TI 2006-058/4
Tinbergen Institute Discussion Paper

Semiparametric Regression with Kernel Error Model

Ao Yuan[1]
Jan G. De Gooijer[2]

[1] Howard University;
[2] Universiteit van Amsterdam.

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Semiparametric Regression with Kernel Error Model

Ao Yuan[1] and Jan G. De Gooijer[2*]

[1] Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC, USA
e-mail: ayuan@howard.edu

[2] Department of Quantitative Economics, University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: j.g.degooijer@uva.nl

ABSTRACT. We propose and study a class of regression models, in which the mean function is specified parametrically as in the existing regression methods, but the residual distribution is modeled nonparametrically by a kernel estimator, without imposing any assumption on its distribution. This specification is different from the existing semiparametric regression models. The asymptotic properties of such likelihood and the maximum likelihood estimate (MLE) under this semiparametric model are studied. We show that under some regularity conditions, the MLE under this model is consistent (as compared to the possibly pseudo consistency of the parameter estimation under the existing parametric regression model), and is asymptotically normal with rate $\sqrt{n}$ and efficient. The nonparametric pseudo-likelihood ratio has the Wilks property as the true likelihood ratio does. Simulated examples are presented to evaluate the accuracy of the proposed semiparametric MLE method.

Key words: information bound, kernel density estimator, maximum likelihood estimate, nonlinear regression, semiparametric model, U-statistic, Wilks property.

Running Heading: Semiparametric regression with kernel errors

[*] Corresponding author

1 Introduction

Consider a general (non)linear regression problem with observations

$$Y_i|(\mathbf{X}_i, \boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*, \mathbf{X}_i) + \epsilon_i, \quad (i = 1, \ldots, n), \tag{1}$$

where $\boldsymbol{\theta}^*$ is the "true" value of the model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset I\!R^k$, and $g(\boldsymbol{\theta}, \boldsymbol{x}) = E(Y_i|(\mathbf{X}_i, \boldsymbol{\theta}))$ is a specified conditional mean function of the k-dimensional parameter vector $\boldsymbol{\theta}$ and the covariates $\boldsymbol{x} \in \mathcal{X} \subset I\!R^p$, and where it is assumed that the $(Y_i, \boldsymbol{X}_i, \epsilon_i)$'s are independent and identically distributed $(i.i.d.)$ realizations from a common random source $(Y, \boldsymbol{X}, \epsilon)$. In the classical parametric situation the ϵ_i's are assumed to have some known distribution $f(\cdot)$, or equivalently $Y_i|(\mathbf{X}_i, \boldsymbol{\theta}) \sim f(\cdot - g(\boldsymbol{\theta}, \mathbf{X}_i))$. Then the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ has the desirable optimal properties when $f(\cdot)$ is the true residual distribution. But, in practice, the true form of $f(\cdot)$ is unknown. Its choice is often based on convenience, and usually restricted to a limited few. In reality, however, any pre-assumed distribution model may not easy to justify and be deviated more or less. Let $q(\boldsymbol{x})$ be the (usually unknown) density of $\boldsymbol{X}$. If it happens that the correct model $f(\cdot)$ is used and if the data are generated at the true parameter $\boldsymbol{\theta}^*$, then it is well-known that the MLE will almost surely (a.s.) converges to

$$\arg \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \int f(y - g(\boldsymbol{\theta}^*, \boldsymbol{x}))q(\mathrm{x}) \log[f(y - g(\boldsymbol{\theta}, \boldsymbol{x}))q(\boldsymbol{x})]dyd\boldsymbol{x},$$

which is achieved by $\boldsymbol{\theta}^*$, the true parameter.

On the other hand, if an incorrect model $f_1(\cdot)$ is specified, it is known (Huber, 1967; Pfanzagl, 1969) that the MLE from the parametric model $Y_i|(\boldsymbol{X}_i, \boldsymbol{\theta}) \sim f_1(\cdot - g(\boldsymbol{\theta}, \boldsymbol{X}_i))$ will a.s. converge to the pseudo-true parameter $\boldsymbol{\Theta}_1$,

$$\boldsymbol{\Theta}_1 = \arg \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \int f(y - g(\boldsymbol{\theta}^*, \boldsymbol{x}))q(\boldsymbol{x}) \log[f_1(y - g(\boldsymbol{\theta}, \boldsymbol{x}))q(\boldsymbol{x})]dyd\boldsymbol{x}. \tag{2}$$

The points in $\boldsymbol{\Theta}_1$ may not necessarily correspond to the "true" parameter(s) generating the data.

As an alternative, several nonparametric approaches may be adopted (Nadaraya, 1964; Watson, 1964; Priestley & Chao, 1972; Gasser & Müller, 1979; Eubank, 1988). These approaches involve estimation of the general functional form $E(Y|\boldsymbol{X} = \boldsymbol{x}) = m(\boldsymbol{x})$ by variants of kernel weighted empirical estimates. The resulting estimates are robust and have lots of optimality properties. But the estimated regression function $\hat{m}(\cdot)$ is, in general, not given in a simple descriptive form. Thus, making it difficult to examine the quantitative effects of the covariates

$\boldsymbol{X}$ on the response variable Y. Also, nonparametric tools for testing certain hypotheses about the covariates are limited. Further, it is well-known that estimates of $m(x)$ are subject to the so-called boundary or edge effects (Gasser & Müller 1979; Rice 1984), in which estimation bias increases near the ends of the estimation interval. Härdle & Mammen (1993) compared the L_2-difference between parametric and nonparametric regressions. Their analysis indicated that the estimation bias is usually significant.

Semiparametric regression methods are a third category of approaches to estimate (1) (Cox, 1975; Cleveland, 1979; Severini & Staniswalis, 1994). Typically, the parametric part involves the modeling of the conditional mean function $g(\cdot,\cdot)$ while the residual distribution function $f(\cdot)$ is specified partially, such as through an estimating equation. But these specifications are equivalent, implicitly, to some known exponential models.

In an attempt to derive a general regression method without imposing too strong subjective model assumptions and, in some situations, make more reasonable inference (in the sense of truly consistent, as compared to the pseudo consistency to a member of $\boldsymbol{\Theta}_1$ given before), we try to take advantage of the parametric regression in characterizing the response-covariates relationship, and that of the nonparametric regression for robustness. Here we consider a semiparametric method, in which for a given mean function, the regression coefficients are modeled parametrically, and $f(\cdot)$ is modeled by a nonparametric kernel density estimator. This intuitively simple method seems not been addressed in the literature, and is the topic of this manuscript. We will show that, under fairly general conditions, the MLE of the regression parameter(s) under this model is consistent, and asymptotically normal with rate $\sqrt{n}$ and efficient.

The paper is organized as follows. In Section 2, we introduce the semiparametric regression method to estimate (1). Also we briefly review some related methods. In Section 3 we study the consistency of the nonparametric likelihood and the MLE based on it. Asymptotic normality of the MLE is considered in Section 4. Further, we introduce a "nonparametric" likelihood ratio (NLR) test statistic and demonstrate that its null distribution follows an asymptotically χ^2 distribution, independent of nuisance parameters. In Section 5 we address asymptotic efficiency. Section 6 contains several numerical results. Section 7 provides a brief discussion on the selection of the mean function. It discusses the basic difference between robust regression and the approach presented here. Proofs of lemmas are relegated to the Appendix.

2 Semiparametric kernel regression

2.1 The method

For a fixed function $g(\cdot,\cdot)$, given the covariate $\boldsymbol{x}$, and given the parameter value $\boldsymbol{\theta}$, the function $f(\cdot)$ can be estimated nonparametrically. A direct approach is the Nadaraya-Watson kernel estimator, given by

$$f_n(\epsilon|\boldsymbol{\theta}) = \frac{1}{nh_n}\sum_{j=1}^{n} K\Big(\frac{\epsilon - Y_j + g(\boldsymbol{\theta}, \mathbf{X}_j)}{h_n}\Big),$$

where $K(\cdot)$ is a probability density function called the kernel, and h_n (the bandwidth) is a positive sequence tending to zero as n tends to infinity. It is known that under various sets of regularity conditions

$$\sup_z |f_n(z) - f(z)| \to 0, \quad a.s. \tag{3}$$

Let $\mathcal{W}_\alpha$ be the class of densities with continuous and bounded α-th derivatives. Stone (1980) has established the well-known optimal rate of convergence of the r-th derivative of $f_n(\cdot)$. The best rate of convergence in probability, uniformly over $\mathcal{W}_\alpha$, is $n^{-(\alpha-r)/(2\alpha+1)}$. On the other hand, for some smooth functionals of $f_n(\cdot)$, it is known that the rate of $n^{-1/2}$ is achievable (Ibragimov *et al.*, 1986; Bickel & Ritov, 1988).

Now, the key idea of our method is to plug in the estimator $f_n(\cdot|\boldsymbol{\theta})$ for the "true" density of the $\epsilon_i = Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)$'s. For $g(\boldsymbol{\theta}, \mathbf{X}) = \boldsymbol{\theta}'\mathbf{X}$ with the constant term θ_0, we assume $f(\cdot)$ is symmetric about the origin as usual, otherwise the problem is not identifiable (Stone, 1975). From the construction of $f_n(\cdot|\boldsymbol{\theta})$, we see that this involves terms of $K(\cdot)$ evaluated at the data-points $Y_i - Y_j - g(\boldsymbol{\theta}, \mathbf{X}_i) + g(\boldsymbol{\theta}, \mathbf{X}_j)$ $(i, j = 1, \ldots, n)$. For some specifications of $g(\cdot,\cdot)$,this will cause the cancellation of some parameters in the difference $-g(\boldsymbol{\theta}, \mathbf{X}_i) + g(\boldsymbol{\theta}, \mathbf{X}_j)$, and thus gives rise to an idenfiability problem. One can overcome this problem by choosing another known nonlinear function $r(\cdot)$ such that there is no parameter cancellation in the difference $r(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)) - r(Y_j - g(\boldsymbol{\theta}, \mathbf{X}_j))$, and this is always possible. For instance, for the linear mean function $g(\boldsymbol{\theta}, \mathbf{X}) = \boldsymbol{\theta}'\mathbf{X}$, we may choose $r(z) = ce^z/(1+e^z)$, for some $c > 0$. This will map the range of $w = r(z)$ into $(0, c)$, and $r(\cdot)$ is one-to-one, strict monotone, bounded differentiable and with inverse $r^{-1}(w) = \log(w/(c-w))$. When there is no parameter cancellation in the difference $-g(\boldsymbol{\theta}, \mathbf{X}_i) + g(\boldsymbol{\theta}, \mathbf{X}_j)$, we just let $r(\cdot)$ to be the identity. Note that, conditional on $(\mathbf{X}_i, \boldsymbol{\theta})$, the models

$$Y_i = g(\boldsymbol{\theta}, \mathbf{X}_i) + \epsilon_i \quad \text{and} \quad r(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)) = r(\epsilon_i)$$

are equivalent. Clearly, for given $\boldsymbol{\theta}$, the random variables $Z_i = r(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i))$ $(i = 1, \ldots, n)$ are *i.i.d.* Now, instead of modeling the distribution of the ϵ_i's, the idea is to model the distribution of the Z_i's. With a slight abuse of notation, the true density of Z_i and its estimate will be denoted by respectively $f(\cdot)$ and $f_n(\cdot)$.

Since all Z_j's are used in the construction of $f_n(\cdot)$ at each Z_i, the nonparametric likelihood specification will contain some unwanted values of $(nh_n)^{-1}K(0)$. This suggests the use of the delete-one version of $f_n(\cdot)$ in the likelihood structure. In particular, the likelihood function of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ given $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is

$$l_n(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X}) = \prod_{i=1}^{n} f_{(n,i)}(Z_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{(n,i)}(Y_i - g(\boldsymbol{\theta}, \boldsymbol{X}_i)), \tag{4}$$

where $f_{(n,i)}(\cdot|\boldsymbol{\theta})$ denotes $f_n(\cdot|\boldsymbol{\theta})$ with the ith data-point Z_i deleted, i.e.

$$f_{(n,i)}(Z_i|\boldsymbol{\theta}) = \frac{1}{(n-1)h_n} \sum_{j \neq i} K\Big(\frac{r(Y_i - g(\boldsymbol{\theta}, \boldsymbol{X}_i)) - r(Y_j - g(\boldsymbol{\theta}, \boldsymbol{X}_j))}{h_n}\Big).$$

Maximizing (4) over $\boldsymbol{\theta}$ yields the MLE $\hat{\boldsymbol{\theta}}_n$.

2.2 Some related methods

At first sight it seems that similar ideas have been proposed earlier in the literature; see Manski (1984) for a comprehensive, but outdated, review. Beran (1974) and Sievers (1978) considered a weighted rank statistic method. Although this method has a few similarities with ours, in essence it is different. Also, some other related methods can be found in, for example, Hall & Marron (1990), Müller *et al.* (2004), and Schick & Wefelmeyer (2004). These methods address different regression problems and the results are different. The semiparametric model in Andrews (1994) is to optimize a criterion function, its application to the regression problem (Section 5) is closely related to the nonparametric regression method, and different from ours. So are the estimation of the least favourable curve in Severini & Wong (1992), and the semiparametric model in Murphy & Van der Vaart (2000). For the estimation of a location parameter with unknown distribution, Van Eden (1973), Stone (1975) and Beran (1978) proposed an efficient estimator. But the problem considered by these authors and the methods used, are different. For instance, Stone's (1975) estimator of the location parameter θ (one-dimensional) has a two-step structure. Given an initial asymptotic normal scale invariant estimator $\bar{\theta}_n$ of θ, he constructed the estimator $\hat{\theta}_n = \bar{\theta}_n - (1/n)\sum_{i=1}^{n} G_n(X_i - \bar{\theta}_n)$, where $G_n(\cdot)$ depends on a normal kernel estimate $\hat{f}_n(\cdot)$ of the unknown density $f(\cdot)$ of the data $X_1, \ldots, X_n$. In our case, we have covariates and their coefficients as parameters. We also plug in $\hat{f}_n(\cdot)$ to estimate $f(\cdot)$, but our estimate has a one-step structure without using any $\bar{\theta}_n$.

More recently, Wolsztynski *et al.* (2005) proposed a minimum-entropy estimation of the regression problem considered here. They choose $\hat{\boldsymbol{\theta}}_n$ to minimize $\hat{H}_n(\boldsymbol{\theta}) = -\int_{-A_n}^{A_n} f_n(u|\boldsymbol{\theta}) \log f_n(u|\boldsymbol{\theta}) du$, where A_n is a suitable (slowly) increasing sequence of positive numbers, $f_n(u|\boldsymbol{\theta}) = (K_n(u|\boldsymbol{\theta}) + K_n(-u|\boldsymbol{\theta}))/2$ is the kernel estimate of the error distribution at fixed $\boldsymbol{\theta}$, and $K_n(u|\boldsymbol{\theta}) = (nh_n)^{-1} \sum_{i=1}^n K(h_n^{-1}(u - (Y_i - g(\boldsymbol{\theta}, X_i))))$. Newey (1988) proposed a method with moment restrictions. Bickel (1982) and Schick (1993) considered asymptotically efficient estimation in semiparametric and general regression models respectively. Their construction of an efficient estimate is based on the fact that any efficient estimation $\hat{H}_n$ of a functional $H(\boldsymbol{\theta})$ must have the form $\hat{H}_n = H(\boldsymbol{\theta}) + n^{-1}\sum_{i=1}^n \psi(Y_i, \boldsymbol{\theta}, f) + o_P(n^{-1/2})$, where $\psi(\cdot, \boldsymbol{\theta}, f)$ is the efficient influence function of the semiparametric model, which can be evaluated in closed form for given $(\boldsymbol{\theta}, f)$. But, since $\boldsymbol{\theta}$ and f are still unknown, they find estimates $\tilde{\boldsymbol{\theta}}$ and $\tilde{f}$ such that $\| H(\tilde{\boldsymbol{\theta}}) - H(\boldsymbol{\theta}) \| = o_P(n^{-1/2})$ and $n^{-1}\sum_{i=1}^n \| \psi(Y_i, \boldsymbol{\theta}, f) - \psi(Y_i, \tilde{\boldsymbol{\theta}}, \tilde{f}) \| = o_P(n^{-1/2})$, where $\| \cdot \|$ denotes the Euclidean norm. The proposed constructions are involved, and some of the necessary conditions are not easy to verify. Also, Efromovich (1996) studied the nonparametric regression problem $Y_i = g(\mathbf{X}_i) + \epsilon_i$, in which the ϵ_i's are *i.i.d.* with known $f(\cdot)$, $g(\cdot)$ is unknown and is the subject of inference, and $Y_i|\mathbf{X}_i \sim f(\cdot - g(\mathbf{X}_i))$. The goal is to estimate the regression function $g(\cdot)$ nonparametrically. Clearly, our setting has some similarity with that in Efromovich, but the problem is different.

