2}\left| \left(\overline{\theta}_1 - \theta\right)^T(\overline{x}_1 - x_i) K''\left(\frac{\theta_*^T(x_* - x_i)}{h_x}\right)(\overline{x}_1 - x_i) \right| \\
& + \frac{1}{h_x^2}\left| \theta^T(\overline{x}_1 - x) K''\left(\frac{\theta_*^T(x_* - x_i)}{h_x}\right)(\overline{x}_1 - x_i) \right| + \frac{1}{h_x}\left| K'\left(\frac{\theta^T(x - x_i)}{h_x}\right)(\overline{x}_1 - x) \right| \\
\leq\ & \frac{\left\|\overline{\theta}_1 - \theta\right\|}{h_x^2}\left| \widetilde{K}_1\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) \right| + \frac{\|\overline{x}_1 - x\|}{h_x^2}\left| \widetilde{K}_2\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) \right| \\
& + \frac{\|\overline{x}_1 - x\|}{h_x}\left| \widetilde{K}_3\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) \right| \\
\leq\ & \left(\frac{\ln n}{n h_x^3}\right)^{1/2} \cdot \left(\sum_{j=1}^{3} \left| \widetilde{K}_j\left(\frac{\overline{\theta}_1^T(\overline{x}_1 - x_i)}{h_x}\right) \right| \right)
\end{aligned}
\tag{30}
$$

Note that the last term is independent of $(\theta, x) \in A_1$. We now define for any $(\theta, x) \in A_1$ and $j = 1, 2, 3$,

$$
\widetilde{f}_{\theta^T X, j}\left(\theta^T x\right) = \frac{1}{n h_x} \sum\nolimits_{i=1}^{n} \widetilde{K}_j\left(\frac{\theta^T(x - x_i)}{h_x}\right).
$$

We have

$$
E\left| \widetilde{f}_{\theta^T X, j}\left(\theta^T x\right) \right| \leq \sup_{(\theta, x) \in \Theta \times \mathbb{S}_X} \left| f_{\theta^T x}\left(\theta^T x\right) \right| \int \left| \widetilde{K}_j(u) \right| du < \infty, \tag{31}
$$

Also, inequality (30) implies

$$
\sup_{(\theta, x) \in A_1} \left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) \right| \leq \left(\frac{\ln n}{n h_x^3}\right)^{1/2} \cdot \left(\sum_{j=1}^{3} \left| \widetilde{f}_{\theta j}\left(\overline{\theta}_1^T \overline{x}_1\right) \right| \right). \tag{32}
$$

Thus, the last three inequalities yield for any $(\theta, x) \in A_1$, for some large enough M, independent on θ_1, x_1 and n,

$$\sup_{(\theta,x)\in A_1} \left| E\left\{ \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right\}\right| \leq M\left(\frac{\ln n}{n h_x^3}\right)^{1/2}. \tag{33}$$

Next, results (31), (32), (33) and the condition that $\frac{\ln n}{n h_x} = o(1)$ give

$$\begin{aligned}
& \sup_{(\theta,x)\in A_1} \left| \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) - E\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right| \\
\leq & \sup_{(\theta,x)\in A_1} \left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right| + \left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - E\nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right)\right| \\
& + \sup_{(\theta,x)\in A_1} \left| E\left\{ \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right\}\right| \\
\leq & \left(\frac{\ln n}{n h_x^3}\right)^{1/2} \sum_{j=1}^{3} \left\{ \left| \widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| + E\left| \widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right)\right|\right\} \\
& + \left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| + M\left(\frac{\ln n}{n h_x^3}\right)^{1/2} \\
\leq & \frac{1}{h_x} \sum_{j=1}^{3} \left| \widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| + \left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| \\
& + 2M\left(\frac{\ln n}{n h_x^3}\right)^{1/2}.
\end{aligned}$$

