A NONPARAMETRIC REGRESSION ESTIMATOR THAT ADAPTS TO ERROR DISTRIBUTION OF UNKNOWN FORM*

Oliver Linton†

The London School of Economics and Political Science

Zhijie Xiao‡

University of Illinois at Urbana-Champaign

October 28, 2004

Abstract

Motivated by efficiency considerations we propose a new estimator for nonparametric regression based on estimated local likelihood estimation. We show that our estimator is asymptotically equivalent to the infeasible local likelihood estimator [Staniswalis (1989), Fan, Farmen, and Gijbels (1998), and Fan and Chen (1999)], which requires specification of the error distribution, and hence our estimator improves on standard nonparametric estimators when the error distribution is not normal. We investigate the finite sample performance of our procedure on simulated data. An empirical application to the IBM transaction data is conducted to study seasonality of high frequency intra-day stock returns.

Keywords: Adaptive Estimation; Efficiency; Local Likelihood Estimation; Local Polynomial Estimation; Nonparametric Regression.

---

*We would like to thank Yanqin Fan, Joel Horowitz, Roger Koenker, Whitney Newey, Jens Perch Nielsen, Peter Phillips, Steve Portnoy, Peter Robinson, and Tom Rothenberg for helpful comments. Financial support from the NSF and the ESRC (UK) is gratefully acknowledged.

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Phone: (0207) 955-7864. Fax: (0207) 831-1860. E-mail address: O.Linton@lse.ac.uk.

‡Department of Economics, University of Illinois at Urbana-Champaign, 1206 South Sixth Street, Champaign, IL 61820, USA. Phone: (217) 333-4520. Fax: (217) 244-6678. E-mail address: zxiao@uiuc.edu.

AMS 1991 subject classifications. 62G07, 62A10

1 Introduction

It has been known for some time that in certain parametric regression models it is possible to 'adapt' to an unknown error distribution by maximizing an estimated likelihood function based on an estimate of the error distribution. The common result is that you do as well in terms of asymptotic variance as if you knew the true error distribution, hence the term adaptive. In estimation problems where a Gaussian assumption on the underlying distribution of the data is inappropriate, adaptive estimation provides an alternative way to the conventional Gaussian maximum likelihood estimator by replacing the Gaussian density function with a nonparametric estimate of the score function of the log-likelihood. It has been proven that an efficiency gain over the least squares methods can be achieved by adaptive estimators in many models.

Adaptive estimation was first studied by Stein (1956) who considered the problem of estimating and testing hypotheses about a parameter in the presence of an infinite dimensional "nuisance" parameter. Beran (1974) and Stone (1975) considered adaptive estimation in the symmetric location model, while Bickel (1982) extended this to linear regression and other models. This latter work provided a starting point for much future work in this area, most of which has exploited the property of Local Asymptotic Normality (LAN) of the class of likelihoods involved. Manski (1984) studied adaptive estimation in non-linear models, Kreiss (1987) considered stationary and invertible autoregressive moving average (ARMA) models, Steigerwald (1992) studied linear regression with ARMA error, and Linton (1993) considered the case of linear regression with autoregressive conditional heteroscedasticity (ARCH), which was extended by Drost and Klaassen (1997) to the GARCH(1,1) case. See Drost, Klaassen, and Werker (1997) for an excellent review and recent development for time series. Jeganathan (1995) extended the theory to nonstationary models with independent and identically distributed (i.i.d.) error, which involves the generalization to Local Asymptotic Mixed Normal (LAMN) likelihoods. Hodgson (1998) further extended this case but with ARMA errors.

We propose a new estimator for nonparametric regression that adapts to the unknown shape of the error term. Consider the following nonparametric regression

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where ε_i is an i.i.d. error term possessing smooth Lebesgue density and satisfying $E(\varepsilon_i|X_i) = 0$

almost surely. The assumption that m is a smooth function implies that for X_i close to x, Y_i contains information about $m(x)$. A popular estimator of $m(x)$ is the local polynomial regression estimator, which can also be obtained from the following weighted least squares criterion

$$\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)(Y_i - \mathcal{P}(\theta, X_i - x))^2. \tag{2}$$

where K is a kernel function, $h = h(n)$ is a bandwidth sequence, and $\mathcal{P}(\theta, X_i - x)$ is a p-th order polynomial of $X_i - x$ with coefficient vector θ. When ε_i are Gaussian, (2) corresponds to the weighted likelihood criterion. In the absence of Gaussianity, asymptotic results of the above estimator generally still hold but this estimator is less efficient than estimators that exploit the distributional information. If the error density f were known, we may replace $(Y_i - \mathcal{P}(\theta, X_i - x))^2$ in (2) by the log likelihood $\log f(Y_i - \mathcal{P}(\theta, X_i - x))$ and obtain a likelihood-based estimator (Fan, Farman and Gijbels, 1998). The local likelihood estimator has lower variance than the least-square based local polynomial estimator as follows by the classical Cramer-Rao inequality; under some conditions, the bias of the local likelihood estimator is the same as the bias of the simple local polynomial estimator, so that the local likelihood estimator has lower mean squared error. In the conventional kernel estimation literature, Tibshirani (1984) introduced the local likelihood estimator in the context of nonparametric generalized linear models. This study was extended in Hastie and Tibshirani (1987). Staniswalis (1989) applied this idea to the estimation of a location parameter θ; she also derived the asymptotic variance of her local constant procedure, although she did not give the bias. Also see Hunsberger (1994), Fan, Heckman and Wand (1995), Aerts and Claeskens (1997), Carroll, Ruppert and Welsh (1998) for related studies. There has been much recent work mostly focusing on density and hazard estimation. In particular see: Copas (1994), Hjort (1993), Hjort and Jones (1996) and Loader (1996). Work in other areas includes Robinson (1989) for a time series regression problem, and the recent paper by Gozalo and Linton (2000) for nonlinear regression models. These procedures are discussed in Härdle and Linton (1994, see p17 especially).

In practice, f is generally unknown and the local likelihood estimator is infeasible. We propose an adaptive procedure for the estimation of (1) based on estimating the score function of the errors, i.e., by replacing f by a nonparametric estimator thereof. We establish the pointwise asymptotic distribution of our estimator of $m(x)$, for interior x, and show that it adapts in the sense that it has the same variance as the infeasible procedure based on knowing the error distribution. This is true regardless of the dimensionality of the regressors. A more complicated implementation of our procedure involves undersmoothing at the first step and will result in an estimator with exactly the same bias as the infeasible local likelihood estimator. In this case, our adaptive estimator achieves the same mean squared error in large samples as the infeasible local likelihood estimator. In this sense, our estimator is efficient. This is not to say that our method has any special properties with regard to minimax risk; as is well known it is not possible to achieve the lower bound here, see Fan (1993). But the pairwise comparison has been used elsewhere, see Linton (1997) and Fan, Härdle, and Mammen (1998). With regard to regularity conditions we make rather strong assumptions about the smoothness of the regression function and error density but impose very few other requirements on the error density: in one set of conditions we allow it to have unbounded support, while in another we look at the case of bounded support. In the latter case, we must also require that the error density approaches zero at the boundary rather fast because otherwise the estimation problem is non-regular, i.e., there exist estimators that converge to the true value more rapidly.

Why is this important in practice? In many data sets the error distribution is likely to be non-normally distributed and perhaps quite far from the normal distribution to such an extent that the local likelihood estimator has much lower variance than the Gaussian-based estimators [and the relative efficiency is unbounded]. As a modelling issue it is hard to believe that we have better information about the error density than about the shape of the main regression effect and so it is quite natural to treat the error density as an unknown parameter just like the regression function. Thus our results are comforting in that they say that we can still use information from the error distribution to improve the performance of our location estimates; this is all the more important in nonparametric regression because the rate of convergence can be so slow.

The principle involved extends to allow for heteroskedasticity of unknown form and to other location models that depend on several functional parameters, like additive regression models. In more general nonparametric models, we may or may not find adaptivity just as in the parametric case, see Bickel, Klaassen, Ritov, and Wellner (1993).

The paper is organized as follows. The model and estimators are given in the next section. Asymptotic results of the estimator are given in Section 3. In section 4 we provide a small Monte Carlo experiment which evaluates the effectiveness of the adaptive regression estimator. An empirical application to IBM transaction data is also performed to study the daily pattern for transaction duration. We give a sketch of the proof of our main result in the appendix.

For notation, we use $f^{(j)}$ to denote the j^{th} derivative of a function f. We also let $\|A\|$ denote the Euclidean norm of the array $A = (a_{i_1,\ldots,i_s})$ defined as $\|A\| = (\sum a^2_{i_1,\ldots,i_s})^{1/2}$. For vectors $\mathbf{k} =$

$(k_1, \ldots, k_d)$ and $x = (x_1, \ldots, x_d)$, we use the following notations

$$\mathbf{k}! = k_1! \times \cdots \times k_d!, \;\; |\mathbf{k}| = \sum_{i=1}^{d} k_i, \;\; x^{\mathbf{k}} = x_1^{k_1} \times \cdots \times x_d^{k_d}, \;\; \sum_{0 \le |\mathbf{k}| \le p} = \sum_{j=0}^{p} \sum_{k_1=0}^{j} \cdots \sum_{\substack{k_d=0 \\ k_1+\cdots+k_d=j}}^{j}. \tag{3}$$

2 The Model and Estimator

2.1 The Model and an Infeasible Estimator

Suppose that $\{Y_i, X_i\}_{i=1}^n$ are i.i.d., where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. We consider the regression model (1) where ε_i is independent of X_i with $E(\varepsilon_i) = 0$ and ε_i has a density function f. We are interested in estimation of the regression function $m(x) = E[Y|X = x]$, where x is an interior point in the support of X_i, and $m(\cdot)$ is assumed to be of unknown form, but smooth. We concentrate on the regression case; instead one could treat other location functionals like the conditional median, in which case moments would not be required, but it is central to identification that the location functional be explicitly specified as this determines the estimation strategy. In view of this we focus on the regression problem.

The approach proposed in this paper may be applied to a wide range of nonparametric estimators, including the widely used Nadaraya-Watson procedure and the local polynomial estimator. In this paper, we give an asymptotic analysis based on the local polynomial procedures. See Fan (1992), and Fan and Gijbels (1996) for discussion on the attractive properties of local polynomials. For a given dataset $\{Y_i, X_i\}_{i=1}^n$, the local polynomial M-estimator of Z_i on X_i [of order p, where p is an integer] can be obtained from minimizing the following criterion:

$$\sum_{i=1}^{n} \varrho\left(Y_i - \mathcal{P}(\theta, X_i - x)\right) K\left(\frac{X_i - x}{h}\right), \tag{4}$$

where: ϱ is a user-determined criterion, $\mathcal{P}(\theta, u) = \sum_{0 \le |\mathbf{k}| \le p} \theta_{\mathbf{k}} \cdot u^{\mathbf{k}}$ is a p^{th} order polynomial in the vector u with coefficients θ, $K(u)$ is a nonnegative weight function on $\mathbb{R}^d$ and h is a bandwidth parameter. Let $\overline{\theta}$ be the vector of minimizing parameters in (4), and let $\overline{m}(x) = \overline{\theta}_0$, where $\overline{\theta}_0$ is the minimizing intercept.

Following the notation of Masry (1996a,b), let $N_\ell = (\ell + d - 1)!/\ell!(d-1)!$ be the number of distinct d-tuples j with $|j| = \ell$. Arrange these N_ℓ d-tuples as a sequence in a lexicographical order and let ϕ_ℓ^{-1} denote this one-to-one map. For each j with $0 \le |j| \le 2p$, let $\mu_j(K) = \int_{\mathbb{R}^d} u^j K(u) du$, $\nu_j(K) = \int_{\mathbb{R}^d} u^j K^2(u) du$, and define the $N \times N$ dimensional matrices M and Γ and $N \times 1$ vector B, where

$N = \sum_{\ell=0}^{p} N_\ell \times 1$, by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & & & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p} \\ \vdots & & & \vdots \\ \Gamma_{p,0} & \Gamma_{p,1} & \cdots & \Gamma_{p,p} \end{bmatrix}, \quad B = \begin{bmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{bmatrix},$$

where $M_{i,j}$ and $\Gamma_{i,j}$ are $N_i \times N_j$ dimensional matrices whose (ℓ, m) element are, respectively, $\mu_{\phi_i(\ell)+\phi_j(m)}$ and $\nu_{\phi_i(\ell)+\phi_j(m)}$. Note that the elements of the matrices $M = M(K)$ and $\Gamma = \Gamma(K)$ are simply multivariate moments of the kernel K and K^2, respectively. In addition, we arrange the N_r elements of the derivatives

$$\frac{1}{r_1! \cdots \cdot r_d!} \frac{\partial^r m(x)}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}, \text{ for } r_1 + \cdots + r_d = r$$

as a column vector $m^{(r)}(x)$.

In this section we shall assume that the density f is known. It is therefore desirable to take $\varrho \propto \log f$ in (4) in which case the estimator has a local likelihood interpretation. When ε_i is normally distributed, $\varrho(z) = z^2$, and the resulting $\overline{m}(x)$ is the standard local polynomial regression smoother, while if ε_i is assumed to be Laplace distributed, then $\overline{m}(x)$ is the standard local polynomial median smoother, see Chaudhuri (1991).

In general, the estimator $\overline{m}(x)$ is only implicitly defined and is a nonlinear function of $Y_1, \ldots, Y_n$. In practice some numerical algorithm must be used to compute the values $\overline{\theta}$, such as the Newton-Raphson method. We shall work with the one-step Newton-Raphson estimator [as analyzed in Bickel (1975) for a class of parametric estimation problems] from a preliminary consistent estimator $\widetilde{\theta}(x)$ [say, the local polynomial least squares estimator that takes $\varrho(z) = z^2$ in (4)], that is, let $\overline{m}_{NR}(x)$ be the first element of $\overline{\theta}_{NR}(x)$, where

$$\overline{\theta}_{NR}(x) = \widetilde{\theta}(x) + \mathcal{I}_n(\widetilde{\theta}(x); f)^{-1} S_n(\widetilde{\theta}(x); f), \tag{5}$$

where $S_n(\theta; f)$ is the smoothed score function

$$S_n(\theta; f) = -\frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{f'}{f}(Y_i - \mathcal{P}(\theta, X_i - x))\widetilde{\mathbf{X}}_i, \tag{6}$$

where

$$\widetilde{\mathbf{X}}_i = \begin{bmatrix} \mathbf{X}_{i,0}(x) \\ \vdots \\ \mathbf{X}_{i,p}(x) \end{bmatrix}$$

and $\mathbf{X}_{i,|j|}(x)$ $(0 \leq |\mathbf{j}| \leq p)$ is a $N_{|j|}$ dimensional subvector whose r-th element is given by $\left[\mathbf{X}_{i,|j|}\right]_r = (x - X_i)^{\phi_{|j|}(r)}$, while $\mathcal{I}_n(\widetilde{\theta}(x); f)$ is a consistent estimate of the information $\mathcal{I}_\theta(x)$ given by

$$\mathcal{I}_n(\widetilde{\theta}(x); f) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \left[\frac{f'}{f}(Y_i - \mathcal{P}(\widetilde{\theta}(x), X_i - x))\right]^2 \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top.$$

We next state the properties of $\overline{\theta}(x)$ (and thus $\overline{m}(x)$). Proposition 1 is essentially Fan and Chen (1999) except that we allow for a more general multivariate random design for the covariates. It gives the asymptotic distribution of $\overline{m}(x)$ and show that it is asymptotically more efficient than the generic estimator $\widehat{m}(x)$.

Define $H = \mathrm{diag}(I_{N_0}, hI_{N_1}, \ldots, h^p I_{N_p})$, where I_{N_j} are N_j-dimensional identity matrices, and

$$\mathcal{I}_\theta(x) = f_X(x) I(f), \tag{7}$$

where $f_X(x)$ is the covariate density, and

$$I(f) = \int \left[\frac{f'(\varepsilon)}{f(\varepsilon)}\right]^2 f(\varepsilon) d\varepsilon \tag{8}$$

is the Fisher information for the error density f.

Proposition 1. *Suppose that our assumptions A1-6 given below hold. Then, there exists a bounded continuous vector function $\overline{\beta}(x)$ such that*

$$\sqrt{nh^d}\left(H[\overline{\theta}(x) - \theta(x)] - h^{p+1}\overline{\beta}(x)\right) \Rightarrow N\left(0, \mathcal{I}_\theta^{-1}(x)\left[M^{-1}\Gamma M^{-1}\right]\right).$$

where the bias term $\overline{\beta}(x) = M^{-1}Bm^{(p+1)}(x)$. In particular,

$$\sqrt{nh^d}\left(\overline{m}(x) - m(x) - h^{p+1}\left[\overline{\beta}(x)\right]_0\right) \Rightarrow N\left(0, \mathcal{I}_\theta^{-1}(x)\left[M^{-1}\Gamma M^{-1}\right]_{0,0}\right),$$

where $[C]_0$ signifies the first element of vector C and $[A]_{0,0}$ signifies the upper-left element of matrix A.

