Sieve Empirical Likelihood and Extensions of the Generalized Least Squares

By Jian Zhang and Irène Gijbels
Academia Sinica, Beijing and Universite Catholique de Louvain

Abstract

In this paper a unified procedure is introduced for estimation of constrained regression models with unspecified distribution of error. The procedure is based on sieve empirical likelihoods. It shows some interesting connections with certain extensions of the generalized least squares approach. A general asymptotic theory is provided. In the parametric regression function setting it is shown that under certain regularity conditions the proposed estimators are asymptotically efficient (or optimal) even if the restriction functions are discontinuous. In the nonparametric regression function setting the convergence rate of the maximum estimator based on sieve empirical likelihoods are given. It is proved that such an estimator can attain the optimal convergence rate in some cases when the size of the regression function space is not very large. In both settings, it is shown that the estimator is adaptive for inhomgeneity of conditional distribution of error with respect to predictor, especially for heteroscedasticity. Several examples are presented to illustrate the flexibility of the procedure.

1. **Introduction**

Regression analysis is usually based on a parametric likelihood. For example, we assume observations $(y_i, x_i), 1 \le i \le n$ are i.i.d with density $f_{Y|X}(y, \theta(x))f(x)$, where Y denotes a $q-$dimensional response, X denotes a $p-$dimensional predictor, the density $f(x)$ of X is unknown and is not related to θ, and the conditional density, $f_{Y|X}(y, \theta(x))$, of Y given X is assumed to be known up to a regression function θ. To estimate θ, we might choose $\hat{\theta}$ to maximize the conditional likelihood $\prod_{i=1}^n f_{Y|X}(y_i, \theta(x_i))$ over a subset of some metric space. The behavior of $\hat{\theta}$ is well studied. It is known that:

(a) When θ belongs to an infinite dimensional space, under appropriate conditions the convergence rate of $\hat{\theta}$ is determined by the entropy of a certain space of score functions or of density functions under some suitable metric. In particular, for the normal or Cauchy likelihood, this rate might be the best in certain sense [see Stone (1982), Wong and Severini (1991), Wong and Shen (1995)].

(b) When θ is restricted to a finite dimensional subspace, under some conditions $\hat{\theta}$ is asymptotically efficient [see Lehmann (1983, p.415)]. For the normal regression model with smooth but unknown scale function, the estimator based on the estimated normal likelihood is adaptive to total lack of knowledge of the functional form of the scale [see Carroll (1982)].

AMS (1991) *subject classification.* 62F12, 62G20, 62G05.
Key words and phrases. Conditional equation, sieve empirical likelihood, generalized least squares, generalized method of moments, semiparametric and nonparametric regressions, asymptotic efficiency.

(c) For some smooth functional $\rho(\theta)$, the plug-in estimator $\rho(\hat{\theta})$ is asymptotic efficient [see, e.g., Lehmann (1983), Wong and Severini (1991) and Shen (1997)].

There are many situations in which we do have some external knowledge about the function form of $f_{Y|X}$ while this knowledge is not sufficient for specifying it fully. In these situations the parametric likelihood inference is unsuitable. This is because $\hat{\theta}$ might be inconsistent or not asymptotically efficient when the parametric assumption is violated. Therefore, it is desirable to develop some alternative nonparametric likelihood method. We hope that it can capture departures from a working parametric assumption and take into account both the sample and auxiliary information to increase efficiency. Then the questions of what kind of nonparametric likelihood should be used and of whether it maintains the basic properties such as (a), (b) and (c) of the above parametric likelihood arise naturally. One difficulty with these questions is how to construct a nonparametric likelihood procedure which can adapt for possible heteroscedasticity in regression models. As pointed out before, a result of this type has been proved by Carroll (1982) for the parametric counterpart with $\theta(x)$ linear in x and smoothness conditions on the scale. Bickel, Klaassen, Ritov and Wellner (1993, p. 112) conjectured that neither the linearity nor the smoothness conditions, nor the requirement of Gaussianity are essential in Carroll's result. It is natural to further conjecture that Carroll's result holds even if the parametric likelihood is replaced by some nonparametric likelihood.

To address these issues, we consider a general regression model with q-dimensional response Y and p-dimensional predictor X. In this model, as a nonparametric nuisance parameter, the joint distribution P of (X, Y) is not fully specified. Suppose that the vector-valued regression function θ and the conditional distribution of Y are constrained by a set of conditional equations, say

$$E[G_k(Y, \theta(X))|X] = 0, \quad k = 1, 2, \cdots, k_o, \tag{1.1}$$

where $G = (G_1, \cdots, G_{k_o})^\tau$ is a predetermined vector-valued restriction function. For example, we can set $G = Y - \theta(X)$ in the ordinary conditional expectation regression model and $G = I(Y \leq \theta(X)) - 1/2$ in the median regression model. Here and hereafter $I(\cdot)$ denotes the indicator function of a set. These equations represent the non-sample information on the model. Usually G is a vector-valued function of *residuals*. Note that any inequality restriction can be converted into an equality restriction by introducing an additional variable. Thus, via the conditional equations, this formulation provides substantial flexibility for using the auxiliary information to solve non-regular estimation problems [see Newey (1993) and Powell (1994)]. Under it, many interesting models are subsumed. In this paper, two typical settings will be considered. One is the parametric regression function setting where θ is a linear function of X, say $X^\tau\beta$ or $(X^\tau\beta, \lambda)$, where λ is some unknown finite dimensional nuisance parameter. In this case, the above model is a special kind of semiparametric model with the finite dimensional parameter β or (β, λ) and the infinite dimensional parameter P. Another is the nonparametric regression function setting where θ is a vector-valued function of X lying in some smooth class of functions. The following are two prototype examples.

Example 1.1 (Linear regression with partial information on the symmetry of the error distribution). Let $0 = s_o < s_1 < \cdots < s_{k_o}$ and $S_k = [s_{k-1}, s_k)$, $1 \leq k \leq k_o$. Consider the linear regression model $Y = X^\tau\beta + e$ with $F_{e|X}(S_k) = F_{e|X}(-S_k)$, $1 \leq k \leq k_o$. Set $G_k(Y, X^\tau\beta) = I(Y - X^\tau\beta \in S_k) - I(Y - X^\tau\beta \in -S_k)$, $1 \leq k \leq k_o$. Then, G satisfies the conditional equations in (1.1).

Example 1.2 (Nonparametric regression with some information on the moments of the error). Consider the regression model $Y = \theta(X) + e$ with $E_{e|X}e^2 = \theta(X)^2 + 1$ and $E_{e|X}e^{2(k-1)+1} = 0$, $1 \le k \le k_o - 1$ (for instance, if, given X, e is symmetrically distributed, then these odd order moment equations hold). Set $G_1 = (Y - \theta(X))^2 - \theta(X)^2 - 1$, $G_k(Y, \theta(X)) = (Y - \theta(X))^{2(k-1)+1}$, $1 \le k \le k_o - 1$. Then, G satisfies the conditional equations in (1.1).

The idea of empirical likelihood (Owen, 1988) holds considerable promise for answering our questions. However, there exist some special difficulties in using Owen's empirical likelihood. For example, the normal likelihood can identify arbitrary dimensional smooth regression function, whereas Owen's is usually employed for a finite dimensional regression function. To estimate an infinite dimensional regression function, an infinite number of constraints are required. However, currently no theory is available for the empirical likelihood when the number of constraints is infinite.

To overcome the difficulties caused by heteroscedasticity and infinite number of constraints, in this paper a kind of (global) nonparametric likelihoods called sieve empirical likelihoods (SEL) for the regression function θ are constructed via the local empirical likelihoods. Here, the local empirical likelihoods are analogous to the local likelihoods of Hjort and Jones (1996) and Loader (1996) in density estimation. The name originates from the idea of using n constraints at observations $x_i, i = 1, 2, \cdots, n$,

$$E[G(Y, \theta(X))|X = x_i) = 0, \quad i = 1, 2, \cdots, n$$

as approximations of the infinite constraints in (1.1). In this sense, SEL empirical likelihoods, which include LeBlanc and Crowley's (1995) likelihood as a special case, can be viewed as some approximations to the empirical likelihood when there exist infinite number of constraints (see Section 2 below). Like LeBlanc and Crowley's likelihood, the key idea behind SEL is building the global likelihood from the local ones. To see the intuitive reason why this idea is promising, we recall one advantage of the local estimation that it can automatically reduce the effect of heteroscedasticity by using the local data [see Fan (1997)]. However, the advantage will disappear when the data structure is approximate linear because the bandwidth should not be made to tend to zero in this case. Indeed, there is a conflict in reducing the effect of heteroscedasticity and achieving the efficiency. Intuitively, hybridizing the local and global estimation techniques may solve this conflict at the cost of a certain computation. Moreover, the other advantages of the local and global estimation methods are expected to hold in the final likelihood. This strategy has been proved effective at least in the semiparametric setting in this paper. The similar strategy has also been used, for instance, in Severini and Staniswalis (1994) and Carroll, Fan, Gijbels and Wang (1997) via quasi-likelihoods.

Base on SEL, a unified framework for simultaneously estimating the regression function and the conditional distribution of the response is developed. As the main contribution of this paper, a first theoretical analysis of the proposed procedures including that of LeBlanc and Crowley (1995) is provided. It is shown that the SEL estimators including that of LeBlanc and Crowley are asymptotically efficient under some regularity conditions in the parametric regression function setting. In particular, these estimators are adaptive for the heteroscedasticity in the model. This property is comparable to property (b) of the parametric maximum likelihood estimator and gives a partial answer to the conjecture of Bickel, Klaassen, Ritov and Wellner (1993) mentioned before.

To give an insight into SEL, we proved that SEL is asymptotically equivalent to some extensions of the generalized least squares (EGLS). EGLS is a general scheme for constructing some objective function reflecting current knowledge of distributional form of the error or a desire for resistence to outlying observations. The resulting optimization solution is expected to be asymptotically efficient in a sense. Such kind of scheme is desirable, for example, in econometrics [see Robinson (1987, p. 883)]. Although Carroll (1982) and Robinson (1987) showed that the generalized least squares (GLS) estimator is asymptotically efficient in the presence of heteroscedasticity, GLS usually requires G to be a linear function with respect to the response [see Carroll and Ruppert (1988)] and seems unsuitable for the situation with a discontinuous or nonlinear G. This shortcoming is overcome in EGLS. Indeed, our theoretical analysis shows that EGLS has the strength to compete with the well-known approach—the generalized method of moments (GMM). For example, in the parametric regression function case, the asymptotic efficient (or optimal) estimator can be constructed via the EGLS method even if G is discontinuous. This is in contrast with the asymptotic efficient (or optimal) GMM estimation theory in which G is usually required to have derivatives [see Newey (1993)]. Unlike GMM, in EGLS no instrumental variables are required to specify and thus the idea can be easily extended to the nonparametric regression function case.

The connections of the proposed methods with the other related methods are revealed in this paper. Generally say, in the parametric regression function setting, if the restriction function G is smooth, the EGLS (therefore SEL) estimator is asymptotically equivalent to some special kind of instrumental variable estimators developed in econometrics, while the optimal GMM estimator is corresponding to the empirical likelihood estimator using the estimated optimal instrumental variable estimating equations. We show that the empirical likelihood estimator based on certain estimated estimating equations has the same asymptotic distribution as that of the optimal GMM estimator.

In the nonparametric regression function setting, it is shown that the SEL based maximum estimator can attain the optimal global convergence rate in some cases. This is similar to property (a) of the parametric likelihood in the above.

The following is a review of some related work and problems. There has been much interest in the estimation of model (1.1) with a parametric regression function in the last two decades, especially, considerable progress has been made with the study of the estimators based on GMM. GMM is called the general esimating equation method in biostatistics. GMM has become a standard tool in studying semiparametric regression models in the econometric and statistical literature. The instrumental variables (the functions of the predictor variable that are known to be independent of the error terms) and the GMM criterion are the main ingredients of this approach. The basic intuition behind GMM can be found in Hansen (1982), Godambe and Kale (1991), Hall (1993), Newey (1990, 1993), Ogaki (1993) and Heyde (1997). Chamberlin (1992) gave a general efficient bound for the estimation of model (1.1) when G has the first derivative. This together with the results in Bickel, Klaassen, Ritov and Wellner (1993) implies that the GMM estimator is asymptotic efficient under some regularity conditions. The GMM estimator has the usual advantage of the moment method over the parametric maximum likelihood that a weaker restriction is imposed on the model. However, it requires a stable estimator of scale. This will lead to a biased estimator in the small sample case [see Kitamura and Stutzer (1997)]. Based on certain estimating equations, Owen (1988, 1990) and Qin and Lawless (1994) developed some empirical likelihood inference procedures for improving the GMM estimator in the small and moderate sample cases. To apply Owen's approach to model (1.1), we need to specify the instrumental variables and the corresponding estimating equations. Usually, some of the estimating equations are required to be estimated from the sample. In such situation, a theoretical analysis on the behavior of Owen's approach seems to be lacking. An alternative nonparametric likelihood procedure was proposed by LeBlanc and Crowley (1995) for a single semiparametric functional where the likelihood is constructed through certain local empirical likelihoods. They showed that their idea can be easily extended to the other models, for example, censored survial data models. Most importantly, they reported that their method seems to work well in some simulated and real data examples. But less is known about the theory of such kinds of estimators. For example, whether their method is efficient relative to the other existing methods? Several other methods have also been suggested and investigated intensively in the literature including quantile regression, generalized least squares, quasi-likelihood, dual likelihood and so on [see, for instance, Koenker and Bassett (1978), Powell (1986), Carroll and Ruppert (1988), Efron (1991), Green and Silverman (1991), McCullagh and Nelder (1991), Mykland (1996)]. In the literature, GMM is mainly employed to estimate model (1.1) in the parametric regression function setting. Recently the idea of GMM has been extended to the nonparametric regression function setting by Carroll, Ruppert and Welsh (1998) via the local moments. Obviously, their method might be inefficient when the nonparametric regression function of interest includes some parametric component.

The remains of this paper proceed as follows. Sections 2 and 3 give the definitions and intuitive illustrations of the proposed sieve likelihoods and estimators. A large sample study of these estimators is presented in Section 4. Section 5 is focused on the connections of the proposed estimators with the GMM and instrumental variable estimators. Some examples can be found in Sections 3 and 4. Some possible extensions are discussed in Section 6. The proofs of the main results are presented in the last section. Throughout this paper, P_n and P stands for the empirical and the underlying distributions of (X, Y), respectively. $P_{Y|X}$ and P_X denote the conditional distribution of Y given X and the distribution of X, respectively. $E_{Y|X}$ and E_X denote the associated expectations. f denotes the Lebesgue density of P_X. Except several specified situations, in the following sections, $||\cdot||$ denotes the Euclidean norm.

2. Sieve empirical likelihood

2.1. Definition

We use the nonparametric likelihood to illustrate the idea behind SEL. Suppose we have n independent observations $\{(x_i, y_i), 1 \le i \le n\}$ from (X, Y). Then, the nonparametric likelihood for the regression is

$$\prod_{i=1}^{n} dP(x_i, y_i)$$

where $dP(x_i, y_i)$ is the mass that P places at point (x_i, y_i). Suppose that P_X is not informative about the regression function. This for example excludes applications to random effect models, but allows for fixed effect models. Under the noninformativeness assumption, the above likelihood of the regression is proportional to the conditional nonparametric likelihood

$$\prod_{i=1}^{n} dP_{Y|X=x_i}(y_i). \tag{2.1}$$

However, this likelihood can not be directly used because $P_{Y|X=x_i}, i = 1, \cdots n$ are unknown. The problem we faced is, given each x_i, only a single observation y_i from $P_{Y|X=x_i}$ is available, which seems impossible to be combined with the auxiliary equations in (1.1) to construct an estimator for $P_{Y|X=x_i}$. This problem can be solved by 'borrowing' information from the nearby observations if we assume that $P_{Y|X=x}$ is continuous with respect to x [see Tibshirani and Hastie (1987) and LeBlanc and Crowley (1995)]: *First, we augment the sample by the local similarity of $P_{Y|X=x}$. For each x_i, we define an neighborhood and pick out all the y_j with x_j lying in this neighborhood. These y_j form a set, say $\{z_{ik}, 1 \le k \le n_i\}$ which can be roughly viewed as observations from $P_{Y|X=x_i}$. Then, with this augmented sample and equations in (1.1), an estimator of $P_{Y|X=x_i}$, say $\hat{P}_{Y|X=x_i}$, is constructed by using the empirical likelihood technique. At the same time a local empirical likelihood for θ is obtained. Here, $\hat{P}_{Y|X=x_i}$ is supported by $\{z_{ik}, 1 \le k \le n_i\}$. Finally, treating sets $\{z_{ik}, 1 \le k \le n_i\}, 1 \le i \le n$ as if they were independent, we define LCEL for θ as the product of these local empirical likelihoods.*

An improved version of the above idea is to include a smoothing weight function, which weighs down the contribution to the $i-$th local empirical likelihood of θ from y_j with x_j being away from x_i. In this setting, for each i, all y_j are augmented into y_i. If we use $w_{ji}, 1 \le j \le n$ with $\sum_{j=1}^{n} w_{ji} = 1$ to weigh the contributions of $y_j, 1 \le j \le n$ to the $i-$th local likelihood, then, the logarithm SEL can be constructed by the following two steps.

Step 1. Given the values of $\theta(x_i), 1 \le i \le n$, for $1 \le i, j \le n$, let $p_{ji} = P_{Y|X=x_i}\{y_j\}$ denote the mass we put on y_j. For each i, we maximize

$$\sum_{j=1}^{n} w_{ji} \log p_{ji}$$

subject to

$$\begin{array}{l} \sum_{j=1}^{n} p_{ji} G(y_j, \theta(x_j)) = 0, \\ \sum_{j=1}^{n} p_{ji} = 1, \quad p_{ji} \ge 0, \quad j = 1, \cdots, n. \end{array}$$

Let $\hat{p}_{ji}, 1 \le j \le n$ be the solution. Then, for $1 \le j \le n$,

$$\hat{p}_{ji} = \frac{w_{ji}}{1 + \alpha_n(x_i, \theta)^\tau G(y_j, \theta(x_j))}$$

where k_o-vector $\alpha_n(x_i, \theta)$ satisfies

$$\sum_{j=1}^{n} w_{ji} \frac{G(y_j, \theta(x_j)}{1 + \alpha_n(x_i, \theta)^\tau G(Y_j, \theta(x_j))} = 0. \tag{2.2}$$

Define

$$l(i,\theta) = \sum_{j=1}^{n} w_{ji} \log \hat{p}_{ji}.$$

Note that $l(i,\theta)$ is an approximation to $\log P_{Y|X=x_i}(y_i)$.