3 Consistency

In this section we study the consistency of the nonparametric likelihood (4) and the MLE based on it. For convenience, when we emphasize the dependence on $\boldsymbol{\theta}$, we use $f(z|\boldsymbol{\theta})$ to denote $f(r(y - g(\boldsymbol{\theta}, \mathbf{x}))$, and $f^{[1]}(\cdot|\cdot)$ to denote its first partial derivative vector $\frac{\partial}{\partial \boldsymbol{\theta}} f(z|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Its Hessian matrix $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} f(z|\boldsymbol{\theta})$ will be denoted by $f^{[2]}(\cdot|\cdot)$. Also we adopt the notation $f^{(k)}(z|\theta)$ to denote its k-th derivative with respect to z. Further, let $\mu_g(\boldsymbol{\theta}) = E[g^{[1]}(\boldsymbol{\theta}, \mathbf{X})]$, $\Omega_g(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}[r^{(1)}(r^{-1}(Z))g^{[1]}(\boldsymbol{\theta}, \mathbf{X})]$, $\tilde{\Omega}_g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[(r^{(1)}(r^{-1}(Z))^2 g^{[1]}(\boldsymbol{\theta}, \mathbf{X}) g^{[1]}(\boldsymbol{\theta}, \mathbf{X})']$ and $\check{\Omega}_g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[r^{(1)}(r^{-1}(Z))g^{[2]}(\boldsymbol{\theta}, \mathbf{X}) - r^{(2)}(r^{-1}(Z))g^{[1]}(\boldsymbol{\theta}, \mathbf{X})g^{[1]}(\boldsymbol{\theta}, \mathbf{X})']$. We also define τ_j as the total variation of $K^{(j)}(\cdot)$ $(j = 0, 1, 2)$, and τ_g as the total variation of $g^{[1]}(\boldsymbol{\theta}, \cdot)$.

To study the asymptotic behavior of $\ell_n(Y|\boldsymbol{\theta}, \boldsymbol{X})$ and $\hat{\boldsymbol{\theta}}_n$, we first list the following regularity assumptions.

(A1) $K^{(i)}(\cdot)$ is bounded $(i = 0, 1, 2)$.

(A2) $h_n \to 0$ and $\sum_{n=1}^{\infty} \exp(-\varepsilon n h_n^8) < \infty$, for all $\varepsilon > 0$.

(A3) $f^{(2)}(\cdot)$ exists on the support of $f(\cdot)$.

(A4) $g^{[2]}(\cdot, \cdot)$ is continuous.

(A5) $f(\cdot)$ has compact support.

(A6) $|f_n(z) - f(z)| = o(f_n(z))$ uniformly in z (a.s.) on the support of $f(\cdot)$.

(A7) $q(\cdot)$ has compact support.

(A8) $0 < \inf_{\boldsymbol{\theta} \in \mathbf{A}} \Omega_g(\boldsymbol{\theta}) < \infty$ componentwise, is nonsingular and continuous on some $\mathbf{A}$.

(A9) $\int u^j K^{(r)}(u)du = 0$ $(0 \leq j \leq r-1)$, $0 \neq \gamma_r := r^{-1} \int u^r K^{(r)}(u)du < \infty$, $(r = 1, 2)$.

(A10) $0 < E(f^{(1)}(Z)/f(Z))^2 < \infty$.

Remark 1: Conditions (A1)–(A10), except (A6), are practical and easy to satisfy. If $K^{(1)}(\cdot)$ is bounded then $K(\cdot)$ is of bounded variation. If we take $K(\cdot)$ to be the truncated normal density with mean 0 and variance σ^2, since $K^{(1)}(u)$ and $uK^{(2)}(u)$ are odd functions of u, we have $\int K^{(1)}(u)du = \int uK^{(2)}(u)du = 0$. Also the condition $\int K^{(2)}(u)du = 0$ is equivalent to $(\sqrt{2\pi}\sigma)^{-1} \int u^2 e^{-u^2/(2\sigma^2)} du = \sigma^2$, which is automatically true since σ^2 is the variance. In most of the proofs, we need (A2) with h_n^4 in the exponent. Only the proof of Lemma 4 needs h_n^8. Nevertheless, we still keep the stronger condition h_n^8 in (A2). Conditions under which (A6) holds will be discussed in Subsection 3.2.

Remark 2: For inference-based parametric models, (A5) will be a serious concern since most commonly used models are of unbounded support. But it is not so serious for nonparametric models, as inferences are based on the estimated "true" model, and in reality it's rare to have a practical true probability model with infinite support.

Entropy estimation has been studied by many authors using the kernel density estimator. Although the kernel density estimator itself is not $\sqrt{n}$-consistent, a functional of it may be, as is the entropy estimation based on it. Györfi & Van der Meulen (1990) studied the strong consistency of such entropy estimators using data in $A_n = \{z : f_n(z) \geq a_n\}$ for some $a_n \to 0$. Eggermont & LaRiccia (1999) obtained the bias of $o(n^{-1/2})$ (a.s.) under relatively simple conditions not involving $\boldsymbol{\theta}$, using the double exponential kernel. Their estimator is best asymptotically normal. We will use their results in our case. Below we state the corresponding conditions.

(A11) $K(\cdot)$ is the double exponential density.

(A12) $h_n = O(n^{-\gamma/(3\gamma+2)})$ for some $\gamma > 2$.

(A13) $E(|Z|^{\lambda}) < \infty$ for some $\lambda > \gamma$.

(A14) $f(\cdot)$ is twice differentiable, and $E(|f^{(r)}(Z)|/f(Z))^2 < \infty$, $(r = 1, 2)$.

3.1 Consistency of the pseudo-likelihood

Let $\boldsymbol{\theta}^*$ be the "true" $\boldsymbol{\theta}$ for the observed data, define an "entropy-like" quantity $L(\boldsymbol{\theta})$ by

$$L(\boldsymbol{\theta}) \quad = \quad \int\int f(y - g(\boldsymbol{\theta}^*, \mathbf{x}))q(\mathbf{x}) \log f(y - g(\boldsymbol{\theta}, \mathbf{x}))dyd\mathbf{x}$$

$$= \int\int f(y - g(\boldsymbol{\theta}^*, \mathbf{x}))q(\mathbf{x}) \log[f(y - g(\boldsymbol{\theta}, \mathbf{x}))q(\mathbf{x})]dyd\mathbf{x} - \int\int f(y - g(\boldsymbol{\theta}^*, \mathbf{x}))q(\mathbf{x}) \log q(\mathbf{x})dyd\mathbf{x}.$$

Note that $L(\cdot)$ is maximized by the maximizer of the first term above, which is $\boldsymbol{\theta}^*$. The empirical version of $L(\boldsymbol{\theta})$ is the average pseudo log-likelihood, defined as

$$L_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f_{(n,i)}(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)).$$

Let

$$\tilde{L}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(Z_i|\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)).$$

By the strong law of large numbers (SLLN), $\tilde{L}_n(\boldsymbol{\theta}) \to L(\boldsymbol{\theta})$ (a.s.). If $L(\cdot)$ can be well approximated by $L_n(\cdot)$, then the MLE from $L_n(\cdot)$ can be expected to be close to the true parameter $\boldsymbol{\theta}^*$, given the smoothness condition on $g(\boldsymbol{\theta}, \boldsymbol{X}_i)$ in (A4). The following results assert the consistency of the nonparametric likelihood.

Theorem 1. *(i) Under condition (A6), for any compact set* $\mathbf{A}$ *of* $\boldsymbol{\theta}$*, we have*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{A}} |L_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \to 0, \qquad a.s.$$

(ii) Under conditions (A11)-(A14), we have

$$|L_n(\boldsymbol{\theta}) - \tilde{L}_n(\boldsymbol{\theta})| = o(n^{-1/2}), \qquad a.s.$$

Consequently, $L_n(\boldsymbol{\theta})$ *is strongly consistent to* $L(\boldsymbol{\theta})$*, and the central limit theory (CLT) and the law of iterated logarithm hold for* $L_n(\boldsymbol{\theta})$ *as for* $\tilde{L}_n(\boldsymbol{\theta})$*.*

Proof: (i) By (A6) we have $\sup_{\boldsymbol{\theta}\in\boldsymbol{A}} |L_n(\boldsymbol{\theta}) - \tilde{L}_n(\boldsymbol{\theta})| = o(1)$ *a.s.* Now we show that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{A}} |\tilde{L}_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \to 0, \quad a.s. \tag{5}$$

Recall the large deviation result (Bahadur & Zabell, 1979; Kotz & Johnson, 1982, p. 32): Under mild regularity conditions, for any *i.i.d.* random variable $V_1, \ldots, V_n$ taking values in some space V, and U, a finite union of open convex non-empty real sets of V, one has

$$\lim_{n\to\infty} n^{-1} \log P(\overline{V}_n \in U) = s(U) := \sup\{\rho(u) : u \in U\}, \quad \text{or} \quad P(\overline{V}_n \in U) \leq Ce^{nCs(U)},$$

for large n and some $0 < C < \infty$, where $-\infty < s(U) \leq 0$, $\overline{V}_n$ is the sample mean of the V_i's, and

$$\rho(u) = \inf_t [-tu + \log \phi(t)], \qquad \phi(t) = E(e^{tV_1}).$$

The regularity conditions (Bahadur and Zabell, 1979) are easily satisfied for most commonly used distributions. In particular, we take $\overline{V}_n = \tilde{L}_n$, $\mu = L$, and for fixed $\epsilon > 0$, we take $U = U(\epsilon) = (-\infty, \mu - \epsilon) \cup (\mu + \epsilon, \infty)$, the sphere centered at $\mu = E(V_1)$ with radius ϵ.

The infimal t in the definition of $\rho(u)$ must satisfy $E(V_1 e^{tV_1}) = uE(e^{tV_1})$, $t \neq 0$ (otherwise $u = E(V_1) \notin U$). Apparently, $\sup\{\rho(u) : u \in U\} \neq 0$, otherwise there is a $t_0 = \inf_t[-tu + \log \phi(t)] < 0$ arbitrarily close to 0, corresponding to a u arbitrarily close to $\mu \notin U$ in the equation $E(V_1 e^{t_0 V_1}) = uE(e^{t_0 V_1})$, which is impossible. Thus, $-\infty < s(\epsilon) = s(U(\epsilon)) < 0$.

Observe that $L(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}^*}[\log f(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i))]$. Hence, for any $\epsilon > 0$, there exists a finite constant $C > 0$ such that $P(|\tilde{L}_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \geq \epsilon) \leq \exp(nCs(\epsilon))$. Since $L(\cdot)$ and $\log f(\cdot|\boldsymbol{\theta})$ are uniformly continuous on $\mathbf{A}$, there exists a finite number of points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J \in \mathbf{A}$, such that $\forall\ \boldsymbol{\theta} \in \mathbf{A}$, $\exists\ \boldsymbol{\theta}_l$, we have

$$|\tilde{L}_n(\boldsymbol{\theta}) - \tilde{L}_n(\boldsymbol{\theta}_l)| < \epsilon/3, \quad \text{and} \quad |L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_j)| < \epsilon/3.$$

Moreover, for all $\boldsymbol{\theta} \in \mathbf{A}$ there exists a j such that

$$|\tilde{L}_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \leq |\tilde{L}_n(\boldsymbol{\theta}) - \tilde{L}_n(\boldsymbol{\theta}_j)| + |\tilde{L}_n(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta}_j)| + |L(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta})|.$$

By the inequality above we have

$$P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} |\tilde{L}_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \geq \epsilon\right) \leq P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} |\tilde{L}_n(\boldsymbol{\theta}) - \tilde{L}_n(\boldsymbol{\theta}_j)| \geq \epsilon/3\right) + P\left(\max_j |\tilde{L}_n(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta}_j)| \geq \epsilon/3\right)$$

$$+ P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} |L(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta})| \geq \epsilon/3\right) = P\left(\max_j |\tilde{L}_n(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta}_j)| \geq \epsilon/3\right)$$

$$\leq \sum_{j=1}^{J} P\left(|\tilde{L}_n(\boldsymbol{\theta}_j) - L(\boldsymbol{\theta}_j)| \geq \epsilon/3\right) \leq J \exp(Cs(\epsilon/3)n).$$

Accordingly, by the Borel-Cantelli lemma, (5) holds.

(ii) This is a direct result from Theorem 2 in Eggermont & LaRiccia (1999). The difference is that, in their case the Z_i's are *i.i.d.* random variables and in our case, $Z_i = Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)$. Hence, the corresponding distribution involves $\boldsymbol{\theta}$, and $L(\cdot)$ involves the marginal distribution of the $\mathbf{X}_i$' s.

3.2 Consistency of the MLE $\hat{\boldsymbol{\theta}}_n$

In this section we study the consistency of the MLE $\hat{\boldsymbol{\theta}}_n$, under the nonparametric likelihood (4). First, note that under conditions (A1), (A2) and (A5), (A3) is the sufficient and necessary condition for (3) to hold (Rao 1983, Theorem 2.1.3). For clarity, in the subsequent proofs, we assume $r(\cdot)$ to be identity. We have

$$f^{[1]}_{(n,i)}(Z_i|\boldsymbol{\theta}) = \frac{1}{(n-1)h_n^2}\sum_{j\neq i} K^{(1)}\Big(\frac{Z_i - Z_j}{h_n}\Big)(g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_i))$$

and

$$\begin{aligned} f^{[2]}_{(n,i)}(Z_i|\boldsymbol{\theta}) = \frac{1}{(n-1)h_n^2}\sum_{j\neq i}\Big(& h_n^{-1}K^{(2)}\Big(\frac{Z_i - Z_j}{h_n}\Big)(g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_i))(g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j))' \\ & + K^{(1)}\Big(\frac{Z_i - Z_j}{h_n}\Big)(g^{[2]}(\boldsymbol{\theta}, \mathbf{X}_j) - g^{[2]}(\boldsymbol{\theta}, \mathbf{X}_i))\Big). \end{aligned}$$

In case $r(\cdot)$ is not the identity function, the $g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j)$'s should be replaced by $r^{(1)}(r^{-1}(Z_j))g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j)$. Similarly, the $g^{[2]}(\boldsymbol{\theta}, \mathbf{X}_j)$'s should be replaced by $r^{(1)}(r^{-1}(Z_j))g^{[2]}(\boldsymbol{\theta}, \mathbf{X}_j) - r^{(2)}(r^{-1}(Z_j))g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j) \times g^{[1]'}(\boldsymbol{\theta}, \mathbf{X}_j)$'s.