Remark 1. The limiting variance of $\overline{m}(x)$ is smaller than the variance of the local polynomial least squares estimator, which is $\sigma^2\left[M^{-1}\Gamma M^{-1}\right]_{0,0}/f_X(x)$, where σ^2 is the error variance. We may also compare $\overline{m}$ with the general class of local $M-$estimators based on (4) where $E[\psi(\varepsilon_i)] = 0$ for $\psi(z) = \varrho'(z)$. In that case, the variance is proportional to $v_\psi(x) = f_X^{-1}(x)E\psi^2(\varepsilon_i)/[E\psi'(\varepsilon_i)]^2$. It follows from the classical Cramér-Rao inequality that $v_\psi(x) \geq \mathcal{I}_\theta^{-1}(x)$ for any such ψ. However, although this estimator may have smaller variance it should be noted that it can be bettered for some combination of (f, m, f_X), see Fan (1993), and its minimax optimality is unproven.

In practice, f is generally unknown and so neither $\overline{\theta}(x)$ nor $\overline{\theta}_{NR}(x)$ are feasible. However, the infeasible procedure defines an efficiency standard against which we should measure our feasible estimator.

2.2 The Feasible Estimator

To obtain a feasible analogue of $\overline{\theta}(x)$, we replace f in (4) or rather (5) by a nonparametric estimate, say $\widetilde{f}$. As in some other applications of kernel regression estimators, the estimator $\widetilde{f}(e_i)$ can be small and may cause technical difficulty since it enters into the estimated score function in the denominator. For this reason, we trim out small $\widetilde{f}(e_i)$ as do Bickel (1982) and Kreiss (1987). The simplest and probably most common trimming is the indicator trimming function: $G_b(x) = I(|x| \geq b)$, where b is the trimming parameter that goes to zero as $n \to \infty$. Instead, we consider the following smoothed trimming, which has been used recently by Andrews (1995) and Ai (1997). Let $g(\cdot)$ be a density function that has support $[0,1]$, $g(0) = g(1) = 0$, and let

$$g_b(x) = \frac{1}{b} g\left(\frac{x}{b} - 1\right),$$

where b is the trimming parameter; then $g_b(x)$ has support on $[b, 2b]$. Letting $G_b(x) = \int_{-\infty}^{x} g_b(z)dz$, we have

$$G_b(x) = \begin{cases} 0, & x < b \\ \int_{-\infty}^{x} g_b(z)dz, & b \leq x \leq 2b \\ 1, & x > 2b. \end{cases}$$

For example, consider the following Beta density $g(z) = B(a+1)^{-1} z^a (1-z)^a, \quad 0 \leq z \leq 1$, for some positive integer a, where $B(a)$ is the beta function defined by $B(a) = \Gamma(a)^2/\Gamma(2a)$, and $\Gamma(a)$ is the Euler gamma function. Then, it can be verified that the function $G_b(x)$ is $(a+1)-$times continuously differentiable on $[0,1]$. This property allows us to use standard Taylor series arguments, whereas indicator function trimming would preclude this. We will suppose that $a \geq 3$.

The proposed estimation algorithm is as follows

1. First obtain a preliminary consistent estimator of the function m and its derivatives by pth order local polynomial smoothing Y_i on X_i with multivariate kernel $\mathcal{K}$ and bandwidth h_1 and call the resulting estimator $\widetilde{\theta}(\cdot)$. Calculate the residuals $\widetilde{\varepsilon}_i = Y_i - \widetilde{m}(X_i)$.

2. Estimate the error density and its derivatives by the leave-one-out kernel estimates:
$$\widetilde{f}(e_i) = \frac{1}{nh_0} \sum_{j \neq i} k\left(\frac{e_i - \widetilde{\varepsilon}_j}{h_0}\right), \ \widetilde{f}'(e_i) = \frac{1}{nh_0^2} \sum_{j \neq i} k'\left(\frac{e_i - \widetilde{\varepsilon}_j}{h_0}\right), \tag{9}$$
with univariate kernel $k(\cdot)$ and bandwidth h_0.

3. Define the trimmed local score function as
$$\widetilde{S}_n(\theta; \widetilde{f}) = -\frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{\widetilde{f}'}{\widetilde{f}}(Y_i - \mathcal{P}(\theta, X_i - x)) G_b(\widetilde{f}(Y_i - \mathcal{P}(\theta, X_i - x))) \widetilde{\mathbf{X}}_i \tag{10}$$

and the trimmed information matrix by

$$\widetilde{\mathcal{I}}_n(\theta;\widetilde{f})=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\left[\frac{\widetilde{f}'}{\widetilde{f}}(Y_i-\mathcal{P}(\theta,X_i-x))\right]^2 G_b(\widetilde{f}(Y_i-\mathcal{P}(\theta,X_i-x)))\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^{\top}. \quad (11)$$

4. Then calculate

$$\widehat{\theta}(x)=\widetilde{\theta}(x)+\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x);\widetilde{f})^{-1}\widetilde{S}_n(\widetilde{\theta}(x);\widetilde{f}). \quad (12)$$

5. Repeat until convergence (optional).

In the next section we will give the asymptotic distribution of the one-step estimator (12). One can expect better performance in practice by iterating (12) a few times or until convergence.

The preliminary estimator $\widetilde{\theta}(x)$, the density estimators $\widetilde{f}$ and $\widetilde{f}'$ in (9), and the estimators of the score $\widetilde{S}_n$ in (10) and the information $\widetilde{\mathcal{I}}_n$ in (11) all involve choices of kernels and bandwidth values. We use univariate kernel $k(\cdot)$ and bandwidth h_0 in estimating the density function $\widetilde{f}$ and $\widetilde{f}'$, use multivariate kernel $\mathcal{K}(\cdot)$ and bandwidth h_1 in the construction of the residuals $\widetilde{\varepsilon}_j$ in step 1, and use kernel $K(\cdot)$ and bandwidth h in the estimation of $\widetilde{\theta}(x)$, $\widetilde{S}_n$, and $\widetilde{\mathcal{I}}_n$. For simplicity of proofs, we assume that $h_1=h_0=h/\log n$ in Assumption A5 (Section 3). A wider range of bandwidth choices can be allowed with more complicated proof (magnitude of additional terms needs to be verified). Another simple approach (with a modified proof to control the order of magnitude of certain terms) is to take $\mathcal{K}=K=k\times\cdots\times k$, and to take $h=h_1=h_0^{1/d}$, which reflects the dimensionality of the estimands. In our theoretical work and presentation of Theorem 1, we do not restrict h, h_0, h_1, but rather assume they initially they are of the same magnitude.

Our analysis is conducted based on the local polynomial estimator. The same idea may be applied to other types of estimators. See an early version of this paper (Linton and Xiao (2001)) for results on the Nadaraya-Watson kernel estimator. This leads to a difference in the bias expression.

In the event that the error density is known to be symmetric, some improvement can be expected by symmetrizing the error density and derivative, thus $\widetilde{f}(e)\mapsto(\widetilde{f}(e)+\widetilde{f}(-e))/2$ and $\widetilde{f}'(e)\mapsto(\widetilde{f}'(e)-\widetilde{f}'(-e))/2$ in (9), although this only affects the higher order terms in our case. Bickel (1982) proposes an estimator that incorporates these restrictions. In his proofs he exploits the symmetry properties of the estimator. We do not impose symmetry on our estimator and our proof technique is quite different from Bickel's. The restriction that the error density is mean zero can be imposed by using centred residuals and a kernel that has mean zero.

3 Main Results

To facilitate the asymptotic analysis, we make the following assumptions on the distribution, the kernel function $k(\cdot)$, the bandwidth parameter h, and on the trimming parameter b in $\widetilde{\theta}(x)$, $\widetilde{f}$, $\widetilde{f}'$, $\widetilde{S}_n$, and $\widetilde{\mathcal{I}}_n$.

A1. *ε_i and X_i are independent and identically distributed (i.i.d.) random variables (vectors), mutually independent, and $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2 < \infty$.*

A2. *ε_i has support $\mathbb{R}$ and has Lebesgue density $f(\varepsilon)$, which has uniformly bounded continuous partial derivatives up to the order $r = q + 1$. Furthermore, $f^{(r)}(\varepsilon)$ is Lipschitz continuous, i.e., there exists $c < \infty$ such that for all $\varepsilon, \varepsilon^*$ on its support, we have*

$$\left|f^{(r)}(\varepsilon) - f^{(r)}(\varepsilon^*)\right| \leq c|\varepsilon - \varepsilon^*|.$$

A3. *Let $\ell(\varepsilon) = \log f(\varepsilon)$, and suppose that $E[(\ell'(\varepsilon))^2] < \infty$, $E[|\ell''(\varepsilon)|] < \infty$ and $E[|\ell'''(\varepsilon)|] < \infty$.*

A4. *The kernel k has support $[-1, 1]$ and is symmetric about zero and satisfies $\int k(u)du = 1$. In addition, $\int u^j k(u)du = 0$, $j = 1, \ldots, q-1$, and $\int u^q k(u)du \neq 0$. Furthermore, k is four times differentiable on its support and $k'(0) = 0$.*

A5 *The kernel k_1 has support $[-1, 1]$ and is symmetric about zero and satisfies $\int k_1(u)du = 1$. The functions $H_j(u) = u^j K(u)$, where $K = k_1 \times \cdots \times k_1 = K$, for all j with $0 \leq |j| \leq 2p+1$, are Lipschitz continuous, i.e., there exists a positive finite constant C such that $|H_j(u) - H_j(v)| \leq C||u - v||$.*

A5. *(a) $nh^{2q+d} \to 0$; (b) $nh^{d+9} \to \infty$; (c) $b = h^\tau$ with $0 < \tau < 1/2$; (d) $h_0 = h_1 = h/\log n$.*

A6. *X_i has Lebesgue density $f_X(x)$ which is bounded away from zero on its support $\mathcal{X}$, a compact subset of $\mathbb{R}^d$. $(D^{\mathbf{k}} f_X)(x)$ and $(D^{\mathbf{k}} m)(x)$ are bounded and uniformly continuous on $\mathcal{X}$, and there exists $C_2 < \infty$ and $C_3 < \infty$ such that, with $|\mathbf{k}| = r$,*

$$|(D^{\mathbf{k}} f_X)(u) - (D^{\mathbf{k}} f_X)(v)| \leq C_2||u - v||, \ |(D^{\mathbf{k}} m)(u) - (D^{\mathbf{k}} m)(v)| \leq C_3||u - v||.$$

In A1 we assume the existence of the variance σ^2 of ε_i, but this is just for the purpose of verifying the properties of the estimator in step 1. If for example the local median smoother were used, then it is possible to make weaker assumptions about the error moments. Assumption A2 includes a wide range of error distributions and includes the normal distribution as a special case. Assumptions A2 and A3 ensure the adaptive property and Taylor expansions of the density function to appropriate orders. By dominated convergence, Assumption A3 also ensures that:

$$\lim_{b\to 0}\int_{f(\varepsilon)<b}[\ell'(\varepsilon)]^2 f(\varepsilon)d\varepsilon = 0, \tag{13}$$

$$\lim_{b\to 0}\int_{f(\varepsilon)<b}\ell''(\varepsilon)f(\varepsilon)d\varepsilon = 0, \tag{14}$$

which guarantee that the trimming effect will be asymptotically ignorable.

We may instead want to work with error densities that have bounded support. In this case, we replace assumptions A2 and A3 by the following assumptions A2$'$ and A3$'$ below, which are sufficient conditions for (13) and (14).

A2$'$. *ε_i has symmetric Lebesgue density $f(\varepsilon)$ which has support* $\mathrm{supp}(f) = [\underline{a}, \overline{a}]$*, where $\underline{a}$ and $\overline{a}$ are unknown boundary parameters that satisfy $-\infty < \underline{a} < \overline{a} < \infty$, and $f(\varepsilon) > 0$ on $(\underline{a}, \overline{a})$. In addition, f has uniformly bounded continuous partial derivatives up to the order r, and $f^{(r)}(\varepsilon)$ is Lipschitz continuous on $(\underline{a}, \overline{a})$, i.e., there exists a constant c such that for all $\varepsilon, \varepsilon^* \in (\underline{a}, \overline{a})$, we have*

$$\left|f^{(r)}(\varepsilon) - f^{(r)}(\varepsilon^*)\right| \leq c|\varepsilon - \varepsilon^*|.$$

In A2$'$ we assume that $f(\varepsilon)$ has bounded support. When f is strictly positive on $[\underline{a}, \overline{a}]$, the situation is non-regular. In some cases, this can lead to inconsistency of solutions of the likelihood score equations, but perhaps the potential for improved rates of convergence for other estimators. Therefore, we shall make an additional assumption

A3$'$. *$f(\varepsilon)$ and its first $s-1$ derivatives vanish at $\underline{a}$ and $\overline{a}$, while $f^{(s)}(\underline{a}) \neq 0$, and $f^{(s)}(\overline{a}) \neq 0$ for some integer s with $2 \leq s \leq r$.*

Assumption A3$'$ guarantees that the density f vanishes at the boundary at a sufficiently fast rate so that the properties of regular estimation holds. See Akahira and Takeuchi (1995) for a discussion of this situation. This assumption also implies that the Fisher information $I(f)$ defined in (8) exists as do various other integrals used below.

Note that because the trimming parameter b is of larger magnitude than h, our estimator will not suffer from boundary bias from the estimation of the density f. Assumption A5(b) is quite strong and implies that we must have $2q > 9$. This assumption is stronger than necessary and arises partly because we have chosen the same bandwidth throughout and partly because of our proof technique.

With a more complicated analysis, it is possible to show that $2q > 6$ is enough. We believe that the result is true even for positive kernels. Certainly, there is no restriction on the dimensionality of the regressors. Assumption A6 is a standard assumption in the literature. We have assumed the same smoothness for f_X, f, and m in A2 and A6.

REMARK 4. The regularity assumptions facilitate our asymptotic analysis. In practice, even when some of these conditions do not hold, if the error distribution is distant from normal, efficiency gain over the conventional kernel estimator may still be found in the adaptive estimator. Also see the Monte Carlo results in Section 4.

We obtain the following result for our adaptive nonparametric regression estimator.

THEOREM 1. *Suppose that Assumptions A1 to A6 [or A1,A2',A3', and A4-A6] hold. Then, as* $n \to \infty$

$$\sqrt{nh^d}\left(H[\widehat{\theta}(x) - \theta(x)] - h^{p+1}\beta(x)\right) \Rightarrow N\left(0, \mathcal{I}_\theta^{-1}(x)\left[M^{-1}\Gamma M^{-1}\right]\right),$$

where:

$$\beta(x) = M^{-1}Bm^{(p+1)}(x) + (h_0/h)^{p+1}\mu_q(k)[\mathcal{I}_\theta(x)M]^{-1}BQ_\varepsilon + (h_1/h)^{p+1}[\mathcal{I}_\theta(x)M]^{-1}AB\mathcal{F}_\varepsilon,$$

$$A = \left[M^{-1}B\right]_0 E\left[\frac{m^{(p+1)}(X)}{f_X(X)}\right], \mathcal{F}_\varepsilon = E\left[\frac{f'(\varepsilon)f''(\varepsilon)}{f(\varepsilon)^2}\right], Q_\varepsilon = E\left[\frac{f'(\varepsilon)f^{(q)}(\varepsilon)}{f(\varepsilon)^2} - \frac{f^{(q+1)}(\varepsilon)}{f(\varepsilon)}\right].$$

In particular,

$$\sqrt{nh^d}\left(\widehat{m}(x) - m(x) - h^{p+1}\left[\beta(x)\right]_0\right) \Rightarrow N\left(0, \mathcal{I}_\theta^{-1}(x)\left[M^{-1}\Gamma M^{-1}\right]_{0,0}\right).$$

Under the additional bandwidth assumption that $h_0/h \to 0$, and $h_1/h \to 0$, $\beta(x) = \overline{\beta}(x)$ and the feasible one-step estimator is asymptotically equivalent to the infeasible estimator.

It may appear obvious that the effect of estimating f should not effect the distribution of the resulting estimator of m, at least when ε is of lower dimension than X i.e., $d > 1$. However, this is a bit misleading because to first estimate f we need to estimate m; our estimator of f, indeed any estimator of f, has convergence rate determined by the preliminary estimation of m. Therefore, our result is quite surprising, especially when compared with other results for nonparametrically generated data [see for example Ahn (1995)], where the dominant term is that due to the higher dimensional estimation problem. On the other hand, under symmetry the location score is orthogonal to the score for the error density in parametric models, which is why one achieves adaptivity in that case. The same is true here. In addition, the estimator $\widehat{\theta}(x)$ depends on an estimate of a functional of the error density rather than the error density at a single point.

The notion of efficiency that is employed here is similar to that used in Fan, Mammen, and Härdle (1998) and Linton (1997) in the context of additive nonparametric regression. We are aiming to do as well as the corresponding estimator one would compute if one knew the error density f. However, any specific estimator can be bettered for some combination of (f, m, f_X), Fan (1993).

In the proof we need a bit more than the mean squared consistency of the score function estimator, as required in Bickel (1982) for the parametric problem. We require uniform rates of convergence on the density and derivative estimates.