Step 2. Define the logarithm SEL by

$$l_s(\theta) = \sum_{i=1}^{n} l(i,\theta).$$

The logarithm of the likelihood ratio between the sieve empirical likelihoods with and without local contraints has the form

$$l_{sr} = l_s(\theta) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ji} \log w_{ji}.$$

Generally, for each i, we can choose $N_i \subset \{1, \cdots, n\}$ and

$$l(i,\theta) = \{\sum_{j \in N_i} w_{ji}\}^{-1} \sum_{j \in N_i} w_{ji} \log \hat{p}_{ji}$$

in the above. This leads to a possible reduction of the degree of correlations among $l(i,\theta), 1 \leq i \leq n$, but on the other hand also to a loss of some information due to dropping $\hat{p}_{ji}, j \notin N_i$, $1 \leq i \leq n$. For simplicity, in the following sections, we consider only the cases with $N_i = \{1, \cdots, n\}, 1 \leq i \leq n$ and $N_i = \{i\}$, $1 \leq i \leq n$, respectively.

Note that when $N_i = \{i\}, 1 \leq i \leq n$, we recover the logarithm of LeBlanc and Crowley's likelihood l_c and likelihood ratio l_{cr} which are defined by

$$l_c(\theta) = \sum_{i=1}^{n} \log \hat{p}_{ii}, \quad l_{cr}(\theta) = l_c(\theta) - \sum_{i=1}^{n} \log w_{ii}.$$

With l_{sr} we can introduce some estimators, called the SEL estimators, for the regression functions as follows.

Parametric regression function. Suppose that $\theta(X) = X^\tau \beta$, where β is the parameter of interest. For the simplicity of symbols, we use $l_{sr}(\beta)$ to denote $l_{sr}(\theta)$. Then, the estimator of β is defined as

$$\hat{\beta} = argmax l_{sr}(\beta).$$

A similar estimator can be defined for the case with $\theta(x) = (\beta^\tau x, \lambda)$.

Nonparametric regression function. Assume θ is lying in an appropriately chosen function space, say Θ. Then, the maximum estimator $\hat{\theta}$ is defined as

$$\hat{\theta} = argmax_{\theta \in \Theta} l_{sr}(\theta).$$

Estimators of conditional distributions. Let $\hat{\theta}$ be estimator defined above. Then, the estimate of $F_{Y|X=x}$, $\hat{F}_{Y|X=x}$ is obtained by replacing θ in (2.2) by $\hat{\theta}$. Note that, it is supported by $\{y_j, 1 \leq j \leq n\}$ with masses $\hat{p}_j(x), 1 \leq j \leq n$.

There are two typical candidates for the weights $w_{ji}, 1 \le i, j \le n$:

The kernel weights with a constant bandwidth: $w_{ji} \propto K(||x_j - x_i||/h), 1 \le i, j \le n$, where $K(|\cdot|)$ is a univariate kernel function.

The kernel weights with a nearest neighbor bandwidth: Let $h(x, m)$ be the mth smallest number among $||x_k - x||$, $1 \le k \le n$. Then, $w_{ji} \propto K(||x_j - x_i||/h(x_i, m))$. Usually we set $K(||t||) = (1 - ||t||^3)_+^3$ [see Cleveland (1979)].

REMARK 2.1. (i) Throughout this paper, we choose the bandwidth h so that $d_o \le h^p n^\eta \le d_1$ for some positive constants d_o, d_1 and η when we decide to use the constant bandwidth; and choose $m = m_n$ so that $d_2 \le m_n n^{\eta-1} \le d_3$ for some positive constants d_2, d_3 and η when we decide to use the nearest neighbor bandwidth.

(ii) By comparing the above two kinds of bandwidths, it is seen from the following extension of a theorem due to Fan and Gijbels (1994) that the nearest neighbor bandwidth is adaptive for the design points while the constant bandwidth is obviously not. However, the computation of the nearest neighbor bandwidth is relatively time consuming.

PROPOSITION 2.1. *Suppose X has a compact support Ω and a continuous positive density f. If $m_n/n \to 0$ and $m_n/\log n \to \infty$, then for any $x \in \Omega$,*

$$h(x, m_n)^p = \frac{1}{f(x)c(p)} \frac{m_n}{n}(1 + o(1))$$

where $c(p)$ is the Lebesgue measure of the $p-$dimensional hypersphere $\{x \in R^p : ||x|| \le 1\}$, and $o(1) \to 0$ almost surely and uniformly in $x \in \Omega$.

2.2. Quadratic approximations

To get a better insight into the above procedure, we first investigate the behavior of the proposed likelihood. Suppose for each i, $\alpha_n(x_i, \theta)$ in (2.2) tends to zero. Then, we have the following informal approximations of $\alpha_n(x_i, \theta)$ and l_r,

$$\begin{aligned} \alpha_n(x_i,\theta) &= [\textstyle\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)) G^\tau(y_j,\theta(x_j))]^{-1} \sum_{j=1}^n w_{ji} G(y_j,\theta(x_j))(1+o_p(1)), \\ & \quad 1 \le i \le n; \\ l_{sr} &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n (\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)))^\tau [\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)) G^\tau(y_j,\theta(x_j))]^{-1} \\ & \quad \times \textstyle\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j))(1+o_p(1)). \end{aligned} \tag{2.3}$$

and we also have the following approximate population version in the neighborhood of the true parameter θ_o,

$$\frac{1}{n} l_{sr} = -\frac{1}{2} E_X\{E_{Y|X} G(Y,\theta(x))^\tau [E_{Y|X} G(Y,\theta(x)) G^\tau(Y,\theta(x))]^{-1} E_{Y|X} G(Y,\theta(x)\} + o_p(1).$$

Applying the same argument, we have

$$\begin{aligned} l_{cr} &= -\textstyle\sum_{i=1}^n \sum_{j=1}^n w_{ji} G^\tau(y_j,\theta(x_j)) (\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)) G^\tau(y_j,\theta(x_j)))^{-1} G(y_j,\theta(x_j)) \\ & \quad + \tfrac{1}{2}\textstyle\sum_{i=1}^n (\sum_{j=1}^n w_{ji} G^\tau(y_j,\theta(x_j))(\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)) G^\tau(y_j,\theta(x_j)))^{-1} \\ & \quad \times G(y_i,\theta(x_i)) G^\tau(y_i,\theta(x_i)) (\textstyle\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j)) G^\tau(y_j,\theta(x_j))^{-1} \\ & \quad \times (\textstyle\sum_{j=1}^n w_{ji} G(y_j,\theta(x_j))(1+o_p(1)) \end{aligned} \tag{2.4}$$

and the same approximate population version as that of l_{sr}. Clearly, the approximation of l_{sr} is quadratic while that of l_{cr} is not.

(2.3) and (2.4) imply that the l_{sr} or l_{cr} based estimators are asymptotically equivalent to certain weighted-least-squares-like estimators if $o_p(1)$ in the these formulae is uniform in x_i.

To formally prove (2.3) and (2.4), we need to calculate the uniform rate of the convergence of some empirical processes indexed by a class of functions, say $\mathcal{F}$, which is equipped with a metric ρ. It is well known that this rate is related to the size of $\mathcal{F}$ measured by its (bracketing) covering number. The covering number $N(\epsilon, \rho, \mathcal{F})$ is defined as the smallest number of $\epsilon-$balls in metric ρ needed to cover $\mathcal{F}$. And the bracketing covering number $N^B(\epsilon, \rho, \mathcal{F})$ is the smallest number m: there exist $f_k^l \leq f_k^u$, $k = 1, \cdots, m$ with $\max_{1 \leq k \leq m} \rho(f_k^u - f_k^l) \leq \epsilon$ such that for any $f \in \mathcal{F}$ there exists a k with $f_k^l \leq f \leq f_k^u$. The commonly used metrics are the L_∞ and $L_r(Q)$, $0 < r < \infty$ with respect to a probability measure Q. The corresponding metric entropies are defined by $H(\epsilon, \rho, \mathcal{F}) = \log N(\epsilon, \rho, \mathcal{F})$ and $H^B(\epsilon, \rho, \mathcal{F}) = \log N^B(\epsilon, \rho, \mathcal{F})$. Usually the covering number is easier to obtain than the bracketing covering number. But the latter is more useful in the nonparametric setting.

Let $K(||t||)$ be a nonnegative kernel function on R^p satisfying the condition (K0) or the condition (K1). (K0) and (K1) are defined by the following way: In both cases, we assume $\int K(||t||)dt = 1$, $\int K(||t||)tdt = 0$; and $\int K(||t||)^{(1+\delta_o)/\delta_o} dt < \infty$ for any $\delta_o > 0$. Furthermore, in (K0), we assume $K(s)$ is non increasing as s goes from 0 to ∞. In (K1), we assume for some positive constant c_o, $|K(s_1) - K(s_2)| \leq c_o |s_1 - s_2|$ for any $s_i \in R^1$, $i = 1, 2$.

Without loss of generality, we suppose $0 \leq K(s) \leq 1$ below. Set

$$w_j(x) = K(||x_j - x||/h) / \sum_{i=1}^n K(||x_k - x||/h) \quad \text{and } w_{ji} = w_j(x_i)$$

for $1 \leq i, j \leq n$, where the bandwith h is defined in the end of Section 2. Note that the dependence of $w_j(x)$ on h is suppressed for notational convenience.

For the predictor variable X, we assume:

(X0) X has a compact and convex support Ω. And f is continuous and bounded above and below.

The condition (X0) is commonly used assumption to avoid some undesirable features of the kernel nonparametric regression when the predictor is unbounded or the density of the predictor is non positive.

In the following we let r_o denote a predetermined positive constant.

For the parametric case (that is, $\theta(x) = x^\tau \beta$ or $\theta(x) = (x^\tau \beta, \lambda)$), we take $\theta(x) = x^\tau \beta$ for example. Let β_o be the true value of β. Set

$$\mathcal{F}_{G_k} = \{G_k(\cdot, \cdot^\tau \beta) : ||\beta - \beta_o|| \leq r_o\}, \mathcal{F}_{G_k G_m} = \{G_k(\cdot, \cdot^\tau \beta) G_m(\cdot, \cdot^\tau \beta) : ||\beta - \beta_o|| \leq r_o\}$$

for $1 \leq k, m \leq k_o$. We assume:

(P1) *(Moments):*

(i) $E_{Y|X}G(Y, x^\tau\beta_o) = 0, x \in \Omega$.

(ii) There exists $\alpha_o > 4$ such that

$$Esup_{||\beta-\beta_o||\le r_o}|G_k(Y, X^\tau\beta)|^{\alpha_o} < \infty, 1 \le k \le k_o.$$

(iii) For $1 \le k \le k_o$,

$$\sup_{x\in\Omega,||\beta-\beta_o||\le r_o} E_{Y|X}G_k^4(Y, x^\tau\beta) < \infty.$$

(iv) For $1 \le k, m \le k_o$, $E_{Y|X}G_k(Y, x^\tau\beta)G_m(Y, x^\tau\beta)$ is finite and continuous with respect to (x, β). And the smallest eigenvalue of $E_{Y|X}G(Y, x^\tau\beta_o)G^\tau(Y, x^\tau\beta_o)$ is positive for $x \in \Omega$.

(v) For $1 \le k \le k_o$,

$$\sup_{x\in\Omega,||\beta-\beta_o||\le r_o} ||\frac{\partial E_{Y|X}G_k(Y, x^\tau\beta)}{\partial\beta}|| < \infty.$$

(P2) *(Covering number):* For $1 \le k \le k_o$,

$$N(\epsilon, L_1(P_n), \mathcal{F}_{G_k}) \le A(P_n)\epsilon^{-w},$$

where $\limsup_n EA(P_n) < \infty$, and w is a positive constant.

Remark 2.2. The condition (P2) holds if $\mathcal{F}_{G_k}$ is a VC-subgraph classes or satisfies the Lipschitz condition [see van der Vaart and Wellner (1996) for the definition of VC-subgraph classes]. Furthermore, under the conditions P1(ii) and (P2),

$$N(\epsilon, L_1(P_n), \mathcal{F}_{G_kG_m}) \le B(P_n)\epsilon^{-2w}$$

where $\limsup_n EB(P_n) < \infty$.

In the nonparametric case (that is, θ belongs to a smooth class Θ and is equipped with the norm $||\theta||^2 = E_X||\theta(X)||^2$), we set

$$\mathcal{F}^*_{G_k} = \{G_k(\cdot, \theta(\cdot)) : ||\theta - \theta_o|| \le r_n\}, \mathcal{F}^*_{G_kG_m} = \{G_k(\cdot, \theta(\cdot))G_m(\cdot, \theta(\cdot)) : ||\theta - \theta_o|| \le r_n\}$$

for $1 \le k, m \le k_o$, and for any positive sequence $\{r_n\}$ with $r_n \to 0$. Let $\theta(\theta, \theta^*, s) \in \Theta, 0 \le s \le 1$ is a path from θ to θ^*. For example, $\theta(\theta, \theta^*, s) = (1-s)\theta + s\theta^*$. For any function g, define

$$\frac{\partial g(\theta(z))}{\partial\theta}[\theta^*(z) - \theta(z)] = \lim_{s\to 0}(g(\theta(\theta, \theta^*, s)(z)) - g(\theta(z)))/s,$$

and

$$||\frac{\partial g(\theta(z))}{\partial\theta}|| = \sup_{\theta^*\in\Theta,\theta^*\ne\theta} |\frac{\partial g(\theta(z))}{\partial\theta}[\theta^*(z) - \theta(z)]|/||\theta^* - \theta||.$$

We define moment conditions N1(i), N1(ii), N1(iii) and N1(v) by replacing $x^\tau\beta$, $X^\tau\beta$ and $||\beta - \beta_o|| \le r_o$ in P1(i), P1(ii) and P1(iii) by $\theta(x)$, $\theta(X)$ and $||\theta - \theta_o|| \le r_o$, $\theta \in \Theta$, respectively.

(N1) *(Moments):*

In addition to N1(i), N1(ii), N1(iii) and N1(v), we assume

(iv) For $1 \le k, m \le k_o$, $D_{km} = G_k G_m$,

$$\sup_{||\theta-\theta_o|| \le r_o, \theta \in \Theta, z \in \Omega} |E_{Y|X} D_{km}(Y, \theta(z+u)) - E_{Y|X} D_{km}(Y, \theta(z))| \to 0$$

as $u \to 0$. The smallest eigenvalue of $E_{Y|X} G(Y, \theta(z)) G^{\tau}(Y, \theta(z))$ is uniformly positive for $||\theta - \theta_o|| \le r_o$, $\theta \in \Theta$ and $z \in \Omega$.

(N2) (Entropy) For $1 \le k, m \le k_o$,

$$H^B(\epsilon, L_2(P), \mathcal{F}_{G_k}) \le A^*(P)(\epsilon/r_n)^{-w^*}, \quad H^B(\epsilon, L_2(P), \mathcal{F}_{G_k G_m}) \le B^*(P)(\epsilon/r_n)^{-w^*},$$

where $A^*(P)$ and $B^*(P)$ are some positive constants.

Remark 2.3. Let $\mathcal{F}_o = \{\theta \in \Theta : ||\theta - \theta_o|| \le r_o\}$. Assume $H^B(\epsilon, ||\cdot||_\infty, \mathcal{F}_o) \le c_o \epsilon^{-w^*}$. *Case 1.* If

$$E \sup_{\mathcal{F}_o} ||\frac{\partial G_k(Y,\theta(X))}{\partial\theta}||^2 < \infty,$$
$$E \sup_{\mathcal{F}_o} ||\frac{\partial (G_k(Y,\theta(X)) G_m(Y,\theta(X)))}{\partial\theta}||^2 < \infty$$

where c_o and w^* are some positive constants, then the condition (N2) holds.

Case 2. Let $G(Y, \theta(X)) = I(Y - \theta(X) \le 0) - 1/2$. Under the condition (X0), it follows from Corollary 2.7.3 of van der Vaart and Wellner (1996, p. 157) that (N2) holds if w^* is replaced by $2w^*$.

Theorem 2.1. (i) (Parametric regression function case): *Under conditions (K0), (X0), (P1) and (P2), (2.3) and (2.4) hold uniformly in β with $||\beta - \beta_o|| \le r_n < n^{-1/\alpha_o}$ if $0 < \eta < (\alpha_o - 2)/(\alpha_o + 2)$.* The similar result holds for the case $\theta(x) = (\beta^\tau x, \lambda)$.

(ii) (Nonparametric regression function case): *Under conditions (K1), (X0), (N1) and (N2), (2.3) and (2.4) hold uniformly in θ with $||\theta - \theta_o|| \le r_n < n^{-1/\alpha_o}$ if*

$$0 < \eta < \begin{cases} \min\{\frac{2(\alpha_o-4)}{2\alpha_o + w^*(\alpha_o-4)}, \frac{2(w^*+\alpha_o-2)}{(2+w^*)\alpha_o}\}, & \text{when } 0 < w^* \le 2; \\ \frac{\alpha_o-4}{(\alpha_o-2)w^*}, & \text{when } w^* > 2. \end{cases}$$

When G is bounded, the results in (i) and (ii) still hold if we replace α_o by ∞.

3. Extensions of the generalized least squares estimators

Several observations can be made from the quadratic approximations of SEL in the last section.

The most important observation is that l_{sr} and l_{cr} can lead to some interesting modifications of the classical weighted least squares criterion. In particular, we have the following criteria:

Population versions,

$$R(\theta) = \frac{1}{2} E_X\{E_{Y|X} G^\tau(Y,\theta(X)) U^{-1}(X,\theta) E_{Y|X} G(Y,\theta(X))\}$$

with $U(X,\theta) = E_{Y|X} G(Y,\theta(X)) G^\tau(Y,\theta(X))$. Sample versions,

$$\begin{aligned}
\hat{R}_s(\theta) &= \tfrac{1}{2}\hat{E}_X\{\hat{E}_{Y|X} G^\tau(Y,\theta(X))\hat{U}^{-1}(X,\theta)\hat{E}_{Y|X} G(Y,\theta(X))\} \\
\hat{R}_c(\theta) &= \hat{E}\{\hat{E}_{Y|X} G^\tau(Y,\theta(X))\hat{U}^{-1}(X,\theta) G(Y,\theta(X))\} - \tfrac{1}{2}\hat{E}\{\hat{E}_{Y|X} G^\tau(Y,\theta(X)) \\
&\quad \times \hat{U}^{-1}(X,\theta) G(Y,\theta(X)) G^\tau(Y,\theta(X))\hat{U}^{-1}(X,\theta)\hat{E}_{Y|X} G(Y,\theta(X))\}
\end{aligned}$$

where $\hat{U}(X,\theta) = \hat{E}_{Y|X} G(Y,\theta(X)) G^\tau(Y,\theta(X))$, $\hat{E}$, $\hat{E}_{Y|X}$ and $\hat{E}_X$ are some estimators of the expectation operators E, $E_{Y|X}$ and E_X, respectively. In practice, we often replace $\hat{U}(X,\theta)$ by $\hat{U}(X,\hat{\theta}_I)$ where $\hat{\theta}_I$ is a consistent initial estimator of θ.