Theorem 2. *Suppose (A1)–(A10) hold and that $L(\cdot)$ has a unique maximizer $\boldsymbol{\theta}^*$. Then for any compact set $\mathbf{A}$ of $\boldsymbol{\theta}$ on which (A8) holds, we have $\sup_{\boldsymbol{\theta}^*\in\mathbf{A}} ||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|| \to 0, a.s.$*

Proof: Given (A1) and (A4), by Taylor's expansion we have $0 = L_n^{[1]}(\hat{\boldsymbol{\theta}}_n) = L_n^{[1]}(\boldsymbol{\theta}^*) + L_n^{[2]}(\boldsymbol{\theta}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}_n$ is an intermediate value between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$. Or $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = -(L_n^{[2]}(\boldsymbol{\theta}_n))^{-1}L_n^{[1]}(\boldsymbol{\theta}^*)$. Let $B(\boldsymbol{\theta}^*, r)$ be the ball centered at $\boldsymbol{\theta}^*$ and with radius r. By (A6), for large n, $f_{(n,i)}(\cdot) \to f(\cdot)$ uniformly, so $\boldsymbol{\theta}_n$ is the maximizer of

$$\frac{1}{n}\sum_{i=1}^n f_{(n,i)}(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)) = \frac{1}{n}\sum_{i=1}^n f(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i)) + o(1).$$

The first term on the right hand side above is maximized by $\tilde{\boldsymbol{\theta}}$ which is consistent, and therefore close to $\boldsymbol{\theta}^*$. Since we assumed $L(\cdot)$ has an unique maximizer, for large n, $\boldsymbol{\theta}_n$ should be close to $\tilde{\boldsymbol{\theta}}$, and hence to $\boldsymbol{\theta}^*$. Thus there is an $r > 0$ such that $\boldsymbol{\theta}_n \in B(\boldsymbol{\theta}^*, r)$ for all large n. Let $abs(|L_n^{[2]}(\cdot)|)$ be the absolute value of the determinant of $L_n^{[2]}(\cdot)$. Then $abs(|L_n^{[2]}(\boldsymbol{\theta}_n)|) \geq \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*, r)} abs(|L_n^{[2]}(\boldsymbol{\theta})|)$. Hence, it remains to prove that

$$\sup_{\boldsymbol{\theta}^*\in\mathbf{A}} L_n^{[1]}(\boldsymbol{\theta}^*) \overset{a.s.}{\to} \mathbf{0} \tag{6}$$

and

$$\lim_n \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*, r)} \inf_{\boldsymbol{\theta}^*\in\mathbf{A}} abs(|L_n^{[2]}(\boldsymbol{\theta})|) = \lim_n \inf_{\boldsymbol{\theta}\in\tilde{\mathbf{A}}} abs(|L_n^{[2]}(\boldsymbol{\theta})|) > 0, \tag{7}$$

where $\tilde{\mathbf{A}} = \{\boldsymbol{\theta} \in I\!R^d : \inf_{\boldsymbol{\alpha}\in\boldsymbol{A}} ||\boldsymbol{\theta} - \boldsymbol{\alpha}|| \leq r\}$ is $\mathbf{A}$ unioning its r-neighborhood. Since $\boldsymbol{A}$ is arbitrarily compact, (7) can be proved with $\tilde{\boldsymbol{A}}$ replaced by $\mathbf{A}$. However by (A10) it suffices to prove that, for any compact set $\mathbf{A}$ satisfying (A8),

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{A}} ||L_n^{[2]}(\boldsymbol{\theta}) - \gamma_1^2 E(f^{(1}(Z)/f(Z))^2\Omega_g(\boldsymbol{\theta})|| \to 0, \quad a.s.$$

Note (A2) implies (B3), the above result follows from Lemma 1 in the Appendix.

We now prove (6). Recall that $\boldsymbol{\theta}^*$ is the only parameter value under which the Z_i's are *i.i.d.* with $f(\cdot)$. Hence, by (A6), we have

$$\frac{1}{f_{(n,i)}(z|\boldsymbol{\theta}^*)} = \frac{1}{f(z|\boldsymbol{\theta}^*)} + \frac{o(f_{(n,i)}(z|\boldsymbol{\theta}^*))}{f(z|\boldsymbol{\theta}^*)f_{(n,i)}(z|\boldsymbol{\theta}^*)} = \frac{1+o(1)}{f(z|\boldsymbol{\theta}^*)}, \quad \text{uniformly in } \ z, \quad a.s.$$

and so

$$L_n^{[1]}(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n \frac{f_{(n,i)}^{[1]}(Z_i|\boldsymbol{\theta}^*)}{f_{(n,i)}(Z_i|\boldsymbol{\theta}^*)} = \frac{1+o(1)}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{K^{(1)}(\frac{Z_i-Z_j}{h_n})[g^{[1]}(\boldsymbol{\theta}^*,\mathbf{X}_i)-g^{[1]}(\boldsymbol{\theta}^*,\mathbf{X}_j)]}{h_n^2 f(Z_i|\boldsymbol{\theta}^*)}, \quad a.s. \tag{8}$$

We first prove the pointwise convergence of

$$\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{K^{(1)}(\frac{Z_i-Z_j}{h_n})[g^{[1]}(\boldsymbol{\theta}^*,\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta}^*,\mathbf{X}_j)]}{h_n^2 f(Z_i|\boldsymbol{\theta}^*)}$$

$$:= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n K_n(Z_i,\mathbf{X}_i;Z_j,\mathbf{X}_j,\boldsymbol{\theta}^*) := U_{n,n}(\boldsymbol{\theta}^*) \to \mathbf{0}, \quad a.s.$$

Note, for each fixed n, the left hand side above is a U-statistic.

Let $Z_{(1)},\ldots,Z_{(n)},\ldots$ be the order statistics of $Z_1,\ldots,Z_n,\ldots$, $\mathbf{X}_{<i>}$ be the associated covariate of $Z_{(i)}$, $\mathcal{F}_n = \sigma((Z_{(i)},\mathbf{X}_{<i>}) : i \leq n; (Z_{n+1},\mathbf{X}_{n+1}),\ldots)$, and define for $n \geq 2$,

$$\tilde{U}_{n,m}(\boldsymbol{\theta}^*) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \tilde{K}_m(Z_i,\mathbf{X}_i;Z_j,\mathbf{X}_j,\boldsymbol{\theta}^*)$$

with $\tilde{K}_m$ to be specified latter. Then for each fixed m, $\{\tilde{U}_{n,m}(\boldsymbol{\theta}^*)\}$ is a sequence of U-statistics, and a reverse martingale with respect to $\{\mathcal{F}_n\}$. Thus, by Theorem 4.3 in Doob (1953, p. 331), $\lim_n \tilde{U}_{n,m}(\boldsymbol{\theta}^*) = \lim_n E(\tilde{U}_{2,m}(\boldsymbol{\theta}^*)|\mathcal{F}_n) \to E(\tilde{U}_{2,m}(\boldsymbol{\theta}^*)|\mathcal{F}_\infty)$ (a.s.). Since $\mathcal{F}_\infty$ is permutable, it is trivial by the Hewitte-Savage 0-1 law, and so by (A4) and (A7) $\lim_n \tilde{U}_{n,m}(\boldsymbol{\theta}^*) = E(\tilde{U}_{2,m}(\boldsymbol{\theta}^*)|\mathcal{F}_\infty) = E(E(\tilde{U}_{2,m}(\boldsymbol{\theta}^*)|\mathcal{F}_\infty)) = E(\tilde{U}_{2,m}(\boldsymbol{\theta}^*)) = \mathbf{0}$. Now

$$U_{n,n} = \tilde{U}_{n,m} + (U_{n,n} - \tilde{U}_{n,n}) + (\tilde{U}_{n,n} - \tilde{U}_{n,m}).$$

By (A10) and $h_n \to 0$, for any $\epsilon > 0$, there is a positive sequence $\{b_n\}$ such that $h_n^{-2}K^{(1)}(b_n/h_n) = \epsilon$ for all n. Note $\{b_n\}$ must satisfy $b_n \to 0$, $b_n/h_n \to \infty$. Now we choose $\tilde{K}_m(\cdot) = K_m(\cdot)\chi(|\cdot| \geq b_m)$ where $\chi(\cdot)$ denotes the indicator function. Note

$$\left|\frac{1}{h_n^2}\tilde{K}_n^{(1)}(\frac{Z_i - Z_j}{h_n}) - \frac{1}{h_m^2}\tilde{K}_m^{(1)}(\frac{Z_i - Z_j}{h_m})\right| \leq 2\epsilon, \quad \text{all} \quad (i,j).$$

Thus,

$$\begin{aligned}(\tilde{U}_{n,n} - \tilde{U}_{n,m})^+ &\leq \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \left|\frac{1}{h_n^2}\tilde{K}_n^{(1)}(\frac{Z_i - Z_j}{h_n}) - \frac{1}{h_m^2}\tilde{K}_m^{(1)}(\frac{Z_i - Z_j}{h_m})\right| \frac{(g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j))^+}{f(Z_i|\boldsymbol{\theta}^*)} \\ &\leq \frac{2\epsilon}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{(g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j))^+}{f(Z_i|\boldsymbol{\theta}^*)}.\end{aligned}$$

Note, the last summation above is a U-statistic. It converges (a.s.) to $2\epsilon E(g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j))^+ < \infty$. This implies that, for large m, $\lim_n(\tilde{U}_{n,n} - \tilde{U}_{n,m})^+$ can be arbitrarily small (a.s.). Similarly, for large m, $\lim_n(\tilde{U}_{n,n} - \tilde{U}_{n,m})^-$ can be arbitrarily small (a.s.).

Also, by definition of $\tilde{K}_n^{(1)}$, and by (A5), $|g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j)|$ is bounded, and the symmetry of $K^{(1)}(\cdot)$ implies $\int_{|u|<b_n} K^{(1)}(u)du = 0$. Let C denote a generic constant. We have

$$\begin{aligned}U_{n,n} - \tilde{U}_{n,n} &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{1}{h_n^2}\left(K^{(1)}(\frac{Z_i - Z_j}{h_n}) - \frac{1}{h_n^2}\tilde{K}_n^{(1)}(\frac{Z_i - Z_j}{h_n})\right)\frac{g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j)}{f(Z_i|\boldsymbol{\theta}^*)} \\ &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{1}{h_n^2}K^{(1)}(\frac{Z_i - Z_j}{h_n})\frac{g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_{\mathbf{i}}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j)}{f(Z_i|\boldsymbol{\theta}^*)}\chi(|Z_i - Z_j| < h_n b_n) \\ &\sim C\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{1}{h_n^2}K^{(1)}(\frac{Z_i - Z_j}{h_n})\frac{1}{f(Z_i|\boldsymbol{\theta}^*)}\chi(|Z_i - Z_j| < h_n b_n)\end{aligned}$$

$$\sim C\frac{1}{h_n^2}\int_{|z_1 - z_2| < h_n b_n}\int K^{(1)}(\frac{z_1 - z_2}{h_n})f(z_2)dz_1dz_2 = C\frac{1}{h_n}\int_{|u|<b_n}\int K^{(1)}(u)f(z_1 - h_n u)dz_1du$$

$$\begin{aligned}&= C\frac{1}{h_n}\int_{|u|<b_n}\int K^{(1)}(u)(f(z_1) - f^{(1)}(z_1 + r_n u)h_n u)dz_1du \\ &= -C\int_{|u|<b_n}\int uK^{(1)}(u)f^{(1)}(z_1 + r_n u)dz_1du \sim -Cb_n \to 0,\end{aligned}$$

where r_n is an intermediate value between 0 and -$h_n u$. We thus have $\lim_n U_{n,n}(\boldsymbol{\theta}^*) \to \mathbf{0}$, a.s.

Since $\mathbf{A}$ is compact, given any $\epsilon > 0$, there is a finite number of points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$ in $\mathbf{A}$, such that $\forall \boldsymbol{\theta}^* \in \mathbf{A}$, $\exists \boldsymbol{\theta}_l \in \mathbf{A}$, with $||\boldsymbol{\theta}^* - \boldsymbol{\theta}_l|| \leq \epsilon$. For some intermediate values $\boldsymbol{\theta}_i$ we have

$$U_{n,n}(\boldsymbol{\theta}^*) - U_{n,n}(\boldsymbol{\theta}_l) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{K^{(1)}(\frac{Z_i - Z_j}{h_n})[g^{[2]}(\boldsymbol{\theta}_i, \mathbf{X}_i) - g^{[2]}(\boldsymbol{\theta}_j, \mathbf{X}_i)](\boldsymbol{\theta}^* - \boldsymbol{\theta}_l)}{h_n^2 f(Z_i|\boldsymbol{\theta}^*)} := \check{U}_{n,n}.$$

Also, we have

$$\sup_{\boldsymbol{\theta}^*\in \boldsymbol{A}} ||U_{n,n}(\boldsymbol{\theta}^*)|| \leq ||\check{U}_{n,n}|| + \max_j ||U_{n,n}(\boldsymbol{\theta}_j)||.$$

Since, by definition of $K_n(\cdots)$, $E(\check{U}_{n,n}) = E(U_{n,n}(\boldsymbol{\theta}^*)) - E(U_{n,n}(\boldsymbol{\theta}_l)) = \mathbf{0}$ and $E(U_{n,n}(\boldsymbol{\theta}_j)) = \mathbf{0}$, and by (A1) and (A7) the kernels of the U-statistics $\check{U}_{n,n}$ and $U_{n,n}(\boldsymbol{\theta}_j)$ are bounded by $\pm||\boldsymbol{\theta}^* - \boldsymbol{\theta}_l||Ch_n^{-2} \subset \pm\epsilon Ch_n^{-2}$ and $\pm Ch_n^{-2}$, for some $0 < C < \infty$. Applying Hoeffding's (1963) inequality for U-statistics (Serfling, 1980, p. 201) componentwise, we have

$$P\Big(\sup_{\boldsymbol{\theta}^*\in \boldsymbol{A}} ||U_{n,n}(\boldsymbol{\theta}^*)|| > \epsilon\Big) \leq P\Big(||\check{U}_{n,n}|| > \epsilon^{-1}/2\Big) + \sum_{j=1}^{J} P\Big(||U_{n,n}(\boldsymbol{\theta}_j)|| > \epsilon/2\Big)$$
$$\leq e^{-C^{-2}nh_n^4} + Je^{-\epsilon^2C^{-2}nh_n^4}.$$

Hence, by (A2) and the Borel-Cantelli lemma, (6) holds.

Remark 3: From (2) we see that the bias of $\hat{\boldsymbol{\theta}}_n$ may be significant if $f(\cdot)$ and $g(\cdot,\cdot)$ do not match the data distribution. But Theorem 2 tells us that under fair conditions, the bias of $\hat{\boldsymbol{\theta}}_n$ is asymptotically negligible. Clearly, the bias is quantifiable as $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = -(L_n^{[2]}(\boldsymbol{\theta}_n))^{-1}L_n^{[1]}(\boldsymbol{\theta}^*)$. Now, by Lemma 4 in the Appendix, $L_n^{[1]}(\boldsymbol{\theta}^*) \overset{a.s.}{\to} \gamma_1 L^{[1]}(\boldsymbol{\theta}^*) = \mathbf{0}$. Also, by Lemma 1 in the Appendix, $L_n^{[2]}(\boldsymbol{\theta}_n) \overset{a.s.}{\to} \gamma_1^2 E[(f^{(1)}(Z)/f(Z))^2]\Omega_g(\boldsymbol{\theta}^*)$. So, provided $L_n^{[2]}(\cdot)$ is non-singular in a neighborhood of $\boldsymbol{\theta}^*$, the bias is asymptotically equal to zero. The convergence property depends on the convergence of $f_n^{[i]}(\cdot)$ to $f^{[i]}(\cdot)$ $(i = 1, 2)$ (cf. Lemma 4).

Now we discuss some conditions under which (A6) holds. Reference will be made to conditions (Bj)s given in Section 4.

Proposition 1. *Under (A1), (A5), (B1)-(B3), and assuming $f(\cdot)$ is continuous,* $\inf_z f(z) > 0$ *on the support of $f(\cdot)$ and $\sum_n \exp(-C_1nh_n^2) < \infty$ for all $C > 0$, then (A6) holds.*

Proof: Under (A1), and (B1)-(B3) and the uniform continuity of $f(\cdot)$, by Theorem 2.1.1 in Rao (1983) we have

$$\sup_z |E(f_n(z)) - f(z)| \to 0.$$

Let $F_n(\cdot)$ and $F(\cdot)$ respectively be the empirical and real distribution functions of Z, since (B2) implies $K(\cdot)$ has bounded variation τ_0, the proof of Theorem 2.1.3 (Rao, 1983) gives that

$$P(\sup_z |f_n(z) - E(f_n(z))| > \epsilon) \leq P(\sup_z |F_n(z) - F(z)| > \epsilon h_n\tau_0^{-1}) \leq C\exp(-C_1nh_n^2),$$

for some $0 < C, C_1 < \infty$. Since $\sum_n \exp(-C_1 n h_n^2) < \infty$, we have

$$\sup_z |f_n(z) - f(z)| = o(1), \quad a.s.$$

Since $\inf_z f(z) > 0$, from the above relationship we have $\inf_z f_n(z) > 0$ for large n, and thus

$$\sup_z |f_n(z) - f(z)| = o(1) f_n^{-1}(z) f_n(z) \leq o(1)(\inf_z f_n(z))^{-1} f_n(z) = o(1) f_n(z), \quad a.s.$$

Remark 4: The assumption $\inf_z f(z) > 0$ on the support of $f(\cdot)$ is used in entropy estimation by a number of authors (Hall 1986; Joe 1989; Van Es 1992; Hall & Morton 1993).