The above result is also useful for deriving nonparametric confidence intervals for the regression function. Under the bandwidth assumptions, the bias term can be ignored and a confidence interval at significance level α can be constructed as follows

$$\left[\widehat{m}(x) - \frac{\widehat{s}(x)}{\sqrt{nh^d}} z_{\alpha/2}, \widehat{m}(x) + \frac{\widehat{s}(x)}{\sqrt{nh^d}} z_{\alpha/2}\right], \tag{15}$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ with $\Phi(\cdot)$ the standard normal distribution, and $\widehat{s}^2$ is a consistent estimate of the asymptotic variance. Suitable estimators include

$$\widehat{s}_1^2(x) = \frac{[M^{-1}\Gamma M^{-1}]_{0,0}}{\widehat{f}_X(x)\widehat{I}(f)}, \text{ with } \widehat{I}(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{f}'(\widetilde{\varepsilon}_i)}{\widetilde{f}(\widetilde{\varepsilon}_i)},$$

or

$$\widehat{s}_2^2(x) = \left[M^{-1}\Gamma M^{-1}\right]_{0,0} \left\{\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\left[\frac{\widetilde{f}'}{\widetilde{f}}(Y_i - \widetilde{m}(x))\right]^2\right\}^{-1}.$$

Assumption A5 gives general bandwidth conditions that are sufficient for the asymptotic analysis. Cross-validation may be used for bandwidth selection, and the method of Hengartner, Wegkamp, and Matzner-Lober (2002) is convenient. In particular, divide the index set $I = \{1, \ldots, n\}$ into two mutually exclusive and exhaustive (large) sets I_1 and I_2 with cardinalities n_1 and n_2, where $n_1 + n_2 = n$, and define the training set $\mathcal{Z}_1 = \{(Y_j, X_j), j \in I_1\}$ and the testing set $\mathcal{Z}_2 = \{(Y_j, X_j), j \in I_2\}$. Then compute $\widehat{m}(\cdot)$ and $\widetilde{f}(\cdot)$ using bandwidths $\mathcal{H} = (h, h_0, h_1)$ with data $\mathcal{Z}_1$ and choose $\mathcal{H}$ to minimize

$$\frac{1}{n_2}\sum_{j\in I_2}(Y_j - \widehat{m}(X_j))^2 \text{ or } \frac{1}{n_2}\sum_{j\in I_2} - \log \widetilde{f}(Y_j - \widehat{m}(X_j)). \tag{16}$$

Alternatively, a (complicated) higher order analysis of the nonparametric estimator may be conducted and a bandwidth choice could be determined based on optimizing the second order effects with respect to the bandwidth. See Fan, Farmen, and Gijbels (2000) for the development of a bandwidth selection method in a local likelihood context.

4 Numerical Results

4.1 Simulations

We conducted a small Monte Carlo simulation to evaluate the finite sample performance of the proposed estimation procedure. The data were generated from (1) with X_i being i.i.d. standard normal. Several specifications of $m(x)$ were investigated in generating the data and we report the results of two cases: DGP1: $m(x) = x^2$; and DGP2: $m(x) = x^4$. Different distributions of ε_i were considered. In particular, we considered cases where ε_i are i.i.d. t-distributions with different degrees of freedom. In particular, we report results for cases with degree of freedom 2 and 4. Notice that in the case of t(2), the regularity conditions do not hold because ε_i has infinite variance. However, the error distribution in this case is distant from normal and, as shown in our result, the adaptive estimator still brings efficiency gain. Two sample sizes were tried, $n = 100, 400$, and qualitatively similar results obtained. We report the results of the case $n = 100$.

The number of replications is 1000 in each case. We investigated both local linear estimation and the third order ($p = 3$) local polynomial estimation with kernel $\mathcal{K}(u) = 0.75(1 - u^2)1(|u| \leq 1)$. For the estimation of the density in step 2, we use the fourth order kernel $k(u) = 15(7u^4 - 10u^2 + 3)1(|u| \leq 1)/32$ when the third order local polynomial estimation was conducted. Given our choice of kernel, it is easy to verify that

$$\mu_j(\mathcal{K}) = \int u^j \mathcal{K}(u)du \ = \begin{cases} \frac{3}{(j+1)(j+3)}, & \text{j even} \\ 0, & \text{j odd} \end{cases}$$

and

$$\nu_j(\mathcal{K}) = \int u^j \mathcal{K}^2(u)du = \begin{cases} \frac{9}{(j+1)(j+3)(j+5)}, & \text{j even} \\ 0, & \text{j odd} \end{cases}$$

Thus the matrices M, B, Γ can be calculated using the above formula. For student-t distribution with k degrees of freedom,

$$\mathcal{I}_\theta(x) = f_X(x)I(f) = \frac{4\left(\frac{1+k}{2}\right)^2 \Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{k+2}{2}\right)}{k\left(\pi\right)^{1/2}\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k+5}{2}\right)}\phi(x),$$

where $\phi(x)$ is the standard normal density.

For comparison purpose, the sampling performance of both the local polynomial estimator and the adaptive local polynomial estimators were examined for each case. In particular, we compared the biases, variances, and mean squared errors of these estimators given different choices of innovation processes and bandwidth value. We estimated $m(x)$ at $x = 0$ and calculated the empirical Bias and Mean Squared Errors (MSE) of these estimates. In each case, different bandwidth values were considered. We consider bandwidth choices $h = d * s_X T^{-1/5}$ and $h_0 = h_1 = h/\ln(n)$ for different choices of d (ranging from 0.5 to 6), where s_X is the sample standard deviation of X. We report some selected representative results in graphics. Figures 1a-1d depict the biases, variances and MSE as a function of d (bandwidth). In all these figures, $m(x) = x^4$. In particular, Figure 1a reports these results of the 3rd order local polynomial estimator for the case with t(4) errors. Figure 1b gives those of the corresponding adaptive (3rd order) local polynomial estimator with the same data. Figure 1c and 1d report those results of the local linear estimator for the case with t(2) errors. Again, Figure 1c is the local linear estimator and Figure 1d corresponds to the adaptive local linear estimator. A comparison between biases and standard errors indicate that the variance effect is relatively large over a range of bandwidth choices, and thus a relatively large bandwidth is needed to minimize the mean square errors.

Figures 2a, 2b - 4a, 4b provide a comparison of the local polynomial estimator and the adaptive estimator in terms of biases, variances and MSE. In particular, Figures 2a 3a, and 4a compare the 3rd order local polynomial estimator and the corresponding adaptive estimator (for data with t(4) errors) in biases (Figure 2a), variances (Figure 3a), and MSE (Figure 4a). Figures 2b 3b, and 4b compare the local linear estimator and the corresponding adaptive estimator (for data with t(2) errors) in biases (Figure 2a), variances (Figure 3a), and MSE (Figure 4a). Other cases being qualitatively similar. For a more detailed comparison, see Tables 1-4 in the Appendix. Table 1 reports the third order local polynomial estimation results when $m(x) = x^2$ and ε_i are i.i.d. t-distributions with 2 degrees of freedom. Table 2 reports the estimation results when $m(x) = x^4$. Table 4 contains results for the case when ε_i are i.i.d. $t(4)$. Results based on local linear estimation and $m(x) = x^4$ are reported in Table 3.

Note that $\overline{\beta}(x) = M^{-1}Bm^{(p+1)}(x)$ and when $m(x) = x^2$, $m^{(j)}(x) = 0$ for all $j \geq 3$. Thus we expect that the estimation bias to be much smaller than the standard errors and a large bandwidth is needed to minimize the mean squared errors since the variance is the main contribution to the MSE. This is confirmed from our empirical result. Since the student-t distribution has a heavy-tail and is different from normal, efficiency gain can be observed from the adaptive procedure. It is clear that the adaptive estimators generally have smaller variances than the conventional estimator. They also have smaller MSE in most cases.

Our simulations indicate that the proposed estimator does a reasonable job for nonparametric estimation.

4.2 Empirical Application

In this section, we apply the proposed adaptive method of nonparametric regression to study seasonality of high frequency [intra-day] stock returns. Recent studies in finance suggest that frequency of transactions should carry information about the state of the market (see, e.g. Easley and O'Hara (1992), Engle and Russell (1998)). Given irregularly spaced transaction data, researchers are particularly interested in duration: the time interval between two transactions. The expected duration, conditional on past calendar time, is called as the time-of-day effect, and is usually the subject of interest to estimate. In this paper, we apply the adaptive nonparametric method to high frequency transaction data from financial markets to estimate the time-of-day effect.

We consider a sequence of arrival time, $\{x_i\}$, such that $x_1 < x_2 < \ldots < x_k < x_{k+1} < \ldots$, the interval between two arrival times, $y_i = x_i - x_{i-1}$, is the duration. We are interested in estimating the conditional expected duration, $m(x_{i-1}) = E(y_i|x_{i-1})$. The data set we consider is IBM transaction data occurred during regular trading hours over the period from November 1, 1990 to January 31, 1991. The regular trading hours (time of the day) are from 9:30AM to 4:00PM. With the rapid development in computing power and storage capacity, transaction data are being collected at higher frequencies. The IBM transaction data that we consider contains near 50,000 observations. We delete the overnight transaction and holidays.

The kernels we use are the same as those we used in Monte Carlo, i.e. $\mathcal{K}(u) = 0.75(1-u^2)1(|u| \leq 1)$ in the local polynomial estimation, and $k(u) = 15(7u^4 - 10u^2 + 3)1(|u| \leq 1)/32$ in estimating the density. Time of the day, x_i, are measured in hours and range from 9.5 (9:20AM) to 16 (4:00PM). Three bandwidth were considered in our applications: $h = 0.05$, 0.1, 0.134, corresponding to windows of length 3, 6, and 8 minutes. Figures 5 provide a normal Q-Q plot of the residuals, indicating there is some departure from normality in the data. Figures 6a -6c gives the nonparametric estimates of the "time-of-day" effect curves using these bandwidth. In particular, the solid lines are the adaptive estimates and the dashed lines are the conventional local polynomial estimates. The estimated daily pattern of transaction durations indicates that the most active time is in the beginning of the day, then the duration increases and reaches the longest period in the middle of the day. Transaction becomes active again before closing.

5 Concluding Remarks

To summarize, we have shown that one can adapt to distributional shape in nonparametric regression. This is true even when there is only one covariate, in which case the error density is as difficult to estimate as the regression function itself. This shows that the source of adaptation is not due to the lower dimensionality of the error density, but rather due to the same sort of orthogonality properties that one finds in the analysis of semiparametric models. Our results are likely to be of practical use in many situations where the error density is far from normal, such as in economics, finance, and insurance.

There are some extensions of the proposed adaptive nonparametric regression estimator that may be of interest. For example, suppose that we allow the error term to be heteroscedastic, i.e.,

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{17}$$

where ε_i are i.i.d. and mean zero with variance one, but $\sigma^2(.)$ is of unknown form. Then the density of $u_i = \sigma(X_i)\varepsilon_i$ also has mean zero but is a multiplicative convolution of the two terms and so our estimator although consistent under some conditions will not be efficient. An improvement can be made by using the standardized residuals $\widetilde{\varepsilon}_j = (Y_j - \widetilde{m}(X_j))/\widetilde{\sigma}(X_j)$, where $\widetilde{\sigma}(X_j)$ is some preliminary estimator of $\sigma(X_j)$, to estimate the density of ε_i. One could also envisage jointly estimating the location and scale parameter efficiently by a multiparameter local likelihood method, and the properties of such an estimator can be derived by similar techniques to the current case.

6 Appendix

We provide a sketch of the proof in this Appendix. For a more detailed proof, the readers are referred to an earlier version of this paper (Linton and Xiao (2001). We denote $\phi(x, y, z, \ldots)$ as a general function whose exact form may change from case to case. For simplicity, we denote $G_b(\widetilde{f}_i)$ and $G_b(f_i)$ as $\widetilde{G}_i$ and G_i. For two random variables X_n, Y_n, we say that $X_n \simeq Y_n$ whenever $X_n = Y_n(1 + o_p(1))$ as $n \to \infty$. Let E_i denote expectation conditional on the sigma field generated by $z_i = (X_i, Y_i)$. In the sequel we drop the argument x of $\theta(x), \theta_0(x)$, $\widetilde{\theta}(x)$, and $\widehat{\theta}(x)$.

6.1 Preliminaries

Minimizing (2) with respect to $\theta_{\mathbf{k}}$ gives an estimate $\widehat{\theta}_{\mathbf{k}}(x)$ and $\widetilde{m}(x) = \widehat{\theta}_0(x) = e_1^\top M_n^{-1}\Psi_n$, where $e_1 = (1, 0, \ldots, 0)^\top$ is the vector with the one in the first position, $M_n(x)$ and $\Psi_n(x)$ are symmetric $N \times N$ ($N = \sum_{\ell=0}^{p} N_\ell \times 1$) matrix and $N \times 1$ dimensional column vector respectively and are defined as

$$M_n(x) = \begin{bmatrix} M_{n,0,0}(x) & \dots & M_{n,0,p}(x) \\ \vdots & \ddots & \vdots \\ M_{n,p,0}(x) & \cdots & M_{n,p,p}(x) \end{bmatrix}, \; \Psi_n(x) = \begin{bmatrix} \Psi_{n,0}(x) \\ \vdots \\ \Psi_{n,p}(x) \end{bmatrix},$$

where $M_{n,|j|,|k|}(x)$ is a $N_{|j|} \times N_{|k|}$ dimensional submatrix with the (l,r) element given by

$$\left[M_{n,|j|,|k|}\right]_{l,r} = \frac{1}{nh^d}\sum_{i=1}^n \left(\frac{x-X_i}{h}\right)^{\phi_{|j|}(l)+\phi_{|k|}(r)} \mathcal{K}\left(\frac{x-X_i}{h}\right),$$

and $\Psi_{n,|j|}(x)$ is a $N_{|j|}$ dimensional subvector whose r-th element is given by

$$\left[\Psi_{n,|j|}\right]_r = \frac{1}{nh^d}\sum_{i=1}^n \left(\frac{x-X_i}{h}\right)^{\phi_{|j|}(r)} \mathcal{K}\left(\frac{x-X_i}{h}\right) Y_i.$$

The estimate of $m(x)$ is given by $\widetilde{m}(x) = e_1 M_n^{-1}\Psi_n$ and its bias and variance effects can be written as $\widetilde{m}(x) - m(x) = e_1^\top M_n^{-1}(x)U_n(x) + e_1^\top M_n^{-1}(x)B_n(x)$. The stochastic term $U_n(x)$ and the bias term $B_n(x)$ are $N \times 1$ vectors

$$U_n(x) = \begin{bmatrix} U_{n,0}(x) \\ \vdots \\ U_{n,p}(x) \end{bmatrix}, \; B_n(x) = \begin{bmatrix} B_{n,0}(x) \\ \vdots \\ B_{n,d}(x) \end{bmatrix},$$

where $U_{n,l}(x)$ and $B_{n,l}(x)$ are defined similarly as $\Psi_{n,l}(x)$ so that $U_{n,|j|}(x)$ and $B_{n,|j|}(x)$ are a $N_{|j|}$ dimensional subvectors whose r-th elements are given by:

$$\left[U_{n,|j|}\right]_r = \frac{1}{nh^d}\sum_{i=1}^n \left(\frac{x-X_i}{h}\right)^{\phi_{|j|}(r)} \mathcal{K}\left(\frac{x-X_i}{h}\right) \varepsilon_i$$

$$\left[B_{n,|j|}\right]_r = \frac{1}{nh^d}\sum_{i=1}^n \left(\frac{x-X_i}{h}\right)^{\phi_{|j|}(r)} \mathcal{K}\left(\frac{x-X_i}{h}\right) \Delta_i(x),$$

where $\Delta_i(x) = m(X_i) - \frac{1}{\mathbf{k}!}\sum_{0\le|\mathbf{k}|\le p}(D^{\mathbf{k}}m)(x)(X_i - x)^{\mathbf{k}}$.

Under assumptions A1-A6 and B1, we have the following results:

$$\sup_{x\in\mathcal{X}} |M_n(x) - f(x)M| = O_p(h + n^{-1/2}h^{-d/2}\log n)$$

$$\sup_{x\in\mathcal{X}} |\widetilde{m}(x) - m(x)| = O_p(h^{p+1} + n^{-1/2}h^{-d/2}\log n), \tag{18}$$

which follow from the results of Masry (1996). For notational convenience, now define $\bar{\varepsilon}_i = Y_i - \mathcal{P}(\theta, X_i - x) = \varepsilon_i + \delta_i$, where $\delta_i = m(X_i) - \mathcal{P}(\theta, X_i - x)$, and define $\breve{\varepsilon}_i = Y_i - \mathcal{P}(\widetilde{\theta}, X_i - x) = \varepsilon_i + \delta_i - \nu = \bar{\varepsilon}_i - \nu$, where $\nu = \nu(x) = \mathcal{P}(\widetilde{\theta}, X_i - x) - \mathcal{P}(\theta, X_i - x)$. Thus $\widetilde{f}(Y_i - \mathcal{P}(\widetilde{\theta}, X_i - x))$ can be written as $\widetilde{f}(\breve{\varepsilon}_i)$. To facilitate asymptotic analysis, we also define the kernel density and derivative estimator based on the unobserved errors:

$$\overline{f}(e_i) = \frac{1}{nh_0}\sum_{j\neq i} k\left(\frac{e_i - \varepsilon_j}{h_0}\right),\ \overline{f}'(e_i) = \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{e_i - \varepsilon_j}{h_0}\right).$$

LEMMA A. *Under our conditions*

$$\widetilde{f}(\breve{\varepsilon}_i) - f(\overline{\varepsilon}_i) = O_p(h_0^q + n^{-1/2}h_0^{-d/2} + h^q + n^{-1/2}h^{-d/2}) \text{ for each } i, \tag{19}$$

$$\max_{1\leq i\leq n}\left|\widetilde{f}(\breve{\varepsilon}_i) - f(\overline{\varepsilon}_i)\right| = O_p(h_0^q + n^{-1/2}h_0^{-d/2-1}\log(n) + \left(h^q + n^{-1/2}h^{-d/2}\log(n)\right)h_0^{-1}). \tag{20}$$

PROOF. First (19). Notice that

$$\begin{aligned}
\widetilde{f}(\breve{\varepsilon}_i) - f(\overline{\varepsilon}_i) &= \left[\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\right] + \left[\widetilde{f}(\breve{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i)\right] \qquad (21)\\
&= \left[\frac{1}{nh_0}\sum_{j\neq i} k\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) - f(\overline{\varepsilon}_i)\right] + \frac{1}{nh_0}\sum_{j\neq i}\left[k\left(\frac{\breve{\varepsilon}_i - \widetilde{\varepsilon}_j}{h_0}\right) - k\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right].
\end{aligned}$$

The first term, $\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)$, is just the conventional density estimator error and satisfies

$$\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i) = O_p(h_0^q + n^{-1/2}h_0^{-1/2}). \tag{22}$$

The second term, $\widetilde{f}(\breve{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i)$, contains the effect coming from the preliminary estimation ν; we show that it is of order $O_p(h_0^q + n^{-1/2}h_0^{-d/2})$. By a second order Taylor expansion, we have

$$\begin{aligned}
\widetilde{f}(\breve{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i) &= \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\nu - \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)[\widetilde{m}(X_j) - m(X_j)]\\
&\quad + \frac{1}{2nh_0^3}\sum_{j\neq i} k''\left(\frac{\varepsilon_i^* - \varepsilon_j^*}{h_0}\right)[\widetilde{m}(X_j) - m(X_j) - \nu]^2\\
&= A_I + A_{II} + A_{III},
\end{aligned}$$

where ε_i^* and ε_j^* are intermediate values.