The $R(\theta)$ based estimation is a two-stage approach: First, make the kernel regressions of $G(Y,\theta(X))$ and $G(Y,\theta(X))G^\tau(Y,\theta(X))$ on X; then, minimize the summation of the local χ^2-square statistics of θ. For example, we consider the ordinary linear regression $Y = X^\tau\beta + e$ with known constant error variance σ^2. The population version of the least squares criterion becomes

$$E(Y - X^\tau\beta)^2 = E(Y - E_{Y|X}Y)^2 + \sigma^2 R(\theta).$$

Obviously, the regression of $G(Y,\theta(X))$ on X is not required in the least squares method. This implies that, similar to the difference between the least absolute deviation regression and the LeBlanc-Crowley likelihood based regression, for the linear regression model, the EGLS estimator may be more variable than the least squares estimator for a small sample although two estimators are asymptotically equivalent [see LeBlanc and Crowley (1995, p.102)]. In fact, there is room for improving the EGLS estimator via the following equality: for any measurable function $H(Y,\theta(X))$,

$$\begin{aligned}
& E_X\{E_{Y|X} G^\tau(Y,\theta(X)) U^{-1}(X,\theta) E_{Y|X} G(Y,\theta(X))\} \\
&= E\{\{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}^\tau U^{-1}(X,\theta) \\
&\quad \times\{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\} \\
&\quad - E\{H(Y,\theta(X)) - E_{Y|X} H(Y,\theta(X))\}^\tau U^{-1}(X,\theta)\{H(Y,\theta(X)) - E_{Y|X} H(Y,\theta(X))\}.
\end{aligned}$$

Suppose that $H(Y,\theta(X)) - E_{Y|X} H(Y,\theta(X))$ is independent of θ. Then the above equality shows that $R(\theta)$ is proportional to

$$\begin{aligned}
R_M(\theta) = \; & \tfrac{1}{2} E\{\{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}^\tau U^{-1}(X,\theta) \\
& \times\{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}.
\end{aligned}$$

The associated sample version:

$$\begin{aligned}
\hat{R}_M(\theta) = \; & \tfrac{1}{2}\hat{E}\{\{\hat{E}_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}^\tau U^{-1}(X,\theta) \\
& \times\{\hat{E}_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}.
\end{aligned}$$

Obviously, compared with $\hat{R}_s(\theta)$, $\hat{R}_M(\theta)$ avoids making an unnecessary regression of $H(Y,\theta(X))$ on X. Similarly, we obtain the following variant of EGLS:

$$\hat{R}_N(\theta) = \frac{1}{2}\hat{E}\{\hat{E}_{Y|X} G^\tau(Y,\theta(X))\hat{U}^{-1}(X,\theta) G(Y,\theta(X))\}$$

by noting that

$$R(\theta) = \frac{1}{2}E_X\{E_{Y|X}G^\tau(Y,\theta(X))U^{-1}(X,\theta)G(Y,\theta(X))\}.$$

It follows from Theorem 2.1 and the results in the next section that the SEL estimator has the same asymptotic distribution as those of $\hat{R}_s(\theta)$, $\hat{R}_M(\theta)$, $\hat{R}_N(\theta)$ and $\hat{R}_c(\theta)$ based estimators. For convenience, here and hereafter we use $R(\beta)$, $\hat{R}_r(\beta)$ and $\hat{R}_c(\beta)$ to denote $R(\theta)$, $\hat{R}_r(\theta)$ and $\hat{R}_c(\theta)$ when $\theta(X) = X^\tau\beta$. Similarly, we define $R(\beta,\lambda)$, $\hat{R}_r(\beta,\lambda)$ and $\hat{R}_c(\beta,\lambda)$.

The following examples demonstrate that SEL and EGLS are more generally applicable than both the generalized least squares and GMM. Let β_o and λ_o denote the true value of β and λ, respectively. Let $F_{e|X}$ and $f_{e|X}$ denote the conditional distribution function and density of e given X.

Example 1.1 (continued). We have

$$R(\beta) = \sum_{k=1}^{k_1} E_X \frac{(F_{e|X}(S_k + X^\tau(\beta-\beta_o)) - F_{e|X}(-S_k + X^\tau(\beta-\beta_o)))^2}{F_{e|X}(S_k + X^\tau(\beta-\beta_o)) + F_{e|X}(-S_k + X^\tau(\beta-\beta_o))}.$$

Example 1.2 (continued). Let $u(X) = (u_i(X))_{1\le i\le k_o}$ be k_o dimensional vector with $u_1(X) = 2\theta_o(X)$ and

$$u_k(X) = \sum_{t=0}^{k-2}\binom{2(k-2)+1}{2t+1}E_{e|X}e^{2(k-2-t)}(\theta_o-\theta)^{2t}, 2\le k\le k_o,$$

where

$$\binom{2(k-2)+1}{2t+1} = \frac{(2(k-2)+1)!}{(2t+1)!(2(k-2-t))!}.$$

Let $v(X) = (v_{ij}(X))_{1\le i,j\le k_o}$ with $v_{11}(X) = E_{e|X}e^4 - (E_{e|X}e^2)^2$, $v_{1k}(X) = v_{k1}(X) = 0, 2\le k\le k_o$, and $v_{ij}(X) = E_{e|X}e^{2(i+j-3)}$ for $2\le i,j\le k_o$. Then, as $||\theta-\theta_o||\to 0$,

$$R(\theta) = \frac{1}{2}E_X(\theta_o(X)-\theta(X))^2u^\tau(X)v^{-1}(X)u(X)(1+o(1)).$$

This shows that l_{sr} and EGLS can weight the observations according to the conditional moments. Note that, the quasi-likelihood allows for only the second moment information and in general fails to produce an efficient estimator when the other moment information is available.

Example 3.1 (Mean regression model). Consider the linear regression model $Y = X^\tau\beta + e$ with unknown parameter β. Suppose $E_{e|X}e = 0$ and $\sigma^2(X) = E_{e|X}e^2 > 0$. Let $G(Y, X^\tau\beta) = Y - X^\tau\beta$, then

$$\begin{aligned} R(\beta) = \ & \tfrac{1}{2}E_X(\beta_o-\beta)^\tau\tfrac{XX^\tau}{\sigma^2(X)}(\beta_o-\beta)\tfrac{1}{1+(\beta_o-\beta)^\tau XX^\tau(\beta_o-\beta)/\sigma^2(X)} \\ R_M(\beta) = \ & \tfrac{1}{2}E\tfrac{1}{\sigma(X)^2}(Y-X^\tau\beta)^2\tfrac{1}{1+(\beta_o-\beta)^\tau XX^\tau(\beta_o-\beta)/\sigma^2(X)}. \end{aligned}$$

The resulting estimators are shown to have the same asymptotic distribution as that of the adaptive weighted least squares estimator investigated by Carroll (1982) and Robinson (1987) (see Section 5).

EXAMPLE 3.2 (AN ENDOGENOUS DUMMY VARIABLE MODEL) (Newey, 1993). Consider the nonlinear model

$$Y = \lambda S + \phi(X, \beta) + e,$$

where $E_{e|X} e = 0$, ϕ is known, S and e are correlated, and (β^τ, λ) is the parameter of interest. This model has many important applications in economics and has been studied by Newey (1993) via GMM. Here we estimate (β^τ, λ) via EGLS. First, we define $Y^* = (Y, S)^\tau$, $G(Y^*, X^\tau\beta, \lambda) = Y - \lambda S - \phi(X, \beta)$. Then, we have

$$R(\beta, \lambda) = \frac{1}{2} E_X \{ \frac{((\lambda_o - \lambda) E_{S|X} I(S = 1) + \phi(X, \beta_o) - \phi(X, \beta))^2}{E_{e|X} e^2 + E_{S|X}((\lambda - \lambda_o)S + \phi(X, \beta) - \phi(X, \beta_o))^2} \}.$$

EXAMPLE 3.3 (REGRESSION QUANTILES) (Koenker and Bassett, 1978). Consider the linear model $Y = X^\tau\beta + e$ with unknown parameter β. Suppose $F_{e|X}(0) = q$, $0 < q < 1$. Let $G(Y, X^\tau\beta) = I(Y \le X^\tau\beta) - q$, where $I(\cdot)$ is indicator function. Then

$$R(\beta) = E_X \frac{(F_{e|X}(X^\tau(\beta - \beta_o)) - q)^2}{F_{e|X}(X^\tau(\beta - \beta_o))(1 - 2q) + q^2}.$$

Note that when $q = 1/2$, $G^2(Y, X^\tau\beta) \equiv 1$. So the classical weighted least squares method completely fails in this situation. When GMM is used, we need to estimate $\partial E_{Y|X} G(Y, x^\tau\beta)/\partial\beta$. Note that G is not differentiable at the points where $Y = X^\tau\beta$. So we usually use a numerical derivative, say

$$\frac{1}{h^*} \sum_{j=1}^{n} w_{ji}(G(y_j, x_j^\tau(\beta - \underline{h}^*)) - G(y_j, x_j^\tau(\beta + \underline{h}^*)))$$

to estimate it, where $\underline{h}^* = (h^*, \cdots, h^*)$ and h^* is another bandwidth. Taking into consideration the difficulty in selecting h^*, we conclude that EGLS is simpler than GMM because in EGLS no partial derivative of $E_{Y|X} G(Y, x_i^\tau\beta)$ is required.

This conclusion is also suitable for the following two examples.

EXAMPLE 3.4 (MULTIPLE QUANTILE REGRESSION). Consider the nonlinear regression model $Y = m(X^\tau\beta) + \sigma(X^\tau\beta)e$ with unknown parameter β. Suppose $F_{e|X}(v_j) = u_j$, $1 \le j \le k_o$, where e is supposed to be independent of X, $v_{k_1} = 0$, $u_{k_1} = 0.5$ for some $1 \le k_1 \le k_o$, v_j, $j \ne k_1$, $1 \le j \le k_o$ are unknown, and $0 < u_1 < u_2 < \cdots < u_{k_o} < 1$ are known constants. Suppose m and σ have some known function forms, for example, $m(X^\tau\beta) = 1/(1 + \exp(-X^\tau\beta))$ and $\sigma(X^\tau\beta) = m(X^\tau\beta)(1 - m(X^\tau\beta))$. Let $\lambda = (v_k)_{1 \le k \ne k_1 \le k_o}$ and

$$G(Y, X^\tau\beta, \lambda) = \begin{pmatrix} I(Y - m(X^\tau\beta) < v_1\sigma(X^\tau\beta)) - u_1 \\ \cdots \\ I(Y - m(X^\tau\beta) < v_{k_o}\sigma(X^\tau\beta)) - u_{k_o} \end{pmatrix}.$$

This model was considered by Welsh, Carroll and Ruppert (1994) via a weighted Czech criterion. Here, we use EGLS to define an alternative estimator. Note that

$$E_{Y|X} G(Y, X^\tau\beta, \lambda) = \begin{pmatrix} F_{e|X}(\frac{m(X^\tau\beta) - m(X^\tau\beta_o) + v_1\sigma(X^\tau\beta)}{\sigma(X^\tau\beta_o)}) - u_1 \\ \cdots \\ F_{e|X}(\frac{m(X^\tau\beta) - m(X^\tau\beta_o) + v_{k_o}\sigma(X^\tau\beta)}{\sigma(X^\tau\beta_o)}) - u_{k_o} \end{pmatrix},$$

and $E_{Y|X}G(Y, X^\tau\beta_o, \lambda_o)G^\tau(Y, X^\tau\beta_o, \lambda_o) =$

$$\begin{pmatrix} u_1(1-u_1) & F_{e|X}(\min\{v_1, v_2\}) - u_1u_2 & \cdots & F_{e|X}(\min\{v_1, v_{k_o}\}) - u_1u_{k_o} \\ \cdots & \cdots & \cdots & \cdots \\ F_{e|X}(\min\{v_{k_o}, v_1\}) - u_{k_o}u_1 & \cdots & \cdots & u_{k_o}(1-u_{k_o}) \end{pmatrix}. \tag{3.1}$$

Then $R(\beta, \lambda)$ can be calculated by the definition.

Note that in Example 3.3, we can develop some instrumental variable estimator based on a characterization of the conditional quantile as the solution to a particular expected loss minimization problem [see Powell, 1994]. However, it seems difficult to find such kind of characterizations in Examples 1.1 and 3.4.

The second observation to be made is when $\theta(X) = X^\tau\beta$ (or $\theta(x) = (x^\tau\beta, \lambda)$), the iterative algorithm of Carroll and Ruppert (1988) can be directly applied here by the following steps:

Step 1: Start off with an initial estimator of β, say $\hat{\beta}_o$.

Step 2: Estimate $E_{Y|X}G(Y, X^\tau\hat{\beta}_o)G^\tau(Y, X^\tau\hat{\beta}_o)$ and $E_{Y|X}G(Y, X^\tau\beta)$ by the kernel regression. Calculate

$$\hat{R}_{s1}(\beta) = \hat{E}_X\{\hat{E}_{Y|X}G(Y, X^\tau\beta)^\tau[\hat{E}_{Y|X}G(Y, X^\tau\beta_o)G(Y, X^\tau\beta_o)^\tau]^{-1}\hat{E}_{Y|X}G(Y, X^\tau\beta)\} \tag{2.5}$$

and optimize $\hat{R}_{s1}(\beta)$ with respect to β. Let $\hat{\beta}_1$ denote the solution.

Step 3. Update $\hat{\beta}_o$ by $\hat{\beta}_1$ and return to *Step 2.*

Repeat *Steps 1 to 3* until the solution $\hat{\beta}_1$ is stable.

Similar algorithms for the $\hat{R}_M$, $\hat{R}_N$ and $\hat{R}_c$ based estimator can be designed.

The last obeservation is similar to LeBlanc and Crowley (1995). That is, the approximation (2.3) suggest the l_{sr} based estimators allow certain adaptive for heteroscedasticity. This is not surprising because, as a hybrid of the local and global modelling techniques, the SEL approach should inherit this feature of the local estimation. To illustrate this point we consider the following example.

Example 3.5. Consider the ordinary nonparametric regression model $Y = \theta(X) + e$ with unknown function $\theta(X)$ and $E_{e|X}e = 0$. Let (y_i, x_i), $i = 1, \cdots, n$ be i.i.d. observations. Write $G(Y, \theta(X), \eta) = Y - \theta(X)$. Then

$$\begin{aligned} l_{sr}(\theta) &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n(\sum_{j=1}^n w_{ji}(y_j - \theta(x_j)))^2[\sum_{j=1}^n w_{ji}(y_j - \theta(x_j))^2]^{-1}(1 + o_p(1)) \\ \hat{R}_M(\theta) &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n(y_i - \theta(x_i))^2[\sum_{j=1}^n w_{ji}(y_j - \theta(x_j))^2]^{-1}. \end{aligned}$$

$l_{sr}(\theta)$ is a summation of n local χ^2-statistics for θ. $l_{sr}(\theta)$ and $\hat{R}_M(\theta)$ are asymptotically adaptive for the unequal variances. Especially, $\hat{R}_M(\theta)$ recovers the objective function for weighting the unequal variances of errors suggested by Silverman (1985) in his smoothing spline estimator. The numerical results in Silverman (1985) support our observation.

4. Asymptotic properties of the proposed estimators

This section presents some general theory on the consistency, convergence rates and asymptotic normalities of the SEL, EGLS, modified EGLS, and LeBlanc and Crowley estimators. Since the assumptions and the corresponding results are almost the same for these estimators, we focus only on the SEL and EGLS estimators in the following subsections. We will adopt the kernel weights, the bandwidth h and η defined in Section 2. Let $||\theta||_\infty = sup_{x\in\Omega}||\theta(x)||$ and $||\theta||^2 = E_X||\theta(X)||^2$. We assume that the true value of the parameter of interest is the inner point of the parameter space.

4.1. **Weak consistency of the SEL estimator**

The weak consistency of the maximum empirical likelihood estimator has proved recently by Kitamura (1997). In this subsection, we show that the SEL estimator is also weakly consistent.

PROPOSITION 4.1. (i) PARAMETRIC REGRESSION FUNCTION CASE. *In addition to the conditions (X0), (K0), P1(i), P1(ii), P1(iv), P2 and* $0 < \eta < (\alpha_o - 4)/\alpha_o$, *we assume that as* $z \to 0$,

$$\begin{aligned}
&\sup_{x\in\Omega,||\beta-\beta_o||\le r_o} |E_{Y|X}G(Y,(x+z)^\tau\beta) - E_{Y|X}G(Y,x^\tau\beta)| \to 0,\\
&\sup_{x\in\Omega,||\beta-\beta_o||\le r_o} |E_{Y|X}G_m(Y,(x+z)^\tau\beta)G_k(Y,(x+z)^\tau\beta) - E_{Y|X}G_m(Y,x^\tau\beta)G_k(Y,x^\tau\beta)| \to 0,\\
&\qquad \text{for} 1 \le m,k \le k_o;\\
&\sup_{||\beta-\beta_o||\ge\rho} ||EG(Y,X^\tau\beta)|| \ne 0, \text{ for any } \quad \rho > 0.
\end{aligned}$$

Then, the SEL estimator is weakly consistent.

(ii) NONPARAMETRIC REGRESSION FUNCTION CASE. *In addition to the conditions (X0), (K1), N1(i), N1(ii), N2 and* $0 < \eta < 2(\alpha_o - 4)/(\alpha_o(2 + w^*))$, *we assume that as* $z \to 0$,

$$\begin{aligned}
&\sup_{x\in\Omega,||\theta-\theta_o||\le r_o} ||E_{Y|X}G(Y,\theta(x+z)) - E_{Y|X}G(Y,\theta(x))|| \to 0,\\
&\sup_{x\in\Omega,||\theta-\theta_o||\le r_o} |E_{Y|X}G_m(Y,\theta(x+z))G_k(Y,\theta(x+z)) - E_{Y|X}G_m(Y,\theta(x))G_k(Y,\theta(x))| \to 0,\\
&\qquad \text{for } 1 \le m,k \le k_o;\\
&\sup_{||\theta-\theta_o||\ge\rho} ||EG(Y,\theta(X))|| \ne 0, \qquad \text{for any } \rho > 0.
\end{aligned}$$

Then, the SEL estimator is weakly consistent.