Proposition 2. *Assume (A5) holds, and* $E \int |f_n(z) - f(z)| dz \to 0$. *Then (A6) holds.*

Proof: The given conditions imply $(f_n(z) - f(z)) \to 0$ (a.s.) for any z. If (A6) is not true, then for a given $\epsilon > 0$, there exists a constant $0 < C < \infty$ and a sequence of sets $\{A_n\}$, with $\mu(A_n) \geq \epsilon$ and $|f_n(z) - f(z)| \geq C f_n(z)$, $\forall z \in A_n$, where $\mu(\cdot)$ is the Lebesque measure on $I\!R$. Let A be the support of $f(\cdot)$, and A^c be the complement of A. The condition $(f_n(z) - f(z)) \to 0$ (a.s.) implies $\mu(A_n \cap A^c) \to 0$. Let $\chi_A(\cdot)$ be the indicator function on A, and $\underline{A} = \inf\lim_n A_n \cap A$. Then $\underline{A} \subset A$ and $\mu(\underline{A}) > 0$. By Fatou's lemma we have,

$$\begin{aligned} E \int |f_n(z) - f(z)| dz &\geq E \int |f_n(z) - f(z)| \chi_{A_n \cap A}(z) dz \geq CE \int f_n(z) \chi_{A_n \cap A}(z) dz \\ &\geq CE \int \inf\lim_n f_n(z) \chi_{\underline{A}}(z) dz = C \int f(z) \chi_{\underline{A}}(z) dz > 0, \end{aligned}$$

which is a contradiction.

For various conditions to ensure $E \int |f_n(z) - f(z)| dz \to 0$, one may consult Devroye & Györfi (1985).

4 Asymptotic Normality and Wilks Property

4.1 Asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}_b$

To study the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$, we impose the following conditions.

(B1) $\int K(y) dy = 1$.

(B2) $K(\cdot)$ has compact support and symmetric around 0.

(B3) $h_n \to 0$, $n h_n^4 \to \infty$.

(B4) $f^{(1)}(\cdot)/f(\cdot)$ is bounded on the support of $f(\cdot)$.

Let $\xrightarrow{D}$ denote convergence in distribution and $\xrightarrow{P}$ convergence in probability. The following asymptotic result holds.

Theorem 3. *Under (A1), (A3)-(A10) and (B2)-(B4), we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{D} N(0, \Omega(\boldsymbol{\theta}^*)),$$

where $\Omega^{-1}(\boldsymbol{\theta}^*) = E[(f^{(1)}(Z)/f(Z))^2]\Omega_g(\boldsymbol{\theta}^*)$.

Note: the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$ does not depend on the kernel $K(\cdot)$, while most of the kernel type estimators do.

Proof: Observe that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = (-L_n^{[2]}(\boldsymbol{\theta}_n))^{-1}\sqrt{n}L_n^{[1]}(\boldsymbol{\theta}^*)$. By Lemma 1, and integrating by parts gives $\gamma_1 = \int uK^{(1)}(u)du = -1$, so we have

$$L_n^{[2]}(\boldsymbol{\theta}_n) = -\gamma_1^2 E\bigg(\frac{f^{(1)}(Z)}{f(Z)}\bigg)^2 \Omega_g(\boldsymbol{\theta}^*) + o_P(1) = -\Omega^{-1}(\boldsymbol{\theta}^*) + o_P(1). \tag{9}$$

Thus we only need to show that

$$\sqrt{n}L_n^{[1]}(\boldsymbol{\theta}^*) \xrightarrow{D} N(\mathbf{0}, \Omega^{-1}(\boldsymbol{\theta}^*)). \tag{10}$$

From (9) and the expression of $L_n^{[1]}(\boldsymbol{\theta}^*)$ as given in (8), it follows that we need to show that

$$\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{K^{(1)}(\frac{Z_i - Z_j}{h_n})[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_i)]}{h_n^2 f(Z_i|\boldsymbol{\theta}^*)} \xrightarrow{D} N(\mathbf{0}, \Omega^{-1}(\boldsymbol{\theta}^*)). \tag{11}$$

Rewriting the left hand side of (11) as a U-statistic gives

$$\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j) := \sqrt{n}U_n, \tag{12}$$

where

$$\begin{aligned} K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j) &= \frac{1}{2}\bigg(\frac{K^{(1)}(\frac{Z_j - Z_i}{h_n})[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_i)]}{h_n^2 f(Z_i|\boldsymbol{\theta}^*)} \\ &\quad + \frac{K^{(1)}(\frac{Z_i - Z_j}{h_n})[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j)]}{h_n^2 f(Z_j|\boldsymbol{\theta}^*)}\bigg) \end{aligned}$$

is a symmetric kernel in the $(Z_i, \mathbf{X}_i)$'s. Let $\overline{K}_n(Z_i, \mathbf{X}_i) = E(K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j)|Z_i, \mathbf{X}_i)$. Then, it follows from (12) that

$$\begin{aligned} U_n &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Big(K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j) - \overline{K}_n(Z_i, \mathbf{X}_i) - \overline{K}_n(Z_j, \mathbf{X}_j)\Big) \\ &\quad + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Big(\overline{K}_n(Z_i, \mathbf{X}_i) + \overline{K}_n(Z_j, \mathbf{X}_j)\Big) \\ &:= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tilde{K}_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j) + \frac{2}{n}\sum_{i=1}^{n}\overline{K}_n(Z_i, \mathbf{X}_i). \end{aligned} \tag{13}$$

Now to prove (11), we only need to show that

$$\sqrt{n}\frac{2}{n}\sum_{i=1}^{n}\overline{K}_n(Z_i,\mathbf{X}_i)\xrightarrow{D}N(\mathbf{0},\Omega^{-1}(\boldsymbol{\theta}^*)),$$

which is given by Lemma 3 in the Appendix, and

$$\sqrt{n}\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\tilde{K}_n(Z_i,\mathbf{X}_i;Z_j,\mathbf{X}_j)\xrightarrow{P}0. \tag{14}$$

Note that the left hand side of (14) is a vector of dimension $dim(\boldsymbol{\theta})$. Thus it suffices to prove that (14) holds componentwise. Since by Lemma 2 in the Appendix, $E(\tilde{K}_n(Z_1,\mathbf{X}_1;Z_2,\mathbf{X}_2))=0$, and the variance of each component on the left hand side in (14) is $C/(nh_n^3)\to 0$ by (B3), for some $0<C<\infty$. Thus, by Chebychev's inequality, we obtain the desired result.

Remark 5: In contrast, the convergence rate of the estimated mode, quantiles, *etc* of the kernel density estimator is typically of $\sqrt{nh_n}$. Here $\hat{\boldsymbol{\theta}}_n$, the estimated mode of the likelihood under the kernel density estimator achieves the $\sqrt{n}$-rate, as it does under a known true model.

Remark 6: A natural estimate of $\Omega^{-1}(\boldsymbol{\theta}^*)$ is given by

$$\hat{\Omega}^{-1}(\hat{\boldsymbol{\theta}}_n)=\Big(\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f_{n,i}^{(1)}(z_i)}{f_{n,i}(z_i)}\Big]^2\Big)\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{g}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{x}_i,z_i)\Big)\Big((\tilde{g}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{x}_i,z_i)\Big)', \tag{15}$$

where $\tilde{g}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{x}_i,z_i)=r^{(1)}(r^{-1}(z_i))g^{[1]}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{x}_i)-(1/n)\sum_{i=1}^{n}r^{(1)}(r^{-1}(z_i))g^{[1]}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{x}_i)$.

The following theorem shows that the semiparametric model MLE $\hat{\boldsymbol{\theta}}_n$ obeys the functional invariance principle enjoyed by the MLE from a known parametric model. For $t\in[0,1]$, let $[nt]$ be the largest integer caped by nt,

$$L_n(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{[nt]}\log f_{(n,i)}(Y_i-g(\boldsymbol{\theta},\mathbf{X}_i)),$$

and $\hat{\boldsymbol{\theta}}_{[nt]}$ be the MLE of $\boldsymbol{\theta}$ under $L_{[nt]}(\cdot)$. Let $\mathbf{W}$ be the k-dimensional standard Brownian motion on $[0,1]$, and $D[0,1]$ be the space of k-dimensional functions of right continuity with left limit at every point.

Theorem 4. *Under conditions of Theorem 3, we have*

$$\sqrt{n}\Omega^{-1/2}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{[nt]}-\boldsymbol{\theta}^*)\xrightarrow{D}\mathbf{W}$$

in the J_1-topology on $D[0,1]$.

Proof: As in the proof of Theorem 3, we have for all $t \in (0,1]$,

$$\sqrt{n}\Omega^{-1/2}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{[nt]} - \boldsymbol{\theta}^*) = (-L^{[2]}_{[nt]}(\boldsymbol{\theta}_{[nt]}))^{-1}\Omega^{-1/2}(\boldsymbol{\theta}^*)\sqrt{n}L^{[1]}_{[nt]}(\boldsymbol{\theta}^*), \tag{16}$$

and

$$L^{[2]}_{[nt]}(\boldsymbol{\theta}_{[nt]}) = -\Omega^{-1}(\boldsymbol{\theta}^*) + o^{(1)}_{P_t}(1), \quad \sqrt{n}L^{[1]}_{[nt]}(\boldsymbol{\theta}^*) = \frac{1+o(1)}{\sqrt{n}}\sum_{i=1}^{[nt]} 2\overline{K}_n(Z_i, \mathbf{X}_i) + o^{(2)}_{P_t}(1),$$

where by (13), (14) and Lemma 2, the $o^{(i)}_{P_t}(1)$s may depend on t, but tend to zero in probability uniformly in t. Let $U_{n,h}(t)$ be the left hand side of (16). The argument for the finite dimensional weak convergence of $U_{n,h}(t)$ is similar as in the proof of Theorem 3, now we only need to prove the tightness of the family of distributions of the $U_{n,h}(t)$'s, indexed by t. For this, let

$$V_{n,h}(t) = n^{-1/2}\sum_{i=1}^{[nt]} 2\Omega^{-1/2}(\boldsymbol{\theta}^*)\overline{K}_n(Z_i, \mathbf{X}_i).$$

For simplicity of proof we assume $\boldsymbol{\theta}$ to be one-dimensional. The multi-dimensional case can be handled similarly, using norm rather than absolute differences to check the tightness condition.

By Theorem 15.6 in Billingsley (1968), we only need to show, $\forall \lambda > 0$, there is a $0 < C < \infty$ such that for all fixed $t_1 < t < t_2$,

$$P\Big(|U_{n,h}(t) - U_{n,h}(t_1)| \geq \lambda, |U_{n,h}(t_2) - U_{n,h}(t)| \geq \lambda\Big) \leq \frac{C}{\lambda^4}(t_2 - t_1)^2, \quad t_1 < t < t_2. \tag{17}$$

In fact, by (16), and the boundedness of $\Omega^{-1}(\boldsymbol{\theta}^*)$, $U_{n,h}(t) = (1 + o^{(1)}_{P_t}(1))V_{n,h}(t) + o^{(2)}_{P_t}(1)$, thus $\forall 0 \leq t_1 < t < t_2 \leq 1$,

$$|U_{n,h}(t) - U_{n,h}(t_1)| \leq |1 + o^{(1)}_{P_{t,t_1}}(1)||V_{n,h}(t) - V_{n,h}(t_1)| + |o^{(2)}_{P_{t,t_1}}(1)|,$$

where $o^{(1)}_{P_{t,t_1}}(1) = o^{(1)}_{P_t}(1)$ if $U_{n,h}(t) \geq U_{n,h}(t_1)$, and $o^{(1)}_{P_{t,t_1}}(1) = o^{(1)}_{P_{t_1}}(1)$ if $U_{n,h}(t) < U_{n,h}(t_1)$; $o^{(2)}_{P_{t,t_1}}(1) = \max\{o^{(2)}_{P_t}(1), o^{(2)}_{P_{t_1}}(1)\}$. From now on let C be a generic constant and n_0 be a generic integer. Since $o^{(2)}_{P_{t,t_1}}(1) \to 0$ uniformly in (t, t_1), there is an n_0 such that $|o^{(2)}_{P_{t,t_1}}(1)| < \lambda/2$ $\forall n \geq n_0$. Thus, there is $0 < C < \infty$ such that, $\forall n \geq n_0$ and all $t_1 < t$, we have

$$|U_{n,h}(t) - U_{n,h}(t_1)| \leq C|V_{n,h}(t) - V_{n,h}(t_1)| + \lambda/2.$$

Similarly, $\forall n \geq n_0$ and all $t < t_2$,

$$|U_{n,h}(t_2) - U_{n,h}(t)| \leq C|V_{n,h}(t_2) - V_{n,h}(t)| + \lambda/2.$$

Now we have

$$P\Big(|U_{n,h}(t) - U_{n,h}(t_1)| \geq \lambda, |U_{n,h}(t_2) - U_{n,h}(t)| \geq \lambda\Big)$$

$$\leq P\Big(|V_{n,h}(t) - V_{n,h}(t_1)| \geq \lambda/(2C), |V_{n,h}(t_2) - V_{n,h}(t)| \geq \lambda/(2C), \Big).$$

Since $V_{n,h}(t) - V_{n,h}(t_1)$ and $V_{n,h}(t_2) - V_{n,h}(t)$ are independent, and in the proof of Lemma 3 for the variance computation, there is $0 < C < \infty$ such that for large n, $E(V_{n,h}(t) - V_{n,h}(t_1))^2 \leq C([nt] - [nt_1])/n$, $\forall t > t_1$. So the right hand side of the previous expression is

$$\begin{aligned} P(|V_{n,h}(t) &- V_{n,h}(t_1)| \geq \lambda/(2C))P(|V_{n,h}(t_2) - V_{n,h}(t)| \geq \lambda/(2C)) \\ &\leq \frac{16C^4}{\lambda^4} E(V_{n,h}(t) - V_{n,h}(t_1))^2 E(V_{n,h}(t_2) - V_{n,h}(t))^2 \\ &\leq \frac{16C}{\lambda^4 n^2}([nt] - [nt_1])([nt_2] - [nt]) \\ &\leq \frac{16C}{\lambda^4}(\frac{[nt_2] - [nt_1]}{n})^2 \leq \frac{C}{\lambda^4}(t_2 - t_1)^2. \end{aligned}$$

4.2 Wilks property

The likelihood ratio statistics plays an important role in statistical hypothesis testing. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$, with $Z_1, \ldots, Z_n$ *i.i.d.* with known common density $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\Theta}_0$ be a proper sub-space of the parameter space $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}$, with $0 \leq r = dim(\boldsymbol{\Theta}_0) < dim(\boldsymbol{\Theta}) = k$, $\hat{\boldsymbol{\theta}}_n$ as before, $\hat{\boldsymbol{\theta}}_{n,0}$ be the MLE of the true data generating $\boldsymbol{\theta}^*$ under the same true distribution, but confined within $\boldsymbol{\Theta}_0$, and $l_n(\mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^n f(Z_i|\boldsymbol{\theta})$ be the likelihood function. Let $\chi^2_{(k-r)}$ be a χ^2 random variable with $k - r$ degrees of freedom. Under the null hypothesis $H_0 :\ \boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, Wilks (1938) proved that the LR statistic $2\log(l_n(\mathbf{Z}|\hat{\boldsymbol{\theta}}_n)/l_n(\mathbf{Z}|\hat{\boldsymbol{\theta}}_{n,0}) \xrightarrow{D} \chi^2_{(k-r)}$. This result was generalized under various settings to various forms (Owen, 1990; Fan *et al.*, 2001). Here, with $f_{(n,i)}(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i))$ and $l_n(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{X})$ as given in (4), we expect that a similar result will hold with $f(Z_i|\boldsymbol{\theta})$ replaced by $f_{(n,i)}(Y_i - g(\boldsymbol{\theta}, \mathbf{X}_i))$ $(i = 1, \ldots, n)$ i.e. we consider asymptotic distribution of the NLR statistic λ_n defined $\lambda_n = \log(l_n(\mathbf{Y}|\hat{\boldsymbol{\theta}}_n, \mathbf{X})/l_n(\mathbf{Y}|\hat{\boldsymbol{\theta}}_{n,0}, \mathbf{X}))$.

