Sieve Empirical Likelihood and Extensions of the Generalized Least Squares

JIAN ZHANG

EURANDOM and Chinese Academy of Sciences

IRÈNE GIJBELS

Université Catholique de Louvain

ABSTRACT. The empirical likelihood cannot be used directly sometimes when an infinite dimensional parameter of interest is involved. To overcome this difficulty, the sieve empirical likelihoods are introduced in this paper. Based on the sieve empirical likelihoods, a unified procedure is developed for estimation of constrained parametric or non-parametric regression models with unspecified error distributions. It shows some interesting connections with certain extensions of the generalized least squares approach. A general asymptotic theory is provided. In the parametric regression setting it is shown that under certain regularity conditions the proposed estimators are asymptotically efficient even if the restriction functions are discontinuous. In the non-parametric regression setting the convergence rate of the maximum estimator based on the sieve empirical likelihood is given. In both settings, it is shown that the estimator is adaptive for the inhomogeneity of conditional error distributions with respect to predictor, especially for heteroscedasticity.

Key words: asymptotic efficiency, conditional equations, generalized least squares, generalized method of moments, semiparametric and non-parametric regressions, sieve empirical likelihood

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 behaviour 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 a sense (Stone, 1982; Wong & Severini, 1991; Wong & Shen, 1995);

(*b*) When θ is restricted to a finite dimensional subspace, under some conditions $\hat{\theta}$ is asymptotically efficient (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 (Carroll, 1982);

(*c*) For some smooth functional $\rho(\theta)$, the plug-in estimator $\rho(\hat{\theta})$ is asymptotic efficient (Lehmann, 1983; Wong & Severini, 1991; 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. A prototype situation is that the regression function θ and the conditional distribution of Y given X are constrained by a set of conditional equations, say

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

where $G = (G_1, \ldots, G_{k_o})^\tau$ 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. Via the conditional equations, this formulation provides substantial flexibility for using the auxiliary information to solve non-regular estimation problems (Newey, 1993; Powell, 1994). Under it, many interesting models are subsumed. In this paper, two typical settings will be considered. One is the parametric regression setting (or the semiparametric 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 setting, 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. The other is the non-parametric regression setting where θ is a vector-valued function of X lying in some smooth class of functions.

In the above situation the parametric likelihood inference is unsuitable. This is because $\hat{\theta}$ might be inconsistent or not asymptotically efficient when the parametric assumption is violated. See Zhang & Liu (2000) for some details. Therefore, it is desirable to develop some alternative non-parametric 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 non-parametric 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 non-parametric 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 *et al.* (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 conjecture further that Carroll's result holds even if the parametric likelihood is replaced by some non-parametric likelihood.

The idea of empirical likelihood (Owen, 1988) holds considerable promise for answering our questions. However, some special difficulties exist in using Owen's empirical likelihood. For example, the normal likelihood can identify an 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 type of (global) non-parametric likelihoods called sieve empirical likelihoods (SEL) for the regression function θ are constructed via the local empirical likelihoods. Unlike LeBlanc & Crowley (1995), we construct SEL for the finite dimensional parameter as well as for the infinite dimensional parameter. The name originates from the ideas of using n constraints at observations $x_i, i = 1, 2, \ldots, n$,

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

as approximations of the infinite constraints in (1) and of using a random sieve approximation of the underlying distribution. In this sense, SEL, which include LeBlanc & Crowley's (1995) likelihood as a special case, can be viewed as some approximation to the empirical likelihood when there exist an infinite number of constraints. 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 (Fan, 1997). However, the advantage will disappear when the data structure is approximately 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.

Base on SEL, a unified framework for simultaneously estimating the regression function and the conditional distribution of the response is developed. As one of the contributions, a first theoretical analysis of the proposed procedures is provided. It is shown that the SEL estimators are asymptotically efficient under some regularity conditions in the parametric regression 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 *et al*. (1993) mentioned before.

To give an intuitive device for analysing SEL, we proved that SEL is asymptotically equivalent to some extensions of the generalized least squares (EGLS). 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 (Carroll & Ruppert, 1988) and seems unsuitable for the situation with a discontinuous or non-linear 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) (Hansen, 1982; Chamberlian, 1992). For example, in the parametric regression setting, 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 (Newey, 1993). Unlike GMM, in EGLS no instrumental variables (the functions of the predictor variable that are known to be independent of the error terms) are required to be specified and thus the idea can be easily extended to the non-parametric regression setting.