REMARK 4.1. Note that in Proposition 4.1, if we further assume that there exist some positive constants c_1 and c_2 such that

$$\sup_{||\beta-\beta_o||\ge\rho} ||EG(Y,X^\tau\beta)|| \ge c_1\rho, \qquad \text{and} \quad \sup_{||\theta-\theta_o||\ge\rho} ||EG(Y,\theta(X))|| \ge c_2\rho$$

respectively for cases (i) and (ii), then similar to Theorem 2.1 we can prove that the SEL estimator has a convergence rate of $O_p(n^{-1/\alpha_o})$.

By substituting the above SEL estimator into $\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(Y,\theta(x))$, we obtain a consistent estimator for $E_{Y|X}G(Y,\theta_o(x))G^\tau(Y,\theta_o(x)), x \in \Omega$. Then, under the above conditions, the EGLS estimator is weakly consistent.

4.2. Asymptotic normality and convergence rates of the SEL and EGLS estimators

Note that the SEL and EGLS estimators have the same asymptotic behavior according to Theorem 2.1 and Remark 4.1. So in this section, we restrict our attention only on certain neighborhoods of the true value of the parameter governed by a sequence of positive constants $r_n \to 0$. We consider the estimator $\hat{\theta}$ defined by minimizing

$$\hat{R}_s(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\sum_{j=1}^{n} w_{ji}G(y_j,\theta(x_j))^\tau[\sum_{j=1}^{n} w_{ji}G(y_j,\theta(x_j))G(y_j,\theta(x_j))^\tau]^{-1}\sum_{j=1}^{n} w_{ji}G(y_j,\theta(x_j)).$$

In the *parametric regression function setting,* we take $\theta(x) = x^\tau\beta$ for example, where β being $p-$vector, and the minimization is performed over $||\beta - \beta_o|| \le r_n$. The corresponding estimator of β is denoted by $\hat{\beta}$. In the *nonparametric regression function* setting, we assume θ is one-dimensional function on $\Omega = C[0,1]$. Let $||\theta||_H = \sup_{x\neq y}|\theta(x) - \theta(y)|/|x-y|^r$, and

$$\Theta = \{\theta \in C^q[0,1] : \theta(0) = \theta(1), \theta^{(1)}(0) = \theta^{(1)}(1), ||\theta^{(j)}||_\infty \le L_j, ||\theta^{(q)}||_H \le L_q, \text{ for } j = 0, \cdots, q\}$$

where $r > 0$, and $L_j, j = 0, \cdots, q$ are fixed constants. We assume the minimization of $\hat{R}(\theta)$ is done over $||\theta - \theta_o|| \le r_n$ and $\theta \in \Theta$. Let "$\xrightarrow{d}$" denote convergence in distribution. By Theorem 2.1, it is easily seen that we only need to establish the asymptotic normality (convergence rate) of $\hat{\beta}$ ($\hat{\theta}$) defined in the above. To this end, in addition to the conditions (K0), (X0), (P1) and (P2), the following conditions are used.

(P3) For each $1 \le k \le k_o$, there exists a positive constant b_1 such that for any $||\beta - \beta_o|| \le r_o$ and $x \in \Omega$,

$$\begin{aligned} E_{Y|X}|G_k(Y, x^\tau\beta) - G_k(Y, x^\tau\beta_o)|^2 &\le b_1||\beta - \beta_o||, \\ |E_{Y|X}[G_k(Y,(x+u)^\tau\beta) - G_k(Y, x^\tau\beta)]| &= o(||\beta - \beta_o||), \text{ as } u \to 0. \end{aligned}$$

(P4) There exist $\alpha_o > 4$ and $0 < \zeta \le 1$ such that for $1 \le k \le k_o$ and any $||\beta - \beta_o|| \le r_o$,

$$|G_k(y, x^\tau\beta) - G_k(y, x^\tau\beta_o)| \le M(y,x)||\beta - \beta_o||^\zeta$$

with

$$\sup EM^{\alpha_o}(Y,X) < \infty, \quad \sup_{z\in\Omega} E_{Y|X}M^4(Y,z) < \infty.$$

(P5) $G(Y, X^\tau\beta)$ has the first three order continuous derivatives such that for $1 \le k, j \le r$, and some constants $r_o > 0$ and $\gamma_i > 2$, $i = 1, 2$,

$$\begin{aligned} &E\sup_{||\beta-\beta_o||\le r_o}|\tfrac{\partial G_k(Y,X^\tau\beta)}{\partial\beta_m}|^{\gamma_1} < \infty, 1 \le m \le p, \\ &E\sup_{||\beta-\beta_o||\le r_o}|\tfrac{\partial^2 G_k(Y,X^\tau\beta)}{\partial\beta_m\partial\beta_u}|^{\gamma_2} < \infty, 1 \le m, u \le p, \\ &E\sup_{||\beta-\beta_o||\le r_o}|\tfrac{\partial^3 G_k(Y,X^\tau\beta)}{\partial\beta_m\partial\beta_u\partial\beta_v}|^2 < \infty, 1 \le m, u, v \le p. \end{aligned}$$

(N3) (i) There exist some measurable function $M(y,x)$ and a positive constant t_o such that for any $\theta_i \in \Theta$, $i=1,2$, and $k=1,\cdots,k_o$,

$$|G_k(y,\theta_1(x)) - G_k(y,\theta_2(x))| \le M(y,x)|\theta_1(x)-\theta_2(x)|$$

and

$$\sup_{x\in\Omega} E_{Y|X} \exp(t_o M(Y,x)) < \infty.$$

(ii) For $k=1,\cdots,k_o$, $E_{Y|X}G_k(Y,\theta(x))$ has a bounded first derivative with respect to θ which satisfies for any $\theta_i \in \Theta$, $i=1,2$, and $x\in\Omega$,

$$|\frac{\partial E_{Y|X}G_k(Y,\theta_1(x))}{\partial\theta} - \frac{\partial E_{Y|X}G_k(Y,\theta_2(x))}{\partial\theta}| \le c|\theta_1(x)-\theta_2(x)|$$

where c is a constant.

(iii) Let $U(x) = E_{Y|X}G(Y,\theta_o(x))G^\tau(Y,\theta_o(x))$. $U(x)$ satisfies $\sup_{x\in\Omega}||U(x+t)-U(x)|| \le c|t|$ and the minimum eigenvalue of $U(x)$ is uniformly positive with respect to $x\in\Omega$.

THEOREM 4.1. (PARAMETRIC REGRESSION FUNCTION CASE). *We suppose that for bounded G, the conditions (K0), (X0), (P1) to (P3) hold; for unbounded G, the conditions (K0), (X0), P1(i), P1(iv), P1(v), (P2) to (P4) hold. If $E_{Y|X=x}G(Y,X^\tau\beta)$ has a continuous derivative with respect to x, $r_n \to 0$ and*

$$0<\eta<\begin{cases} 1/2, & G \text{ is bounded},\\ \min\{\frac{\alpha_o-2}{\alpha_o+2}, \frac{\alpha_o-1}{\alpha_o+1}(\frac{1}{2}+\frac{\zeta_1-1}{\alpha_o-1})\}, & G \text{ is unbounded},\end{cases}$$

then

$$n^{1/2}(\hat\beta-\beta_o) \xrightarrow{d} N(0,V)$$

with

$$V^{-1} = E_X\left\{[\frac{\partial E_{Y|X}G^\tau(Y,X^\tau\beta_o)}{\partial\beta}][E_{Y|X}G(Y,X^\tau\beta_o)G^\tau(Y,X^\tau\beta_o)]^{-1}[\frac{\partial E_{Y|X}G(Y,X^\tau\beta_o)}{\partial\beta}]\right\}.$$

Here, we assume V is positive definite.

For the smooth function G, a slightly better result can be obtained by using the condition (P5).

THEOREM 4.2. (PARAMETRIC REGRESSION FUNCTION CASE). *Under the conditions (X0), (K1), P1(i), P1(ii), P1(iv), (P5), if*

$$0<\eta<\min\{\frac{\alpha_o-2}{\alpha_o+2}, \frac{\gamma_1-1}{\gamma_1+1}, \frac{\gamma_2-1}{\gamma_2+1}\},$$

then $\sqrt{n}(\hat\beta-\beta_o)$ has the same asymptotic distribution as that in Theorem 2.2.

REMARK 4.2. Taking $\hat\theta$ defined in Proposition 4.1 as an initial estimator, we conclude from Theorems 4.1 and 4.2 that under certain regularity conditions, the SEL and EGLS estimators asymptotically have the same distributions as the GMM estimators based on the optimal instrumental variables. Note that it follows from Bickel, Klaassen, Ritov and Wellner (1993), the SEL and EGLS estimators are asymptotically efficient in the sense of Hansen (1982) and Bickel, Klaassen, Ritov and Wellner (1993). It also follows from Hansen (1982) that $\partial E_{Y|X} G(Y, X^\tau \beta)/\partial \beta U^{-1}(X)$ is the optimal instrumental variable even if G is discontinuous. When G is discontinuous, it is still unknown whether the GMM estimator is asymptotically efficient.

The next theorem gives the convergence rate of the SEL (or EGLS) estimator. Let $w^* = 1/(q+r) < 1$, and $q \geq 1$, where q and r are defined at the beginning of this subsection. Let $b_1 = 2/(2+w^*)$, $b_2 = 2(1-w^*)/(2-w^*)$, $1 \leq a_1 \leq 2$ and

$$\begin{aligned}\pi^* = \ & \min\{\tfrac{1}{3(2/(1-w^*)-b_2-1)+2b_2-a_1b_1}, \ \tfrac{2}{(6+w^*)(2/(1-w^*)-b_2-1)+4b_2-2w^*-a_1b_1(2-w^*)}, \\ & \tfrac{1}{3(a_1b_1-b_2)+2b_2-a_1b_1}, \ \tfrac{2}{(6+w^*)(a_1b_1-b_2)+4b_2-2w^*-a_1b_1(2-w^*)}\}.\end{aligned}$$

For any $\pi_o < \pi^*$, write

$$\begin{aligned}\eta_1(\pi_o) &= \min\{\pi_o(a_1b_1 - b_2), \tfrac{2\pi_o}{1-w^*} - \pi_o(b_2+1)\}, \\ \eta_2(\pi_o) &= \min\{\tfrac{1}{3}(1 - 2b_2\pi_o + a_1b_1\pi_o), \tfrac{2}{6+w^*}(1 - 2b_2\pi_o + w^*\pi_o + \tfrac{1}{2}a_1b_1(2-w^*)\pi_o)\}.\end{aligned}$$

THEOREM 4.3. (NONPARAMETRIC REGRESSION FUNCTION CASE). *Suppose $\hat\theta$ is weakly consistent or there exists a weakly consistent initial estimator of θ_o. Then, under the conditions (X0), (K1), N1(i) and (N3), for any $0 \leq \pi_o < \pi^*$, $\eta_1(\pi_o) < \eta < \eta_2(\pi_o)$, we have*

$$||\hat\theta - \theta_o|| = O_p(\max\{n^{-1/(2+w^*)}, n^{-\pi_o/(1-w^*)}\}), \quad ||\hat\theta^{(1)} - \theta_o^{(1)}|| = O_p(\max\{n^{-(1-w^*)/(2+w^*)}, n^{-\pi_o}\}).$$

REMARK 4.3. (i) Obviously, Theorem 4.3 can be directly extended to the $p-$dimension. Also the exponential order moment condition N3(i) can be relaxed via the technique used in Theorem 2.1.

(ii) It follows from Theorem 5.3 that when $w^* \leq 1/4$,

$$||\hat\theta - \theta_o|| = O_p(n^{-1/(2+w^*)}), \quad ||\hat\theta^{(1)} - \theta_o^{(1)}|| = O_p(n^{-(1-w^*)/(2+w^*)})$$

which are the optimal convergence rates in the ordinary nonparametric regression case (see Example 3.5 and Stone (1982)). When $w^* = 1/2, 1/3$, $\pi^*/(1-w^*) = 0.276$ and 0.411 which have not attained the corresponding optimal values 0.4 and 4/9 in the ordinary nonparametric regression case. This may be due to the inaccuracy estimation of the convergence rate of $\hat\theta^{(1)}$ in terms of the uniform norm by using the interpolation inequality. Note that in the ordinary nonparametric regression case, Stone (1982) showed that the optimal convergence rates of the regression function in terms of the L_2 norm and the uniform norm, respectively, differ from each other by only a factor $\log n$.

4.3. **Examples**

In what follows, we assume that $\hat\beta$ and $\hat\theta$ are the SEL (or EGLS) estimators based on i.i.d. observations (x_i, y_i), $i = 1, \cdots, n$. Set $f_{e|X}(z, x) = \frac{\partial F_{e|X=x}(z)}{\partial z}$.

EXAMPLE 1.1 (CONTINUED). In addition to the assumptions made in Section 3, we assume $f_{e|X}(z,x)$ is continuous and bounded with respect to (z,x). Suppose that $||\beta_o||$ is bounded by a known constant, and that the conditions (K0), (X0) and $0 < \eta < 1/2$ hold. Then it follows from Theorems 2.1 and 4.1 that

$$n^{1/2}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, V)$$

with

$$V^{-1} = \sum_{k=1}^{k_o} E_X \left(\frac{(f_{e|X}(s_k, X) + f_{e|X}(-s_k, X) - f_{e|X}(s_{k-1}, X) - f_{e|X}(-s_{k-1}, X))^2}{2(F_{e|X}(s_k) - F_{e|X}(s_{k-1}))} XX^\tau \right).$$

This directly implies that if $f_{e|X}(z,x)$ is of the form $g_o(z/s(x))$ with g_o known and $s(x)$ unknown, then $\hat{\beta}$ is adaptive for heteroscedasticity. Moreover, we can show that, under certain smoothness and symmetry conditions on $f_{e|X}$, if $\max_{1\le k\le k_o}(s_k - s_{k-1}) \to 0$, $s_{k_o} \to \infty$, then

$$V^{-1} \to E_X\{E_{e|X}(\frac{\partial \log f_{e|X}(z,X)}{\partial z})^2 XX^\tau\}.$$

Thus, when the underlying probability of observations is symmetric, an approximately effcient and adaptive estimator can be obtained by refining the sequence of $\{S_k\}$ [see Bickel, Klaassen, Ritov and Wellner (1993)].

EXAMPLE 1.2 (CONTINUED). In addition to the assumptions made before, we assume $\Omega = [0,1]$, $E_{e|X}\exp(t_o|e|) < \infty$ for some positive constant t_o, $U(x) = E_{Y|X}G(Y,\theta_o(x))G^\tau(Y,\theta_o(x))$ is uniformly positive definite with respect to $x \in \Omega$, and that $\sup_{x\in\Omega}||U(x+t) - U(x)|| \le c|t|$ for some positive constant c. Then, Theorem 4.3 holds.

EXAMPLE 3.1 (CONTINUED). Assume that the error e satisfies the following conditions: $E_{e|X}e = 0$, $\sup_{x\in\Omega} E_{Y|X=x}|e|^{\alpha_o} < \infty$, and $\sup_{x\in\Omega} E_{e|X=x}e^2 > 0$ with $\alpha_o > 4$. Taking the least squares estimator as the initial estimator of β. Then, under the conditions (K0), (X0) and $0 < \eta \le 1/3$, it follows from Theorems 2.1 and 4.2 that

$$n^{1/2}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, V)$$

with $V^{-1} = E_X(XX^\tau/E_{e|X}e^2)$. Thus, $\hat{\beta}$ is asymptotically as efficient as Carroll's estimator (see Carroll (1982)). Since the estimator of LeBlanc and Crowley (1995) has the same distribution, the open problem in their paper mentioned in Section 1 is solved.

EXAMPLE 3.2 (CONTINUED). Under the conditions similar to those in Example 3.1, we have

$$n^{1/2}\left(\begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix} - \begin{pmatrix} \beta_o \\ \lambda_o \end{pmatrix}\right) \xrightarrow{d} N(0, V)$$

with

$$V^{-1} = E_X(E_{e|X}e^2)^{-1}TT^\tau, \quad T = (\frac{\partial\phi(X,\beta_o)}{\partial\beta}, E_{S|X}I(S=1))^\tau.$$

EXAMPLE 3.3 (CONTINUED). In addition to the assumptions made in Section 3, we assume $f_{e|X}(z,x)$ is continuous and bounded with respect to (z,x). Note that $f_{e|X}(0,x)$ may depend on

x as pointed out by Jung (1996). Suppose that $||\beta_o||$ is bounded by a known constant, and that the conditions (K0), (X0) and $0 < \eta < 1/2$ hold. Then it follows from Theorems 2.1 and 4.1 that

$$n^{1/2}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, V)$$

with $V^{-1} = \frac{1}{q(1-q)} E_X((f_{e|X}(0, X))^2 X X^\tau)$. Thus, $\hat{\beta}$ has the same asymptotic distribution as the quasi-likelihood based regression quantile estimator investigated by Jung (1996). In particular, it is adaptive to the unequal behavior of conditional densities of the error given different values of the predictor.

EXAMPLE 3.4 (CONTINUED). Note that by the assumption $f_{e|X}$ is independent of X. For simplicity, we denote $f_{e|X}(z, x)$ by $f_e(z)$. In addition to the assumptions made in Section 3, we assume $f_e(z)$ is continuous and bounded with respect to z. Suppose that $||\beta_o^*||$ is bounded by a known constant, and that the conditions (K0), (X0) and $0 < \eta < 1/2$ hold. Then it follows from Theorems 2.1 and 4.1 that

$$n^{1/2}\left(\begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix} - \begin{pmatrix} \beta_o \\ \lambda_o \end{pmatrix}\right) \xrightarrow{d} N(0, V)$$

where V can be calculated by using the definition in Theorem 4.1, formula (3.1), and

$$\partial E_{Y|X} G^\tau(Y, X^\tau\beta, \lambda)/\partial \begin{pmatrix} \beta \\ \lambda \end{pmatrix} = (f_e(v_k)\psi_k)_{1 \le k \ne k_1 \le k_o}$$

with

$$\psi_k = \begin{pmatrix} (\frac{\partial m(X,\beta)}{\partial\beta} + v_k \frac{\partial\sigma(X,\beta)}{\partial\beta})\sigma^{-1}(X, \beta) \\ \phi_k \end{pmatrix};$$

and ϕ_{k_1} is a $(k_o - 1)-$dimensional zero vector, and for $1 \le k \ne k_1 \le k_o$, ϕ_k is a $(k_o - 1)-$dimensional vector with a one in kth position and zeroes elsewhere.