The first term,

$$A_I = \nu\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) = O_p(h^q + n^{-1/2}h^{-d/2}),$$

since $\nu = O_p(h^q + n^{-1/2}h^{-d/2})$, $q = p + 1$ and

$$\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) = f'(\overline{\varepsilon}_i) + o_p(1)$$

by standard kernel theory.

Regarding A_{III}, notice that $\max_{1\le j\le n}|\widetilde{m}(X_j)-m(X_j)| = O_p(h_1^q + n^{-1/2}h_1^{-d/2}\log n)$, so that

$$
\begin{aligned}
& \frac{1}{nh_0^3}\sum_{j\neq i} k''\left(\frac{\varepsilon_i^*-\varepsilon_j^*}{h_0}\right)[\widetilde{m}(X_j)-m(X_j)-\nu(x)]^2 \\
\le & \left(\max_{1\le j\le n}|\widetilde{m}(X_j)-m(X_j)| + \sup_{x\in\mathcal{X}}|\nu(x)|\right)^2 \frac{1}{nh_0^3}\sum_{j\neq i}\left|k''\left(\frac{\varepsilon_i^*-\varepsilon_j^*}{h_0}\right)\right| \\
\le & \left(\max_{1\le j\le n}|\widetilde{m}(X_j)-m(X_j)| + \sup_{x\in\mathcal{X}}|\nu(x)|\right)^2 \frac{1}{h_0^3}\sup_u |k''(u)| \\
= & O_p(h_1^q + n^{-1/2}h_1^{-d/2}\log n + h^q + n^{-1/2}h^{-d/2}\log n)^2 h_0^{-3}
\end{aligned}
$$

Under the bandwidth condition A5, the remainder term is of order $O_p(h^q + n^{-1/2}h^{-d/2})$.

We now turn to the proof of the magnitude of A_{II}. This calculation is quite long. The general strategy is to expand out the random denominator of $\widetilde{m}(X_j)-m(X_j)$ around its probability limit and then calculate the moments of the resulting degenerate U-statistics term by term. Notice that $\widetilde{m}(X_j)-m(X_j) = e_1^\top M_n^{-1}(X_j)U_n(X_j) + e_1^\top M_n^{-1}(X_j)B_n(X_j)$, so we have

$$
\begin{aligned}
A_{II} &= \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top M_n^{-1}(X_j)U_n(X_j) + \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top M_n^{-1}(X_j)B_n(X_j) \\
&= A_{IIA} + A_{IIB}.
\end{aligned}
$$

We expand $M_n^{-1}(X_j)$ around its limit $(Mf(X_j))^{-1}$ and get

$$
\begin{aligned}
A_{IIA} &= \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top [Mf_X(X_j)]^{-1} U_n(X_j) \\
&+ \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) \sum_{\rho=1}^{p+1} e_1^\top \left\{[Mf_X(X_j)]^{-1}[M_n(X_j)-Mf_X(X_j)]\right\}^\rho [Mf_X(X_j)]^{-1} U_n(X_j) \\
&+ \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left\{[Mf_X(X_j)]^{-1}[M_n(X_j)-Mf_X(X_j)]\right\}^{p+2} M_n^{-1}(X_j)U_n(X_j).
\end{aligned}
\tag{23}
$$

Since $M^{0,l}$ are $1\times N_l$ row vectors, we have

$$
\begin{aligned}
& \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top [Mf_X(X_j)]^{-1} U_n(X_j) \\
= & \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) f_X(X_j)^{-1}\sum_\kappa \omega^{0,\kappa}\left(\frac{1}{nh_1^d}\sum_{l=1}^n \left(\frac{X_j-X_l}{h_1}\right)^\kappa \mathcal{K}\left(\frac{X_j-X_l}{h_1}\right)\varepsilon_l\right),
\end{aligned}
$$

where $\omega^{0,\kappa}$ are elements in the first row of M^{-1} and the sum over κ is over a finite index set. Thus

$$\begin{aligned}
&\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[M f_X(X_j)\right]^{-1} U_n(X_j) \qquad (24)\\
=\;& \sum_{j\neq i}\sum_{\kappa}\omega^{0,\kappa}\frac{1}{n^2h_0^2h_1^d} f_X(X_j)^{-1}k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\mathcal{K}\left(\frac{X_j-X_i}{h_1}\right)\left(\frac{X_j-X_i}{h_1}\right)^{\kappa}\varepsilon_i\\
&+\sum_{j\neq i}\sum_{l\neq j,l\neq i}\sum_{\kappa}\omega^{0,\kappa}\frac{1}{n^2h_0^2h_1^d} f_X(X_j)^{-1}k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\mathcal{K}\left(\frac{X_j-X_l}{h_1}\right)\left(\frac{X_j-X_l}{h_1}\right)^{\kappa}\varepsilon_l.
\end{aligned}$$

By the i.i.d. assumption, we have

$$\frac{1}{n}\sum_{j\neq i}\sum_{\kappa}\omega^{0,\kappa}\frac{1}{nh_0^2h_1^d} f_X(X_j)^{-1}k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\mathcal{K}\left(\frac{X_j-X_i}{h_1}\right)\left(\frac{X_j-X_i}{h_1}\right)^{\kappa}\varepsilon_i = O_p(n^{-1}),$$

since [using integration by parts, a change of variables and dominated convergence]:

$$E_i k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) = h_0^2\int k(u)f'(\overline{\varepsilon}_i-uh_0)du,$$

and

$$E_i\frac{1}{f_X(X_j)}\mathcal{K}\left(\frac{X_j-X_i}{h_1}\right)\left(\frac{X_j-X_i}{h_1}\right)^{\kappa} = h_1^d\int K(u)u^\kappa du.$$

For the second term in (24), if we denote

$$\phi_1(i,j,l) = \frac{1}{n^2h_0^2h_1^d} f_X(X_j)^{-1}k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\mathcal{K}\left(\frac{X_j-X_l}{h_1}\right)\left(\frac{X_j-X_l}{h_1}\right)^{\kappa}\varepsilon_l,$$

we only need to verify the magnitude of

$$\sum_{j\neq i}\sum_{l\neq j,l\neq i}\phi_1(i,j,l). \qquad (25)$$

Notice that (25) is mean zero with variance

$$O(n^2)E[\phi_1(i,j,l)^2+\phi_1(i,j,l)\phi_1(i,l,j)] + O(n^3)E[\phi_1(i,j,l)\phi_1(i,r,l)].$$

The orders of $E[\phi_1(i,j,l)^2]$, $E\phi_1(i,j,l)\phi_1(i,l,j)]$, $E[\phi_1(i,j,l)\phi_1(i,r,l)]$ can be verified. Under our assumptions, we have: $E_ik'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)^2 = O_p(h_0)$, $E_ik'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\varepsilon_j k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_l}{h_0}\right)\varepsilon_l = O_p(h_0^4)$, $Ek'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_r}{h_0}\right)\varepsilon_l^2 = O_p(h_0^4)$, $E_i\left[\frac{1}{f_X(X_j)}\mathcal{K}\left(\frac{X_j-X_i}{h_1}\right)\left(\frac{X_j-X_i}{h_1}\right)^{\kappa}\right]^2 = O_p(h_1^d)$, and

$$Ef_X(X_j)^{-1}f_X(X_l)^{-1}\mathcal{K}\left(\frac{X_j-X_l}{h_1}\right)\mathcal{K}\left(\frac{X_l-X_j}{h_1}\right)\left(\frac{X_j-X_l}{h_1}\right)^{\kappa}\left(\frac{X_l-X_j}{h_1}\right)^{\kappa} = O(h_1^d),$$

$$E_l\left[f_X(X_j)^{-1}\mathcal{K}\left(\frac{X_j-X_l}{h_1}\right)\left(\frac{X_j-X_l}{h_1}\right)^{\kappa}\right]E_l\left[f_X(X_r)^{-1}\mathcal{K}\left(\frac{X_r-X_l}{h_1}\right)\left(\frac{X_r-X_l}{h_1}\right)^{\kappa}\right] = O(h_1^{2d}).$$

Therefore, it can be verified that: $E[\phi_1(i,j,l)^2] = O(n^{-4}h_0^{-3}h_1^{-d})$, $E[\phi_1(i,j,l)\phi_1(i,l,j)] = O(n^{-4}h_1^{-d})$, $E[\phi_1(i,j,l)\phi_1(i,r,l)] = O(n^{-4})$. Thus,

$$\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[Mf_X(X_j)\right]^{-1} U_n(X_j) = O_p(n^{-1/2}).$$

Similarly, we can verify that the higher order terms in A_{IIA}

$$\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left\{\left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right]\right\}^\rho \left[Mf_X(X_j)\right]^{-1} U_n(X_j)$$

are of small order of magnitude.

For

$$\begin{aligned}
&\left|\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h}\right) e_1^\top \left\{\left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right]\right\}^{p+2} M_n^{-1}(X_j)U_n(X_j)\right| \\
\leq\ & \max_{1\leq j\leq n}\left|e_1^\top \left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right]^{p+2}\left[Mf_X(X_j)\right]^{-1}\right. \\
&\left.\times\left[M_n(X_j)-Mf_X(X_j)\right] M_n^{-1}(X_j)U_n(X_j)\right| \frac{1}{nh_0^2}\sum_{j\neq i}\left|k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right| \\
=\ & O_p\left(h_1^{p+2}\times h_1\times\frac{1}{nh_0^2}\times n\right) = O_p(h_1^{p+3}h_0^{-2}).
\end{aligned}$$

For the bias term A_{IIB},

$$\begin{aligned}
A_{IIB} =\ & \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[Mf_X(X_j)\right]^{-1} B_n(X_j) \\
&+\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right] M_n^{-1}(X_j)B_n(X_j).
\end{aligned}$$

Similarly, by a verification on the moments of U-statistics, we have

$$\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[Mf_X(X_j)\right]^{-1} B_n(X_j) = O_p(h_1^{p+1})$$

and

$$\begin{aligned}
&\left|\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right) e_1^\top \left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right] M_n^{-1}(X_j)B_n(X_j)\right| \\
\leq\ & \max_{1\leq j\leq n}\left|e_1^\top \left[Mf_X(X_j)\right]^{-1}\left[M_n(X_j)-Mf_X(X_j)\right] M_n^{-1}(X_j)B_n(X_j)\right| \frac{1}{nh_0^2}\sum_{j\neq i}\left|k'\left(\frac{\overline{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right| \\
=\ & O_p(h_1^{p+2}).
\end{aligned}$$

In conclusion, $A_{II} = o_p(n^{-1/2}h^{-d/2}+h^q)$.

Now to the proof of the uniform convergence result (20). We decompose $\widetilde{f}(\breve{\varepsilon}_i) - f(\overline{\varepsilon}_i)$ as in (21). Notice that the first term, $\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)$, which is just the conventional density estimator error, satisfies $\max_{1\leq i\leq n} |\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)| = O_p(h_0^q + n^{-1/2}h_0^{-1/2}\log(n))$. The second term in (21) can be further decomposed as

$$\widetilde{f}(\breve{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i) = \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\nu(x) - \frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)[\widetilde{m}(X_j) - m(X_j)] \quad (26)$$
$$+\frac{1}{nh_0^3}\sum_{j\neq i} k''\left(\frac{\varepsilon_i^* - \varepsilon_j^*}{h_0}\right)[\widetilde{m}(X_j) - m(X_j) - \nu(x)]^2,$$

where ε_i^* and ε_j^* are intermediate values. Notice that $\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)$ is the conventional kernel estimator of $f'(\overline{\varepsilon}_i)$, we have denoted as $\overline{f}'(\overline{\varepsilon}_i)$, and write the first term in (26) as $\nu(x)\overline{f}'(\overline{\varepsilon}_i) = \nu(x)f'(\overline{\varepsilon}_i) + \nu(x)[\overline{f}'(\overline{\varepsilon}_i) - f'(\overline{\varepsilon}_i)]$. By Assumptions A2 that $f'(\cdot)$ is uniformly bounded and $\sup_{x\in\mathcal{X}}\nu(x) = O_p(h^q + n^{-1/2}h^{-d/2}\log(n))$, and $\max_{1\leq i\leq n}|\overline{f}'(\overline{\varepsilon}_i) - f'(\overline{\varepsilon}_i)| = O_p(h_0^q + n^{-1/2}h_0^{-3/2}\log(n)) = o_p(1)$, we have

$$\max_{1\leq i\leq n}\left|\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\nu(x)\right| = O_p(h_0^q + n^{-1/2}h_0^{-d/2}\log(n)).$$

For the second term in (26),

$$\left|\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)[\widetilde{m}(X_j) - m(X_j)]\right|$$
$$\leq \max_{1\leq j\leq n}|\widetilde{m}(X_j) - m(X_j)|\frac{1}{nh_0^2}\sum_{j\neq i}\left|k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right|.$$

By i.i.d. property and conditional on ε_i and X_i,

$$\frac{1}{nh_0^2}\sum_{j\neq i}\left|k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right| \simeq E_i\frac{1}{h_0^2}\left|k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right|,$$

where

$$E_i\frac{1}{h_0^2}\left|k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right| = \frac{1}{h_0}\int |k'(u)|\, f(\overline{\varepsilon}_i - h_0 u)du.$$

Again, notice that $f(\overline{\varepsilon}_i - h_0 u)$ is uniformly bounded, thus $\max_{1\leq i\leq n}|\int |k'(u)|\, f(\overline{\varepsilon}_i - h_0 u)du| = O_p(1)$. In addition, notice that $\max_{1\leq j\leq n}|\widetilde{m}(X_j) - m(X_j)| = O_p(h_1^q + n^{-1/2}h_1^{-d/2}\log(n))$, so that

$$\max_{1\leq i\leq n}\left|\frac{1}{nh_0^2}\sum_{j\neq i} k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)[\widetilde{m}(X_j) - m(X_j)]\right| = O_p(h_1^q + n^{-1/2}h_1^{-d/2}\log(n))h_0^{-1}.$$

For the third term in (26)

$$\begin{aligned}
&\frac{1}{nh_0^3}\sum_{j\neq i} k''\left(\frac{\varepsilon_i^*-\varepsilon_j^*}{h_0}\right)[\widetilde{m}(X_j)-m(X_j)-(\widetilde{m}(x)-m(x))]^2 \\
\leq\ & \left(\max_{1\leq j\leq n}|\widetilde{m}(X_j)-m(X_j)|+\sup_{x\in\mathcal{X}}|\nu(x)|\right)^2\frac{1}{h_0^3}\sup_u|k''(u)| \\
=\ & o_p(h^{q-1}+n^{-1/2}h^{-d/2-1}\log(n)).
\end{aligned}$$

Thus,

$$\max_{1\leq i\leq n}\left|\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)\right|\leq\max_{1\leq i\leq n}\left|\widetilde{f}(\breve{\varepsilon}_i)-\overline{f}(\bar{\varepsilon}_i)\right|+\max_{1\leq i\leq n}\left|\overline{f}(\bar{\varepsilon}_i)-f(\bar{\varepsilon}_i)\right|=O_p(h^{q-1}+n^{-1/2}h^{-d/2-1}\log(n))$$

as required. ■

Remark A1. The uniform rate given by (20) is not the best result, but suffices our purpose of proofs in this paper. In fact, following a similar analysis as Masry (1996), a better rate - $O_p(h^q+n^{-1/2}h^{-d/2}\log(n))$ - could be obtained with substantially more complicated analysis.

Remark A2. The above results can be extended to estimates of derivatives based on similar analysis. The above results also hold for any $\breve{\varepsilon}_i$ in a small neighborhood of $\bar{\varepsilon}_i$, say, $\breve{\varepsilon}_i-\bar{\varepsilon}_i=O_p(n^{-1/2}h^{-d/2})$.