Theorem 5. *Under the null hypothesis H_0: $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, and the conditions of Theorem 3, then the NLR statistic $2\lambda_n \xrightarrow{D} \chi^2_{(k-r)}$.*

Proof: For simplicity, we rewrite $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ with $dim(\boldsymbol{\theta}_0) = r$, and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_1^*)$. Recall the notation $L_n(\boldsymbol{\theta})$ introduced in Subsection 3.1. By Taylor's expansion we have, for some intermediate value $\boldsymbol{\theta}_n$ between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$,

$$\log l_n(\mathbf{Y}|\hat{\boldsymbol{\theta}}_n, \mathbf{X}) = nL_n(\hat{\boldsymbol{\theta}}_n) = nL_n(\boldsymbol{\theta}^*) + \frac{1}{2}\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_n)\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big).$$

By (A.7), Theorem 3, and the Slutsky theorem, we have

$$\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_n)\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big) \xrightarrow{D} \mathbf{V}'\Omega^{-1}(\boldsymbol{\theta}^*)\mathbf{V},$$

where $\Omega^{-1}(\boldsymbol{\theta}^*) = \gamma_1^2 E[(f^{(1)}(Z)/f(Z))^2]\Omega_g(\boldsymbol{\theta}^*)$, and $\mathbf{V} \sim N(\mathbf{0}, \Omega(\boldsymbol{\theta}^*))$. For ease of notation we omit the $\boldsymbol{\theta}^*$ in Ω and Ω_2. Let $\mathbf{U} = \Omega^{-1/2}\mathbf{V}$, then $\mathbf{U} = (u_1, \dots, u_k)' \sim N(\mathbf{0}, \mathbf{I}_k)$, and

$$\mathbf{V}'\Omega^{-1}\mathbf{V} = \mathbf{U}'(\Omega^{1/2})'\Omega^{-1}\Omega^{1/2}\mathbf{U} = u_1^2 + \ldots + u_k^2 \sim \chi_k^2. \tag{18}$$

Let Ω_0 and Ω_0^{-1} be the upper left $r \times r$ block of Ω and Ω^{-1}. Similarly we have

$$\log l_n(\mathbf{Y}|\hat{\boldsymbol{\theta}}_{n,0}, \mathbf{X}) = nL_n(\boldsymbol{\theta}_{n,0}) = nL_n(\boldsymbol{\theta}_0^*) + \frac{1}{2}\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \boldsymbol{\theta}_0^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_{n,0})\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \boldsymbol{\theta}_0^*)\Big),$$

with

$$\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \ \boldsymbol{\theta}_0^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_{n,0})\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \ \boldsymbol{\theta}_0^*)\Big) \xrightarrow{D} \mathbf{V}_0'\Omega_0^{-1}\mathbf{V}_0,$$

and $\mathbf{V}_0 \sim N(\mathbf{0}, \Omega_0)$. Let $\boldsymbol{U}_0 = \Omega_0^{-1}\boldsymbol{V}_0$, then $\boldsymbol{V}_0'\Omega_0^{-1}\boldsymbol{V}_0 = \boldsymbol{U}_0'\boldsymbol{U}_0 = u_1^2 + \ldots + u_r^2 \sim \chi_r^2$. This, together with the fact that under H_0 $nL_n(\boldsymbol{\theta}^*) = nL_n(\boldsymbol{\theta}_0^*)$ and (18), gives

$$\begin{aligned} 2\lambda_n &= 2(nL_n(\boldsymbol{\theta}^*) - nL_n(\boldsymbol{\theta}_0^*)) + \Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_n)\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\Big) \\ &\quad - \Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \boldsymbol{\theta}_0^*)\Big)' L_n^{[2]}(\boldsymbol{\theta}_{n,0})\Big(\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,0} - \boldsymbol{\theta}_0^*)\Big) \\ &\xrightarrow{D} \boldsymbol{U}'\boldsymbol{U} - \boldsymbol{U}_0'\boldsymbol{U}_0 = u_{r+1}^2 + \ldots + u_k^2 \sim \chi_{k-r}^2. \end{aligned}$$

5 Efficiency

Let $\mathcal{F}$ be a collection of distributions, $T(F)$ be a functional of the data distribution $F \in \mathcal{F}$. Further, let T_n be any estimator of $T(F)$, which depends only on the data, such that $\sqrt{n}(T_n - T) \xrightarrow{D} Z$ for some random variable Z with $Var[Z] = \Omega$, and let $I^{-1}(F|T, \mathcal{F})$ be the information bound for this problem. It is well-known by the Hájek convolution theorem (e.g. Bickel *et al.*, 1993), that under fairly general conditions, $\Omega \geq I^{-1}(F|T, \mathcal{F})$ in the sense of matrix non-negative definiteness. When the equality holds, T_n is called an efficient estimator of T. For our problem, when $f(\cdot)$ is known, the information bound is the Fisher information given by

$$\begin{aligned} I(\boldsymbol{\theta}) &= E\Big(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y - g(\boldsymbol{\theta}, \mathbf{X}))\Big)\Big(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y - g(\boldsymbol{\theta}, \mathbf{X}))\Big)' \\ &= -E\Big(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(Y - g(\boldsymbol{\theta}, \mathbf{X}))\Big) = \tilde{\Omega}_g(\boldsymbol{\theta}) E\Big(\frac{f^{(1)}(Z)}{f(Z)}\Big)^2, \end{aligned}$$

where $\tilde{\Omega}_g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[(r^{(1)}(r^{-1}(Z))^2 g^{[1]}(\boldsymbol{\theta}, \mathbf{X}) g^{[1]}(\boldsymbol{\theta}, \mathbf{X})']$. When $f(\cdot)$ is unknown, the information bound is bigger. From Example 1, p. 105, Bickel *et al.* (1993), the efficient score function for this problem is given by

$$I^* = (g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}) - E(g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X})))f^{(1)}(Z)/f(Z).$$

Also, under fair conditions (Corollary 1, p. 72, Bickel *et al.*, 1993), the efficient influence function is given by $\tilde{I} = ||I^*||^{-2}I^*$, while the information bound can be written as

$$I^{-1}(F|T,\mathcal{F}) = ||\tilde{I}||^2 := E_{\boldsymbol{\theta}^*}(\tilde{I}\tilde{I}'), \quad \text{or} \quad I(F|T,\mathcal{F}) = E[(f^{(1)}(Z)/f(Z))^2]\Omega_g(\boldsymbol{\theta}^*).$$

By Theorem 3 the asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}_n$ under our semiparametric model is given by $\Omega(\boldsymbol{\theta}^*) = I^{-1}(F|T,\mathcal{F})$, which does not depend on the kernel $K(\cdot)$, achieves the information lower bound for this problem and thus $\hat{\boldsymbol{\theta}}_n$ is efficient. In contrast, the information lower bound may not hold for many estimators which depend on some other structure than the data. For instance, many kernel type estimators have an asymptotic variance with $\int K^2(u)du$ as a factor in it (Fan *et al.*, 1994; Fan & Gijbels, 1994), which can take an arbitrarily value in $(0,\infty)$ by choosing $K(\cdot)$. It is easy to see that $I^{-1}(F|T,\mathcal{F}) \geq I^{-1}(\boldsymbol{\theta})$, when the equality holds, and if there exists an estimator which achieves the Fisher information lower bound, such an estimate is called *adaptive*. Clearly, adaptation is a property of the model parametrization. Begun *et al.* (1983) gave necessary conditions (Corollary 3.1) for adaptation in the case of semiparametric estimation.

6 Numerical studies

In this section, we present the results of two Monte Carlo experiments to show the finite sample behaviour of the MLE $\hat{\boldsymbol{\theta}}_n$ vis-*à*-vis the linear least squares (LS) estimator, and the nonlinear least squares (NLS) estimator through the use of various performance measures. We also evaluate the NLR statistic λ_n defined in Subsection 4.2. Throughout the simulations, we use the biweight kernel $K(u) = \frac{15}{16}(1-u^2)^2\chi(|u| \leq 1)$. The experiments were carried out in Fortran using IMSL-Fortran subroutines UMINF (minimization of a function of k variables using a quasi-Newton method and a finite difference gradient), RNLIN (fitting a nonlinear regression model), RCOVB (computing the estimated asymptotic covariance matrix of the estimated NLR parameters), RNNOR (generating pseudorandom numbers from a standard normal distribution using an inverse CDF method), and RNGAM (generating pseudorandom numbers from a standard gamma distribution).

6.1 Bandwidth selection

Implementing the MLE $\hat{\boldsymbol{\theta}}_n$ requires a method for choosing the value of the bandwidth h_n. For the kernel density estimator $f_n(\cdot)$, there is a vast literature on this topic, ranging from simple to involved methods. But none of the methods has overall advantage (Turlach, 2006). For the problem under study, a good estimator of the regression parameters relies on a good estimator of the error density. A simple and convenient method to use in this case is the so-called *rule of thumb* bandwidth selector of Deheuvels (1977), which is given by $h_n = 1.06\hat{\sigma}_n n^{-1/5}$ where $\hat{\sigma}_n$ is the standard deviation of the data. This choice of h_n does not affect the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$, and satisfies conditions (A2), (A12) and (B3).

Now, let $\boldsymbol{\theta}^0$ be an initial estimate of $\boldsymbol{\theta}$, e.g. the least squares estimate. Then $\hat{\sigma}_n^2$ can be computed as $\hat{\sigma}_n^2 = \sum_{i=1}^n (Z_i^* - \overline{Z^*})^2/(n-1)$, where $Z_i^* = Y_i - g(\boldsymbol{\theta}^0, \mathbf{X}_i)$, and $\overline{Z^*} = (1/n)\sum_{i=1}^n Z_i^*$. Given this result, the bandwidth h_n can be computed prior to the maximization of the pseudo-likelihood (4). Next, in each step of the numerical optimization routine the value of h_n can be adjusted, given intermediate estimates of the true parameter $\boldsymbol{\theta}^*$. This choice of the bandwidth is based mainly on convenience. It may not be optimal, and as such a topic of future study.

6.2 Results

Example 1. Consider the regression model (1) with

$$E(Y_i|X_i, \boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*, X_i) = \theta_0^* + \theta_1^* X_i, \tag{19}$$

where $\boldsymbol{\theta}^* = (\theta_0^*,\ \theta_1^*)'$. In a parametric setting, this model corresponds with $Y_i = \theta_0 + \theta_1 X_i + \epsilon_i$. In the case of (19) the errors ϵ_i's must be distributed symmetrically about 0, as pointed out in Section 1. Here we take $\epsilon_i \overset{\text{i.i.d.}}{\sim} (0.5N(-1,1) + 0.5N(1,1))\chi(-10,10)$, the truncated normal mixture. For the linear regression model we assume $Z_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, with σ^2 unknown. To avoid the identifiability problem we employ the nonlinear function $r(v) = 10e^v/(1+e^v)$. For each simulation, we draw 1000 random samples of size $n = 500$ of (Y_i, X_i) with $X_i \overset{\text{i.i.d.}}{\sim} N(1,1)\chi(-11,10)$.

Table 1 about here

Averaged over all replications, Table 1 shows the MLE $\hat{\boldsymbol{\theta}}_n$ and the LS squares estimates, denoted by $\overline{\boldsymbol{\theta}}_n$, for some selected values of $\boldsymbol{\theta}^*$. Further we present values of the empirical mean-squared errors (MSEs) for each component of the estimators. We observe that the MLE $\hat{\boldsymbol{\theta}}_n$ and the LS estimator $\overline{\boldsymbol{\theta}}_n$ perform equally well.

Table 2 about here

Example 2. Consider the regression model (1) with the following two specifications for the conditional mean function

$$E(Y_i|X_i,\boldsymbol{\theta}_1^*) \;=\; g(\boldsymbol{\theta}_1^*,X_i)=\theta_1^*\exp(\theta_2^*X_{1i}), \tag{20}$$

$$E(Y_i|\boldsymbol{X}_i,\boldsymbol{\theta}_2^*) \;=\; g(\boldsymbol{\theta}_2^*,\boldsymbol{X}_i)=\theta_1^*\exp(\theta_2^*X_{1i})+\theta_3^*\exp(\theta_4^*X_{2i}), \tag{21}$$

where $\boldsymbol{\theta}_1^*=(\theta_1^*,\,\theta_2^*))'$ and $\boldsymbol{\theta}_2^*=(\theta_1^*,\,\theta_2^*,\,\theta_3^*,\,\theta_4^*)'$. Model (20) was fitted to empirical data by Neter *et al.* (1983, pp. 475-478). Model (20) may be considered as a generalization of (21). We sample the Z_i's from a standard gamma distribution with density function $f(z)=z\exp(-z)$, $z>0$. Since the functions $g(\cdot,\cdot)$ have no constant term, we don't have the identifiability problem. So we can take $r(\cdot)$ to be the identity function. Furthermore, for each simulation we draw random samples of size n of $(Y_i,\boldsymbol{X}_i)$ with, in the case of (20), $X_{1i}\overset{\text{i.i.d.}}{\sim}N((1.5,1)\chi(-8\le x_1\le 5)$, and, in the case of (21), $\boldsymbol{X}_i\overset{\text{i.i.d.}}{\sim}N((1.5,4.6)',\boldsymbol{I}_2)\chi(-8\le x_1\le 5;-3\le x_2\le 10)$. The true parameters are taken as $\boldsymbol{\theta}_1^*=(60,-0.03)'$ and $\boldsymbol{\theta}_2^*=(1,-1.4,1,0.8)'$.

Figure 1 about here

Averaged over all 1000 replications, Table 2 shows the empirical means and MSEs of $\hat{\boldsymbol{\theta}}_n$ for $n=100,300$, and 500. And, with the same conditional mean functions $g(\cdot,\cdot)$, the means and MSEs of the NLS estimate $\tilde{\boldsymbol{\theta}}_n$. These latter results are based on $Z_i\overset{\text{i.i.d.}}{\sim}N(0,\sigma^2)$ with σ^2 unknown. For $\boldsymbol{\theta}_1^*$ we see that, in terms of mean parameter values, the performance of the MLE $\hat{\boldsymbol{\theta}}_n$ and the NLS estimator $\tilde{\boldsymbol{\theta}}_n$ is very similar. However, the NLS estimator has slightly lower MSE values for θ_1^*. On the other hand, for θ_2^*, the lowest MSE values are obtained by the MLE estimator. Figure 1 shows plots of the mean estimated standard error of $\hat{\boldsymbol{\theta}}_n$ (solid lines) obtained from the 1000 simulations, and the asymptotic standard errors (dashed lines) of the estimated parameters, using (15), for sample sizes $n=100,150,\ldots,500$. Asymptotic results appear to take effect at about $n=200$.

Figure 2 about here

Note that for $\boldsymbol{\theta}_2^*$ the performance of the MLE and NLS estimators is quite different. We see that the semiparametric MLE $\hat{\boldsymbol{\theta}}_n$ is performing well, and gives much more accurate estimates of the parameters $\boldsymbol{\theta}^*$ than the NLS estimator $\tilde{\boldsymbol{\theta}}_n$. This is confirmed by the reported MSE values. To verify the accuracy of our estimator of the asymptotic variance-covariance matrix $\Omega(\boldsymbol{\theta}^*)$, Figure 2 shows plots of the mean estimated standard error of $\hat{\boldsymbol{\theta}}_n$ (solid lines) and the asymptotic standard errors (dashed lines) of the estimated parameters, for sample sizes $n=100,150,\ldots,500$. Clearly, in this case, our asymptotic results can be employed at about $n = 400$. These results are typical for other parameter vectors $\boldsymbol{\theta}_2^*$, and sample sizes.

Example 3. To gain insight in the performance of the test statistic λ_n, we consider the mean function

$$E(Y_i|\mathbf{X}_i, \boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*, \mathbf{X}_i) = \theta_1^* \exp(\theta_2^* X_{1i}) + \theta_3^* X_{2i} \exp(\theta_4^* X_{2i}).$$

Similar to Example 2, the Z_i's are sampled from a standard gamma distribution. The null hypothesis $H_0 : \theta_4^* = 0$ will be investigated versus the alternative $H_1 : \theta_4^* \neq 0$. The parameter vector of interest is $\boldsymbol{\theta}^* = (1, -1.4, 1, 0)'$. According to Theorem 5, the distribution of λ_n should be asymptotically χ_1^2-distributed. To verify this empirically, we plot the quantile of the 1000 computed statistics against the quantile of the χ_1^2-distribution. Figure 3 shows the Q-Q plots for $n = 100$ and $n = 200$. The reference lines are chosen to pass through the 25% and 75% data quantiles. The plots depict the λ_n statistic closely following the χ_1^2-distribution, with better results for the case $n = 200$ than for $n = 100$; this is consistent with the asymptotic theory.