EXAMPLE 3.5 (CONTINUED). In addition to the assumptions made before, we assume $\Omega = [0, 1]$, $E_{e|X} \exp(t_o|e|) < \infty$ for some positive constant t_o, $\min_{x\in\Omega} \sigma(x) > 0$, and $|\sigma^2(x+t) - \sigma^2(x)| \le c|t|$ for all $x \in \Omega$, where $\sigma^2(x) = E_{e|X=x} e^2$ and $c > 0$ is a constant. Then, a result similar to Example 1.2 holds. In particular, when $w^* \le 1/4$, the SEL (or EGLS) estimator attains the optimal global convergence rate in L_2 norm.

5. **Other related estimators**

For a *smooth function G*, the EGLS estimators can be viewed as some special kinds of the instrumental variable estimators discussed in Newey (1990). This can be seen by calculating the derivative of $\hat{R}_{s1}(\beta)$ in (2.5) and setting the expression to equal zero,

$$0 = \frac{2}{n} \sum_{j=1}^n [\sum_{i=1}^n w_{ji} \sum_{k=1}^n w_{ki} \frac{\partial G^\tau(y_k, x_k^\tau\beta)}{\partial\beta} \hat{U}^{-1}(x_i)] G(y_j, x_j^\tau\beta)$$

with

$$\hat{U}(x_i) = \sum_{j=1}^n w_{ji} G(y_j, x_j^\tau\hat{\beta}_o) G^\tau(y_j, x_j^\tau\hat{\beta}_o),$$

for $1 \leq i \leq n$. This is just the instrumental estimating equation with estimated instrumental variables

$$\sum_{i=1}^{n} w_{ji} \sum_{k=1}^{n} w_{ki} \frac{\partial G^{\tau}(y_k, x_k^{\tau}\beta)}{\partial \beta} \hat{U}^{-1}(x_i), 1 \leq j \leq n.$$

Analogously, the $\hat{R}_N$ based estimator can be regarded as the approximate solution to the instrumental estimating equation with estimated instrumental variables

$$\sum_{i=1}^{n} w_{ji} \frac{\partial G^{\tau}(y_j, x_j^{\tau}\beta)}{\partial \beta} \hat{U}^{-1}(x_i), 1 \leq j \leq n.$$

The $\hat{R}_c$ based estimator is the approximate solution to the instrumental estimating equation with the estimated instrumental variable

$$\begin{array}{l} 2\sum_{i=1}^{n} w_{ji} \sum_{k=1}^{n} w_{ki} \frac{\partial G^{\tau}(y_k, x_k^{\tau}\beta)}{\partial \beta} \hat{U}^{-1}(x_i) \\ -\sum_{i=1}^{n} w_{ji} (\sum_{k=1}^{n} w_{ki} \frac{\partial G^{\tau}(y_k, x_k^{\tau}\beta)}{\partial \beta}) \hat{U}^{-1}(x_i) G(y_i, x_i^{\tau}\hat{\beta}_o) G(y_i, x_i^{\tau}\hat{\beta}_o) \hat{U}^{-1}(x_i), 1 \leq j \leq n. \end{array}$$

Moreover, we can directly construct the global empirical likelihood via the estimated instrumental variables as follows. Suppose we have an initial estimator $\hat{\beta}_o$. Usually, $\frac{\partial G(y, x^{\tau}\beta)}{\partial \beta}$ can be decomposed into two parts: $D(x)$ and $\frac{\partial G(y, x^{\tau}\beta)}{\partial \beta} - D(x)$, where $D(x)$ depends only on x. Then, we have the estimated instrumental variables

$$IV(x_i, \beta) = [D(x_i) + \sum_{j=1}^{n} w_{ji} (\frac{\partial G(y_j, x_j^{\tau}\beta)}{\partial \beta} - D(x_j))] \hat{U}^{-1}(x_i), 1 \leq i \leq n.$$

The empirical likelihood, namely l_a, can be constructed by

$$\max \sum_{i=1}^{n} \log p_i$$

subject to

$$\sum_{i=1}^{n} p_i = 1, p_i \geq 0, 1 \leq i \leq n, \text{ and } \sum_{i=1}^{n} p_i IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta) = 0.$$

The final solution is

$$l_a(\beta) = -\sum_{i=1} \log(n(1 + \eta^{\tau}(\beta) IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta)))$$

where $\eta(\beta)$ satisfies

$$\sum_{i=1}^{n} \frac{IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta)}{n(1 + \eta^{\tau}(\beta) IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta))} = 0.$$

It is easy to show that under certain regularity conditions the GMM criterion can be used as a local quadratic approximation of l_a:

$$\begin{array}{rl} l_a(\beta) & = -n \log n - \frac{1}{2} [\sum_{i=1}^{n} IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta)]^{\tau} [E_X IV(X, \beta_o) U(X) IV^{\tau}(X, \beta_o)]^{-1} \\ & \times [\sum_{i=1}^{n} IV(x_i, \hat{\beta}_o) G(y_i, x_i^{\tau}\beta)] (1 + o_p(1)). \end{array}$$

6. **Discussions**

In this paper, we proposed the SEL and EGLS approaches to the estimation of model (1.1). The large sample study shows such methods are promising. Like the parametric likelihood based maximum estimators [see, for example, Lehmann (1983) and Wong and Severini (1991)], the SEL based maximum estimators including that of LeBlanc and Crowley (1995) are asymptotically optimal in the parametric regression function cases and can achieve the optimal global convergence rate in some nonparametric regression function cases. It is shown that the SEL procedure is adaptive for heteroscedasticity in the model. Furthermore, the SEL ratio statistic for a finite dimensional parameter is asymptotically χ^2 distributed [see LeBlanc and Crowley (1995)]. Recently this theorem has been extended to the case of an infinite dimensional parameter [see Fan, Zhang and Zhang (1999) and Fan, Liu and Zhang (1999)]. Some issues on whether the SEL estimator has property (c) of the parametric likelihood mentioned in Section 1, how to select the bandwidth for a finite sample in practice and how to compute the associated estimators have not been studied yet. In a separate paper, we will compare via simulations the various estimators defined in Sections 2.

In the nonparametric regression function case, as a generalization of Fan, Heckman and Wang (1995), the local polynomial empirical likelihood provides an alternative estimator. Its behavior is the subject of work in progress.

Extensions of the SEL and EGLS methods to the model with θ being composed of a finite dimensional parameter of interest and an infinite-dimensional nuisance parameter are straightforward. In the partial linear model, we used the penalized SEL approach to give some new estimator of the parameter of interest. The study of the behavior of such kind of estimators is undertaking.

Like LeBlanc and Crowley (1995), the SEL can be constructed for censored survival data and for random effect regression model. However, their properties are still unknown.

7. **Proofs of the main theorems**

Throughout the remains of the paper, we denote $f_{max} = \max_x f(x)$ and $f_{min} = \min_x f(x)$. The proofs of the lemmas are deferred to the Appendix.

Proof of Proposition 2.1. Let Q_n be the empirical distribution based on $x_i, 1 \le i \le n$, and $Q = P_X$. Let $s_n = 2m_n f_{min} c(p)/n$ with $f_{min} = \min_x f(x)$. Consider the empirical processes

$$\{(Q_n - Q)I(||\cdot - x|| \le s) = \frac{1}{n}\sum_{k=1}^{n} I(||x_k - x|| \le s) - EI(||X - x|| \le s) : x \in \Omega, s \in [0, s_n]\}$$

where $I(\cdot)$ stands for the indicator function of a set. By the uniform continuity of f and the assumption on m_n, as $n \to \infty$,

$$\begin{aligned}\sup_{x\in\Omega, 0\le s\le s_n} Var(I(||X - x|| \le s) &\le \sup_{x\in\Omega, 0\le s\le s_n} EI(||X - x|| \le s)\\ &= \sup_{x\in\Omega, 0\le s\le s_n} \int_{||t||\le s}(f(t + x) - f(x))dt\\ &\quad + f(x)s^p c(p)\\ &\le (1 + o(1)) \max_{x\in\Omega} f(x)s_n^p c(p).\end{aligned}$$

Invoking that $ns_n^p/\log n \to \infty$ and using Theorem 37 of Pollard (1984, p.34), we have

$$\sup_{x\in\Omega} |(Q_n - Q)I(||\cdot - x|| \le s)| = o(s^p)$$

provided $0 \le s \le s_n$. This implies as $n \to \infty$,

$$Q_n I(||\cdot - x|| \le s) = (1 + o(1))f(x)s^p c(p) \tag{7.1}$$

where $o(1) \to 0$ almost surely and uniformly in $x \in \Omega$ and $0 \le s \le s_n$. Now for any $0 < \delta < 1$ and $x \in \Omega$, let

$$h_\delta(x, m_n) = \frac{(1+\delta)m_n}{nf(x)c(p)}, \quad h_{-\delta}(x, m_n) = \frac{(1-\delta)m_n}{nf(x)c(p)}.$$

Note that $0 < h_{-\delta}(x, m_n) < h_{-\delta}(x, m_n) < s_n$. Using (7.1), we have

$$\begin{aligned} \tfrac{m_n}{n} &> (1-\delta)(1+o(1))\tfrac{m_n}{n} \\ &= Q_n I(||\cdot - x|| \le h_{-\delta}(x, m_n)) \\ &\le Q_n I(||\cdot - x|| \le h_\delta(x, m_n)) \\ &= (1+\delta)(1+o(1))\tfrac{m_n}{n} \\ &> \tfrac{m_n}{n} \end{aligned}$$

uniformly in $x \in \Omega$ when n is sufficiently large. This together with the definition of $h(x, m_n)$ shows

$$h_{-\delta}(x, m_n) \le h(x, m_n) \le h_\delta(x, m_n)$$

uniformly for $x \in \Omega$ when n is sufficiently large. The proof is completed.

Define

$$\begin{aligned} S_n(x, \theta) &= \textstyle\sum_{j=1}^n w_j(x) G(y_j, \theta(x_j)) G^\tau(y_j, \theta(x_j)), \\ Z_n(\theta) &= \max_{1\le j\le n} ||G(y_j, \theta(x_j))||, \\ A_n(x, \theta) &= \textstyle\sum_{j=1}^n w_j(x) G(y_j, \theta(x_j)). \end{aligned}$$

Let $e_n(x, \theta)$ be the minimum eigenvalue of $S_n(x, \theta)$.

Lemma 7.1. If

$$\sup_{\theta\in\Theta, ||\theta-\theta_o||\le r_n, x\in\Omega} \frac{(1 + ||Z_n(\theta)||)||A_n(x, \theta)||}{e_n(x, \theta))} = o_p(1),$$

then (2.3) and (2.4) hold uniformly in $\theta \in \Theta, ||\theta - \theta_o|| \le r_n$, and $x \in \Omega$.

The following lemma directly follows from the proof of Theorem 37 of Pollard (1984).

Lemma 7.2. Let $\mathcal{F}_n$ be a class of functions with envelop 1. Let υ_n denote the empirical process

$$\upsilon_n(g) = n^{-1/2} \sum_{j=1}^n (g(y_j, x_j) - Eg(Y, X))$$

for $g \in \mathcal{F}_n$. Suppose that for some constants v and w,

$$\sup_{g\in\mathcal{F}_n} Var(g) \le v, \quad N(\epsilon, L_1(P_n), \mathcal{F}_n) \le A(P_n)\epsilon^{-w}$$

where $\limsup EA(P_n)$ is bounded by some positive constant $A(P)$. Then, for $M > 0$, when n is sufficiently large,

$$P(\sup_{\mathcal{F}_n} |\upsilon_n(g)| > M) \le c_1(P)(\sqrt{n}/M)^w \exp(-M^2/(2 \times 64^2 \times v)) + c_2(P)v^{-w}\exp(-nv)$$

where $c_1(P)$ and $c_2(P)$ are two positive constants.

Lemma 7.3 (Shen and Wong (1994)). Let $\mathcal{F}_n$ be a class of functions with envelop 1 and $Eg(Z) = 0$. Suppose that for some positive constants A_o, a, r, and d,

$$H^B(u, L_2(P), \mathcal{F}_n) \le \begin{cases} A_o(n^{2d}u)^{-r}, & u \in (0, a], \\ A_o, & u > a. \end{cases}$$

Let $\upsilon_n(g) = n^{-1/2}\sum_{i=1}^n g(z_i)$ and $v \ge \sup_{\mathcal{F}} Var(g)$. Then, there exist some positive constants c_i, $i = 1, \cdots, 4$ such that

$$P^*(\sup_{\mathcal{F}_n} |\upsilon_n(g)| > M) \le 6\exp(-(1-\epsilon)M^2/(10v))$$

provided $M \le \epsilon n^{1/2}v/4$, $v \le 1$, and

$$M \ge \begin{cases} \max(c_1 n^{[r-2-8dr]/(2r+4)}, c_2 n^{-dr}v^{(2-r)/4}), & 0 < r < 2, \\ c_3 n^{-dr}\log n, & r = 2, \\ c_4 n^{-2d} n^{(r-2)/(2r)}, & r > 2, \\ \max(c_1 n^{-1/2}, c_2 v^{1/2})\max((\log(1/(n^{2d}v^{1/2})))^{1/2}, 2), & r = 0^+; \end{cases}$$

where u^{0^+} deontes $\log u$.

For convenience, in the following we use $A_n(x,\beta)$, $Z_n(\beta)$, $S_n(x,\beta)$ and $e_n(x,\beta)$ to denote the symbols $A_n(x,\theta)$, $Z_n(\theta)$, $S_n(x,\theta)$ and $e_n(x,\theta)$ defined in Lemma 7.1. Lemmas 7.4∼7.7 are key steps in proving Theorem 2.1.

Lemma 7.4. Suppose $\theta(X) = X^\tau\beta$. Then, under the conditions (K0), (X0), P1(i), P1(ii), P1(iii), P1(v) and (P2), $0 < \eta < (\alpha_o - 2)/\alpha_o$ and $r_n \le n^{-1/\alpha_o}$, we have

$$\sup(1 + Z_n(\beta))|A_n(x,\beta)| = o_p(1)$$

where the supremum is with respect to (x, β, h^p) with $\{x \in \Omega, ||\beta - \beta_o|| \le r_n, h^p_{nl} \le h^p \le h^p_{nu}\}$.

Lemma 7.5. Suppose $\theta(X) = X^\tau\beta$. Then, under the conditions (K0), (X0), P1(i), P1(ii), P1(iii), P1(iv), $0 < \eta < (\alpha_o - 2)/(\alpha_o + 2)$ and $r_n \to 0$, we have

$$\begin{aligned} S_n(x,\beta) &= E_{Y|X}G(Y, x^\tau\beta)G^\tau(Y, x^\tau\beta) + o_p(1) \\ &= E_{Y|X}G(Y, x^\tau\beta_o)G^\tau(Y, x^\tau\beta_o) + o_p(1) \end{aligned}$$

where $o_p(1)$ is uniform in $x \in \Omega, ||\beta - \beta_o|| \le r_n$ and $h^p_{nl} \le h^p \le h^p_{nu}$.

Lemma 7.6. Suppose θ belongs to Θ. Under the conditions (K1), (X0), N1(i), N1(ii), N1(iii), N1(v), (N2), $r_n < n^{-1/\alpha_o}$, and

$$0 < \eta < \begin{cases} \min\{\frac{\alpha_o - 2}{\alpha_o}, \frac{2(\alpha_o - 2 + w^*)}{(2+w^*)\alpha_o}\}, & 0 < w^* \le 2, \\ \min\{\frac{\alpha_o-2}{\alpha_o}, \frac{1}{w^*}\}, & w^* > 2, \end{cases}$$

we have

$$\sup(1+Z_n(\theta))|A_n(x,\theta)| = o_p(1)$$

where the supremum is respect to (x, θ, h^p) with $x \in \Omega$, $||\theta-\theta_o|| \le r_n, \theta \in \Theta$ and $h^p_{nl} \le h^p \le h^p_{nu}$.

LEMMA 7.7. Suppose that θ belongs to Θ. Then, under the conditions (K1), (X0), N1(i), N1(ii), N1(iii), N1(iv), (N2), $r_n \to 0$, and

$$0 < \eta < \begin{cases} \frac{2(\alpha_o-4)}{2\alpha_o+w^*(\alpha_o-4)}, & 0 < w^* \le 2, \\ \frac{\alpha_o-4}{(\alpha_o-2)w^*}, & w^* > 2, \end{cases}$$

we have

$$\begin{aligned} S_n(x,\theta) &= E_{Y|X} G(Y,\theta(x)) G^\tau(Y,\theta(x)) + o_p(1) \\ &= E_{Y|X} G(Y,\theta_o(x)) G^\tau(Y,\theta_o(x)) + o_p(1) \end{aligned}$$

where $o_p(1)$ is uniform in $x \in \Omega$, $||\theta-\theta_o|| \le r_n$, $\theta \in \Theta$, and $h^p_{nl} \le h^p \le h^p_{nu}$, and θ_o is the true value of θ.

PROOF OF THEOREM 2.1. Since the proofs for (2.3) and (2.4) are similar, we only prove (2.3) below. First, we note that under condition (X0), by Proposition 2.1, there exist two positive constants d^*_o and d^*_1 independent of x such that when n is large,

$$h^p_{nl} = d_o n^{-\eta} \le h^p \le d_1 n^{-\eta} = h^p_{nu}$$

for both the constant bandwidth and the nearest neighbor bandwidth $h = h(x, m_n)$. Thus, it suffice to show that the condition in Lemma 7.1 holds uniformly for $h^p_{nl} \le h^p \le h^p_{nu}$. The details of such verifications are summarized in Lemmas 7.4 and 7.5 for the parametric regression function case, and in Lemmas 7.6 and 7.7 for the nonparametric regression function case. The proof is completed.