As a result we get

$$\begin{aligned}
& P\left(\sup_{(\theta,x)\in A_k} \left| \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) - E\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right| > 5M\left(\frac{\ln n}{n h_x^3}\right)^{1/2}\right) \qquad (34) \\
\leq & P\left(\left| \nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\nabla \widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| > M\left(\frac{\ln n}{n h_x^3}\right)^{1/2}\right) + \\
& + \sum_{j=1}^{3} P\left(\left| \widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right) - E\widetilde{f}_{\theta^T X,j}\left(\overline{\theta}_1^T \overline{x}_1\right)\right| > M\left(\frac{\ln n}{n h_x^3}\right)^{1/2}\right).
\end{aligned}$$

We now bound the four terms in the RHS of (34) using the same argument, as all kernels used in the construction of $\widetilde{f}_{\theta^T X,j}$ and $\widehat{f}_{\theta^T X}$ all bounded and compactly supported. We therefore prove the bound only for the term $\left| \nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) - E\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right|$.

Set $m = \left(\frac{nh_x}{\ln n}\right)^{1/2}$, and note that for n sufficiently large, $m < \max\left(n, \frac{\varepsilon}{4b}\right)$ where $b = 2\sup_{\Theta\times\mathbb{S}_X}\left|\|x\|\frac{\partial}{\partial u}K(u)\right| < \infty$, and $\varepsilon = M(nh_x \ln n)^{1/2}$. Define for $(\theta, x) \in A_1$,

$$Z_i = (x - x_i)\left\{\left.\frac{\partial}{\partial t}\right|_{t=\frac{\theta^T(x-x_i)}{h_x}} K(t) - E\left(\left.\frac{\partial}{\partial t}\right|_{t=\frac{\theta^T(x-x_i)}{h_x}} K(t)\right)\right\}, \quad i = 1, ..., m.$$

Now, notice that $|Z_i| \leq b$, and by Theorem 1 of Hansen (2008),

$$\sigma^2(m) \equiv \sup_{(\theta,x)\in A_1} E\left|\sum_{i=1}^{\lfloor m\rfloor} Z_i\right|^2 \leq Cmh_x$$

for some large enough $C > 0$. By Theorem 2.1 of Liebscher (1996) we obtain

$$\begin{aligned}
& P\left(\left|\nabla\widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T\overline{x}_1\right) - E\nabla\widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T\overline{x}_1\right)\right| > M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right) \\
= \; & P\left(\left|\sum_{i=1}^n Z_i\right| > \varepsilon\right) \\
\leq \; & 4\exp\left(-\frac{\varepsilon^2}{64\frac{n}{m}\sigma^2(m) + \frac{8}{3}\varepsilon mb}\right) + 4\frac{n}{m}\alpha_m \\
\leq \; & 4\exp\left(-\frac{M^2(nh_x\ln n)}{64Cnh_x + 3Mnh_xb}\right) + 4An\left(\frac{nh_x}{\ln n}\right)^{-(\eta_2+1)/2} \\
\leq \; & 4\exp\left(-\frac{M^2\ln n}{64C + 3Mb}\right) + 4An\left(\frac{\ln n}{nh_x}\right)^{(\eta_2+1)/2} \\
\leq \; & 4n^{-M/(64+3b)} + 4An\left(\frac{\ln n}{nh_x}\right)^{(\eta_2+1)/2}, \qquad (35)
\end{aligned}$$

where the last inequality is justified by taking $M \geq C$. Now, we have for the first term of (35), $n^{-M/(64+3b)} = o\left(\frac{h_x\ln n}{n}\right)$ for sufficiently large M. Also, recall that $\frac{\ln n}{n^{\phi_2}h_x^2} = o(1)$ with $\phi_2 = (\eta_2 - 2)/(\eta_2 + 2) \in (0, 1)$. Thus, using $\frac{\ln n}{nh_x} = o(h_x n^{\phi_2 - 1}) = o(h_x n^{-4/(\eta_2+2)})$ we get for the the second term of (35), $4An\left(\frac{\ln n}{nh_x}\right)^{(\eta_2+1)/2} = o\left(h_x n^{-1}\right)$. Hence (29) is established. This completes the proof of Lemma 6. ■

The next Lemma is Lemma C.2 of Gao and King (2004) that gives a bound

for the stochastic order of second- and third-order degenerate U-statistics of strong mixing stochastic process.