Figure 3 about here

7 Discussion

7.1 Mean function

Specification of the mean function is an important issue for any parametric/semiparametric regression method. Since the error distribution is model free, a mis-specification of the mean function seems to be less serious as in the parametric case. However, it may still result in inferior inference as for any parametric/semiparametric regression method. How to select the mean function $g(\cdot, \cdot)$ is a practical issue. A good mean function should fit the data well and be as simple (smooth) as possible. By fitting the data well, we mean the corresponding log semiparametric likelihood, evaluated at the MLE, is relatively large. But typically, this will favor a sophisticated mean function. So there is a trade-off between goodness-of-fit and the complexity of the model. Let $\mathcal{G} = \{g(\cdot, \cdot)\}$ be a finite set of available optional mean functions we are to use for the problem under study. Different g_i's in $\mathcal{G}$ may have the same parametric dimension but different parametrizations, or different parametric dimensions and the corresponding model not necessarily be nested. So our case is different from the common model selection, which chooses the optimal parameter dimension among a set of nested models. In our case, we suggest to select the optimal $g \in \mathcal{G}$ in the sense of balancing the goodness-of-fit and its complexity. The commonly used AIC or BIC do not apply, since here the models under consideration may not be nested. Rather we suggest the use of Rissanen's (1996) minimal description length criterion to choose the mean function. A full investigation of this criterion awaits further research.

7.2 Robustness

Although in Subsection 2.2 we discussed some related estimation methods, the literature on robust statistics was not addressed. The reason is that robust parametric statistics tend to rely on replacing the normal distribution by the t-distribution, with low degrees of freedom (high kurtosis). Our method is more general in this respect. Also, the focus and treatment are different between robust regression and the semiparametric regression method presented here. The former is more focused on treatment of breakdown points (the proportion of "incorrect" observations), using the empirical influence function and sensitivity curves to evaluate the behavior of the estimator(s). It also deals with the construction of robust algorithms, usually called M-estimator, which needs the specification of a ρ function to replace the likelihood function. For our method, we use a kind of "empirical likelihood" to replace the subjective specification of the likelihood, with the main focus on consistent estimates of the parameters, no matter what the error specification is, and robust with respect to uncertainty on $f(\cdot)$.

7.3 Concluding remarks

The semiparametric MLE method studied in this paper, allows for a great deal of generality in the error specification. The method is ready to use because of the Wilks property. It is powerful since it achieves optimal rates of convergence. The MLE achieves the lower bound for the class of semiparametric estimators, and hence is efficient in this class. This information lower bound is generally bigger than or equal to the inverse of the Fisher information. When the information lower bound equals the inverse of the Fisher information under some model specification, the MLE is "adaptive" by definition. But in general adaptation is *not* possible.

Based on our simulated examples, the semiparametric MLE method accurately estimates the true parameters, while the traditional NLS may fail (see, e.g., Example 2). However, the suggested naive estimate of the asymptotic variance is not satisfactory in Example 2 for $n < 400$. This may due to the difficulty in estimating the factor $E[f^{(1)}(Z)/f(Z)]^2$ in the asymptotic variance. Recall, in Remark 6, we suggested the estimate $(1/n)\sum_{i=1}^{n}[f_{n,i}^{(1)}(z_i)/f_{n,i}(z_i)]^2$. It is known that $f_n(\cdot)$ is consistent for $f(\cdot)$ does not necessarily mean that $f_n^{(1)}(\cdot)$ is consistent for

$f^{(1)}(\cdot)$. Generally, $f_n^{(1)}(z) \sim h_n^{-1} \int K^{(1)}(u)du - f^{(1)}(z) \int uK^{(1)}(u)du + o(1)$, and $f_n^{(1)}(z) \to f^{(1)}(z)$ only if $\int K^{(1)}(u)du = 0$ and $\int uK^{(1)}(u)du = -1$. The first condition above is asserted by (A9), the second is also true (see 2nd line in the proof of Theorem 3). So the conditions imply $f_n^{(1)}(z) \to f^{(1)}(z)$ (a.s.). But we still don't know if this convergence is uniform, or the convergence speed (usually slower than that of $f_n(\cdot)$ to $f(\cdot)$). Here we require more, i.e. we need the uniform convergence of $[f_n^{(1)}(\cdot)/f_n(\cdot)]^2$ to $[f^{(1)}(\cdot)/f(\cdot)]^2$, and this has not been fully investigated.

Acknowledgments: The authors are grateful to Professor Ali Gannoun, CNAM-Paris, for useful comments on an earlier version of this paper. They also thank the Editor and two referees, whose comments greatly improved the presentation of the results.

References

Andrews, D.W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **62**, 43-72.

Bahadur, R.R. & Zabell, S.L. (1979). Large deviations of the sample mean in general vector spaces. *Ann. Probab.* **7**, 587-621.

Begun, J.M., Hall, W.J., Huang W.-M. & Wellner, J.A.(1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-452.

Beran, R. (1974). Asymptotically efficient rank estimates in location models. *Ann. Statist.* **2**, 63-74.

Beran, R. (1978). An efficient and robust adaptive estimator of location, *Ann. Statist.* **6**, 292-313.

Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Bickel, P.J. & Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā*, Ser. A **50**, 381-393.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins University Press, Baltimore, Maryland.

Billingsley, P. (1968). *Convergence in probability measures*, Wiley, New York.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.*, **74**, 823-836.

Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269-276.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés, *Rev. Statist. Appl.* **35**, 5-42.

Devroye, L. & Györfi, L. (1985). *Nonparametric density estimation, the L_1 view.* Wiley, New York.

Doob, J.L. (1953). *Stochastic processes.* Wiley, New York.

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classic multinomial estimator. *Ann. Math. Statist.* **27**, 642-669.

Efromovich, S. (1996). On nonparametric regression for IID observations in a general setting. *Ann. Statist.* **24**, 1126-1144.

Eggermont, P.P.B. & LaRiccia, V.N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inf. Theory* **45**, 1321-1326.

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression.* Marcel Dekker, New York.

Fan, J., Hu, T. & Truong, Y.K. (1994). Robust nonparametric function estimation. *Scand. J. Statist.* **21**, 433-446.

Fan, J. & Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89**, 560-570.

Fan, J., Zhang, C. & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193. Correction **30**, 1811.

Gasser, T. & Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation* (eds. T. Gasser and M. Rosenblatt), Springer, New York, 23-68.

Györfi, L. & Van der Meulen E.C. (1990). On the nonparametric estimation of the entropy functional. In *Nonparametric functional estimation and related topics* (ed. George Roussas), pp. 81-95, NATO ASI Series, Ser. C: Mathematical and Physical Science, Vol. 335, Kluwer Academic Publishers Dordrecht/Boston/London.

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41**, 683-697.

Hall, P. (1986). On powerful distributional tests based on sample spacings. *J. Multivariate Statist.* **19**, 201-225.

Hall, P. & Marron, J.S. (1990). On variance estimation in nonparametric regression, *Biometrika* **77**, 521-528.

Hall, P. & Morton, S.C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45**, 69-88.

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits, *Ann. Statist.* **21**, 1926-1947.

Hoeffding, J.B.S. (1963). Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58**, 13-30.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1**, 221-233.

Ibragimov, I.A., Nemirovskii, A.S. & Khas'minskii, R.Z. (1986). Some problems on nonparametric estimation in Gaussian white noise. *Theory of Probab. and Its Appl.* **31**, 391-406.

Kotz, S. & Johnson, N.L. (1982). *Encyclopedia of statistical sciences*, Vol. 6. Wiley, New York.

Manski, C. (1984). Adaptive estimation of nonlinear regression models. *Econometric Reviews* **3(2)**, 145-194.

Murphy, S.A. & Van der Vaart, A.W. (2000). On profile likelihood, *J. Amer. Statist. Assoc.* **95**, 449-485.

Müller, U.U., Schick, A. & Wefelmeyer, W. (2004). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *manuscript.*

Nadaraya, E.A. (1964). On estimating regression, *Theory of Probab. and Its Appl.* **9**, 141-142.

Neter, J., Wasserman, W. & Kutner, M.H. (1983). *Applied linear regression models*, Irwin, Homewoods.

Newey, W.K. (1988). Adaptive estimation of regression models via moment restrictions. *J. of Economics* **38**, 301-339.

Owen, A. (1991). Empirical likelihood for linear models, *Ann. Statist.* **19**, 1725-1747.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimators, *Metrika* **14**, 249-272.

Priestley, M.B. & Chao, M.T. (1972). Non-parametric function fitting, *J. Roy. Stat. Soc.*, Ser. B **34**, 385-392.

Rao, B.L.S. (1983). *Nonparametric functional estimation.* Academic Press: Orlando, Florida.

Rice, J. (1984). Boundary modification for kernel regression, *Commun. Statist. Theor. Meth.* **13**, 893-900.

Rissanen, J. (1996). Fisher information and stochastic complexity, *IEEE Transactions on Information Theory* **42**, 40-47.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486-1521.

Schick, A. & Wefelmeyer, W. (2004). Root n consistent and optimal density estimators for moving average processes. *Scand. J. Statist.* **31**, 63-78.

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

Severini, T.A. & Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models, *J. Amer. Statist. Assoc.* **89**, 501-511.

Severini, T.A. & Wong, W.H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768-1802.

Sievers, G.I. (1978). Weighted rank statistics for simple linear regression, *J. Amer. Statist. Assoc.* **73**, 628-631.

Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.* **3**, 267-284.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8**, 1348-1360.

Turlach, B.A. (2006). Bandwidth selection in kernel density estimation: A review, http://citeseer.ist.psu.edu/214125.html

Van Eden, C. (1973). Efficient-robust estimation of location, *Ann. Math. Statist.* **41**, 172-181.

Van Es, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings, *Scand. J. Statist.* **19**, 61-72.

Watson, G.S. (1964). Smooth regression analysis, *Sankhyā*, Ser. A **26**, 359-386.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann Math. Statist.* **9**, 60-62.

Wolsztynski, E.W., Thierry, E. & Pronzato, L. (2005). Minimum-entropy estimation in semi-parametric models, *Signal Processing* **85**, 937-949.

Jan G. De Gooijer, Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands.
E-mail: j.g.degooijer@uva.nl

Appendix

Lemma 1. *Assume (B3), (A3), (A4), (A7), (A9) and (A10) hold. Then for any compact set* $\mathbf{A}$ *satisfying (A8),*

$$\sup_{\boldsymbol{\theta}\in\mathbf{A}} ||L_n^{[2]}(\boldsymbol{\theta}) - \gamma_1^2 E[(f^{(1)}(Z)/f(Z))^2]\Omega_g(\boldsymbol{\theta})|| \to 0, \quad a.s.$$

Proof: Plugging in the expressions for $f_{(n,i)}^{[1]}(\cdot)$ and $f_{(n,i)}^{[2]}(\cdot)$, we have

$$\begin{aligned} L_n^{[2]}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1} \frac{f_{(n,i)}^{[2]}(Z_i|\boldsymbol{\theta})}{f_{(n,i)}(Z_i|\boldsymbol{\theta})} - \frac{1}{n}\sum_{i=1} \frac{f_{(n,i)}^{[1]}(Z_i|\boldsymbol{\theta})(f_{(n,i)}^{[1]}(Z_i|\boldsymbol{\theta}))'}{f_{(n,i)}^2(Z_i|\boldsymbol{\theta})} \\ &= \frac{1+o(1)}{n}\sum_{i=1} \frac{f_{(n,i)}^{[2]}(Z_i|\boldsymbol{\theta})}{f(Z_i|\boldsymbol{\theta})} - \frac{1+o(1)}{n}\sum_{i=1} \frac{f_{(n,i)}^{[1]}(Z_i|\boldsymbol{\theta})(f_{(n,i)}^{[1]}(Z_i|\boldsymbol{\theta}))'}{f^2(Z_i|\boldsymbol{\theta})} \end{aligned}$$

$$= (1+o(1))\sum_{i=1}\sum_{j\neq i} \frac{1}{n(n-1)h_n^3} \frac{K^{(2)}(\frac{Z_i-Z_j}{h_n})(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_j))(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_j))'}{f(Z_i|\boldsymbol{\theta})}$$

$$+(1+o(1))\sum_{i=1}\sum_{j\neq i} \frac{1}{n(n-1)h_n^2} \frac{K^{(1)}(\frac{Z_i-Z_j}{h_n})(g^{[2]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[2]}(\boldsymbol{\theta},\mathbf{X}_j))}{f(Z_i|\boldsymbol{\theta})}$$

$$-\sum_{i=1}\sum_{j\neq i}\sum_{l\neq i,j} \frac{1+o(1)}{n(n-1)^2h_n^4} \frac{K^{(1)}(\frac{Z_i-Z_j}{h_n})K^{(1)}(\frac{Z_i-Z_l}{h_n})(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_j))(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_l))'}{f^2(Z_i|\boldsymbol{\theta})}$$

$$-\sum_{i=1}\sum_{j\neq i} \frac{1+o(1)}{n(n-1)^2h_n^4} \frac{K^{(1)^2}(\frac{Z_i-Z_j}{h_n})(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_j))(g^{[1]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[1]}(\boldsymbol{\theta},\mathbf{X}_j))'}{f^2(Z_i|\boldsymbol{\theta})}$$

$$:= (1+o(1))U_{n,n}^{(1)}(\boldsymbol{\theta}) + (1+o(1))U_{n,n}^{(2)}(\boldsymbol{\theta}) - (1+o(1))\frac{n-2}{n-1}U_{n,n}^{(3)}(\boldsymbol{\theta}) - R_n(\boldsymbol{\theta}),$$

where $R_n(\boldsymbol{\theta})$ is the last term in the previous equation. As in the proof of Theorem 2, for each fixed n, the $U_{n,m}^{(k)}(\boldsymbol{\theta})$'s are U-statistics and reverse martingales with respect to $\{\mathcal{F}_m\}$, and so

$$\begin{aligned} U_{n,n}^{(1)}(\boldsymbol{\theta}) &\overset{a.s.}{\to} 2\Omega_g(\boldsymbol{\theta})\lim_n \int\int h_n^{-3}K^{(2)}(\frac{z_i-z_j}{h_n})f(z_j)dz_i dz_j \\ &= 2\Omega_g(\boldsymbol{\theta})\int\int u^2K^{(2)}(u)f^{(2)}(v)dudv = 2\Omega_g(\boldsymbol{\theta})\gamma_2\int f^{(2)}(v)dv = \mathbf{0}, \\ U_{n,n}^{(2)}(\boldsymbol{\theta}) &\overset{a.s.}{\to} E[g^{[2]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[2]}(\boldsymbol{\theta},\mathbf{X}_j)]\int\int h_n^{-2}K^{(1)}(\frac{z_i-z_j}{h_n})f(z_j)dz_i dz_j \\ &= -E[g^{[2]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[2]}(\boldsymbol{\theta},\mathbf{X}_j)]\int\int uK^{(1)}(u)f^{(1)}(v)dudv \\ &\quad -\gamma_1\int f^{(1)}(v)dvE[g^{[2]}(\boldsymbol{\theta},\mathbf{X}_i) - g^{[2]}(\boldsymbol{\theta},\mathbf{X}_j)] = \mathbf{0}. \end{aligned}$$

In the above we used (A4), (A7) and the fact that, by (A3) and (A9),

$$\int\int h_n^{-2}K^{(1)}(\frac{z_1-z_2}{h_n})f(z_2)dz_1dz_2 = \int\int h_n^{-1}K^{(1)}(u)f(z-h_nu)dzdu$$

$$= \int\int h_n^{-1} K^{(1)}(u)[f(z) - f^{(1)}(z)h_n u + O(h_n^2) f^{(2)}(z)u^2] dz du = -\gamma_1 \int f^{(1)}(v)dv + O(h_n).$$

Similarly, by (A10)

$$U_{n,n}^{(3)}(\boldsymbol{\theta}) \overset{a.s.}{\to} \Omega_g(\boldsymbol{\theta}) \lim_n \int\int\int h_n^{-4} \frac{K^{(1)}(\frac{z_1-z_2}{h_n})K^{(1)}(\frac{z_1-z_3}{h_n})f(z_2)f(z_3)}{f(z_1)} dz_1 dz_2 dz_3$$

$$= \Omega_g(\boldsymbol{\theta})\gamma_1^2 E[(f^{(1)}(Z)/f(Z))^2].$$

Also, since there are $n(n-1)$ terms in $R_n(\boldsymbol{\theta})$, with a dividing factor $n(n-1)^2 h_n^4$, and by (B3), $(n-1)h_n^4 \to \infty$, so we have $\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} R_n(\boldsymbol{\theta}) \overset{a.s.}{\to} \mathbf{0}$. Now we centralize the $U_{n,n}^{(i)}(\boldsymbol{\theta})$'s, and by the same way as in the proof of (6), we can prove

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} ||U_{n,n}^{(i)}(\boldsymbol{\theta}) - E(U_{n,n}^{(i)}(\boldsymbol{\theta}))|| \to 0, \quad a.s.$$

Thus, collecting terms, we complete the proof.