6.2 Proof of Theorem 1

For notational simplicity, we write $\widetilde{\mathcal{I}}_n(\theta;\widetilde{f})$ as $\widetilde{\mathcal{I}}_n(\theta)$, and $\widetilde{S}_n(\theta;\widetilde{f})$ as $\widetilde{S}_n(\theta)$. The estimator $\widehat{\theta}(x)$ can be written, after standardization by $\sqrt{nh^d}$, as

$$\sqrt{nh^d}H(\widehat{\theta}(x)-\theta(x))=\sqrt{nh^d}H(\widetilde{\theta}(x)-\theta(x))+\sqrt{nh^d}H\widetilde{\mathcal{I}}_n(\widetilde{\theta})^{-1}\widetilde{S}_n(\widetilde{\theta}). \tag{27}$$

Let

$$\widetilde{s}_i(\widetilde{\theta})=\frac{\widetilde{f}'}{\widetilde{f}}(Y_i-\mathcal{P}(\widetilde{\theta},X_i-x)),\ \text{then}\ \widetilde{S}_n(\widetilde{\theta})=-\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\widetilde{s}_i(\widetilde{\theta})\widetilde{\mathbf{X}}_i.$$

We expand $\widetilde{s}_i(\widetilde{\theta})$ around the true value of θ and obtain

$$\widetilde{s}_i(\widetilde{\theta})=\widetilde{s}_i(\theta)+\frac{\partial\widetilde{s}_i(\theta)}{\partial\theta}(\widetilde{\theta}-\theta)+\frac{1}{2}(\widetilde{\theta}-\theta)^\top\frac{\partial^2\widetilde{s}_i(\theta^*)}{\partial\theta\partial\theta^\top}(\widetilde{\theta}-\theta), \tag{28}$$

where θ^* is an intermediate value between θ and $\widetilde{\theta}$. Based on (28), we obtain the complete expansion of $\widetilde{S}_n(\widetilde{\theta})$. The preliminary estimator $\widetilde{\theta}(x)$ is consistent and indeed satisfies $\widetilde{\theta}(x)-\theta(x)=O_p(n^{-1/2}h^{-d/2}+h^q)=O_p(n^{-1/2}h^{-d/2})$. Under the given assumptions, we have

$$\widetilde{S}_n(\widetilde{\theta})=\widetilde{S}_n(\theta)-\widetilde{\mathcal{I}}_n(\theta_0)(\widetilde{\theta}-\theta)+R(\theta)(\widetilde{\theta}-\theta)+o_p(||R(\theta)(\widetilde{\theta}-\theta)||), \tag{29}$$

where

$$R(\theta) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\frac{\widetilde{f}''(Y_i-\mathcal{P}(\theta,X_i-x))}{\widetilde{f}(Y_i-\mathcal{P}(\theta,X_i-x))}G_b(\widetilde{f}_i)\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top$$
$$+\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\frac{f'}{f}(Y_i-\mathcal{P}(\theta,X_i-x))\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top g_b(\widetilde{f}_i)\widetilde{f}'(Y_i-\mathcal{P}(\theta,X_i-x)).$$

Notice that $[H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta})H^{-1}]^{-1} = O_p(1)$ and substituting (29) into (27) we may express $\sqrt{nh^d}H(\widehat{\theta}(x)-\theta(x))$ as

$$\sqrt{nh^d}\left[H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x))H^{-1}\right]^{-1}H^{-1}\widetilde{S}_n(\theta(x)) +$$
$$\left[\sqrt{nh^d}H(\widetilde{\theta}(x)-\theta(x)) - \left[H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x))H^{-1}\right]^{-1}\left[H^{-1}\widetilde{\mathcal{I}}_n(\theta(x))H^{-1}\right]\sqrt{nh^d}H(\widetilde{\theta}(x)-\theta(x))\right] + o_p(1),$$

since $\widetilde{\theta}(x)$ is a $\sqrt{nh^d}H$-consistent estimator of $\theta(x)$ under our assumptions. Furthermore, it can be shown that $H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x))H^{-1} - H^{-1}\widetilde{\mathcal{I}}_n(\theta_0(x))H^{-1} \xrightarrow{p} 0$ and $H^{-1}\widetilde{\mathcal{I}}_n(\theta_0(x))H^{-1} \xrightarrow{p} \mathcal{I}_\theta(x)M$. Therefore,

$$\sqrt{nh^d}H(\widetilde{\theta}(x)-\theta(x)) - \left[H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x))H^{-1}\right]^{-1}\left[H^{-1}\widetilde{\mathcal{I}}_n(\theta(x))H^{-1}\right]\sqrt{nh^d}H(\widetilde{\theta}(x)-\theta(x)) \xrightarrow{p} 0, \quad (30)$$

and in fact

$$\sqrt{nh^d}H(\widehat{\theta}(x)-\theta_0(x)) = [\mathcal{I}_\theta(x)M]^{-1}\sqrt{nh^d}H^{-1}\widetilde{S}_n(\theta_0(x)) + o_p(1).$$

Finally, we show that

$$\sqrt{nh^d}\left[H^{-1}\widetilde{S}_n(\theta_0(x)) - \beta(x)\right] \Rightarrow N(0,\mathcal{I}_\theta(x)\Gamma). \quad (31)$$

In the following three subsections, we derive the asymptotic results for the Hessian $\widetilde{\mathcal{I}}_n(\widetilde{\theta})$, the score $\widetilde{S}_n(\theta_0)$, and the remainder term $R(\theta_0)$.

6.2.1 The Hessian

We want to show that

$$H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta}(x))H^{-1} \xrightarrow{p} f_X(x)I(f)\mathbf{M}.$$

We write $\widetilde{f}(Y_i-\mathcal{P}(\widetilde{\theta},X_i-x))$ and $\widetilde{f}'(Y_i-\mathcal{P}(\widetilde{\theta},X_i-x))$ as $\widetilde{f}(\breve{\varepsilon}_i)$ and $\widetilde{f}'(\breve{\varepsilon}_i)$ respectively. Thus

$$\widetilde{\mathcal{I}}_n(\widetilde{\theta}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\left[\frac{\widetilde{f}'(\breve{\varepsilon}_i)}{\widetilde{f}(\breve{\varepsilon}_i)}\right]^2 G_iH^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1}.$$

We decompose $\widetilde{f}(\breve{\varepsilon}_i)$ as $f(\bar{\varepsilon}_i)+[\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)]$, and $\widetilde{f}'(\breve{\varepsilon}_i)$ as $f'(\bar{\varepsilon}_i)+[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)]$. By a geometric expansion, $\widetilde{f}(\breve{\varepsilon}_i)^{-2}$ can be written as

$$f(\bar{\varepsilon}_i)^{-2} - f(\bar{\varepsilon}_i)^{-4}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)) + f(\bar{\varepsilon}_i)^{-4}\widetilde{f}(\breve{\varepsilon}_i)^{-2}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))^2(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))^2.$$

Thus we have the following decomposition of $\widetilde{\mathcal{I}}_n(\widetilde{\theta})$:

$$H^{-1}\widetilde{\mathcal{I}}_n(\widetilde{\theta})H^{-1} = \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3 - \mathcal{J}_4 + \mathcal{J}_5 + \mathcal{J}_6 - \mathcal{J}_7 - \mathcal{J}_8 + \mathcal{J}_9 + \mathcal{J}_{10},$$

where:

$$\begin{aligned}
\mathcal{J}_1 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left[\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\right]^2 H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_2 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left[\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\right]^2 [1-G_b(\widetilde{f}_i)]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_3 &= \frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^2}G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_4 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)^2}{f(\bar{\varepsilon}_i)^4}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1}
\end{aligned}$$

$$\begin{aligned}
\mathcal{J}_5 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]^2}{f(\bar{\varepsilon}_i)^2}G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_6 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)^2}{f(\bar{\varepsilon}_i)^4\widetilde{f}(\breve{\varepsilon}_i)^2}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))^2(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))^2G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_7 &= \frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^4}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_8 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]^2}{f(\bar{\varepsilon}_i)^4}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_9 &= \frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^4\widetilde{f}(\breve{\varepsilon}_i)^2}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))^2(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))^2G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
\mathcal{J}_{10} &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{\left[\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]^2}{f(\bar{\varepsilon}_i)^4\widetilde{f}(\breve{\varepsilon}_i)^2}(\widetilde{f}(\breve{\varepsilon}_i)+f(\bar{\varepsilon}_i))^2(\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i))^2G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1}.
\end{aligned}$$

We show that

$$\mathcal{J}_1 \xrightarrow{p} f_X(x)I(f)\mathbf{M},$$

and

$$\mathcal{J}_j = o_p(1),\ j = 2,\ldots,10.$$

This is carried out in a series of lemmas given below.

LEMMA H1. *Under our conditions,* $\mathcal{J}_1 \xrightarrow{p} f_X(x)I(f)\mathbf{M}$.

PROOF. $\mathcal{J}_1$ contains the errors coming from the local deviation from X_i to x in terms of $\delta_i = \bar{\varepsilon}_i - \varepsilon_i$, which is largely determined by the smoothness property of $m(\cdot)$. For simplicity of exposition, we denote f'/f as ψ. Thus,

$$\begin{aligned}\mathcal{J}_1 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)^2\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top + \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left[\psi(\bar{\varepsilon}_i)^2 - \psi(\varepsilon_i)^2\right]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\ &= \mathcal{J}_{11} + \mathcal{J}_{12}.\end{aligned}$$

Since $\{\varepsilon_i\}$ and $\{X_i\}$ are i.i.d. and are mutually independent, for each d-tuples j with $0 \le |j| \le 2p$

$$E\left[K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)^2\left(\frac{x-X_i}{h}\right)^j\right] \simeq h^d f_X(x)I(f)\mu_j(K),$$

by a change of variables and dominated convergence. By a law of large numbers for independent random variables we have

$$\mathcal{J}_{11} = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)^2\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top \xrightarrow{p} f_X(x)I(f)\mathbf{M}.$$

Now we examine $\mathcal{J}_{12}$. Define $\rho(\delta) = \int\left[\psi(\varepsilon+\delta)^2 - \psi(\varepsilon)^2\right]f(\varepsilon)d\varepsilon$ for any δ. This quantity is finite, differentiable, and satisfies $\rho(0) = 0$. By independence of ε_i and X_i we have

$$\mathcal{J}_{12} \simeq E\left[\frac{1}{h^d}K\left(\frac{x-X_i}{h}\right)\rho(\delta_i)H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1}\right].$$

By dominated convergence, each element in this expectation is zero because as $h \to 0$, $\max_{i:|X_i-x|\le h}|m(X_i) - m(x)| = \max_{i:|X_i-x|\le h}|\delta_i| \to 0$ by the differentiability of m at x. ■

LEMMA H2. *Under our conditions,* $\mathcal{J}_2 = o_p(1)$.

PROOF. Under our conditions $G_b(\widetilde{f}_i) = G_b(f_i) + o_p(1)$. In particular, notice that under bandwidth assumption A5,

$$\max_{1\le i\le n}\left|\widetilde{f}_i - f_i\right| = o_p(b), \tag{32}$$

and thus $\max_{1\le i\le n}|G_b(\widetilde{f}_i) - G_b(f_i)| = o_p(1)$. Let $G_i = G_b(f_i)$. Then,

$$\begin{aligned}&\left|\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^j\left[\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\right]^2\left[G_i - G_b(\widetilde{f}_i)\right]\right| \\ \le\;& \frac{1}{nh^d}\sum_{i=1}^n\left|K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^j\right|\psi(\bar{\varepsilon}_i)^2\cdot\max_{1\le i\le n}\left|G_b(\widetilde{f}_i) - G_b(f_i)\right| = o_p(1).\end{aligned}$$

Therefore, we can ignore the estimation errors in trimming effect. Furthermore, making an expansion of $\psi(\bar{\varepsilon}_i)$ we obtain

$$\begin{aligned}\mathcal{J}_2 &= \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)^2[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\ &+\frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)\psi'(\varepsilon_i)\delta_i[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i'H^{-1} \\ &+\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi'(\varepsilon_i)^2\delta_i^2[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1}+o_p(1).\end{aligned}$$

We verify the orders of magnitude for these terms. It can be shown that the leading trimming term is determined by

$$\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)^2[1-G_b(f(\varepsilon_i))]H^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1},$$

with elements

$$\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^j\psi(\varepsilon_i)^2[1-G_b(f(\varepsilon_i))].$$

Notice that

$$E\left[\psi(\varepsilon_i)^2\{1-G_b(f(\varepsilon_i))\}\right]=\int_{0<f(\varepsilon)<b}\psi(\varepsilon_i)^2 f(\varepsilon)d\varepsilon+\int_{b<f(\varepsilon)<2b}\psi(\varepsilon_i)^2\left(\int_{f(\varepsilon)}^\infty g_b(z)dz\right)f(\varepsilon)d\varepsilon,$$

under Assumption A3 (or A3′), we can show that

$$\int_{0<f(\varepsilon)<b}\psi(\varepsilon_i)^2 f(\varepsilon)d\varepsilon+\int_{b<f(\varepsilon)<2b}\psi(\varepsilon_i)^2\left(\int_{f(\varepsilon)}^\infty g_b(z)dz\right)f(\varepsilon)d\varepsilon=o(1).$$

The precise order of magnitude of the trimming effect will depend on the tail behavior of f. In addition, $E[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^j]=h^d f_X(x)\mu_j(K)$. By i.i.d. assumption, we have

$$\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^j\psi(\varepsilon_i)^2[1-G_b(f(\varepsilon_i))]=o_p(1).$$

Other terms can be analyzed similarly. ■

For $j=3,4,\ldots,10$, terms $\mathcal{J}_j$ are functions of $\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$ and/or $\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)$. To facilitate asymptotic analysis, we decompose $\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$ into the sum of $\widetilde{f}'(\breve{\varepsilon}_i)-\overline{f}'(\bar{\varepsilon}_i)$ and $\overline{f}'(\bar{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$ (and $\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)$ into the sum of $\widetilde{f}(\breve{\varepsilon}_i)-\overline{f}(\bar{\varepsilon}_i)$ and $\overline{f}(\bar{\varepsilon}_i)-f(\bar{\varepsilon}_i)$). The first term, $\widetilde{f}'(\breve{\varepsilon}_i)-\overline{f}'(\bar{\varepsilon}_i)$, contains errors coming from preliminary estimation $\widetilde{m}(X_i)-m(X_i)$ and $\widetilde{m}(x)-m(x)$, and the second term, $\overline{f}'(\bar{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$, contains the errors coming from nonparametric kernel smoothing.

Lemma H3. *Under our conditions,* $\mathcal{J}_3=o_p(1)$.

PROOF. We decompose $\widetilde{f}'(\breve{\varepsilon}_i) - f'(\bar{\varepsilon}_i)$ and obtain:

$$\begin{aligned}
\mathcal{J}_3 &= \frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\widetilde{f}'(\breve{\varepsilon}_i)-\overline{f}'(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^2}G_iH^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
&\quad +\frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)\left[\overline{f}'(\bar{\varepsilon}_i)-f'(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^2}G_iH^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
&= \mathcal{J}_{31}+\mathcal{J}_{32}.
\end{aligned}$$

$$\begin{aligned}
\mathcal{J}_{31} &\simeq \nu(x)\frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\frac{\frac{1}{nh_0^3}\sum_{j\neq i}k''\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)}{f(\bar{\varepsilon}_i)}G_iH^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
&\quad +\frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\frac{\frac{1}{nh_0^3}\sum_{j\neq i}k''\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)[\widetilde{m}(X_j)-m(X_j)]}{f(\bar{\varepsilon}_i)}G_iH^{-1}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top H^{-1} \\
&= \mathcal{J}_{31A}+\mathcal{J}_{31B}.
\end{aligned}$$

Since $|\nu(x)| = O_p(h^{p+1}+n^{-1/2}h^{-d/2})$, to show that $\mathcal{J}_{31A} = o_p(1)$, we show that, for each d-tuples τ with $0 \le |\tau| \le 2p$,

$$\frac{2}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^\tau\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\frac{\frac{1}{nh_0^3}\sum_{j\neq i}k''\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)}{f(\bar{\varepsilon}_i)}G_i = o_p\left(1/(h^{p+1}+n^{-1/2}h^{-d/2})\right).$$

Let

$$\varphi_{An}(z_i,z_j) = \frac{2}{n^2h^dh_0^3}K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^\tau\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}\frac{G_i}{f(\bar{\varepsilon}_i)}k''\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right).$$

the above result can be verified that the first two moments of $\sum_{i=1}^n\sum_{j\neq i}\varphi_{An}(z_i,z_j)$ are $o\left(1/(h^q+\frac{1}{\sqrt{nh_n^d}})\right)$. For $\mathcal{J}_{31B}$, since $\widetilde{m}(X_j)-m(X_j) = e_1^\top M_n^{-1}(X_j)U_n(X_j)+e_1^\top M_n^{-1}(X_j)B_n(X_j)$, we expand $M_n^{-1}(X_j)$ around its limit $(Mf(X_j))^{-1}$ and get $M_n^{-1}(X_j) = (Mf(X_j))^{-1}-(Mf(X_j))^{-1}\times$ $[M_n(X_j)-Mf_X(X_j)]M_n^{-1}(X_j)$, then we show that each of the terms in the expansion is $o_p(1)$.

In order to show $\mathcal{J}_{32} = o_p(1)$, we decompose $\overline{f}'(\bar{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$ into a bias effect $B'(\bar{\varepsilon}_i)$ and a variance effect $V'(\bar{\varepsilon}_i)$, and, again, analyze second order U-statistics with degeneracy.