PROOF OF PROPOSITION 4.1. We take the parametric regression case to show the SEL estimator is consistent. The proofs for the other cases are similar. Like Kitamura (1997), we employ a variant of the classical argument developed by Wald (1949) to show that for any $\rho > 0$, as $n \to \infty$,

$$\begin{aligned} P(\sup_{||\beta-\beta_o|| \ge \rho} \ & \tfrac{1}{n} \textstyle\sum_{i=1}^n \sum_{j=1}^n w_{ji} [-\log(1+\alpha_n^\tau(x_i,\beta) G(y_j, x_j^\tau \beta))] \\ & < \tfrac{1}{n} \textstyle\sum_{i=1}^n \sum_{j=1}^n w_{ji} [-\log(1+\alpha_n^\tau(x_i,\beta_o) G(y_j, x_j^\tau \beta_o))]) \to 1. \end{aligned} \tag{7.2}$$

First, we set $A_n(x,\beta) = \sum_{j=1}^n w_j(x) G(y_j, x_j^\tau \beta)$ and prove under the assumptions,

$$\begin{aligned} 0 \ & \ge \tfrac{1}{n} \textstyle\sum_{i=1}^n \sum_{j=1}^n w_{ji} [-\log(1+\alpha_n^\tau(x_i,\beta_o) G(y_j, x_j^\tau \beta_o))] \\ & \ge -\tfrac{1}{n} \textstyle\sum_{i=1}^n \alpha_n^\tau(x_i,\beta_o) A_n(x_i,\beta_o) = O_p(n^{-2/\alpha_o}). \end{aligned} \tag{7.3}$$

This follows from the following claim: under the conditions (X0), (K0), P1(i), P1(ii) and P2, if $0 < \eta < (\alpha_o - 2)/(\alpha_o + 2)$, then

$$\begin{aligned} A_n(x,\beta_o) = &\ O_p(n^{-1/\alpha_o}), \\ \alpha_n(x,\beta_o) = &\ [E_{Y|X} G(Y, x^\tau \beta_o) G^\tau(Y, x^\tau \beta_o)]^{-1} A_n(x,\beta_o)(1+o_p(1)) \end{aligned}$$

which can be proved by the same argument used in the proof of Theorem 2.1.

Secondly, we set $|G(y,x)|^* = \sup_{||\beta-\beta_o||\le r_o} |G(y, x^\tau\beta)|$ and prove that under the assumptions, uniformly in $||\beta-\beta_o|| \le r_o$,

$$\begin{array}{l} \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n w_{ji}G(y_j, x_j^\tau\beta)I(|G(y_j,x_j)|^* \le n^{1/\alpha_o}) \to EG(Y, X^\tau\beta), \\ \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n w_{ji}G(y_j, x_j^\tau\beta)G^\tau(y_j, x_j^\tau\beta)I(|G(y_j,x_j)|^* \le n^{1/\alpha_o}) \to EG(Y, X^\tau\beta)G^\tau(Y, X^\tau\beta). \end{array} \tag{7.4}$$

To show the first equality in (7.4), it suffices to show under the assumptions,

$$\begin{array}{l} \frac{1}{nh^pf(x)}\sum_{j=1}^n K(\frac{||x_j-x||}{h})G(y_j, x_j^\tau\beta)I(|G(y_j,x_j)|^* \le n^{1/\alpha_o}) \\ \qquad -\frac{1}{h^pf(x)}EK(\frac{||X-x||}{h})G(Y, X^\tau\beta)I(|G(y_j,x_j)|^* \le n^{1/\alpha_o}) = o_p(1), \\ \frac{1}{f(x)}\int K(||t||)E_{Y|X}G(Y,(x+th)^\tau\beta)f(x+th)dt = E_{Y|X}G(Y, x^\tau\beta) + o(1), \\ \frac{1}{h^pf(x)}EK(\frac{||X-x||}{h})G(Y, X^\tau\beta)I(|G(Y,X)|^* > n^{1/\alpha_o}) = o(1) \end{array}$$

which follow from the same argument used in the proof of Theorem 2.1. The second equality in (7.4) can be shown by a similar way.

It follows from (7.4) that for $u(\beta) = EG(Y, X^\tau\beta)/||EG(Y, X^\tau\beta)||$ and some $\delta > 0$,

$$\begin{array}{l} P(\sup_{||\beta-\beta_o||\ge\rho}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n w_{ji}[-\log(1+\alpha_n^\tau(x_i,\beta)G(y_i, x_i^\tau\beta))] \ge -|0_p(n^{-2/\alpha_o})|) \\ \quad \le P(\sup_{||\beta-\beta_o||\ge\rho}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n w_{ji}[-\log(1+\alpha_n^\tau(x_i,\beta)G(y_i, x_i^\tau\beta))] \quad \ge -|O_p(n^{-2/\alpha_o})|, \\ \qquad \max_i|G(y_i,x_i)|^* \le n^{1/\alpha_o})) + P(\max_i|G(y_i,x_i)|^* > n^{1/\alpha_o}) \\ \quad \le P(\sup_{||\beta-\beta_o||\ge\rho}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n w_{ji}[-\log(1+n^{-1/\alpha_o-\delta}u(\beta)^\tau G(y_j, x_j^\tau\beta) \\ \qquad \times I(|G(y_j,x_j)|^* \le n^{1/\alpha_o})] \ge -|0_p(n^{-2/\alpha_o})|) + o_p(1) \\ \quad = P(-n^{-1/\alpha_o-\delta}\inf_{||\beta-\beta_o||\ge\rho}||EG(Y, X^\tau\beta)||(1+o_p(1)) + 0_p(n^{-(2/\alpha_o+2\delta)}) \ge -|0_p(n^{-2/\alpha_o})|) \\ \qquad + o_p(1) \\ \quad = o_p(1). \end{array}$$

This together with (7.3) implies (7.2). The proof is completed.

Let $U(x) = E_{Y|X}G(Y, x^\tau\beta_o)G^\tau(Y, x^\tau\beta_o)$ below. Lemmas 7.8 to 7.11 below will be used in the proof of Theorem 4.1.

LEMMA 7.8. We suppose that for the bounded G, the conditions (K0), (X0), P1(i), P1(iv), P1(v), (P2) and (P3) hold; and for the unbounded G, the conditions (K0), (X0), P1(i), P1(ii) and (P2) to (P4) hold. If $r_n \to 0^+$ and

$$0 < \eta < \left\{ \begin{array}{cl} 1/2, & G \text{ is bounded}, \\ (\alpha_o - 1)/(2(\alpha_o+1)), & G \text{ is unbounded}, \end{array} \right.$$

then

$$\sum_{k=1}^n w_k(x)(G(y_k, x_k^\tau\beta) - G(y_k, x_k^\tau\beta_o))^\tau - E_{Y|X}[G^\tau(Y, x^\tau\beta)] = o_p(\max\{||\beta-\beta_o||, n^{-1/2}\})$$

where o_p is uniform in (x, β, h) with $x \in \Omega, ||\beta-\beta_o|| \le r_n$ and $h_{nl}^p \le h^p \le h_{nu}^p$, and h_{nl}^p and h_{nu}^p are defined as in the proof of Theorem 2.1.

LEMMA 7.9. Under the same conditions in Lemma 7.8, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n w_i(x)E_{Y|X}[G^\tau(Y,x_i^\tau\beta)]U^{-1}(x_i)) &= E_{Y|X}[G^\tau(Y,x^\tau\beta)U^{-1}(x)]\\ &+o_p(\max\{||\beta-\beta_o||,n^{-1/2}\})\end{aligned}$$

where $o_p(1)$ is uniform in $x\in\Omega$, $||\beta-\beta_o||\le r_n$ and $h_{nl}^p\le h^p\le h_{nu}^p$.

LEMMA 7.10. Suppose that when G is bounded, (P2), P1(i), P1(iii), P1(iv), P1(v) and (P3) hold; when G is unbounded, (P1) to (P4) hold. Then

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^n(G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o)-EG(Y,X^\tau\beta)) &= o_p(\max\{||\beta-\beta_o||,n^{-1/2}\}),\\ EG(Y,X^\tau\beta) &= O_p(||\beta-\beta_o||).\end{aligned}$$

LEMMA 7.11. Suppose that when G is bounded, (P2), P1(i), P1(iii), P1(iv), P1(v) and (P3) hold; when G is unbounded, (P1), (P2), (P3) and (P4) hold. Then

$$\begin{aligned}\frac{1}{n}\ &\sum_{j=1}^n E_{Y|X}[G^\tau(Y,x_j^\tau\beta)]U^{-1}(x_j)(G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o)-E_{Y|X}[G(Y,x_j^\tau\beta]))\\ &=o_p(\max\{||\beta-\beta_o||^2,n^{-1}\}).\end{aligned}$$

PROOF OF THEOREM 4.1. To begin with, we let o_p below denote a number tending to zero uniformly in $x\in\Omega$, $||\beta-\beta_o||\le r_n$ and $h_{nl}^p\le h^p\le h_{nu}^p$.

It follows from the conditions (P1) to (P4), and the argument in the proof of Theorem 2.1 that as $0<\eta<(\alpha_o-2)/(\alpha_o+2)$ and $r_n\to 0$,

$$\begin{aligned}\sum_{j=1}^n w_j(x)G(y_j,x_j^\tau\beta)G(y_j,x_j^\tau\beta)) &= \sum_{j=1}^n w_j(x)G(y_j,x_j^\tau\beta_o)G(y_j,x_j^\tau\beta_o))+o_p(1)\\ &=U(x)(1+o_p(1)).\end{aligned}$$

Substituting it into $\hat{R}_s(\beta)$, we have

$$\hat{R}_s(\beta)=\frac{1}{2n}\sum_{i=1}^n(\sum_{j=1}^n w_{ji}G^\tau(y_j,x_j^\tau\beta))U(x_i)^{-1}(\sum_{j=1}^n w_{ji}G(y_j,x_j^\tau\beta))(1+o_p(1)).$$

Without loss of generality, we drop the factor $1+o_p(1)$. Observe that

$$\hat{R}_s(\beta)-\hat{R}_s(\beta_o)=\frac{1}{2n}\sum_{j=1}^n(\sum_{i=1}^n w_{ji}H_n(\beta,x_i))(G(y_j,x_j^\tau\beta)+G(y_j,x_j^\tau\beta_o)) \tag{7.5}$$

where

$$H_n(\beta,x)=\sum_{k=1}^n w_k(x)(G(y_k,x_k^\tau\beta)-G(y_k,x_k^\tau\beta_o))^\tau U(x)^{-1}.$$

Appealing to Lemma 7.8, we have

$$H_n(\beta,x)=o_p(\max\{||\beta-\beta_o||,n^{-1/2}\})+E_{Y|X}[G^\tau(Y,x^\tau\beta)]U^{-1}(x)(1+o_p(1)).$$

This together with Lemma 7.9 leads to

$$\begin{aligned}\sum_{i=1}^n w_i(x)H_n(\beta,x_i) &= o_p(\max\{||\beta-\beta_o||,n^{-1/2}\})\\ &\quad+\sum_{i=1}^n w_i(x)E_{Y|X}[G^\tau(Y,x_i^\tau\beta)]U(x_i)^{-1}\\ &=o_p(\max\{||\beta-\beta_o||,n^{-1/2}\})+E_{Y|X}[G^\tau(Y,x^\tau\beta)]U(x)^{-1}.\end{aligned}$$

Combining this with (7.5), Lemmas 7.10 and 7.11, we obtain

$$\begin{aligned}\hat{R}_s(\beta)-\hat{R}_s(\beta_o) &= \tfrac{1}{n}\textstyle\sum_{j=1}^n E_{Y|X}[G^\tau(Y,x_j^\tau\beta)]U^{-1}(x_j)G(y_j,x_j^\tau\beta_o)\\ &+\tfrac{1}{2n}\textstyle\sum_{j=1}^n E_{Y|X}[G^\tau(Y,x_j^\tau\beta)]U^{-1}(x_j)(G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o))\\ &+o_p(\max\{||\beta-\beta_o||,n^{-1/2}\})[\textstyle\sum_{j=1}^n(G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o))\\ &+\tfrac{2}{n}\textstyle\sum_{j=1}^n G(y_j,x_j^\tau\beta_o)]\\ &= \tfrac{1}{n}\textstyle\sum_{j=1}^n E_{Y|X=x_j}[G^\tau(Y,X^\tau\beta)]U^{-1}(x_j)G(y_j,x_j^\tau\beta_o)\\ &+\tfrac{1}{2n}\textstyle\sum_{j=1}^n E_{Y|X=x_j}[G^\tau(Y,X^\tau\beta)]U^{-1}(x_j)E_{Y|X=x_j}[G(Y,X^\tau\beta)]\\ &+o_p(\max\{||\beta-\beta_o||^2,n^{-1}\}).\end{aligned}$$

Since $E_{Y|X}[G(Y,x^\tau\beta)]$ has a bounded continuous derivative with respect to β, we have

$$\begin{aligned}\hat{R}_s(\beta)-\hat{R}_s(\beta_o) &= (\beta-\beta_o)^\tau\tfrac{1}{n}\textstyle\sum_{j=1}^n\frac{\partial E_{Y|X}[G^\tau(Y,x_j^\tau\beta_o)]}{\partial\beta}U^{-1}(x_j)G(y_j,x_j^\tau\beta_o)\\ &+\tfrac{1}{2}(\beta-\beta_o)^\tau(\tfrac{1}{n}\textstyle\sum_{j=1}^n\frac{\partial E_{Y|X}[G^\tau(Y,x_j^\tau\beta_o)]}{\partial\beta}U^{-1}(x_j)\\ &\cdot\frac{\partial E_{Y|X}[G^\tau(Y,x_j^\tau\beta_o)]}{\partial\beta})(\beta-\beta_o)+o_p(\max\{||\beta-\beta_o||^2,n^{-1}\}).\end{aligned}$$

Then via the same argument and symbols of Theorem 5 of Pollard (1984, p.141), we have

$$\Delta=\Delta(y,x)=\frac{\partial E_{Y|X}[G^\tau(Y,x^\tau\beta_o)]}{\partial\beta}U^{-1}(x)G(y,x^\tau\beta_o)$$

and

$$\begin{aligned}E\Delta &= 0,\\ E\Delta\Delta^\tau &= E\{[\tfrac{E_{Y|X}G(Y,X^\tau\beta_o)}{\partial\beta}]^\tau U^{-1}(X)[\tfrac{E_{Y|X}G(Y,X^\tau\beta_o)}{\partial\beta}]\},\\ V &= E\Delta\Delta^\tau.\end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\beta}-\beta_o)\xrightarrow{d}N(0,V).$$

The proof is completed.

Proof of Theorem 4.2. The proof is very similar to the classical argument on the maximum likelihood estimator [see Lehmann (1983)]. The details are omitted.

Lemmas 7.12 to 7.15 below are employed to prove Theorem 4.3.

Lemma 7.12. Under the conditions (X0), (K1), N1(i), N3(i) and N3(ii), if $\pi_1\ge 0$, $q\ge 1$, and

$$\begin{aligned}\eta &\ge \pi_1(1-b_2),\\ \eta &> \pi_1(a_1b_1-b_2),\\ \eta &< \tfrac{1}{3}(1-2b_2\pi_1+a_1b_1\pi_1),\\ \eta &< \tfrac{2}{6+w^*}(1-2b_2\pi_1+w^*\pi_1+a_1b_1(1-\tfrac{w^*}{2})),\end{aligned}$$

then

$$\sum_{j=1}^n w_j(x)(G(y_j,\theta(x_j))-G(y_j,\theta_o(x_j))-E_{Y|X}G(Y,\theta(x))=O_p(n^{\eta+b_2\pi_1})$$

where O_p is uniform in θ and x with $||\theta-\theta_o||\le cn^{-\pi_1}$, $||\theta^{(1)}-\theta_o^{(1)}||\le c_1n^{-\pi_1}$ and $x\in\Omega$, and c_1 is any fixed positive constant.

Lemma 7.13. In addition to the assumptions in Lemma 7.12, we assume that N3(iii) holds. Then, we have

$$\sum_{i=1}^{n} w_i(x) E_{Y|X} G^\tau(Y, \theta(x_i)) U^{-1}(x_i) - E_{Y|X} G^\tau(Y, \theta(x)) U^{-1}(x) = O_p(n^{-(\eta + b_2 \pi_1)})$$

where O_p is uniform in θ and x with $||\theta - \theta_o|| \le c_1 n^{-\pi_1}$, $\theta \in \Theta$, and $x \in \Omega$.

Lemma 7.14. Under the conditions (X0), N1(i), N3(i) and N3(ii), if $\pi_1 < 1$, then

$$\frac{1}{n} \sum_{j=1}^{n} (G(y_j, \theta(x_j)) + G(y_j, \theta_o(x_j))) = O_p(n^{-\pi_1})$$

uniformly in θ with $||\theta - \theta_o|| \le c_1 n^{-\pi_1}$ and $\theta \in \Theta$.

Lemma 7.15. Under the conditions (X0), N1(i) and (N3), if $0 < \pi_1 < 1/(2 + w^*)$ and $0 < w^* < 2$. Then

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^{n} \quad & E_{Y|X} G^\tau(Y, \theta(x_j)) U^{-1}(x_j) (G(y_j, \theta(x_j)) - G(y_j, \theta_o(x_j)) - E_{Y|X} G(Y, \theta(x))) \\ & = O_p(n^{-\pi_1 - 1/(2 + w^*)}) \end{aligned}$$

uniformly in θ with $||\theta - \theta_o|| \le c_1 n^{-\pi_1}$ and $\theta \in \Theta$.

Proof of Theorem 4.3. First of all, we claim O_p and o_p below are uniform with respect to $\theta \in \Theta$. Note that, applying the arguments in the proof of Theorem 2.1, we obtain as $||\theta - \theta_o|| \le r_n \to 0$,

$$\hat{R}_s(\theta) - \hat{R}_s(\theta_o) = -\frac{1}{2} \sum_{i=1}^{n} (\sum_{j=1}^{n} w_{ji} G(y_j, \theta(x_j)))^\tau U^{-1}(x) \sum_{j=1}^{n} w_{ji} G(y_j, \theta(x_j))(1 + o_p(1)). \qquad (7.6)$$

The strategy adopted in the remainder of the proof, similar to Shen and Wong (1994), is to improve the rate iteratively by obtaining increasingly faster uniform approximation rate of the objection function $\hat{R}_s(\theta)$ in a sequence of shrinking neighborhoods.

Interation: Let the initial value of π_1 is zero. If at the previous step we have obtained

$$||\hat{\theta} - \theta_o|| = O_p(n^{-\pi_1}), \quad ||\hat{\theta}^{(1)} - \theta_o^{(1)}|| = O_p(n^{-\pi_1}),$$

and if π_1 satisfy

$$\begin{aligned} \eta \ge & \ \ \pi_1(1 - b_2), \\ \eta > & \ \ \pi_1(a_1 b_1 - b_2), \\ \eta < & \ \ \tfrac{1}{3}(1 - 2b_2\pi_1 + a_1 b_1 \pi_1), \\ \eta < & \ \ \tfrac{2}{6 + w^*}(1 - 2b_2\pi_1 + w^*\pi_1 + a_1 b_1(1 - \tfrac{w^*}{2})), \end{aligned}$$

then in the next step we can show

$$\begin{aligned} ||\hat{\theta} - \theta_o|| = & \ \ O_p(n^{-(\pi_1 + \pi_2)/2}) + O_p(n^{-(\pi_1/2 + 1/(2 + w^*))}, \\ ||\hat{\theta}^{(1)} - \theta_o^{(1)}|| = & \ \ O_p(n^{-(\pi_1 + \pi_2)(1 - w^*)/2}) + O_p(n^{-(\pi_1/2 + 1/(2 + w^*))(1 - w^*)}), \end{aligned} \qquad (7.7)$$

and replace π_1 by $(1 - w^*) \min\{(\pi_1 + \pi_2)/2, (\pi_1/2 + 1/(2 + w^*))\}$.