Lemma 2. *Let $\tilde{K}_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j)$ as given by (12) and (13), and $U_n = 2(n(n-1))^{-1} \sum_{1 \le i < j \le n} \tilde{K}_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j)$. Let $U_{n,r}$ be the r-th component of U_n. Under conditions (B2)-(B4), (A1) and (A4)-(A7) it holds that, for some $0 < C < \infty$, $Var(U_{n,r}) = C(n^2 h_n^3)^{-1} \quad (r = 1, \dots, dim(\boldsymbol{\theta}))$.*

Proof: Since $U_{n,r}$ is U_n with $g^{[1]}(\cdot, \cdot)$ replaced by $g_r^{[1]}(\cdot, \cdot)$, the r-th component of $g^{[1]}(\cdot, \cdot)$. The proof is the same by assuming $\boldsymbol{\theta}$ as one-dimensional, and so for $g^{[1]}(\cdot, \cdot)$ in the rest proof. Recall

$$\tilde{K}_n(Z_i, \mathbf{X}_i, Z_j, \mathbf{X}_j) = K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j) - \overline{K}_n(Z_i, \mathbf{X}_i) - \overline{K}_n(Z_j, \mathbf{X}_j)$$

and $\overline{K}_n(Z_i, \mathbf{X}_i) = E(K_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j)|Z_i, \mathbf{X}_i)$. Let $H_{n1}(Z_1, \mathbf{X}_1) = E(\tilde{K}_n(Z_1, \mathbf{X}_1; Z_2, \mathbf{X}_2)|Z_1, \mathbf{X}_1)$, $\eta_{n1} = Var(H_{n1}(Z_1, \mathbf{X}_1))$ and $\eta_{n2} = Var(\tilde{K}_n(Z_i, \mathbf{X}_i; Z_j, \mathbf{X}_j))$. By Lemma A in Serfling (1980, p. 183) we have

$$Var(U_n) = \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum_{r=1}^{2} \begin{pmatrix} 2 \\ r \end{pmatrix} \begin{pmatrix} n-2 \\ 2-r \end{pmatrix} \eta_{nr}.$$

By (A4) and (A7),

$$\begin{aligned} E(\overline{K}_n(Z_i, \mathbf{X}_i)) = E(U_n) \quad &= \quad \frac{1}{h_n^2} \int\int\int\int \frac{K^{(1)}(\frac{z_j - z_i}{h_n})[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{x}_i) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{x}_j)]}{f(z_i)} \\ &\times f(z_i) q(\mathbf{x}_i) f(z_j) q(\mathbf{x}_j) dz_i d\mathbf{x}_i dz_j d\mathbf{x}_j = 0, \end{aligned}$$

so $H_{n1}(Z_1, \mathbf{X}_1) = E(\overline{K}_n(Z_2, \mathbf{X}_2)) \equiv 0$, $\eta_{n1} = 0$, and

$$\eta_{n2} = E\bigg(\tilde{K}_n(Z_1, \mathbf{X}_1; Z_2, \mathbf{X}_2)\bigg)^2 = E(K_n^2(Z_1, \mathbf{X}_1; Z_2, \mathbf{X}_2)) + 2E(\overline{K}_n^2(Z_1, \mathbf{X}_1))$$

$$-4E(K_n(Z_1,\mathbf{X}_1;Z_2,\mathbf{X}_2)\overline{K}'_n(Z_1,\mathbf{X}_1))+2E(\overline{K}_n(Z_1,\mathbf{X}_1)\overline{K}'_n(Z_2,\mathbf{X}_2)). \qquad (22)$$

By (B2) and (A5), $K(\cdot)$ and $f(\cdot)$ have compact support. So by the continuity of $f(\cdot)$ (which is implied by (A3)) and the dominated convergence theorem we have

$$\begin{aligned}E(K_n^2(Z_1,\mathbf{X}_1;Z_2,\mathbf{X}_2)) &= \frac{2\Omega_g(\boldsymbol{\theta}^*)}{h_n^3}\int\int\left(K^{(1)}(u)\right)^2\frac{f(z+uh_n)}{f(z)}dudz\\ &= O(1)\frac{2\Omega_g(\boldsymbol{\theta}^*)}{h_n^3}\int\int\left(K^{(1)}(u)\right)^2dudz.\end{aligned}$$

Since (A1) and (B2) imply $\int\left(K^{(1)}(u)\right)^2du<\infty$, we get for some $0<C<\infty$,

$$E(K_n^2(Z_1,\mathbf{X}_1;Z_2,\mathbf{X}_2))\le Ch_n^{-3}.$$

Also,

$$\overline{K}_n(Z_1,\mathbf{X}_1)=\frac{[g^{[1]}(\boldsymbol{\theta}^*,\mathbf{X}_1)-\mu_g(\boldsymbol{\theta}^*)]}{2h_n}\int K^{(1)}(u)\frac{f(Z_1+h_nu)-f(Z_1)}{f(Z_1)}du.$$

Thus

$$\begin{aligned}E\Big(K_n(Z_1,\mathbf{X}_1;Z_2,\mathbf{X}_2)\overline{K}_n(Z_1,\mathbf{X}_1)\Big) &= \frac{\Omega_g(\boldsymbol{\theta}^*)}{4h_n^3}\Big(\int\int K^{(1)}(\frac{z_2-z_1}{h_n})\frac{f(z_2)}{f(z_1)}\int K^{(1)}(u)[f(z_1+h_nu)\\ &\quad -f(z_1)]dudz_1dz_2-\int\int K^{(1)}(\frac{z_1-z_2}{h_n})\int K^{(1)}(u)[f(z_1+h_nu)-f(z_1)]dudz_1dz_2\Big)\\ &= \frac{\Omega_g(\boldsymbol{\theta}^*)}{4h_n^2}\int\int\int K^{(1)}(v)K^{(1)}(u)[f(z+h_nu)-f(z)][\frac{f(z+h_nv)+f(z)}{f(z)}]dvdudz\\ &\sim \frac{\Omega_g(\boldsymbol{\theta}^*)}{4h_n}\int\int\int K^{(1)}(v)uK^{(1)}(u)(f^{(1)}(z)/f(z))[f(z+h_nv)+f(z)]dvdudz.\end{aligned}$$

By the compactness of the supports of $K(\cdot)$ and $f(\cdot)$, and (B4), the above integration is $O(1)h_n^{-1}$. Thus there is a $0<C<\infty$, such that for large n, the absolute value of the above term is bounded by Ch_n^{-1}. Similarly,

$$E[\overline{K}_n^2(Z_1,\mathbf{X}_1)]=\frac{1}{4h_n^2}\Omega_g(\boldsymbol{\theta}^*)\int\left(\int K^{(1)}(u)\frac{f(z+h_nu)-f(z)}{f(z)}du\right)^2f(z)dz=o(1)h_n^{-2},$$

and there is a $0<C<\infty$, such that for large n,

$$|E[\overline{K}_n^2(Z_1,\mathbf{X}_1)]|\le Ch_n^{-2}.$$

Thus, the leading term in (22) is of order h_n^{-3}. So we have, for some $0<C<\infty$,

$$Var(U_n)=\begin{pmatrix} n\\ 2\end{pmatrix}^{-1}\eta_{n2}=\frac{C}{n^2h_n^3}.$$

Lemma 3. *Let $\overline{K}_n(Z_i, \mathbf{X}_i)$ be as given prior to (13). Under conditions (B2), (B3), (A3), (A4), (A7), (A8)-(A10), we have* $\sqrt{n}\frac{2}{n}\sum_{i=1}^n \overline{K}_n(Z_i, \mathbf{X}_i) \xrightarrow{D} N(\mathbf{0}, \Omega^{-1}(\boldsymbol{\theta}^*))$.

Proof: We have

$$\overline{K}_n(Z_i, \mathbf{X}_i) = -\frac{[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_i) - \mu_g(\boldsymbol{\theta}^*)]}{2h_n f(Z_i)} \int K^{(1)}(u)[f(Z_i + h_n u) - f(Z_i)]du.$$

We need to show that for any constant vector $\boldsymbol{a} = (a_1, \ldots, a_{k+1})'$,

$$\sqrt{n}\frac{2}{n}\sum_{i=1}^n \boldsymbol{a}'\overline{K}_n(Z_i, \mathbf{X}_i) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{a}'\Omega^{-1}(\boldsymbol{\theta}^*)\boldsymbol{a}).$$

Rewrite the above double array average as (omitting the negative sign in $\overline{K}_n$)

$$\frac{2}{n}\sum_{i=1}^n \boldsymbol{a}'\overline{K}_n(Z_i, \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^n V_{n,i}, \quad V_{n,i} = \frac{\boldsymbol{a}'[g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_i) - \mu_g(\boldsymbol{\theta}^*)]}{h_n f(Z_i)} \int K^{(1)}(u)[f(Z_i+h_n u)-f(Z_i)]du.$$

For each fixed n, the $V_{n,i}$'s are *i.i.d.* random variables, with $E(V_{n,i}) = 0$. We only need to check the Lindeberg condition for $\{V_{n,i}\}$ (e.g. Serfling 1980, pp. 31-32). By (B2), (B3), (A3), (A4), (A7), (A8) and (A10), and the fact that $h_n \to 0$, we have

$$\begin{aligned} Var(V_{n,1}) &= E(V_{n,1}^2) = \frac{\boldsymbol{a}'\Omega_g(\boldsymbol{\theta}^*)\boldsymbol{a}}{h_n^2} \int \frac{1}{f(z)}\Big(\int K^{(1)}(u)[f(z + h_n u) - f(z)]du\Big)^2 dz \\ &= \boldsymbol{a}'\Omega_g(\boldsymbol{\theta}^*)\boldsymbol{a} \int \frac{1}{f(z)}\Big(\int uK^{(1)}(u) f^{(1)}(z + \xi_n u)du\Big)^2 dz \to \boldsymbol{a}'\Omega_g(\boldsymbol{\theta}^*)\boldsymbol{a}\gamma_1^2 E(\frac{f^{(1)}(Z)}{f(Z)})^2, \end{aligned}$$

where in the above ξ_n is an intermediate value between 0 and h_n, which may depend on z. By (B2) and (A5), $K(\cdot)$ and $f^{(1)}(\cdot)$ have compact supports. Then, by the boundedness of $f^{(1)}(\cdot)$, we can use the dominated convergence theorem to get the limit. So,

$$B_n^2 := Var(\sum_{i=1}^n V_{n,i}) = \sum_{i=1}^n Var(V_{n,i}) = n(\boldsymbol{a}'\Omega^{-1}(\boldsymbol{\theta}^*)\boldsymbol{a} + o(1)).$$

Since the $V_{n,i}$'s are *i.i.d.* with $E(V_{n,i}) = 0$, the Lindeberg condition is

$$\frac{1}{B_n^2}\sum_{i=1}^n \int_{|v|>\epsilon B_n} v^2 dF_{n,i}(v) = Var(V_{n,1}\chi(|V_{n,1}| > \epsilon\sqrt{n(\boldsymbol{a}'\Omega^{-1}(\boldsymbol{\theta}^*)\boldsymbol{a} + o(1))})) \to 0, \quad \forall\ \epsilon > 0,$$

since $Var(V_{n,1}) < \infty$, where $F_{n,i}$ is the distribution function of $V_{n,i}$. Thus

$$B_n^{-1}\sum_{i=1}^n V_{n,i} \xrightarrow{D} N(0, 1)$$

.

In the following, let $\tilde{E}f^{[0]}(z) = f(z)$, $\tilde{E}f^{[1]}(z) = \gamma_1 f^{[1]}(z) = \gamma_1[g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - \mu_g(\boldsymbol{\theta})]f^{(1)}(z)$, and for some X *i.i.d.* with the X_i's,

$$\begin{aligned}\tilde{E}f^{[2]}(z) &= \gamma_2 E\Big(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X})\Big)\Big(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X})\Big)' f^{(2)}(z)\\ &\quad + \gamma_1(g^{[2]}(\boldsymbol{\theta}, \mathbf{x}) - \check{\Omega}_g(\boldsymbol{\theta}))f^{(1)}(z),\end{aligned}$$

where the expectation E is with respect to $\mathbf{X}$.

Lemma 4. *Suppose conditions (B2), (A1), (A2), (A4), and (A7) hold, and assume that $f^{(j)}(\cdot)$ is uniformly continuous and $\sup_z |f^{(j)}(z)| < \infty$ ($j = 1, 2$). Then, as $n \to \infty$, $\sup_{\mathbf{x},z} ||f_n^{[j]}(z) - \tilde{E}f^{[j]}(z)|| \to 0$, (a.s.) ($j = 0, 1, 2$).*

Proof: Note

$$f_n^{[1]}(z) = \frac{1}{nh_n^2}\sum_{j=1}^n K^{(1)}(\frac{z - Z_j}{h_n})\Big(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j)\Big)$$

and

$$\begin{aligned}f_n^{[2]}(z) &= \frac{1}{nh_n^3}\sum_{j=1}^n K^{(2)}(\frac{z - Z_j}{h_n})\Big(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{X}_j)\Big)\Big(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}^*, \mathbf{X}_j)\Big)'\\ &\quad + \frac{1}{nh_n^2}\sum_{j=1}^n K^{(1)}(\frac{z - Z_j}{h_n})(g^{[2]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[2]}(\boldsymbol{\theta}^*, \mathbf{X}_j)).\end{aligned}$$

The case of $j = 0$ is implied in Theorem 2.1.3 in Rao (1983). Here we only prove the case $j = 1$. The case $j = 2$ is similar. We have

$$\sup_{\mathbf{x},z} ||f_n^{[1]}(z) - \tilde{E}f^{[1]}(z)|| \le \sup_{\mathbf{x},z} ||f_n^{[1]}(z) - Ef_n^{[1]}(z)|| + \sup_{\mathbf{x},z} ||Ef_n^{[1]}(z) - \tilde{E}f^{[1]}(z)|| := V_{n,1} + V_{n,2}.$$

Let $F_{n1}(\cdot)$ and $F_1(\cdot)$ respectively be the empirical and true distributions of the Z_i's, and $F_{n2}(\cdot)$ and $F_2(\cdot)$ be those for the $\mathbf{X}_i$'s. Note that (A1) implies that $K^{(j)}(\cdot)$ has bounded variation τ_j ($j = 0, 1, 2$), (A4) and (A7) imply that $g^{[j]}(\cdot)$ has bounded variation $\tau_{g,j}$ ($j = 1, 2$), and we have

$$\begin{aligned}V_{n,1} &= \frac{1}{h_n^2}\sup_{\mathbf{x},z} || \int K^{(1)}(\frac{z - y}{h_n})(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{v}))dF_{n1}(y)dF_{n2}(\mathbf{v})\\ &\quad - \int K^{(1)}(\frac{z - y}{h_n})(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{v}))dF_1(y)dF_2(\mathbf{v})||\end{aligned}$$

$$\begin{aligned}&\le \frac{1}{h_n^2}\sup_{\mathbf{x},z}\int |F_{n1}(y)F_{n2}(\mathbf{v}) - F_1(y)F_2(\mathbf{v})||dK^{(1)}(\frac{z - y}{h_n})| \times ||d[g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{v}])||\\ &= \frac{1}{h_n^2}\sup_{z}\int |F_{n1}(y)F_{n2}(\mathbf{v}) - F_1(y)F_2(\mathbf{v})||dK^{(1)}(\frac{z - y}{h_n})| \times ||dg^{[1]}(\boldsymbol{\theta}, \mathbf{v})||\\ &\le \frac{1}{h_n^2}\tau_1\tau_{g,1}\sup_{\mathbf{v},y}|F_{n1}(y)F_{n2}(\mathbf{v}) - F_1(y)F_2(\mathbf{v})|.\end{aligned}$$

By the result on large deviation in Dvoretzky *et al.* (1956), there are positive constants C_r and $0 < \alpha_r \leq 2$ $(r = 1, 2)$, such that

$$P\bigg(\sup_y |F_{n1}(y) - F_1(y)| > \epsilon\bigg) \leq C_1 \exp(-\alpha_1 \epsilon^2 n), \quad \forall\, \epsilon > 0,$$

and

$$P\bigg(\sup_{\mathbf{x}} |F_{n2}(\mathbf{x}) - F_2(\mathbf{x})| > \epsilon\bigg) \leq C_2 \exp(-\alpha_2 \epsilon^2 n), \quad \forall\, \epsilon > 0.$$