Lemma 7 (Gao and King, 2004) *(i) Let $\psi(\cdot,\cdot,\cdot)$ be a symmetric Borel function defined on $\mathbb{R}^r \times \mathbb{R}^r \times \mathbb{R}^r$, and let the process ξ_i be an r-dimensional strictly stationary and strong mixing stochastic process. Assume that for any fixed $x, y \in \mathbb{R}^r$, $E[\psi(\xi_1, x, y)] = 0$. Then*

$$E\left\{\sum_{1\leq i<j<k\leq T}\psi\left(\xi_i,\xi_j,\xi_k\right)\right\}^2 \leq CT^3M^{1/(1+\delta)},$$

where $0<\delta<1$ is a small constant, $C>0$ is a constant independent of T and the function ψ, $M=\max\{M_1,M_2,M_3\}$, and

$$\begin{aligned}
M_1 &= \max_{1\leq i<j\leq T}\max\left\{E\left|\psi\left(\xi_1,\xi_i,\xi_j\right)\right|^{2+\delta}, \int\left|\psi\left(\xi_1,\xi_i,\xi_j\right)\right|^{2+\delta} dP(\xi_1)\, dP\left(\xi_i,\xi_j\right)\right\},\\
M_2 &= \max_{1\leq i<j\leq T}\max\left\{\int\left|\psi\left(\xi_1,\xi_j,\xi_k\right)\right|^{2+\delta} dP(\xi_i)\, dP\left(\xi_1,\xi_j\right)\right\},\\
M_3 &= \max_{1\leq i<j\leq T}\max\left\{\int\left|\psi\left(\xi_1,\xi_j,\xi_k\right)\right|^{2+\delta} dP(\xi_1)\, dP(\xi_i)\, dP\left(\xi_j\right)\right\}.
\end{aligned}$$

(ii) Let $\phi(\cdot,\cdot)$ be a symmetric Borel function defined on $\mathbb{R}^r \times \mathbb{R}^r$, and let the process ξ_i be defined as in part (i). Assume that for any fixed $x \in \mathbb{R}^r$, $E[\phi(\xi_1, x)] = 0$. Then

$$E\left\{\sum_{1\leq i<j<k\leq T}\phi\left(\xi_i,\xi_j\right)\right\}^2 \leq CT^2M_4^{1/(1+\delta)},$$

where $0<\delta<1$ is a small constant, $C>0$ is a constant independent of T and the function ϕ, and

$$M_4 = \max_{1\leq i\leq T}\max\left\{E\left|\phi(\xi_1,\xi_i)\right|^{2+\delta}, \int\left|\phi(\xi_1,\xi_i)\right|^{2+\delta} dP(\xi_1)\, dP(\xi_i)\right\}.$$

We conclude the appendix by proving that the trimming term $\widehat{\rho}_i^{\theta}$, defined in (4), is eventually equals to 1 for any sufficiently large n with probability 1.

Lemma 8 *Let (A1)-(A4) hold and*

$$I_{n,\theta}^{i} = \begin{cases} \mathbf{1}, & \text{if } \min\left\{\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right), \widehat{f}_{\theta^T X}^{-i}\left(\theta^T x_i\right)\right\} > a_0 n^{-c}, \\ 0, & \text{otherwise,} \end{cases}$$

for some small constants $a_0, c > 0$ such that $n^c\left(h_y^2 + h_x^2\right) = o(1)$ and $n^{1-2c} h_y h_x \to \infty$. Then eventually for any sufficiently large n

$$\max_{1 \le i \le n} \sup_{\theta \in \Theta} \left| I_{n,\theta}^{i} - 1 \right| = 0$$

with probability 1.