Similar analysis can be applied to other terms that involve $\widetilde{f}'(\breve{\varepsilon}_i)-f'(\bar{\varepsilon}_i)$ and $\widetilde{f}(\breve{\varepsilon}_i)-f(\bar{\varepsilon}_i)$.

The analysis for $\mathcal{J}_4$, $\mathcal{J}_5$, $\mathcal{J}_7$ and $\mathcal{J}_8$ is similar. The terms $\mathcal{J}_6$, $\mathcal{J}_9$, and $\mathcal{J}_{10}$ contain $\widetilde{f}(\breve{\varepsilon}_i)$ in the denominator. Under our assumptions, for each d-tuples τ with $0 \leq |\tau| \leq 2p$,

$$\begin{aligned}
& \left| \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \left(\frac{x - X_i}{h}\right)^{\tau} \frac{f'(\bar{\varepsilon}_i)^2}{f(\bar{\varepsilon}_i)^4 \widetilde{f}(\breve{\varepsilon}_i)^2} (\widetilde{f}(\breve{\varepsilon}_i) + f(\bar{\varepsilon}_i))^2 (\widetilde{f}(\breve{\varepsilon}_i) - f(\bar{\varepsilon}_i))^2 G_b(\widetilde{f}_i) \right| \\
\leq & \max_{1 \leq i \leq n} \left| (\widetilde{f}(\breve{\varepsilon}_i) - f(\bar{\varepsilon}_i))^2 \right| \left| \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \left(\frac{x - X_i}{h}\right)^{\tau} \frac{f'(\bar{\varepsilon}_i)^2}{f(\bar{\varepsilon}_i)^4 \widetilde{f}(\breve{\varepsilon}_i)^2} (\widetilde{f}(\breve{\varepsilon}_i) + f(\bar{\varepsilon}_i))^2 G_b(\widetilde{f}_i) \right| \\
= & O_p(h^{2q-2} + n^{-1}h^{-d-2})b^{-2},
\end{aligned}$$

thus $|\mathcal{J}_6| = o_p(1)$. We have used Lemma A here. Similar analysis can be conducted for $\mathcal{J}_9$ and $\mathcal{J}_{10}$. $\blacksquare$

6.2.2 The Score Function

We want to show that

$$\sqrt{nh^d}\left[H^{-1}\widetilde{S}_n(\theta_0) - \beta(x)\right] \Rightarrow N(0, f_X(x)I(f)\Gamma).$$

By definition

$$\sqrt{nh^d} H^{-1} \widetilde{S}_n(\theta) = -\frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{\widetilde{f}'}{\widetilde{f}} (Y_i - \mathcal{P}(\theta, X_i - x)) G_b(\widetilde{f}_i) H^{-1} \widetilde{\mathbf{X}}_i.$$

For the denominator, $\widetilde{f}^{-1}(\bar{\varepsilon}_i) = f^{-1}(\bar{\varepsilon}_i) - \{\widetilde{f}(\bar{\varepsilon}_i) - f(\bar{\varepsilon}_i)\}/f(\bar{\varepsilon}_i)^2 +$ R_2, where $R_2 = \{\widetilde{f}(\bar{\varepsilon}_i) - f(\bar{\varepsilon}_i)\}^2 / f^2(\bar{\varepsilon}_i)\widetilde{f}(\bar{\varepsilon}_i)$. Therefore,

$$\begin{aligned}
\sqrt{nh^d} H^{-1} \widetilde{S}_n(\theta) = & -\frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)} H^{-1} \widetilde{\mathbf{X}}_i \\
& + \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)} [1 - G_b(\widetilde{f}_i)] H^{-1} \widetilde{\mathbf{X}}_i \\
& - \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{\widetilde{f}'(\bar{\varepsilon}_i) - f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)} G_b(\widetilde{f}_i) H^{-1} \widetilde{\mathbf{X}}_i \\
& + \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{\widetilde{f}'(\bar{\varepsilon}_i) \left[\widetilde{f}(\bar{\varepsilon}_i) - f(\bar{\varepsilon}_i)\right]}{f(\bar{\varepsilon}_i)^2} G_b(\widetilde{f}_i) H^{-1} \widetilde{\mathbf{X}}_i \\
& + \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) G_b(\widetilde{f}_i) \widetilde{f}'(\bar{\varepsilon}_i) H^{-1} \widetilde{\mathbf{X}}_i R_2 \\
= & \; T_1 + T_2 - T_3 + T_4 + T_5.
\end{aligned}$$

We analyze the asymptotic behavior of T_j, $j = 1, 2, \ldots, 5$.

LEMMA S1. *Under our conditions*

$$T_1 - \sqrt{nh^d}\beta_1(x) \Rightarrow N(0, \mathcal{I}_\theta(x)\Gamma).$$

where $\beta_1(x) = h^{p+1}\mathcal{I}_\theta(x)Bm^{(p+1)}(x)$.

PROOF. Notice that $\bar{\varepsilon}_i = \varepsilon_i + m(X_i) - \mathcal{P}(\theta, X_i - x) = \varepsilon_i + \delta_i$ and

$$\begin{aligned}
T_1 &= -\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\varepsilon_i)}{f(\varepsilon_i)}H^{-1}\widetilde{\mathbf{X}}_i \\
&\quad -\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\varepsilon_i+\delta_i)-f'(\varepsilon_i)}{f(\varepsilon_i)}H^{-1}\widetilde{\mathbf{X}}_i \\
&\quad +\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\varepsilon_i)[f(\varepsilon_i+\delta_i)-f(\varepsilon_i)]}{f(\varepsilon_i)f(\varepsilon_i+\delta_i)}H^{-1}\widetilde{\mathbf{X}}_i \\
&\quad -\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{[f'(\varepsilon_i+\delta_i)-f'(\varepsilon_i)]\,[f(\varepsilon_i+\delta_i)-f(\varepsilon_i)]}{f(\varepsilon_i)f(\varepsilon_i+\delta_i)}H^{-1}\widetilde{\mathbf{X}}_i \\
&= T_{11}+T_{12}+T_{13}+T_{14}.
\end{aligned}$$

We verify each of these terms. First, by a central limit theorem for i.i.d. random vectors,

$$T_{11} = -\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\varepsilon_i)}{f(\varepsilon_i)}H^{-1}\widetilde{\mathbf{X}}_i \Rightarrow N(0, f_X(x)I(f)\Gamma).$$

For the second term,

$$T_{12} \simeq -\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f''(\varepsilon_i)\delta_i}{f(\varepsilon_i)}H^{-1}\widetilde{\mathbf{X}}_i$$

For each d-tuples τ with $0 \le |\tau| \le 2p$,

$$\begin{aligned}
&\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f''(\varepsilon_i)\delta_i}{f(\varepsilon_i)}\left(\frac{x-X_i}{h}\right)^\tau \\
\approx\ & h^{p+1}\sqrt{nh^d}f_X(x)m^{(p+1)}(x)\int f''(\varepsilon)d\varepsilon\int K(u)\,u^{\tau+q}du
\end{aligned}$$

Therefore,

$$T_{12} \simeq -h^{p+1}\sqrt{nh^d}f_X(x)m^{(p+1)}(x)BE\left[\frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\right]$$

Similarly, by calculation of moments, it is easy to verify that under our conditions,

$$T_{13} \simeq h^{p+1}\sqrt{nh^d}f_X(x)m^{(p+1)}(x)BE\left[\frac{f'(\varepsilon_i)}{f(\varepsilon_i)}\right]^2,$$

and T_{14} are $o_p(h^{p+1}\sqrt{nh^d})$. ■

LEMMA S2. *Under our conditions,* $T_2 \xrightarrow{p} 0$.

PROOF. First, we expand the trimming function to the second order,

$$G_b(\widetilde{f}_i) - G_b(f_i) = g_b(f_i)(\widetilde{f}_i - f_i) + \frac{1}{2}g_b'(f_i^*)(\widetilde{f}_i - f_i)^2,$$

where f_i^* is an intermediate point between $\widetilde{f}_i$ and f_i. Then

$$\begin{aligned}\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}[G_i - G_b(\widetilde{f}_i)] &= \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)(\widetilde{f}_i - f_i) + \\ & \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\frac{1}{2}g_b'(f_i^*)(\widetilde{f}_i - f_i)^2.\end{aligned}$$

We use a crude bound on the last term and show that it is $o_p(1)$. Therefore, we must analyze the term

$$\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)(\widetilde{f}_i - f_i)H^{-1}\widetilde{\mathbf{X}}_i. \tag{33}$$

This again involves some further U-statistic calculation. Similar to the previous analysis, we decompose (33) into $T_{21}+T_{22}$, where

$$\begin{aligned}T_{21} &= \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)(\widetilde{f}(\breve{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i))H^{-1}\widetilde{\mathbf{X}}_i, \\ T_{22} &= \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)(\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i))H^{-1}\widetilde{\mathbf{X}}_i.\end{aligned}$$

Further,

$$\begin{aligned}T_{21} &= [\nu(x)]\,\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)\frac{1}{nh_0^2}\sum_{j\neq i}k'\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)H^{-1}\widetilde{\mathbf{X}}_i \\ &\quad + \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)\frac{1}{nh_0^2}\sum_{j\neq i}k'\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)[\widetilde{m}(X_j)-m(X_j)]\,H^{-1}\widetilde{\mathbf{X}}_i + o_p(1) \\ &\simeq T_{21A}+T_{21B}.\end{aligned}$$

For the leading term T_{21A}, since $\nu(x) = O_p(h^q + n^{-1/2}h^{-d/2})$, we need to show that for each $0\le|\tau|\le 2p$,

$$\sum_{i=1}^n \frac{1}{\sqrt{nh^d}}K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)\left[\frac{1}{nh_0^2}\sum_{j\neq i}k'\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right] = o_p\left(1/(h^q+n^{-1/2}h^{-d/2})\right).$$

Notice that

$$\begin{aligned}&\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)\left[\frac{1}{nh_0^2}\sum_{j\neq i}k'\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right] \\ \simeq\;& \frac{n}{\sqrt{nh^d}}E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}g_b(f_i)f'(\overline{\varepsilon}_i)\right].\end{aligned}$$

We need to show that

$$E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)f'(\bar{\varepsilon}_i)\right]=o\left(h^d\right).$$

Conditional on X_i,

$$\begin{aligned}
& E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)f'(\bar{\varepsilon}_i)|X_i\right] \\
= & K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}\frac{1}{b}g\left(\frac{f(v)}{b}-1\right)f(v-\delta_i)dv
\end{aligned}$$

Notice that $g(\cdot)\ge 0$ is bounded, say, $g(\cdot)\le C$, then

$$\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}\frac{1}{b}g\left(\frac{f(v)}{b}-1\right)f(v-\delta_i)dv\le\frac{C}{b}\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}f(v-\delta_i)dv.$$

Thus

$$\begin{aligned}
& E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)f'(\bar{\varepsilon}_i)\right] \\
\le & \frac{C}{b}\times h^d\int\int_{b\le f(v)\le 2b}K(U)U^{\tau}\frac{f'(v)^2}{f(v)}f(v+m(x-hU)-\mathcal{P}(\theta,hU))f_X(x-hU)dvdU.
\end{aligned}$$

By Assumption A3, we have

$$\int_{b\le f(v)\le 2b}f'(v)^2dv=\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}f(v)dv\le\sup_{b\le f(v)\le 2b}f(v)\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}dv=o(b)$$

by dominated convergence, since $\int_{b\le f(v)\le 2b}\frac{f'(v)^2}{f(v)}dv\le I(f)<\infty$. Therefore,

$$E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)f'(\bar{\varepsilon}_i)\right]\le\frac{C}{b}\times h^d\times o(b)=o(h^d).$$

Consequently, under our bandwidth conditions,

$$\sum_{i=1}^{n}\frac{1}{\sqrt{nh^d}}K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\left[\frac{1}{nh_0^2}\sum_{j\ne i}k'\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right]=o_p\left(1/(h^q+n^{-1/2}h^{-d/2})\right).$$

Thus $T_{21A}=o_p(1)$.For T_{21B}, following similar analysis as those in the proof of Lemma H3, we can verify that each of these terms is $o_p(1)$.

The second component of errors comes from standard kernel estimation and can be analyzed similarly. We decompose $\overline{f}(\bar{\varepsilon}_i)-f(\bar{\varepsilon}_i)$ into a bias effect $\mathcal{B}(\bar{\varepsilon}_i)$ and a variance effect $\mathcal{V}(\bar{\varepsilon}_i)$ thus

$$\begin{aligned}
T_{22} &= \frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\mathcal{V}(\bar{\varepsilon}_i)H^{-1}\widetilde{\mathbf{X}}_i \\
&\quad+\frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\mathcal{B}(\bar{\varepsilon}_i)H^{-1}\widetilde{\mathbf{X}}_i \\
&= T_{22A}+T_{22B}.
\end{aligned}$$

The variance term is

$$T_{22A} = \sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{nh_0\sqrt{nh^d}}K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\left\{\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)-E_i\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right]\right]\right\}H^{-1}\widetilde{\mathbf{X}}_i.$$

Denote, for each $0\leq|\tau|\leq 2p$,

$$\gamma_{3n}(z_i,z_j)=\frac{1}{nh_0\sqrt{nh^d}}K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\left\{\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)-E_i\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right]\right]\right\},$$

again, this is a second order U-statistic with degeneracy and

$$\begin{aligned} & E\left[\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{nh_0\sqrt{nh^d}}K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\left\{\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)-E_i\left[k\left(\frac{\bar{\varepsilon}_i-\varepsilon_j}{h_0}\right)\right]\right]\right\}\right]^2 \\ = & \; O(n^2)\left[E\gamma_{3n}(z_i,z_j)^2+E\left[\gamma_{3n}(z_i,z_j)\gamma_{3n}(z_j,z_i)\right]\right]+O(n^3)E\left[\gamma_{3n}(z_i,z_j)\gamma_{3n}(z_l,z_j)\right].\end{aligned}$$

We show $T_{22A}=o_p(1)$ by verifying the orders of magnitude of $E[\gamma_{3n}(z_i,z_j)^2]$, $E[\gamma_{3n}(z_i,z_j)\gamma_{3n}(z_j,z_i)]$, and $E[\gamma_{3n}(z_i,z_j)\gamma_{3n}(z_l,z_j)]$.

The leading bias term is

$$T_{22B}=\frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}g_b(f_i)\mathcal{B}(\bar{\varepsilon}_i)H^{-1}\widetilde{\mathbf{X}}_i=o_p(\sqrt{nh^d}h_0^q)=o_p(1).$$

Thus $T_{22}=o_p(1)$. Consequently the term (33) $=o_p(1)$.

Finally we turn to the leading term of T_2,

$$\frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\frac{f'(\bar{\varepsilon}_i)}{f(\bar{\varepsilon}_i)}[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i,$$

again, we denote f'/f as ψ and under given assumptions we obtain a Taylor expansion that $\psi(\bar{\varepsilon}_i)\simeq\psi(\varepsilon_i)+\psi'(\varepsilon_i)\delta_i$. Thus, the above term can be approximated by

$$\frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\psi(\varepsilon_i)[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i \tag{34}$$

$$+\frac{1}{\sqrt{nh^d}}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\psi'(\varepsilon_i)\delta_i[1-G_i]H^{-1}\widetilde{\mathbf{X}}_i. \tag{35}$$

To verify the order of magnitude of (34), we calculate the first two moments. Notice that under our assumptions, $\psi(\varepsilon_i)[1-G_i]$ are i.i.d. and

$$E[\psi(\varepsilon_i)[1-G_i]]=\int_{f(\varepsilon_i)\leq b}f'(\varepsilon_i)d\varepsilon_i+\int_{b\leq f(\varepsilon_i)\leq 2b}f'(\varepsilon_i)[1-\int_{-\infty}^{f(\varepsilon_i)}g_b(z)dz]d\varepsilon_i.$$

For b small enough, there exists $-\infty < \underline{a}_1 < \underline{a}_2 < \overline{a}_2 < \overline{a}_1 < \infty$ such that:

$$\{\varepsilon_i : f(\varepsilon_i) \leq b\} = \{-\infty < \varepsilon_i \leq \underline{a}_1\} \cup \{\overline{a}_1 \leq \varepsilon_i \leq \infty\},$$

$$\{\varepsilon_i : b \leq f(\varepsilon_i) \leq 2b\} = \{\underline{a}_1 < \varepsilon_i \leq \underline{a}_2\} \cup \{\overline{a}_2 \leq \varepsilon_i \leq \overline{a}_1\},$$

where $f(\underline{a}_1) = f(\overline{a}_1) = b$ and $f(\underline{a}_2) = f(\overline{a}_2) = 2b$. Therefore,

$$\begin{aligned} E[\psi(\varepsilon_i)[1-G_i]] &= \int_{-\infty}^{\underline{a}_1} f'(\varepsilon_i)d\varepsilon_i + \int_{\overline{a}_1}^{\infty} f'(\varepsilon_i)d\varepsilon_i \\ &\quad + \int_{\underline{a}_1}^{\underline{a}_2} f'(\varepsilon_i)[1 - \int_{-\infty}^{f(\varepsilon_i)} g_b(z)dz]d\varepsilon_i + \int_{\overline{a}_2}^{\overline{a}_1} f'(\varepsilon_i)[1 - \int_{-\infty}^{f(\varepsilon_i)} g_b(z)dz]d\varepsilon_i. \end{aligned}$$

Notice that $\int_{-\infty}^{\underline{a}_1} f'(\varepsilon_i)d\varepsilon_i + \int_{\overline{a}_1}^{\infty} f'(\varepsilon_i)d\varepsilon_i = [f(\underline{a}_1) - 0] + [0 - f(\overline{a}_1)] = 0$ and $\int_{\underline{a}_1}^{\underline{a}_2} f'(\varepsilon_i)[1 - \int_{-\infty}^{f(\varepsilon_i)} g_b(z)dz]d\varepsilon_i + \int_{\overline{a}_2}^{\overline{a}_1} f'(\varepsilon_i)[1 - \int_{-\infty}^{f(\varepsilon_i)} g_b(z)dz]d\varepsilon_i = 0$, thus $\psi(\varepsilon_i)[1-G_i]$ are mean zero. Thus we just need to verify the second moment of (34). Notice that:

$$E\left[K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^{\tau}\right]^2 = O(h^d), \tag{36}$$

$$E[\psi(\varepsilon_i)^2[1-G_i]] = \int \left[\frac{f'}{f}(\varepsilon_i)\right]^2 f(\varepsilon_i)[1-G_i]d\varepsilon_i = o(1) \tag{37}$$

under assumption A3 (or A3′). Thus, combining (36) and (37), we can show that the order of (34) is $o_p(1)$. If we assume that Assumption A3′ holds, it can be shown that

$$E[\psi(\varepsilon_i)^2(1-G_i)] = O(b^{(s-1)/s}). \tag{38}$$

and the order of magnitude of (34) is $O_p(b^{(s-1)/(2s)}) = o_p(1)$.