Since

$$|F_{n1}(y)F_{n2}(\mathbf{x}) - F_1(y)F_2(\mathbf{x})| \leq F_{n1}(y)|F_{n2}(\mathbf{x}) - F_2(\mathbf{x})| + F_2(\mathbf{x})|F_{n1}(y) - F_1(y)|$$

$$\leq |F_{n2}(\mathbf{x}) - F_2(\mathbf{x})| + |F_{n1}(y) - F_1(y)|,$$

we have

$$\begin{aligned} P\bigg(V_{n,1} > \epsilon\bigg) &\leq P\bigg(\sup_{\mathbf{x},y} |F_{n1}(y)F_{n2}(\mathbf{x}) - F_1(y)F_2(\mathbf{x})| > \epsilon h_n^2 \tau_1^{-1}\bigg) \\ &\leq P\bigg(\sup_y |F_{n1}(y) - F_1(y)| > \frac{\epsilon}{2} h_n^2 \tau_1^{-1}\bigg) + P\bigg(\sup_{\mathbf{x}} |F_{n2}(\mathbf{x}) - F_2(\mathbf{x})| > \frac{\epsilon}{2} h_n^2 \tau_1^{-1}\bigg) \\ &\leq \sum_{k=1}^{2} C_k \exp(-\alpha_k \frac{\epsilon^2}{4} \tau_1^{-2} n h_n^4). \end{aligned}$$

Since by (A2), $\sum_{n=1}^{\infty} e^{-\alpha_k \frac{\epsilon^2}{4} \tau_1^{-2} n h_n^4} < \infty$, by the Borel-Cantelli lemma we have $V_{n,1} \to 0$ (a.s). Note that, for some $0 \leq \theta_n \leq 1$,

$$\begin{aligned} Ef_n^{[1]}(z) &= \frac{1}{h_n^2} \int K^{(1)}(\frac{z-y}{h_n})(g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - g^{[1]}(\boldsymbol{\theta}, \mathbf{v})) f_1(y) f_2(\mathbf{v}) dy d\mathbf{v} \\ &= (g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - \mu_g(\boldsymbol{\theta})) \int u K^{(1)}(u) f^{(1)}(z + \theta_n h_n u) du. \end{aligned}$$

So

$$\begin{aligned} ||Ef_n^{[1]}(z) - \tilde{E}f^{[1]}(z)|| &\leq ||g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - \mu_g(\boldsymbol{\theta})|| \bigg(\sup_{|v| \leq \delta} |f^{(1)}(z+v) - f^{(1)}(z)| \int |uK^{(1)}(u)| du \\ &\quad + \sup_{|v| > \delta h_n^{-1}} |vK^{(1)}(v)| \delta^{-1} \int f^{(1)}(z) dz + \sup_z |f^{(1)}(z)| \int_{|u| > \delta h_n^{-1}} |uK^{(1)}(u)| du \bigg) \\ &= ||g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - \mu_g(\boldsymbol{\theta})|| \bigg(\sup_{|v| \leq \delta} |f^{(1)}(z+v) - f^{(1)}(z)| \int |uK^{(1)}(u)| du \\ &\quad + \sup_z |f^{(1)}(z)| \int_{|u| > \delta h_n^{-1}} |uK^{(1)}(u)| du \bigg). \end{aligned}$$

Since (A2) implies $h_n \to 0$ and $nh_n^8 \to \infty$, by the uniform continuity of $f^{(1)}(\cdot)$, the first term above can be arbitrarily small uniformly in z as δ does. Conditions (A4) and (A7) imply

$\sup_{\mathbf{x}} ||g^{[1]}(\boldsymbol{\theta}, \mathbf{x}) - \mu_g(\boldsymbol{\theta})|| < \infty$. Also, by assumption, $\sup_z |f^{(1)}(z)|$ is finite, so the second term in brackets above is zero for large n by (B2). Thus, $V_{n,2} \to 0$. This completes the proof for $j = 1$. For $j = 2$, $f_n^{[j]}(z)$ has two terms. The proof is an easy modification of the case $j = 1$ with $K^{(1)}(\cdot)$ replaced by $K^{(2)}(\cdot)$, γ_1 by γ_2, $\mu_g(\boldsymbol{\theta})$ by $\check{\Omega}(\boldsymbol{\theta})$, $u|K^{(1)}(u)|$ by $u^2|K^{(2)}(u)|$, and nh_n^4 by nh_n^6. Hence, the proof has been omitted.

Table 1: Mean and MSE values of $\hat{\boldsymbol{\theta}}_n$ and $\overline{\boldsymbol{\theta}}_n$ for the estimates of the parameters in (19).

	Mean		MSE	
$\boldsymbol{\theta}^*$	$\hat{\boldsymbol{\theta}}_n$	$\overline{\boldsymbol{\theta}}_n$	$\hat{\boldsymbol{\theta}}_n$	$\overline{\boldsymbol{\theta}}_n$
(1, -0.15)	(0.9985, -0.1473)	(0.9960, -0.1472)	(0.0032, 0.0013)	(0.0041, 0.0013)
(1, 0.15)	(0.9983, 0.1524)	(0.9960, 0.1528)	(0.0032, 0.0013)	(0.0041, 0.0013)

Table 2: Means, and MSEs of $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ for the estimates of the parameters in (20) and (21).

		n	Parameter $\boldsymbol{\theta}_1^*$		Parameter $\boldsymbol{\theta}_2^*$			
			60	-0.03	1	-1.4	1	0.8
Mean	$\hat{\boldsymbol{\theta}}_n$	100	62.0212	-0.0291	1.0211	-1.4076	0.9985	0.8003
	$\tilde{\boldsymbol{\theta}}_n$		62.0160	-0.0292	2.7386	-0.7044	1.0597	0.7927
MSE	$\hat{\boldsymbol{\theta}}_n$		4.1577	0.0715×10^{-4}	0.0328	0.0264	0.0003	0.0000
	$\tilde{\boldsymbol{\theta}}_n$		4.1374	0.0693×10^{-4}	2.9017	0.4706	0.0059	0.0001
Mean	$\hat{\boldsymbol{\theta}}_n$	300	62.0072	-0.0290	1.0065	-1.4110	0.9996	0.8001
	$\tilde{\boldsymbol{\theta}}_n$		62.0029	-0.0290	2.6835	-0.7605	1.0707	0.7912
MSE	$\hat{\boldsymbol{\theta}}_n$		4.0503	0.0236×10^{-4}	0.0267	0.0159	0.0002	0.0000
	$\tilde{\boldsymbol{\theta}}_n$		4.0330	0.0285×10^{-4}	2.9006	0.4368	0.0056	0.0001
Mean	$\hat{\boldsymbol{\theta}}_n$	500	62.0069	-0.0289	1.0063	-1.4052	1.0001	0.8000
	$\tilde{\boldsymbol{\theta}}_n$		62.0021	-0.0290	2.6899	-0.7835	1.0727	0.7909
MSE	$\hat{\boldsymbol{\theta}}_n$		4.0409	0.0191×10^{-4}	0.0129	0.0064	0.0001	0.0000
	$\tilde{\boldsymbol{\theta}}_n$		4.0221	0.0214×10^{-4}	2.9027	0.3959	0.0056	0.0001

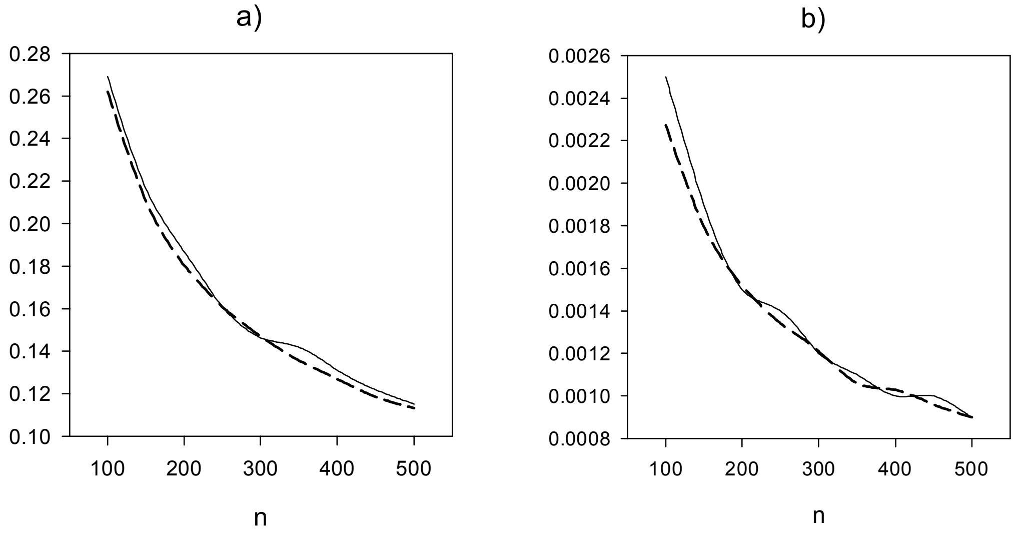

Figure 1: *Mean estimated standard errors of* $\hat{\boldsymbol{\theta}}_n$ *(solid lines) based on 1000 simulations, and asymptotic standard errors (dashed lines) for sample sizes* $n = 100, 150, \ldots, 500$*; conditional mean specification (20) with a)* $\theta_1^* = 60$ *and b)* $\theta_2^* = -0.03$.

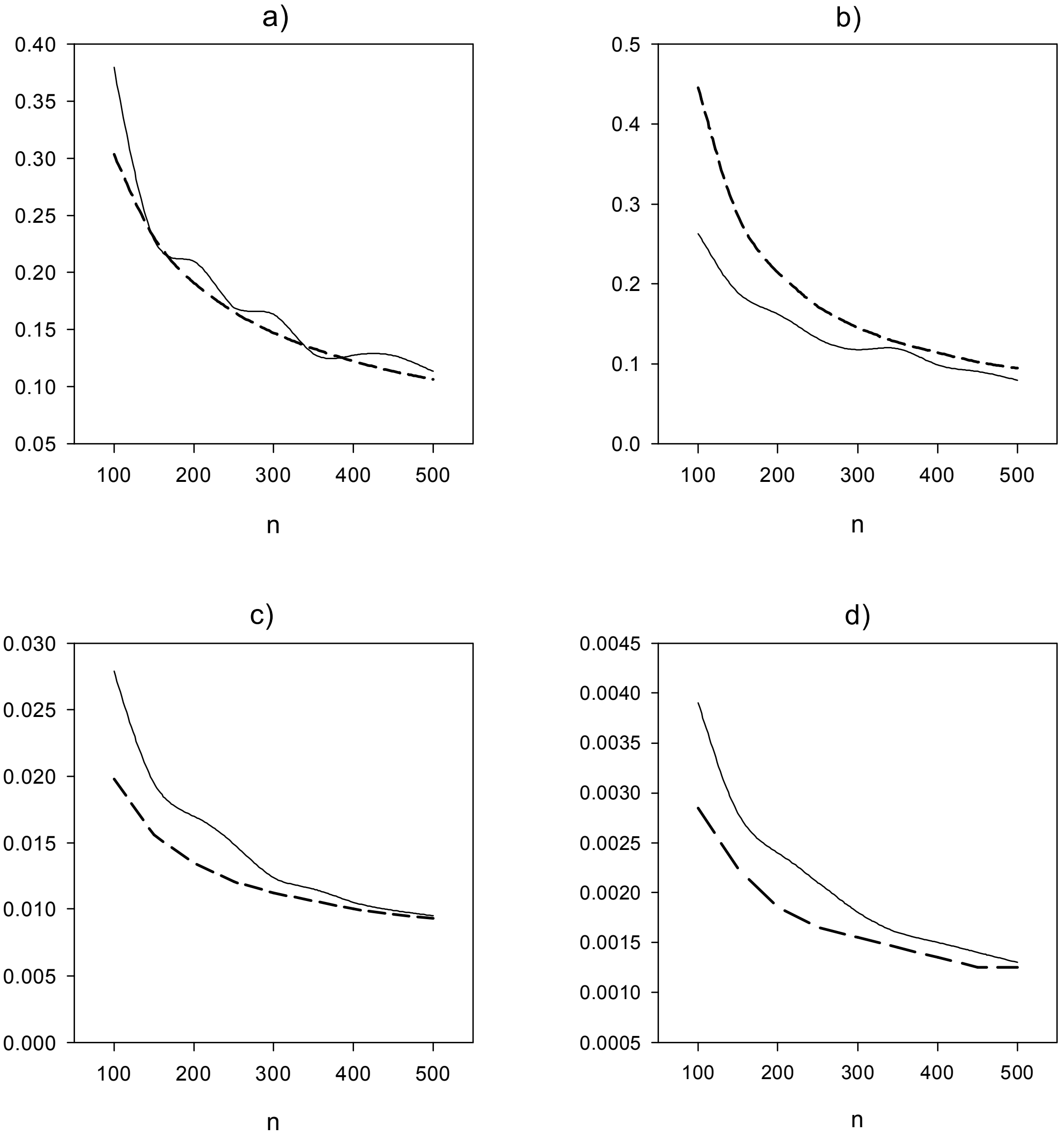

Figure 2: *Mean estimated standard errors of* $\hat{\boldsymbol{\theta}}_n$ *(solid lines) based on 1000 simulations, and asymptotic standard errors (dashed lines) for sample sizes* $n = 100, 150, \ldots, 500$*; conditional mean specification (21) with a)* $\theta_1^* = 1$*, b)* $\theta_2^* = -1.4$*, c)* $\theta_3^* = 1$*, and d)* $\theta_4^* = 0.8$.

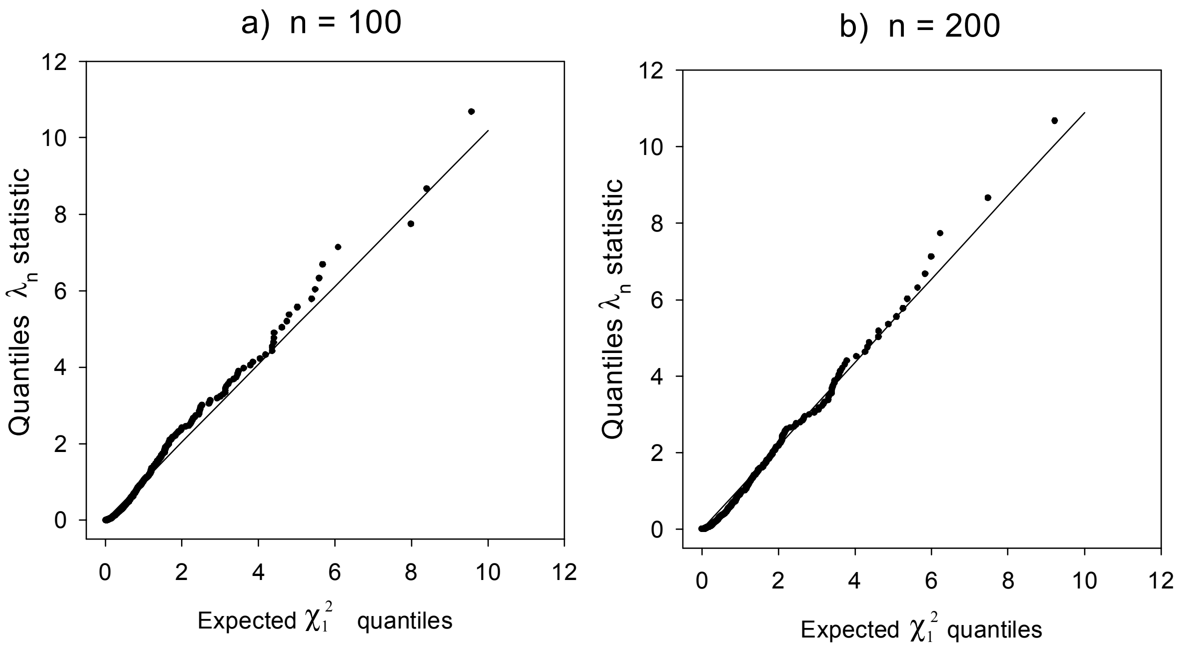

Figure 3: *Q-Q plots for the λ_n statistic for the case $n = 100$ and $n = 200$.*