Proof. Define

$$\mathbb{T}_\theta = \left\{(y,x) \in \mathbb{R}^{1+d} : \min\left\{f_{Y,\theta^T X}\left(y, \theta^T x\right), f_{\theta^T X}\left(\theta^T x\right)\right\} > 2a_0 n^{-c}\right\}.$$

It is trivial now to show that

$$\sup_{\theta \in \Theta} \left| I_{n,\theta}^{i} - 1 \right| \le \sup_{\theta \in \Theta} I_{\{(y_i,x_i) \notin \mathbb{T}_\theta\}} + I_{\{Z_n^i > a_0 n^{-c}\}},$$

where

$$Z_n^i = \sup_{\theta \in \Theta} \max\left\{\left|\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right) - f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)\right|, \left|\widehat{f}_{\theta^T X}^{-i}\left(\theta^T x_i\right) - f_{\theta^T X}\left(\theta^T x_i\right)\right|\right\}.$$

By definition of $\mathbb{S}$ there exists some large N such that for any $n \ge N$, we have that $\mathbb{S} \subseteq \bigcap_{\theta \in \Theta} \mathbb{T}_\theta$, and as $(y_i, x_i) \in \mathbb{S}$, we get $\sup_{\theta \in \Theta} I_{\{(y_i,x_i) \notin \mathbb{T}_\theta\}} = 0$ for any $1 \le i \le n$. We now show that

$$P\left(\limsup_{n \to \infty} \left\{\bigcup_{i=1}^{n} \left\{Z_n > a_0 n^{-c}\right\}\right\}\right) = 0. \tag{36}$$

For sake of brevity, we prove here only that

$$\sum_{n=1}^{\infty} P\left(\bigcup_{i=1}^{n}\left\{\sup_{\theta\in\Theta}\left|\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i,\theta^T x_i\right)-f_{Y,\theta^T X}\left(y_i,\theta^T x_i\right)\right|>a_0 n^{-c}\right\}\right)<\infty, \tag{37}$$

from which (36) follows by the Borel-Cantelli lemma. The second term of Z_n^i can be handled in the same way.

For some $C_1, C_2 > 0$ independent of n, we have

$$\sup_{\theta\in\Theta}\left|\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i,\theta^T x_i\right)-\widehat{f}_{Y,\theta^T X}\left(y_i,\theta^T x_i\right)\right|\leq\frac{C_1}{nh_y h_x},$$

and from the proof of Lemma 5,

$$\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|E\widehat{f}_{Y,\theta^T X}\left(y_i,\theta^T x_i\right)-f_{Y,\theta^T X}\left(y,\theta^T x\right)\right|\leq C_2\left(h_y^2+h_x^2\right).$$

where $z=(y,x)$. The last two results imply that for n large enough,

$$\begin{aligned}&\sum_{n=1}^{\infty} P\left(\bigcup_{i=1}^{n}\left\{\sup_{\theta\in\Theta}\left|\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i,\theta^T x_i\right)-f_{Y,\theta^T X}\left(y_i,\theta^T x_i\right)\right|>a_0 n^{-c}\right\}\right)\\ \leq\ &\sum_{n=1}^{\infty} P\left(\sup_{z\in\mathbb{S}}\sup_{\theta\in\Theta}\left|\widehat{f}_{Y,\theta^T X}\left(y,\theta^T x\right)-E\widehat{f}_{Y,\theta^T X}\left(y,\theta^T x\right)\right|>an^{-c}\right),\end{aligned} \tag{38}$$

for some $0<a<a_0$. We can continue to bound the last term as in the proof of Lemma 6. Let $\{A_k\}_{k=1}^{J}$ form a cover of subspace $\Theta\times\mathbb{S}$, with $J(n)=O\left(h_y^{-1}h_x^{-1}n^{2c}\right)$, and

$$A_k=\left\{\theta,x,y:\left\|\theta-\overline{\theta}_k\right\|\leq\left(h_x n^{-c}\right)^{1/2},\left\|x-\overline{x}_k\right\|\leq\left(h_x n^{-c}\right)^{1/2},\left\|y-\overline{y}_k\right\|\leq h_x n^{-c}\right\},$$