Write

$$a(x,\theta(x)) = U^{-1/2}(x)E_{Y|X}G(Y,\theta(x)), \quad W_j = -U^{-1/2}(x_j)G(y_j,\theta_o(x_j)), \quad j=1,\cdots,n.$$

Then, the combination of (7.6) and Lemmas 7.12 to 7.15 gives if

$$\begin{array}{rl} 1/(2+w^*) \geq & \pi_1 \geq 0, \quad w^* \leq 1, \\ \eta > & \pi_1(a_1b_1 - b_2), \\ \eta < & \frac{1}{3}(1 - 2b_2\pi_1 + a_1b_1\pi_1), \\ \eta < & \frac{2}{6+w^*}(1 - 2b_2\pi_1 + w^*\pi_1 + a_1b_1(1 - \frac{w^*}{2})), \end{array} \tag{7.8}$$

then

$$\begin{array}{rl} \hat{R}_s(\theta) - \hat{R}_s(\theta_o) & = -\frac{1}{n}\sum_{j=1}^n a^\tau(x_j,\theta(x_j))W_j + \frac{1}{2n}\sum_{j=1}^n ||a(x_j,\theta(x_j))||^2 \\ & +O_p(n^{-(\pi_2+\pi_1)}) + O_p(n^{-(\pi_1+2/(2+w^*))}) \end{array} \tag{7.9}$$

with $\pi_2 = \eta + b_2\pi_1$ and $b_2 = 2(1-w^*)/(2-w^*)$. It follows from Theorems 2.2 and 2.3 of Mammen and van de Geer (1997) that

$$\begin{array}{rl} \frac{1}{n}\sum_{j=1}^n W_j a(x_j,\theta(x_j)) = & (\max\{||a||, n^{-1/(2+w^*)}\})^{1-w^*/2}O_p(n^{-1/2}) \\ \frac{1}{n}\sum_{j=1}^n ||a(x_j,\theta(x_j))||^2 = & \max\{||a||^2, O_p(n^{-2/(2+w^*)})\} \end{array}$$

where $||a||^2 = E|a(X,\theta(X))|^2$ and O_p is uniform with respect to $\theta \in \Theta$. Thus, by (7.9) and the definition of $\hat{\theta}$, we find

$$\begin{array}{l} -\max\{||a(\cdot,\hat{\theta}(\cdot))||^{1-w^*/2}, n^{-(2-w^*)/(2(2+w^*))}\}O_p(n^{-1/2}) \\ \quad +\frac{1}{2}\max\{||a(\cdot,\hat{\theta}(\cdot))||^2, O_p(n^{-2/(2+w^*)})\} \\ \quad \leq O_p(n^{-(\pi_2+\pi_1)}) + O_p(n^{-(\pi_1+2/(2+w^*))}) \end{array}$$

which together with the assumptions implies the first equality in (7.7). Invoke the Sobolev interpolation inequality, there exists a constant b, for any $0 < \rho < 1$,

$$||\hat{\theta}^{(1)} - \theta_o^{(1)}||^2 \leq b(\rho^{-2}||\hat{\theta} - \theta_o||^2 + \rho^{2(q+r-1)}||\hat{\theta}^{(q)} - \theta_o^{(q)}||_H^2)$$

where $||\cdot||_H$ is the Hölder norm. Let $\rho = c_1||\hat{\theta} - \theta_o||^{w^*}$ with $w^* = 1/(q+r)$. Then, we derive the second equality in (7.7).

The above iteration continues if (7.8) holds, $\pi_1 < (1-w^*)/(2-w^*)$, $\eta \geq \pi_1(1-b_2)$, and $(\eta + 2b_2\pi_1 + \pi_1)(1-w^*) > 2\pi_1$.

Now the proof is completed by some simple calculations.

Acknowledgements. The first author has greatly benefited from several long discussions with Professor Wing Hung Wong, Department of Statistics, UCLA. Example 1.1 is borrowed from him. The work was partially supported by 'Project d'Actions de Recherche Concertées' (No. 93/98-164) from the Belgian government and by the National Natural Science Foundation of China. It was partially conducted while the first author was visiting Department of Statistics, the Chinese University of Hong Kong, and Institut de Statistique, Université Catholique de Louvain. The first author is grateful to both the Department and the Institute for their hosptitality and support.

REFERENCES

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

BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. AND WELLNER, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins University Press, Baltimore and London.

CARROLL, R.J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224-1233.

CARROLL, R.J., FAN, J., GIJBELS, I. AND WANG, M. (1997). Generalized partially linear single-index models. *Jour. Ameri. Statist. Asso.* **92** 477-489.

CARROLL R.J. AND RUPPERT, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

CARROLL, R.J., RUPPERT, D. AND WELSH, A.H. (1998). Local Estimating Equations. *Jour. Ameri. Statist. Asso.* **93** 214-227.

CHAMBERLAIN, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* **60** 567-596.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Jour. Ameri. Statist. Asso.* **74** 829-836.

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

FAN, J. (1997). Discussion on "Polynomial splines and their tensor products in extended linear modeling" by Stone, C., Hansen, M.H., Kooperberg, C. and Truong, Y. *Ann. Statist.* **25** 1425-1432.

FAN, J. AND GIJBELS, I. (1994). Censored regression: local linear approximations and their applications. *Jour. Ameri. Statist. Asso.* **89** 560-570.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

FAN, J. AND HECKMAN, N. AND WANG, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood function. *Jour. Ameri. Statist. Assoc.* **90** 141-150.

FAN, J. AND LIU, A. AND ZHANG, J. (1999). Sieve empirical likelihood ratio in the infinite dimensional parameter space. *manuscript.*

FAN, J. AND ZHANG, C. AND ZHANG, J. (1999). Sieve likelihood ratio statistics and Wilks phenomenon. *manuscript.*

GODAMBE, V.P. AND KALE, B.K. (1991). Estimating functions: an overview. in *Estimating Functions*, edited by V.P. Godambe, Clarendon Press, Oxford, p. 3-20.

GREEN, P.J. AND SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* Chapman and Hall, London.

HALL, A. (1993). Some aspects of generalized method of moments estimation. In *Handbook of Statistics, Vol.* **11**, Maddala, G.S., C.R. Rao and H.D. Vinod, eds, Elsevier Science Publishers B.V., Amsterdam, p. 393-417.

HEYDE, C. C. (1997). *Quasi-likelihood and its Application.* Springer, New York.

HJORT, N.L. AND JONES, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619-1647.

JUNG, S. (1996). Quasi-likelihood for median regression models. *Jour. Ameri. Statist. Asso.* **91** 251-257.

KITAMURA, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25** 2084-2102.

KITAMURA, Y. AND STUTZER, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65** 861-874.

KOENKER, R. AND BASSETT, G. (1978). Regression quantitles. *Econometrica* **46** 33-50.

KOLMOGOROV, A. AND TIHOMIROV, V. (1959). $\epsilon-$capacity of sets in function spaces. *Uspekhi Mat. Nauk.* **14** 3-86.

LEBLANC M. AND CROWLEY, J. (1995). Semiparametric regression functionals. *Jour. Ameri. Statist. Asso.* **90** 95-105.

LEHMANN, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

LOADER, C.R. (1996). Local likelihood density estimation. *Ann. Statist.* **24** 1602-1618.

MAMMEN, E. AND VAN DE GEER, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014-1035.

MCCULLAGH, P. AND NELDER, J.A. (1991). *Generalized Linear Models.* Chapman and Hall, London.

MYKLAND, P.A. (1995). Dual likelihood. *Ann. Statist.* **23** 396-421.

NEWEY, W.K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica* **58** 809-837.

NEWEY, W.K. (1993). Efficient Estimation of models with conditional moment restrictions. In *Handbook of Statistics, Vol.* **11**, Maddala, G.S., C.R. Rao and H.D. Vinod, eds, Elsevier Science Publishers B.V., Amsterdam, p. 419-453.

Ogaki, M. (1993). Generalized method of moments: econometric applica tion. In *Handbook of Statistics, Vol.* **11**, Maddala, G.S., C.R. Rao and H.D. Vinod, eds, Elsevier Science Publishers B.V., p. 455-488.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237-249.

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

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Powell, J.L. (1986). Censored regression quantiles. *Jour. Econometrics* **32** 143-155.

Powell, J.L. (1994). Estimation of semiparametric models. In *Handbook of Econometrics, Vol.* **IV**, Engle, R.F. and McFadden, D.L. eds, Elsevier Science Publishers B.V., Amsterdam, p. 2443-2521.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300-325.

Robinson, P.M. (1987). Asymptotically efficient estimation in the pr esence of heteroskedasticity. *Econometrica* **55** 875-891.

Shen, X. and Wong, W.H. (1994). Convergence rate of sieve estimates. *Ann. Statist.,* **22** 580-615.

Shen, X. (1997). On methods of sieves and penalization. *Ann. Statist.* **25** 2555-2591.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussions). *J. Royal Statist. Soc. B* **47** , 1-52.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.

Tibshirani, R. and Hastie, T.J. (1987). Local likelihood estimation. *Jour. Ameri. Statist. Asso.* **82**, 559-567.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595-601.

Welsh, A.H., Carroll and Ruppert, D. (1994). Fitting heteroscedastic regression models. *Jour. Ameri. Statist. Asso.* **89**, 100-116.

Wong, W.H. and Severini, T.A. (1991). On maximum likelihood estimation in infinite dimensional parameter space. *Ann. Statist.* **19** 603-632.

Wong, W.H. and Shen, X. (1995). Probability inequalities for likelih ood ratios and convergence rates of sieve MLES. *Ann. Statist.* **23** 339-362.

Appendix

In the following, let r_o denote a predetermined positive constant.

Proof of Lemma 7.1. It follows from the same argument as Owen (1988). The details are omitted.

Proof of Lemma 7.4. For convenience, in what follows we use c to denote a constant which may have different value in the different situations. For $1 \le k \le k_o$, and positive constants M_n and r_n, we set $|G_k|^* = |G_k(y,x)|^* = \sup_{||\beta-\beta_o||\le r_o} |G_k(y, x^\tau\beta)|$, and

$$\begin{aligned}\mathcal{F}_n = \quad & \{K(||\cdot - x||/h)G_k(\cdot, \cdot^\tau\beta)I(|G_k|^* \le M_n)/M_n : x\in\Omega, ||\beta-\beta_o|| \le r_n,\\ & h_{nl}^p \le h^p \le h_{nu}^p\}.\end{aligned}$$

Observe that

$$A_n(x,\beta) = (1+o_p(1))(A_{n1}(x,\beta) + A_{n2}(x,\beta) + A_{n3}(x,\beta))$$

with

$$\begin{aligned}A_{n1}(x,\beta) \quad &= \tfrac{1}{nh^pf(x)}\textstyle\sum_{j=1}^n K(\tfrac{||x_j-x||}{h})G_k(y_j, x_j^\tau\beta)I(|G_k(y_j,x_j)|^* \le M_n)\\ &\quad -EK(\tfrac{||\cdot-x||}{h})G_k(Y, X^\tau\beta)I(|G_k(Y,X)|^* \le M_n),\\ A_{n2}(x,\beta) \quad &= \tfrac{1}{nh^pf(x)}\textstyle\sum_{j=1}^n K(\tfrac{||x_j-x||}{h})G_k(y_j, x_j^\tau\beta)I(|G_k(y_j,x_j)|^* > M_n)\\ &\quad -EK(\tfrac{||\cdot-x||}{h})G_k(Y, X^\tau\beta)I(|G_k(Y,X)|^* > M_n),\\ A_{n3}(x,\beta) \quad &= \tfrac{1}{h^pf(x)}EK(\tfrac{||X-x||}{h})G_k(Y, X^\tau\beta).\end{aligned}$$

Let $\delta_o > 0$ be small enough such that $\delta_1 = \alpha_o - 1 - \delta_o > 1 + \delta_o$. Then, by conditions (K0), (X0), P1(i) and P1(ii), and by the Hölder inequality, we have

$$|A_{n2}(x,\beta)| \le c\frac{1}{(h^pM_n^{\delta_1})^{1/(\delta_o+1)}}. \tag{8.1}$$

Under P1(i), P1(v) and (X0), we have

$$\begin{aligned}|A_{n3}(x,\beta)| \quad &\le \tfrac{1}{f(x)}\textstyle\int K(||t||)|E_{Y|X}G_k(Y,(x+th)^\tau\beta) - E_{Y|X}G_k(Y,(x+th)^\tau\beta_o)|\\ &\quad \times f(x+th)dt\\ &\le c||\beta-\beta_o|| \le cr_n.\end{aligned} \tag{8.2}$$

It follows from the condition P1(ii) and the same argument as Owen (1990) that

$$\sup_{||\beta-\beta_o||\le r_o} Z_n(\beta) = o_p(n^{1/\alpha_o}).$$

It follows from (8.1) and (8.2) that to complete the proof, it suffices to show that for $0 < \eta < (\alpha_o - 2)/(\alpha_o + 2)$, $h_{nl}^p \le h^p \le h_{nu}^p$ and $0 \le r_n < n^{-1/\alpha_o}$, there exists $M_n > 0$, such that

$$\begin{aligned}&\sup\nolimits_{x\in\Omega, ||\beta-\beta_o||\le r_o} |A_{n1}(x,\beta)|n^{1/\alpha_o} = O_p(1),\\ &M_n^{-\delta_1/(\delta_o+1)}h^{-p/(\delta_o+1)}n^{1/\alpha_o} = O(1),\\ &r_n n^{1/\alpha_o} = O(1).\end{aligned} \tag{8.3}$$

For this purpose, we define the following empirical process

$$\upsilon_n(g) = n^{-1/2}\sum_{j=1}^{n}(g(y_j, x_j, \beta, x) - Eg(Y, X, \beta, x)), g \in \mathcal{F}_n.$$

Obviously, by the condition P1(iii), for any $g \in \mathcal{F}_n$, $|g| \leq 1$, and

$$\begin{aligned} Var(g(Y,X,\beta,x)) & \leq Eg^2(Y,X,\beta,x) \\ & \leq \tfrac{h^p}{M_n^2}\int K^2(||t||)E_{Y|X}G_k^2(Y,(x+th)^\tau\beta)f(x+th)dt \\ & \leq cn^{-\eta}M_n^{-2}\sup_{z\in\Omega,||\beta-\beta_o||\leq r_o} E_{Y|X}G_k^2(Y,z^\tau\beta) \\ & = cn^{-\eta}M_n^{-2}. \end{aligned} \tag{8.4}$$

By the Hölder inequality and the conditions (K0) and P1(ii), we derive

$$N(\epsilon, L_1(P_n), \mathcal{F}_n) \leq A^*(P_n)(M_n\epsilon)^{-w^*}$$

where $A^*(P_n) = c^*(E_{P_n}|G_k|^{*2})^w$, and c^* and w^* are some positive constants independent of P_n and n. Obviously, $EA^*(P_n)$ tends to $EA^*(P)$.

Now let $M_n = n^{s_o}$, $M = M_o f_{min} n^{1/2-1/\alpha_o-\eta-s_o}$. Let

$$\mathcal{G} = \{x \in \Omega, ||\beta - \beta_o|| \leq r_n, h_{nl}^p \leq h^p \leq h_{nu}^p\}.$$

Via (8.4), $v = cn^{-\eta-2s_o}$. Now applying Lemma 7.2, we have

$$\begin{aligned} P(\sup_{\mathcal{G}}|A_{n1}(x,\beta)|n^{1/\alpha_o} > M_o) &= P(\sup_{\mathcal{F}_n}|\upsilon(g)| > M) \\ & \leq c_1(P)(n^{1/2}M_o n^{-1/2+\eta+1/\alpha_o})^w \exp(-\tfrac{(M_o f_{min})^2}{2\times 64^2}n^{1-\eta-2/\alpha_o}) \\ & \quad +c_2(P)c^{-w}n^{\eta+2s_o}\exp(-cn^{1-\eta-2s_o}). \end{aligned}$$

This shows that the sufficient condition for (8.3) is that there exists $s_o > 0$ such that

$$\begin{aligned} & n^{1-\eta-2/\alpha_o}/\log n \to \infty, \\ & n^{1-\eta-2s_o}/\log n \to \infty, \\ & n^{1/\alpha_o+(\eta-\delta_1 s_o)/(\delta_o+1)} = O(1), \\ & r_n \leq n^{-1/\alpha_o}. \end{aligned}$$

A simple calculation show that they are true if

$$0 < \eta \leq \frac{1}{\delta_1+2}(\delta_1 - 2\frac{\delta_o+1}{\alpha_o}), \quad r_n \leq n^{-1/\alpha_o}.$$

Let δ_o be small enough, then it only requires

$$0 < \eta < \frac{1}{\alpha_o+1}(\alpha_o - 1 - 2/\alpha_o) = (\alpha_o-2)/\alpha_o, \quad r_n \leq n^{-1/\alpha_o}.$$

Thus, the desired result is proved.

Proof of Lemma 7.5. The proof is similar to that for Lemma 7.4. The details are omitted. The proof is completed.

PROOF OF LEMMA 7.6. The proof is similar to the parametric case. However, Lemma 7.3, instead of Lemma 7.2, is used. As before we need to check the condition in Lemma 7.1. The details are omitted.

PROOF OF LEMMA 7.7. The proof is similar to that of Lemma 7.6. The details are omitted.