Now we look at the term (35), notice that $E[K\left(\frac{x-X_i}{h}\right)\delta_i] \simeq h^{d+q}\phi(K, m, f_X)$, where $\phi(K, m, f_X)$ is a function of the kernel and derivatives of m and f_X evaluated at x. And $E\left[\psi'(\varepsilon_i)(1-G_i)\right] = o(1)$. Consequently, by calculation of moments we have, under bandwidth assumption A5, (35) $= o_p(1)$. ■

LEMMA S3. *Under our conditions*

$$T_3 = \sqrt{nh^d}\left(\beta_2 + o_p(h_0^q)\right), \text{ where}$$
$$\beta_2 = h_0^q \mu_q(K) E\left[\frac{f^{(q+1)}(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\right] B.$$

PROOF. The analysis of T_3 is similar to that of $\mathcal{J}_3$.

$$\begin{aligned} T_3 &= \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{\left[\widetilde{f}'(\overline{\varepsilon}_i) - \overline{f}'(\overline{\varepsilon}_i)\right]}{f(\overline{\varepsilon}_i)} G_i H^{-1}\widetilde{\mathbf{X}}_i \\ &\quad + \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{\left[\overline{f}'(\overline{\varepsilon}_i) - f'(\overline{\varepsilon}_i)\right]}{f(\overline{\varepsilon}_i)} G_i H^{-1}\widetilde{\mathbf{X}}_i \\ &= T_{31} + T_{32} \end{aligned}$$

and

$$
\begin{aligned}
T_{31} &\simeq \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{1}{f(\overline{\varepsilon}_i)} \frac{1}{nh_0^3} \sum_{j\neq i} k''\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) \left[e_1^\top M_n^{-1}(X_j) U_n(X_j)\right] G_i H^{-1} \widetilde{\mathbf{X}}_i \\
&+ \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{1}{f(\overline{\varepsilon}_i)} \frac{1}{nh_0^3} \sum_{j\neq i} k''\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) \left[e_1^\top M_n^{-1}(X_j) B_n(X_j)\right] G_i H^{-1} \widetilde{\mathbf{X}}_i,
\end{aligned}
$$

analysis of these terms will then be the similar to the previous analysis. In particular, we can show that the variance term is $o_p(1)$ and, for the bias term, for each d-tuples τ with $0 \le |\tau| \le 2p$,

$$
\begin{aligned}
& \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \left(\frac{x-X_i}{h}\right)^\tau \frac{1}{f(\overline{\varepsilon}_i)} \frac{1}{nh_0^3} \sum_{j\neq i} k''\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) \left[e_1^\top M_n^{-1}(X_j) B_n(X_j)\right] \\
\approx\ & \sqrt{nh^d} h_1^{p+1} E\left[M^{-1} B \frac{m^{(p+1)}(X_j)}{f_X(X_j)}\right]_0 E\left[\frac{f''(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\right] \mu_{q+\tau}(K).
\end{aligned}
$$

thus

$$
\begin{aligned}
& \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{1}{f(\overline{\varepsilon}_i)} \frac{1}{nh_0^3} \sum_{j\neq i} k''\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) \left[e_1^\top M_n^{-1}(X_j) B_n(X_j)\right] G_i H^{-1} \widetilde{\mathbf{X}}_i \\
\approx\ & \sqrt{nh^d} h_1^{p+1} E\left[M^{-1} B \frac{m^{(p+1)}(X_j)}{f_X(X_j)}\right]_0 E\left[\frac{f''(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\right] B.
\end{aligned}
$$

$$
\begin{aligned}
T_{32} &= \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{\frac{1}{nh_0^2} \sum_{j\neq i} \left[k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right) - E_i\left[k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right]\right]}{f(\overline{\varepsilon}_i)} G_i H^{-1} \widetilde{\mathbf{X}}_i \\
&+ \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{\frac{1}{nh_0^2} \sum_{j\neq i} E_i\left[k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right] - f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)} G_i H^{-1} \widetilde{\mathbf{X}}_i \\
&= T_{32A} + T_{32B}.
\end{aligned}
$$

For T_{32A}, it is again a second order U-statistic with first order degeneracy and can be analyzed in a similar way to the proof of T_{22A}.

For T_{32B}, notice that, conditional on $X_1, \ldots, X_n$,

$$
E_i\left[\frac{1}{nh_0^2} \sum_{j\neq i} E_i\left[k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right] - f'(\overline{\varepsilon}_i)\right] = h_0^q f^{(q+1)}(\overline{\varepsilon}_i) \int k(u) u^q du
$$

it is easy to show that

$$
\begin{aligned}
& \frac{1}{\sqrt{nh^d}} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \frac{\frac{1}{nh_0^2} \sum_{j\neq i} E_i\left[k'\left(\frac{\overline{\varepsilon}_i - \varepsilon_j}{h_0}\right)\right] - f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)} G_i H^{-1} \widetilde{\mathbf{X}}_i \\
\approx\ & \sqrt{nh^d} h_0^q \mu_q(K) E\left[\frac{f^{(q+1)}(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\right] B.
\end{aligned}
$$

$\blacksquare$

LEMMA S4. *Under our conditions*

$$T_4 = \sqrt{nh^d}\left(\beta_3 + o_p(h_1^{p+1}) + o_p(h_0^{p+1})\right), \text{ where}$$

$$\beta_3 = h_1^{p+1} AE\left[\frac{f'(\varepsilon)f''(\varepsilon)}{f(\varepsilon)^2}\right] + h_0^{p+1}\mu_q(K)E\left[\frac{f'(\varepsilon)f^{(q)}(\varepsilon)}{f(\varepsilon)^2}\right]B.$$

PROOF. The proof is similar to that of Lemma S3.

$$\begin{aligned}
& \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{\widetilde{f}'(\overline{\varepsilon}_i)\left[\widetilde{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\right]}{f(\overline{\varepsilon}_i)^2} G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i \\
= & \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{\widetilde{f}'(\overline{\varepsilon}_i)\left[\widetilde{f}(\overline{\varepsilon}_i) - \overline{f}(\overline{\varepsilon}_i)\right]}{f(\overline{\varepsilon}_i)^2} G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i \\
& + \frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{\widetilde{f}'(\overline{\varepsilon}_i)\left[\overline{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\right]}{f(\overline{\varepsilon}_i)^2} G_b(\widetilde{f}_i)H^{-1}\widetilde{\mathbf{X}}_i
\end{aligned}$$

the leading bias terms are

$$\sqrt{nh^d}h_1^{p+1}E\left[M^{-1}B\frac{m^{(p+1)}(X)}{f_X(X)}\right]_0 E\left[\frac{f'(\varepsilon)f''(\varepsilon)}{f(\varepsilon)^2}\right]B$$

and

$$\sqrt{nh^d}h_0^q\mu_q(K)E\left[\frac{f'(\varepsilon)f^{(q)}(\varepsilon)}{f(\varepsilon)^2}\right]B.$$

and the variance terms are $o_p(1)$. $\blacksquare$

LEMMA S5. *Under our conditions,* $T_5 \xrightarrow{p} 0$.

PROOF. For the remainder terms T_5, we need to show that

$$\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{f'(Y_i-\theta_0)\{\widetilde{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\}^2}{f^2(\overline{\varepsilon}_i)\widetilde{f}(\overline{\varepsilon}_i)} G_i H^{-1}\widetilde{\mathbf{X}}_i$$

goes to zero as $n \to \infty$. Notice that for each $0 \le |\tau| \le 2p$,

$$\begin{aligned}
& \left|\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^\tau \frac{f'(Y_i-\theta_0)\{\widetilde{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\}^2}{f^2(\overline{\varepsilon}_i)\widetilde{f}(\overline{\varepsilon}_i)} G_i\right| \\
\le & \max_{1\le i\le n}\left|\{\widetilde{f}(\overline{\varepsilon}_i) - f(\overline{\varepsilon}_i)\}^2\right| \times \left|\frac{1}{\sqrt{nh^d}}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^\tau \frac{f'(\overline{\varepsilon}_i)}{f(\overline{\varepsilon}_i)}\frac{G_i}{f(\overline{\varepsilon}_i)\widetilde{f}(\overline{\varepsilon}_i)}\right|.
\end{aligned}$$

Since $|\widetilde{f}_i| > b$, and using the result (20), the remainder term is of order $O_p(n^{1/2}h^{d/2})O_p(b^{-2})O_p(h_0^{2q-2} + n^{-1}h_0^{-d-2}\log^2(n))$, which is $o_p(1)$ under our bandwidth conditions.

6.2.3 Remainder term

It suffices to show that

$$R(\theta_0) = o_p(1)$$

and

$$\sup_{|\theta-\theta_0|\leq cn^{-1/2}h^{-d/2}} \left\| \frac{\partial^2 \widetilde{S}_n}{\partial\theta\partial\theta^\top}(\theta) \right\| = o_p(\sqrt{nh^d}) \tag{39}$$

because $nh^d(\widetilde{\theta}-\theta_0)^2 = O_p(1)$. The expression for $R_n(\theta^*)$ is quite complicated but its analysis is similar to that of the Hessian except that: (a) we require only bounds in probability that are quite weak; (b) those bounds, however, have to be locally uniform in the argument θ. By an analysis similar to that of the Hessian, it can be shown that $\partial^2\widetilde{S}_n(\theta)/\partial\theta\partial\theta^\top = O_p(1)$ for any given θ. The extension to local uniform over $\{\theta : ||\theta-\theta_0|| \leq n^{-1/2}h^{-d/2}\}$ follows from the smoothness properties on the kernel that we have imposed. We just examine a single key term

$$R_{n1}(\theta) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\left(\frac{x-X_i}{h}\right)^\tau \frac{\widetilde{f}'''}{\widetilde{f}}(\overline{\varepsilon}_i)G_b(\widetilde{f}_i).$$

By construction $G_b(\widetilde{f}_i)/\widetilde{f}_i \leq b^{-1}$, while $\sup_{t\in\mathbb{R}}|\widetilde{f}'''(t)| \leq h_0^{-4}\sup_u |k'''(u)|$. Therefore, $|R^*_{n1}(\theta)| = O_p(h_0^{-4}b^{-1})$. Therefore, (39) is satisfied because $h_0^{-4}b^{-1}/\sqrt{nh^d}$ is assumed to go to zero under our condition. By expanding out we can obtain better results but with considerably more calculations.

For $R(\theta_0)$, by an analysis similar to that of the score and also use the result of Lemma A, we can show that it is $o_p(1)$. ■

REFERENCES

Aerts, M. and G. Claeskens, 1997, local polynomial estimation in multiparameters likelihood models, Journal of the American Statistical Association, 92, 1536-1545.

Ahn, H. (1995). Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty, Journal of Econometrics 67, 337-378.

Ai, C., (1997). A Semiparametric Maximum Likelihood Estimator. Econometrica, 65, 933-963.

Andrews, D.W.K., (1995). Nonparametric Kernel Estimation for Semiparametric Models. Econometric Theory, 11,560-596.

Beran, R., (1974). Asymptotically efficient adaptive rank estimates in location models. Annals of Statistics 2, 248-266.

Bickel, P.J., (1975), One-step Huber estimates in the linear model, Journal of the American Statistical Association 70, 428-434.

Bickel, P. J., (1982). On Adaptive Estimation. Annals of Statistics, 10, 647-671.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993). Efficient and Adaptive Estimation for Semiparametric Models. Springer: Berlin.

Carroll, R., D. Ruppert,and A. Welsh, 1998, Local estimating equations, Journal of the American Statistical Association, 93, 214-227.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. Annals of Statistics 19, 760-777.

Copas, J.B. (1994). Local likelihood based on kernel censoring. Journal of the Royal Statistical Society, Series B, 57, 221-235.

Drost, F., and C.A.J. Klaassen (1997). Efficient estimation in semiparametric GARCH models. Journal of Econometrics 81, 193-221.

Drost, F., C.A.J. Klaassen, and B.J.M. Werker (1997). Adaptive Estimation in Time Series Models. Annals of Statistics 25, 786-817.

Fan, J. (1992). Design-Adaptive Nonparametric Regression. Journal of the American Statistical Association, 87 , 998-1004.

Fan, J. (1993). Local Linear Regression Smoothers and their Minimax Efficiencies. The Annals of Statistics, 21, 196-216.

Fan, J. N. Heckman, and M. Wand, 1995, Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, Journal of the American Statistical Association, 90, 141-150.

Fan, J., and J. Chen (1999). One-step Local Quasi-Likelihood Estimation. Journal of the Royal Statistical Society, Series B. 61, 927-943.

Fan, J., M. Farmen, and I. Gijbels (2000). Local Maximum Likelihood Estimation and Inference. Forthcoming in Journal of the American Statistical Association.

Fan, J. and I. Gijbels (1996). Local Polynomial Modelling and Its Applications, Chapman & Hall.

Fan, J., W. Härdle, and E. Mammen (1998). Direct estimation of low-dimensional components in additive models. The Annals of Statistics 26, 943-971.

Gonzalez-Manteiga, W. (1990). Asymptotic normality of Generalized Functional Estimators dependent on Covariates. Journal of Statistical Planning and Inference. 24, 377-390.

Gozalo, P. and O. Linton. (2000) Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression. Journal of Econometrics 99, 63-106.

Härdle, W., and O. B. Linton (1994), Applied Nonparametric Methods, The Handbook of Econometrics, Vol. IV, 2295-2339, Eds. D. McFadden and R. Engle, North Holland.

Hastie T. and Tibshirani, R. (1987). Local likelihood Estimation. Journal of the American Statistical Association, 82, 559-567.

Hengartner, N.W., M.H. Wegkamp, and E. Matzner-Lober (2002). Bandwidth selection for local linear regression smoothers. Journal of the Royal Statistical Society, Series B 64, 791-804.

Hjort, N.L. (1993). Dynamic likelihood hazard estimation. Biometrika. To appear.

Hjort, N.L., and M.C. Jones (1996). Locally parametric nonparametric density estimation. Annals of Statistics 24, 1619-1647.

Hodgson, D., (1998). Adaptive Estimation of Cointegrating Regressions with ARMA Errors. Journal of Econometrics 85, 231-268.

Hunsberger, 1994, Semiparametric estimation in likelihood based models, Journal of the American Statistical Association, Vol 89, 1354-1365.

Jeganathan, P. (1995). Some Aspects of Asymptotic Theory with Applications to Time Series Models. Econometric Theory, 11, 818-887.

Kreiss, J., (1987). On Adaptive Estimation in Stationary ARMA process. Annals of Statistics, 15, 112-133.

Linton, O.B. (1993). Adaptive estimation in ARCH Models. Econometric Theory 9, 539-569.

Linton, O.B. (1997). Efficient estimation of additive nonparametric regression models. Biometrika, 84, 469-474.

Linton, O.B., and Z. Xiao (2001). Adaptive Nonparametric Regression, working paper.

Loader, C.R. (1996). Local likelihood density estimation. Annals of Statistics 24, 1602-1618.

Manski, C.F., (1984). Adaptive Estimation of Nonlinear Regression Models. Econometric Reviews, 3, 187-208.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. Journal of Time Series Analysis 17, 571-599.

Robinson, P.M. (1989). Time varying nonlinear regression. in Statistical Analysis and Forecasting of Economic Structural Change. Eds. P. Hackl and A. Westland. Springer-Verlag.

Staniswalis, J.G. (1989). The Kernel Estimate of a Regression Function in Likelihood Based Models. Journal of the American Statistical Association, 84, 276-283.

Steigerwald, D., (1992). Adaptive Estimation in Time Series Regression Models. Journal of Econometrics, 54, 251-276.

Stein, C. (1956). Efficient nonparametric testing and estimation. in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, ed. by J. Neyman. University of California Press.

Stone, C., (1975). Adaptive maximum likelihood estimation of a location parameter. Annals of Statistics, 3, 267-284.