Define for $\left(\overline{\theta}_k,\overline{y}_k,\overline{x}_k\right)$,

$$Z_i=K\left(\frac{\overline{\theta}_k^T\left(\overline{x}_k-x_i\right)}{h_x}\right)K\left(\frac{\overline{\theta}_k^T\left(\overline{y}_k-y_i\right)}{h_y}\right)-E\left(K\left(\frac{\overline{\theta}_k^T\left(\overline{x}_k-x_i\right)}{h_x}\right)K\left(\frac{\overline{\theta}_k^T\left(\overline{y}_k-y_i\right)}{h_y}\right)\right).$$

Now, notice that $|Z_i| \leq b \equiv 2\sup_{\Theta\times\mathbb{S}_X} |K(u)| < \infty$, and by Theorem 1 of Hansen (2008), for any $1 \leq m \leq n$,

$$\sigma^2(m) \equiv \sup_{(\theta,x)} E\left|\sum_{i=1}^{\lfloor m\rfloor} Z_i\right|^2 \leq Cmh_yh_x$$

for some large enough $C > 0$. Set $m = Cn^{1-2c}h_yh_x/a_1$ and $\varepsilon = a_1n^{1-c}h_yh_x$, and note that $4bm < \varepsilon$ for any sufficiently large n. By Theorem 2.1 of Liebscher (1996) we obtain

$$\begin{aligned}
& P\left(\left|\widehat{f}_{Y,\theta^TX}\left(\overline{y}_k,\overline{\theta}_k^T\overline{x}_k\right) - E\widehat{f}_{Y,\theta^TX}\left(\overline{y}_k,\overline{\theta}_k^T\overline{x}_k\right)\right| > a_1n^{-c}\right) \\
= \;& P\left(Z_i > \varepsilon\right) \\
\leq \;& 4\exp\left(-\frac{\varepsilon^2}{64\frac{n}{m}\sigma^2(m) + \frac{8}{3}\varepsilon mb}\right) + 4\frac{n}{m}\alpha_m \\
\leq \;& 4\exp\left(-\frac{a_1^2n^{2-2c}h_y^2h_x^2}{64Cn^1h_yh_x + \frac{8}{3}Cn^{2-3c}h_y^2h_x^2b}\right) + 4An^{-\eta_1}h_y^{-1}h_x^{-1}n^{2c} \\
\leq \;& 4\exp\left(-\frac{a_1^2n^c}{C(64+3b)}\right) + 4AJ(n)\,n^{-\eta_1},
\end{aligned}$$

where $J(n) = h_y^{-1}h_x^{-1}n^{2c}$. Thus, we have

$$\sum_{n=1}^{\infty} J(n)\left(\sup_{z\in\mathbb{S}}\sup_{\theta\in\Theta}\left|\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right) - E\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)\right| > a_1n^{-c}\right) < \infty, \qquad (39)$$

and (37) is established with (38), (39), and the arguments in the proof of Lemma 6.

■

Acknowledgements

The author is most grateful to Professor Q. Yao for helpful discussions and comments.

References

[1] Aït-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. The Journal of Finance, 54, 1361–1395.

[2] An, H.Z., Huang, F.C. (1996). The geometric ergodicity of nonlinear autoregressive models. Statistica Sinica, 6, 943-956.

[3] Carrasco, M., Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. Econometric Theory, 18, 17-39.

[4] Davis, R.A., Mikosch, T. (2009). Probabilistic Properties of Stochastic Volatility Models. In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (eds.): Handbook of Financial Time Series, 255-267. Springer, New York.