PROOF OF LEMMA 7.8. For $1 \leq m \leq r$, set

$$\begin{aligned} ||G(y,x)||^* &= \sup_{||\beta-\beta_o||\leq r_o} ||G(y,x^\tau\beta)||, \\ \mathcal{G}_{n1m} &= \{\tfrac{1}{M_n}K(\tfrac{||\cdot-x||}{h})(G(\cdot,\cdot^\tau\beta)-G(\cdot,\cdot^\tau\beta_o))I(||G||^*\leq M_n): \\ &\quad ||\beta-\beta_o||\leq n^{-1/2}, x\in\Omega, h_{nl}^p\leq h^p\leq h_{nu}^p\}, \\ \mathcal{G}_{n2m} &= \{K(\tfrac{||\cdot-x||}{h})\tfrac{(G(\cdot,\cdot^\tau\beta)-G(\cdot,\cdot^\tau\beta_o))}{||\beta-\beta_o||^{1/2}n^{1/4}M_n}I(||G||^*\leq M_n): \\ &\quad ||\beta-\beta_o||> n^{-1/2}, x\in\Omega, h_{nl}^p\leq h^p\leq h_{nu}^p\}. \end{aligned}$$

Let b denote an arbitrary constant. Note that

$$\begin{aligned} E\ & |K(\tfrac{||X-x_1||}{h_1})(G_m(Y,X^\tau\beta_1)-G_m(Y,X^\tau\beta_o))-K(||\tfrac{||X-x_2||}{h_2})(G(Y,X^\tau\beta_2)-G(Y,X^\tau\beta_o))| \\ & \leq (E(K(\tfrac{||X-x_1||}{h_1})-K(\tfrac{||X-x_2||}{h_2}))^2)^{1/2}(E||G(Y,X)||^{*2})^{1/2} \\ & +E|G_m(Y,X^\tau\beta_1)-G_m(Y,X^\tau\beta_2)|. \end{aligned}$$

For bounded G, we use the conditions (P2), P1(i), P1(iv) and (P3), and for unbounded G, we use the conditions (P2), P1(i), P1(ii), P1(iv), (P3) and (P4). Then we have

$$N(\epsilon, L_1(P_n), \mathcal{G}_{n1m}) \leq A(P_n)(M_n\epsilon)^{-w_3}$$

where $\limsup EA(P_n)$ is bounded by a positive consatnt $A(P)$ and w_3 is a positive constant;

$$\begin{aligned} & EK^2(\tfrac{||X-x||}{h})(G_m(Y,X^\tau\beta)-G_m(Y,X^\tau\beta_o)^2\tfrac{1}{M_n^2} \\ & \leq bEK^2(\tfrac{||X-x||}{h})||\beta-\beta_o||\tfrac{1}{M_n^2} \\ & \leq bn^{-1/2-\eta}M_n^{-2}, \end{aligned}$$

when $||\beta-\beta_o||\leq n^{-1/2}$; and

$$\frac{1}{f(x)h^p}EK(\frac{||X-x||}{h})G^\tau(Y,X^\tau\beta)-E_{Y|X}G^\tau(Y,x^\tau\beta)=o_p(||\beta-\beta_o||).$$

In addition, for unbounded G, we

$$\begin{aligned} & \tfrac{1}{nh^p}\textstyle\sum_{j=1}^n K(\tfrac{||x_j-x||}{h})||(G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o))||I(||G(y_j,x_j)||^*>M_n) \\ & \quad +\tfrac{1}{h^p}EK(\tfrac{||X-x||}{h})||G(Y,X^\tau\beta)-G(Y,X^\tau\beta_o)||I(||G(Y,X)||^*>M_n) \\ & \leq \tfrac{1}{nh^p}\textstyle\sum_{j=1}^n K(\tfrac{||x_j-x||}{h})M(y_j,x_j)I(M(y_j,x_j)>M_n)||\beta-\beta_o||^\zeta \\ & \quad +\tfrac{1}{h^p}EK(\tfrac{||X-x||}{h})M(Y,X)I(M(Y,X)>M_n)||\beta-\beta_o||^\zeta \\ & \leq b(n^{-\eta}M_n^{\delta_1})^{-1/(\delta_o+1)}||\beta-\beta_o||^\zeta \end{aligned}$$

where $0<\delta_o<\alpha_o/2-1$ and $\delta_1=\alpha_o-1-\delta_o$, and b may depend on δ_o. Choosing $M_n=n^{s_o}$, $v=b_4n^{-\eta-1/2-2s_o}$ and $M=M_on^{-\eta-s_o}$, and using the same argument as that of Lemma 7.5, we find that

$$0<\eta<\begin{cases} 1/2, & \text{for bounded } G, \\ \frac{\alpha_o-1}{\alpha_o+1}(\frac{1}{2}+\frac{\zeta-1}{\alpha_o-1}), & \text{for unbounded } G \end{cases} \tag{8.5}$$

is the sufficient condition for

$$\sum_{j=1}^{n} w_j(x)(G(y_j, x_j^\tau \beta) - G(y_j, x_j^\tau \beta_o))^\tau - E_{Y|X}[G^\tau(Y, x^\tau \beta)^\tau] = o_p(n^{-1/2}) \tag{8.6}$$

where o_p is uniform in $x \in \Omega$ and $||\beta - \beta_o|| \leq n^{-1/2}$.

Similarly, for $||\beta_i - \beta_o|| > n^{-1/2}$, $x_i \in \Omega$, $h_{nl}^p \leq h^p \leq h_{nu}^p$, $i = 1, 2$,

$$\begin{array}{l} E||K(\frac{||X-x_1||}{h_1})\frac{(G(Y,X^\tau\beta_1)-G(Y,X^\tau\beta_o))^\tau}{||\beta_1-\beta_o||^{1/2}n^{1/4}} - K(\frac{||X-x_2||}{h_2})\frac{(G(Y,X^\tau\beta_2)-G(Y,X^\tau\beta_o))}{||\beta_1-\beta_o||^{1/2}n^{1/4}}|| \\ \leq b(E|K(\frac{||X-x_1||}{h_1}) - K(\frac{||X-x_2||}{h_2})|^2)^{1/2} + bE||G(Y, X^\tau\beta_1) - G(Y, X^\tau\beta_2)|| \\ \quad +b||\beta_1 - \beta_2||n^{1/2}, \\ E||K(\frac{||X-x_1||}{h_1})\frac{(G(Y,X^\tau\beta_1)-G(Y,X^\tau\beta_o))^\tau}{||\beta_1-\beta_o||^{1/2}n^{1/4}}||^2 \leq bh^p n^{-1/2} \end{array}$$

where b may depend on $EM(Y, X)^2$. This leads to

$$N(\epsilon, L_1(P_n), \mathcal{G}_{n2m}) \leq B(P_n)(M_n\epsilon)^{-w_4} n^{p/2}$$

with $\limsup EB(P_n)$ is bounded a positive constant $B(P)$. Now using the same argument as that for (8.6), we obtain that (8.5) is the sufficient condition for

$$\begin{array}{ll} \sum_{j=1}^{n} & w_j(x)[(G(y_j, x_j^\tau\beta) - G(y_j, x_j^\tau\beta_o) - E_{Y|X}[G^\tau(Y, x^\tau\beta)^\tau]]\frac{1}{||\beta-\beta_o||^{1/2}n^{-1/4}} \\ & = o_p(n^{-1/2}) \end{array}$$

where o_p is uniform in (x, β) with $x \in \Omega$ and $||\beta - \beta_o|| > n^{-1/2}$. Combining this with (8.6) leads to the desired result. The proof is completed.

Proof of Lemma 7.9. First of all, using the Jensen inequality, we obtain the following inequalities:

$$\begin{array}{l} E||K(\frac{||X-x_1||}{h_1})E_{Y|X}(G(Y, X^\tau\beta_1) - G(Y, X^\tau\beta_o))^\tau U^{-1}(X) \\ \quad -K(\frac{||X-x_2||}{h_2})E_{Y|X}(G(Y, X^\tau\beta_2) - G(Y, X^\tau\beta_o))^\tau U^{-1}(X)|| \\ \quad \leq (E||G(Y, X^\tau\beta) - G(Y, X^\tau\beta_o)||^2)^{1/2}\frac{1}{e_o^2}(E|K(\frac{||X-x_1||}{h_1}) - K(\frac{||X-x_2||}{h_2})|^2)^{1/2} \\ \quad +\frac{1}{e_o^2}E||G(Y, X^\tau\beta_1) - G(Y, X^\tau\beta_2)||^2; \\ EK^2(\frac{||X-x||}{h})||E_{Y|X}(G(Y, X^\tau\beta) - G(Y, X^\tau\beta_o))^\tau U^{-1}(X)||^2 \\ \quad \leq bEK^2(\frac{||X-x||}{h})E_{Y|X}||G(Y, X^\tau\beta) - G(Y, X^\tau\beta_o)||^2, \end{array}$$

where e_o is the lower bound to the smallest eigenvalues of $U(x)$. We also have

$$\begin{array}{l} \frac{1}{h^p f(x)}EK(||X - x||h)G(Y, X^\tau\beta)^\tau U^{-1}(X) - E_{Y|X}G(Y, x^\tau\beta)^\tau U^{-1}(x) \\ \quad = \frac{1}{f(x)}\int K(||t||)E_{Y|X}(G(Y, (x + th)^\tau\beta) - G(Y, x^\tau\beta))^\tau U^{-1}(x + th)f(x + th)dt \\ \quad +\frac{1}{f(x)}\int K(||t||)E_{Y|X}G(Y, x^\tau\beta)(U^{-1}(x + th) - U^{-1}(x))f(x + th)dt \\ \quad +\frac{1}{f(x)}\int K(||t||)E_{Y|X}G(Y, x^\tau\beta)^\tau U^{-1}(x)(f(x + th) - f(x))dt \\ \quad = o(||\beta - \beta_o||) \end{array}$$

Now with the assumptions and the same argument in the proof of Lemma 7.4, we can obtain the desired result. The proof is completed.

PROOF OF LEMMA 7.10. The proof is similar to that of Lemma 7.8 and is omitted.

PROOF OF LEMMA 7.11. For simplicity, we assume G is bouned. Set

$$\begin{array}{ll} g_{n1}(y,x,\beta) & = n^{1/2}E_{Y|X}G^\tau(Y,x^\tau\beta)U^{-1}(x)(G(y,x^\tau\beta)-G(y,x^\tau\beta_o)) \\ g_{n2}(y,x,\beta) & = \frac{E_{Y|X}G(Y,x_j^\tau\beta)}{||\beta-\beta_o||}U^{-1}(x_j)\frac{G(y_j,x_j^\tau\beta)-G(y_j,x_j^\tau\beta_o)}{||\beta-\beta_o||^{1/2}n^{1/4}}. \end{array}$$

As in the proof of Lemma 7.8, we only need to show

$$\frac{1}{n}\sum_{j=1} g_{n1}(y_j,x_j,\beta) = o_p(n^{-1/2}) \tag{8.7}$$

uniformly in $||\beta-\beta_o|| \le n^{-1/2}$. And

$$\frac{1}{n}\sum_{j=1}^{n} g_{n2}(y_j,x_j,\beta) = o_p(1) \tag{8.8}$$

uniformly in $||\beta-\beta_o|| > n^{-1/2}$. To this end, we first evaluate the covering numbers of

$$\begin{array}{ll} \mathcal{G}_{n1} & = \{\theta_{n1}(\cdot,\cdot,\beta) : ||\beta-\beta_o|| \le n^{-1/2}\} \\ \mathcal{G}_{n2} & = \{\theta_{n1}(\cdot,\cdot,\beta) : r_o \ge ||\beta-\beta_o|| > n^{-1/2}\}. \end{array}$$

Note that, using the Jensen inequality, we obtain

$$E|g(Y,X,\beta_1)-g(Y,X,\beta_2)| \le b_o(E||G(Y,X^\tau\beta_1)-G(Y,X^\tau\beta_2)||^2)^{1/2}$$

where b_o is a positive constant. Also we have

$$|\frac{1}{||\beta_1-\beta_o||^{3/2}} - \frac{1}{||\beta_2-\beta_o||^{3/2}}|n^{-1/4} \le b_2 n||\beta_1-\beta_2||$$

for $||\beta_i-\beta_o|| > n^{-1/2}$, $i=1,2$, where b_2 is a positive constant. These inequalities together with the condition (P2) yield

$$N(\epsilon, L_1(Q), \mathcal{G}_{n1}) \le b_1 n^{w/2}\epsilon^{-w_1}, \quad N(\epsilon, L_2(Q), \mathcal{G}_{n2}) \le b_2 n^{w}\epsilon^{-w_2}$$

where w, b_i and w_i, $i=1,2$ are positive constants independent of the probability Q. Using the conditions P1(iv) and P1(v), we derive that

$$\begin{array}{ll} E|g_{n1}(Y,X,\beta)|^2 & \le b_3 n^{-1/2}, \text{ for } ||\beta-\beta_o|| \le n^{-1/2}; \\ E|g_{n2}(Y,X,\beta)|^2 & \le b_3 n^{-1/2}, \text{ for } ||\beta-\beta_o|| > n^{-1/2} \end{array}$$

where b_3 is a positive constant. Now using Lemma 7.2, (8.7) and (8.8) are proved.

PROOF OF LEMMA 7.12. For simplicity, we assume G is bounded. Let c_1 denote any constant. Set

$$\mathcal{G}_k = \{K(\frac{|\cdot - x|}{h})(G_k(\cdot,\theta(\cdot)) - G_k(\cdot,\theta_o(\cdot))) : ||\theta-\theta_o|| \le c_1 n^{-\pi_1}, x\in\Omega, h_{nl}\le h\le h_{nu}\}.$$

Note that it follows from Kolmogorov and Tihomirov (1959) that $H^B(\epsilon, ||\cdot||, \Theta) \le H(\epsilon, ||\cdot||_\infty) \le c\epsilon^{-w^*}$. This together with the conditions (X0), (K1) and N3(i) gives

$$H^B(\epsilon, ||\cdot||, \mathcal{G}_k) \le A(P)(\epsilon n^{\pi_1})^{-w^*} + B(P)\log(n^{-\eta}\epsilon)$$

where $A(P)$ and $B(P)$ are two positive constants independent of k.

Under the conditions (X0), (K1) and N3(i), by the interpolation inequality of Shen (1997, Proposition 6), we obtain

$$EK^2(\frac{|X-x|}{h})(G_k(Y,\theta(X)) - G_k(Y,\theta_o(X)))^2 \le c_1 h||\theta-\theta_o||_\infty^2 \le c_1 h||\theta-\theta_o||^{2b_1}.$$

Under the conditions (X0), (K1), N1(i) and N3(ii), by the conditional equation in (1.1) and the interpolation inequality of Shen (1997), we have

$$\begin{aligned}
\tfrac{1}{f(x)h} \quad & EK(\tfrac{|X-x|}{h})G_k(Y,\theta(X)) - E_{Y|X}G_k(Y,\theta(x)) \\
&= \tfrac{1}{f(x)}\int K(|t|)E_{Y|X}[G_k(Y,\theta(x+th)) - G_k(Y,\theta(x)) - G_k(Y,\theta_o(x+th)) + G_k(Y,\theta_o(x))] \\
&\quad \times f(x+th)dt + \tfrac{1}{f(x)}\int K(|t|)E_{Y|X}G_k(Y,\theta(x))(f(x+th)-f(x))dt \\
&= \tfrac{1}{f(x)}\int K(|t|)[\tfrac{\partial E_{Y|X}G_k(Y,\theta(x^*))}{\partial\theta} - \tfrac{\partial E_{Y|X}G_k(Y,\theta_o(x^*))}{\partial\theta})\theta^{(1)}(x^*) \\
&\quad + \tfrac{\partial G_k(Y,\theta_o(x^*))}{\partial\theta}(\theta^{(1)}(x^*) - \theta_o^{(1)}(x^*)]thf(x+th)dt + O(h||\theta-\theta_o||_\infty) \\
&\le c_1 h[||\theta-\theta_o||_\infty(||\theta^{(1)}||_\infty + 1) + ||\theta^{(1)} - \theta_o^{(1)}||_\infty] \\
&\le c_1 h[||\theta-\theta_o||^{b_1} + ||\theta^{(1)} - \theta_o^{(1)}||^{b_2}].
\end{aligned}$$

Let $\pi_2 = \eta + b_2\pi_1$, $M = M_o n^{1/2-\pi_2-\eta}$ and $v = c_1 n^{-\eta-a_1b_1\pi_2}$. Invoking Lemma 7.3, we have

$$\sum_{j=1}^n w_j(x)(G(y_j,\theta(x_j)) - G(y_j,\theta_o(x_j)) - E_{Y|X}G(Y,\theta(x)) = O_p(n^{\eta+b_2\pi_1})$$

provided

$$\begin{aligned}
&\pi_2 > \pi_2 a_1 b_1, \quad 1-\pi_2 > \eta, \quad 1 - 2\pi_2 + \pi_1 a_1 b_1 > \eta, \\
&\tfrac{1}{2} - \pi_2 + \tfrac{\pi_1 w^*}{2} + \pi_1 a_1 b_1 \tfrac{2-w^*}{4} > \tfrac{2+w^*}{4}\eta.
\end{aligned}$$

These conditions follow from the assumptions in the lemma. The proof is completed.

Proof of Lemma 7.13. It is similar to that of Lemma 7.12 and thus omitted.

Proof of Lemma 7.14. It is similar to that of Lemma 7.12 and thus omitted.

Proof of Lemma 7.15. Without loss of generality, we assume $||G|| \le 1$ and let c_1 denote any positive constant. Define

$$\mathcal{F}_n = \{E_{Y|X}G^\tau(\cdot,\theta(\cdot))U^{-1}(\cdot)(G(\cdot,\theta(\cdot)) - G(\cdot,\theta_o(\cdot))) : ||\theta-\theta_o|| \le cn^{-\pi_1}\}.$$

It follows from the assumptions that

$$H^B(\epsilon, L_2(P), \mathcal{F}_n) \le c_1(\epsilon n^{\pi_1})^{-w^*}$$

with $w^* = 1/(q+r)$. And via the interpolation inequality of Shen (1997),

$$\begin{aligned}
&E(E_{Y|X}G^{\tau}(Y,\theta(X))U^{-1}(X)(G(Y,\theta(X)) - G(Y,\theta_o(X)) - E_{Y|X}G(Y,\theta(X)))^2 \\
&\leq c_1 E||E_{Y|X}G(Y,\theta(X))||^2||G(Y,\theta(X)) - G(Y,\theta_o(X))||^2 \\
&\leq c_1 E|\theta(X) - \theta_o(X)|^3 \\
&\leq c_1||\theta - \theta_o||_{\infty}||\theta - \theta_o||^2 \\
&\leq c_1||\theta - \theta_o||^{2+b_1}
\end{aligned}$$

with $b_1 = 2/(2+w^*)$. Letting $M = M_o n^{1/2-\pi_1-1/(2+w^*)}$ and $v = c_1 n^{-\pi_1(2+b_1)}$ in Lemma 7.3, the desired result is proved.

Jian Zhang
Institute of Systems Science
Academia Sinica
Beijing 100080, P.R. China
e-mail : jzhang@iss01.iss.ac.cn
Irène Gijbels
Institute of Statistics
Universite Catholique de Louvain
Voie du Roman Pays 20
1348 Lovain-La-Neuve, Belgium
e-mail: gijbels@stat.ucl.ac.be