Tibshirani, R. (1984). Local Likelihood estimation. PhD Thesis, Stanford University.

Wand, M.P., and M.C. Jones (1995). Kernel Smoothing. Chapman and Hall: London.

7 Tables and Figures

TABLE 1: Third order local polynomial; Error Term = $t(2)$, $n = 100$, $m = x^2$

	Local Polynomial			Adaptive Local Polynomial		
Bandwidth	Bias	Variance	MSE	Bias	Variance	MSE
1.06	-0.1663	40.1462	40.1738	-0.1098	18.0255	18.0376
2	-0.1235	19.6882	19.7034	-0.0614	8.7485	8.7523
3	-0.0889	8.2348	8.2426	0.0394	3.6289	3.6304
4	-0.0605	4.9709	4.9745	0.1521	2.1800	2.2031
5	-0.0472	3.0709	3.0731	0.2427	1.3392	1.3981
6	-0.0402	2.1461	2.1477	0.3033	0.9357	1.0277
7	-0.0356	1.7086	1.7098	0.3433	0.7467	0.8645
8	-0.0352	1.4075	1.4288	0.3687	0.6257	0.7616
9	-0.0358	1.2068	1.2081	0.3782	0.5253	0.6683
9.3	-0.0378	1.2050	1.2064	0.4000	0.5259	0.6859

TABLE 2: Third order local polynomial; Error Term = $t(2)$, $n = 100$, $m(x) = x^4$

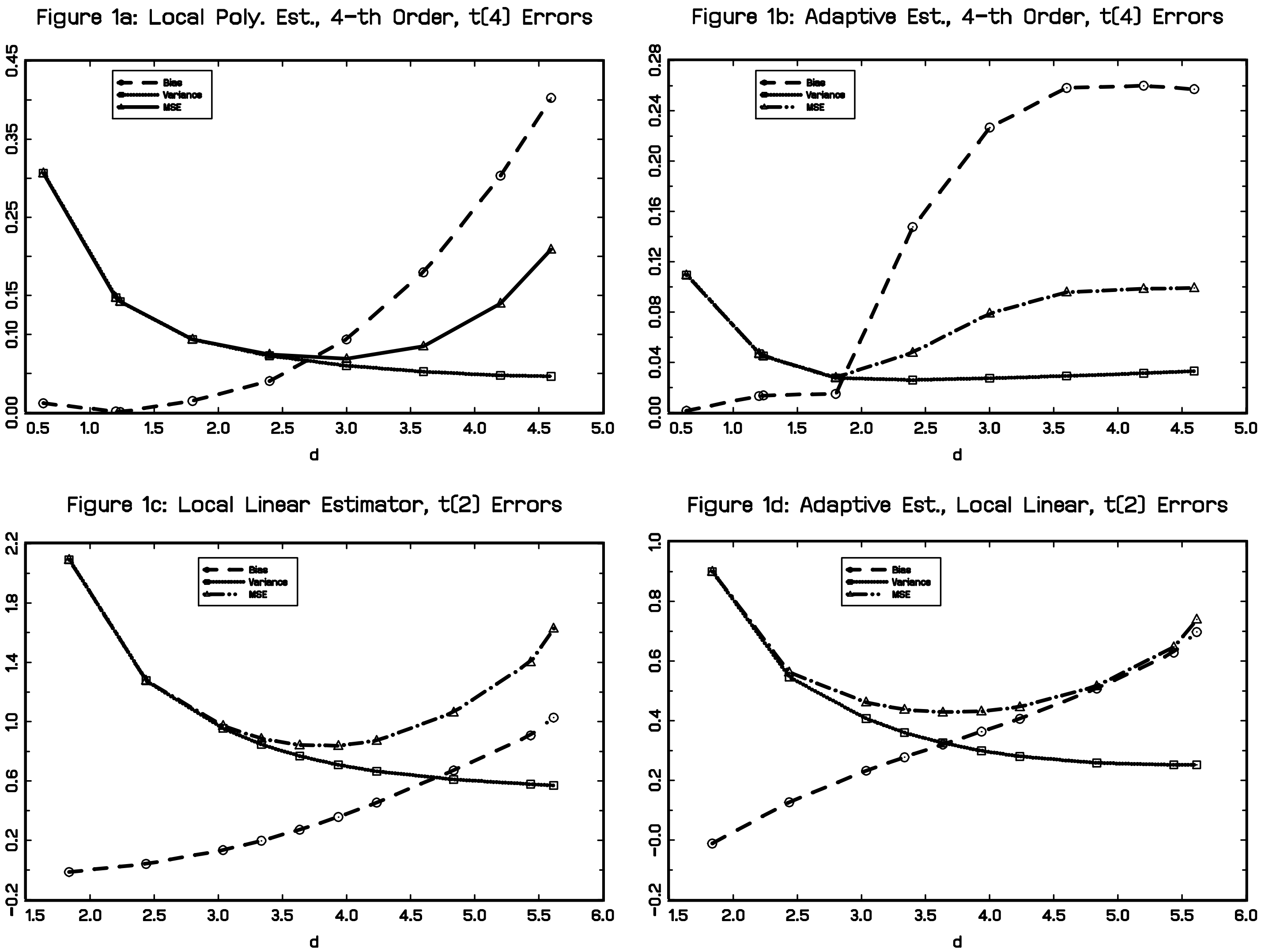
	Local Polynomial			Adaptive Local Polynomial		
Bandwidth	Bias	Variance	MSE	Bias	Variance	MSE
4	-0.1032	5.1264	5.1371	0.0689	2.2501	2.2548
4.66	-0.1250	3.7735	3.7892	0.1264	1.6534	1.6694
5.66	-0.1901	2.4019	2.4381	0.1693	1.0503	1.0789
6.66	-0.2953	1.8322	1.9195	0.1690	0.8069	0.8354
7.66	-0.4404	1.5479	1.7419	0.1307	0.6922	0.7093
8.06	-0.5055	1.4340	1.6895	0.1065	0.6469	0.6582
8.86	-0.6507	1.2684	1.6918	0.0397	0.5838	0.5853
9.06	-0.6896	1.2429	1.7184	0.0190	0.5750	0.5753
9.16	-0.7093	1.2322	1.7353	0.0817	0.5715	0.5782
9.26	-0.7292	1.2224	1.7542	-0.1308	0.5695	0.5866
9.36	-0.7493	1.2136	1.7750	-0.1458	0.5686	0.5899

TABLE 3: Local linear; Error Term = $t(2)$, $n = 100$, $m = x^4$

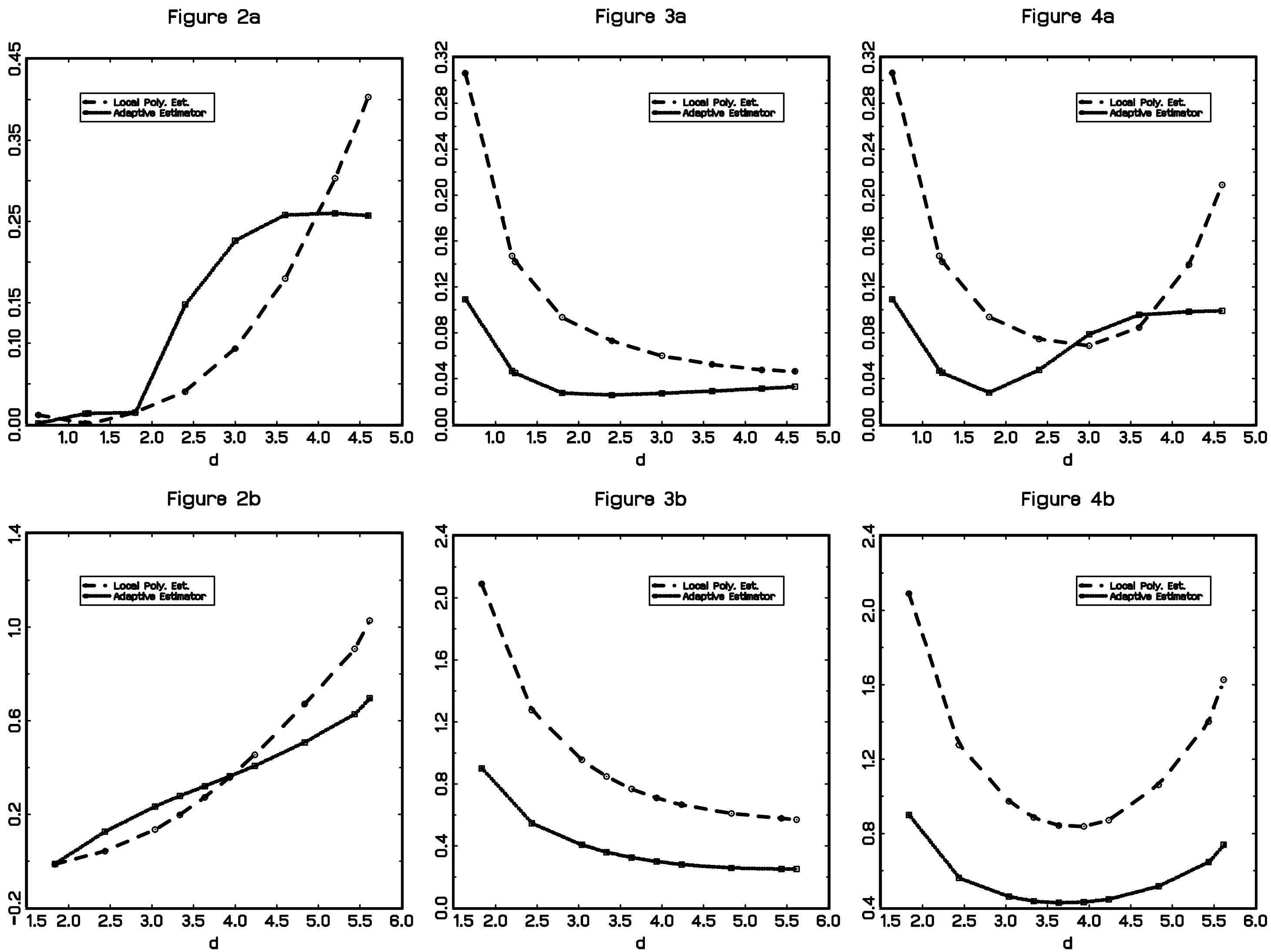
	Local Linear			Adaptive Local Linear		
Bandwidth	Bias	Variance	MSE	Bias	Variance	MSE
3.06	-0.0130	2.0904	2.0906	-0.0116	0.9000	0.9000
4.06	0.0427	1.2752	1.2770	0.1262	0.5463	0.5622
5.06	0.1345	0.9553	0.9733	0.2329	0.4078	0.4620
5.56	0.1976	0.8470	0.8861	0.2780	0.3600	0.4373
6.06	0.2723	0.7690	0.8431	0.3207	0.3255	0.4283
6.56	0.3579	0.7096	0.8377	0.3629	0.2993	0.4310
7.06	0.4540	0.6662	0.8723	0.4068	0.2805	0.4460
8.06	0.6721	0.6107	1.0624	0.5075	0.2588	0.5163
9.06	0.9078	0.5781	1.4022	0.6284	0.2522	0.6471
9.36	1.0281	0.5700	1.6270	0.6970	0.2524	0.7395

TABLE 4: Third order local polynomial; Error Term = $t(4)$, $n = 100$, $m(x) = x^4$

	Local Polynomial			Adaptive Local Polynomial		
Bandwidth	Bias	Variance	MSE	Bias	Variance	MSE
1.06	0.0120	0.3063	0.3065	-0.0014	0.1092	0.1092
2	0.0017	0.1470	0.1470	-0.0132	0.0468	0.0470
2.06	0.0093	0.1420	0.1420	-0.0136	0.0449	0.0451
3	-0.0149	0.0935	0.0937	0.0149	0.0276	0.0278
4	-0.0405	0.0729	0.0745	0.1477	0.0258	0.0476
5	-0.0936	0.0599	0.0686	0.2265	0.0273	0.0786
6	-0.1798	0.0524	0.08472	0.2581	0.0291	0.0957
7	-0.3030	0.0476	0.1394	0.2591	0.0313	0.0985
7.66	-0.4029	0.0465	0.2088	0.2572	0.0330	0.0992



Figures 1a-1d. Shows Bias, Variance and MSE of local polynomial and the adaptive local polynomial estimators for t(4) and t(2) errors as a function of bandwidth constant d.



Figures 2ab, 3ab, and 4ab. Compares local polynomial and adaptive local polynomial estimator in terms of (2) bias, (3) variance, and (4) mse for different bandwidth constants d. In Figures (a) the error distribution is t(4), while in Figures (b) the error distribution is t(2).

Figure 5: Normal QQ Plot

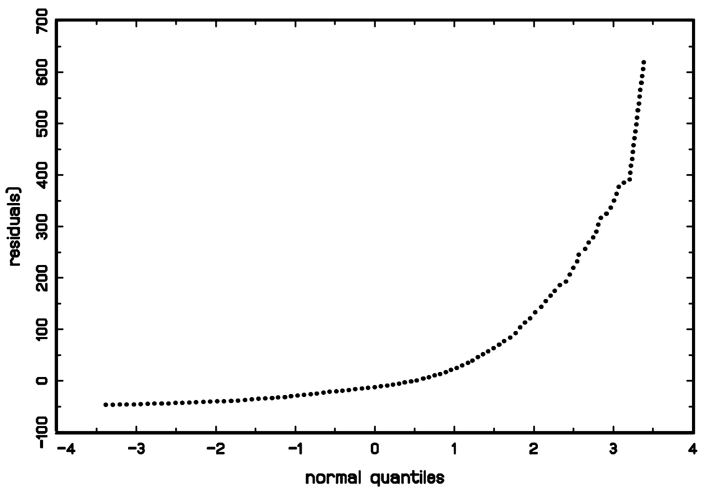


Figure 6a: Time of Day Effect, h1

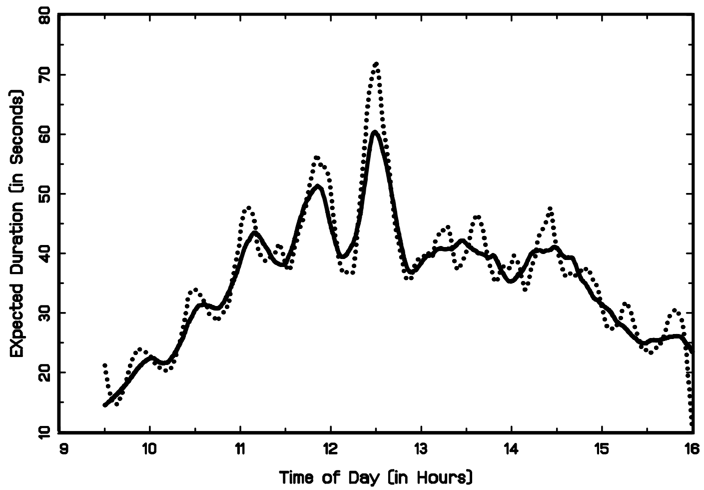


Figure 6b: Time of Day Effect, h2

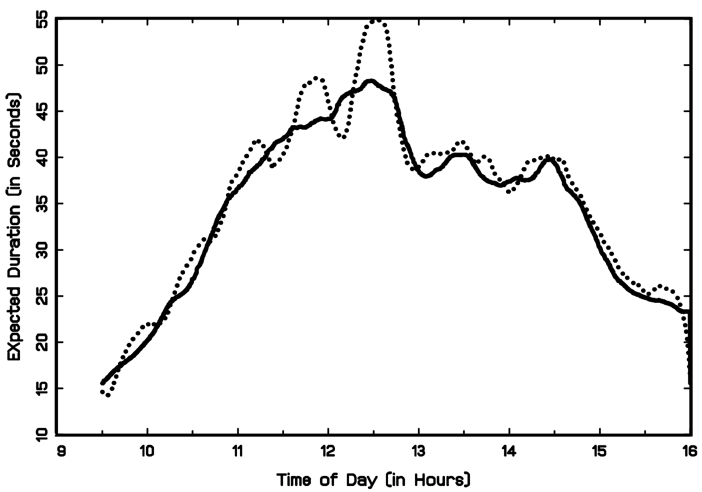
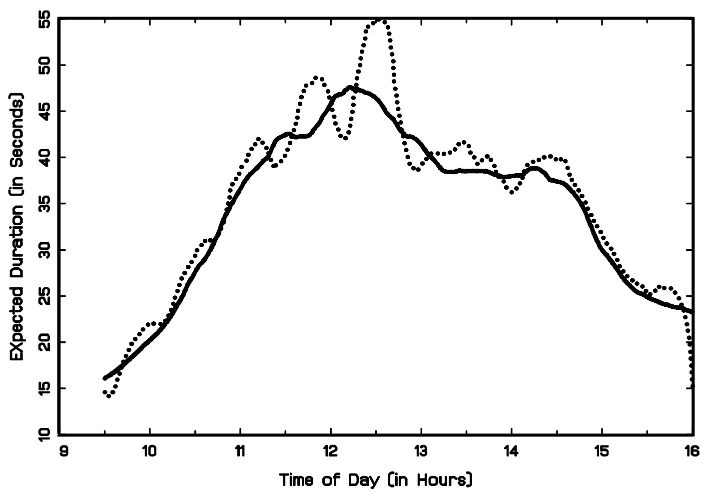


Figure 6c: Time of Day Effect, h3



Figures 5 shows the QQ plot of the residuals. Figures 6abc give the estimated time of day effect curves using different bandwidths..