[5] Delecroix, M., Härdle, W., Hristache, M. (2003). Efficient estimation in conditional single-index regression. Journal of Multivariate Analysis, 86, 213–226.

[6] Delecroix, M., Hristache, M., Patilea, V. (2006). On semiparametric M-estimation in single-index regression., Journal of Statistical Planning and Inference, 136, 730–769.

[7] Engle, R.F. (2001). Financial econometrics a new discipline with new methods. Journal of Econometrics, 100, 53-56.

[8] Engle, R. F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics, 22, 367–381.

[9] Fan, J., Peng, L., Yao, Q., Zhang, W. (2009). Approximating conditional density functions using dimension reduction. Acta Mathematica Applicatae Sinica, 25, 445-456.

[10] Fan J., Yao, Q. (2003). Nonlinear time series: nonparametric and parametric methods. Springer-Verlag.

[11] Fan, J., Yao, Q., Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika, 83, 189-206.

[12] Gao, J., King, M.L. (2004). Adaptive testing in continuous-time models. Econometric Theory, 20 , 844-882.

[13] Hall, P. (1989). On projection pursuit regression. Annals of Statistics, 17, 573–588.

[14] Hall, P., Yao, Q. (2005). Estimation for conditional distribution functions via dimension reduction. The Annals of Statistics, 33, 1404-1421.

[15] Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. Econometric Theory, 24 , 726-748.

[16] Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. Annals of Statistics, 21, 157–178.

[17] Härdle, W., Stoker, T.M. (1989). Investigating smooth multiple regression by method of average derivatives. Journal of the American Statistical Association, 84, 986-995.

[18] Hyndman, R.J. (1995). Highest density forecast regions for non-linear and non-normal time series models. Journal of Forecasting, 14, 431–441.

[19] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics, 58, 71-120

[20] Ichimura, H., Todd, P.E. (2006). Implementing nonparametric and semiparametric estimators. In Handbook of Econometrics, Volume 6.

[21] Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. Stochastic Processes and their Applications, 65, 69-80.

[22] Marron, J.S. (1992). Graphical understanding of higher order kernels. North Carolina Inst. Statistics Mimeo Series 2082.

[23] Marron, J.S., Wand M.P. (1992). Exact mean integrated squared error. Annals of Statistics, 20, 712-736.

[24] Müller, H.G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. Annals of Statistics, 12, 766–774.

[25] Pham, D.T., Tran, L.T. (1985). Some mixing properties of time series models. Stochastic Processes and their Applications, 19, 297–303.

[26] Scott, D.W. (1992). Multivariate density estimation: Theory, Practice, and Visualization. John Wiley and Sons, New York, Chichester.

[27] Silverman, B.W. (1986). Density estimation. London: Chapman and Hall.

[28] White, H. (1984). Asymptotic theory for econometricians. Orlando, Academic Press, Inc.

[29] Wu, W., Yu, K., Mitra, G. (2008). Kernel Conditional Quantile Estimation for Stationary Processes with Application to Conditional Value-at-Risk. Journal of Financial Econometrics, 6, 253-270.

[30] Xia Y., Härdle, W., Linton, O. (2012). Optimal Smoothing for a Computationally and Statistically Efficient Single Index Estimator, within "Exploring Research Frontiers in Contemporary Statistics and Econometrics" edited by Van Keilegom I. and Wilson P.W. Springer-Verlag, Berlin.

[31] Xia Y., Tong, H., Li, W.K. (1999). On extended partially linear single-index models. Biometrika, 86, 831–842.

[32] Yao, Q., Tong, H. (1994). On prediction and chaos in stochastic systems. Philosophical Transactions of the Royal Society A, 348, 357-369.

[33] Yin, X., Cook, R.D. (2002). Dimension reduction for the conditional k-th moment in regression. Journal of the Royal Statistical Society Series B, 64, 159–175.

[34] Yin, X., Cook, R.D. (2005). Direction estimation in single-index regressions. Biometrika, 92, 371–384.