In the non-parametric regression 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 above.

The following is a brief review of some related works and problems. 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 (Kitamura, 1997; Kitamura & Stutzer, 1997). LeBlanc & Crowley (1995) proposed an alternative non-parametric likelihood procedure 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 survival 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?

The remains of this paper proceed as follows. Section 2 gives the definitions and intuitive illustrations of the proposed likelihoods and estimators. A large sample study of these estimators is presented in section 3. Some examples can be found in section 4. Some possible extensions are discussed in section 5. The technical conditions used in this paper and the proofs of the main results are presented in the appendix. Throughout this paper, P_n and P stand for the empirical and 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 for several specified cases, in the following sections, $||\cdot||$ denotes the Euclidean norm. Let "$\xrightarrow{\mathcal{D}}$" denote convergence in distribution.

2. Sieve empirical likelihood

2.1. Definition

We use the non-parametric likelihood to illustrate the idea behind SEL. Suppose we have n independent observations $\{(x_i, y_i)\colon 1 \le i \le n\}$ from (X, Y). Let $\mathcal{F}$ be a space of distributions for (X, Y). Following Shen *et al.* (1999), to construct a sieve likelihood we first make a (random) sieve approximation of $\mathcal{F}$ by $\mathcal{F}_n$, a class of distributions with a finite support, say S_n. Then we focus on the sieve likelihood for $Q \in \mathcal{F}_n$:

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

where $q(x_i, y_i) = [dQ/d\mu_n](x_i, y_i)$ is the mass that Q places at point (x_i, y_i) and μ_n is the counting measure on S_n. Unlike Shen *et al.* (1999), we take $S_n = \{(x_i, y_j)\colon 1 \le i \le n, 1 \le j \le n\}$, rather than the sample $\{(x_i, y_i)\colon 1 \le i \le n\}$, as the support. We will see that such an overparametrization provides a flexible device to exploit the structural information of the underlying distribution in defining a profile sieve likelihood—SEL. 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 non-informativeness assumption, the above likelihood of the regression is proportional to the conditional non-parametric likelihood

$$\prod_{i=1}^{n} q_{Y|X=x_i}(y_i) \tag{2}$$

where

$$q_{Y|X=x_i}(\cdot) = \frac{q(x_i, \cdot)}{\sum_{j=1}^{n} q(x_i, y_j)}$$

with the support $\{y_j\colon 1 \le j \le n\}$. Obviously, without any auxiliary information, the non-parametric maximum likelihood estimator (MLE) of P is of the form $q(x_i, y_i) = 1/n$, $\quad q(x_i, y_j) = 0$, $j \ne i, 1 \le j \le n, 1 \le i \le n$. That is, we put the equal mass $1/n$ on each sample point (x_i, y_i). The corresponding non-parametric MLEs of $P_{Y|X=x_i}, 1 \le i \le n$ satisfy $q_{Y|X=x_i}(y_i) = 1$, $q_{Y|X=x_i}(y_j) = 0$, for $j \ne i$, $1 \le j \le n$, $1 \le i \le n$. So the profile conditional log-likelihood

$$\sup_{Q\in\mathcal{F}_n}\sum_{i=1}^{n}\log q_{Y|X=x_i}(y_i)=0. \tag{3}$$

Of course the above estimators are not very useful. To get certain meaningful estimators, we need some auxiliary information of $P_{Y|X}$. However, if we add the auxiliary information such as (1) to the likelihood (2), then the corresponding profile likelihood may not exist. The problem we faced is, given each x_i, only a single observation y_i from $P_{Y|X=x_i}$ is available. This hampers the use of the auxiliary information (1). This problem can be solved by "borrowing" information from the nearby observations if we assume that both $P_{Y|X=x}$ and $\theta(x)$ are continuous with respect to x (Tibshirani & Hastie, 1987; LeBlanc & Crowley, 1995). In other words, in a neighbourhood around x_i, we approximate $P_{Y|X=x}$ and $\theta(x)$ by

$$P_{Y|X=x}\approx P_{Y|X=x_i},\quad \theta(x)\approx\theta(x_i),\quad \text{for } x\approx x_i.$$

This implies that all the y_j with x_j lying in this neighbourhood can be roughly viewed as observations from $P_{Y|X=x_i}$. These y_j form a set, say $\{z_{ik}\colon 1\le k\le n_i\}$. Then, with this augmented sample, the auxiliary information (1) can be used via the empirical likelihood technique. This yields an estimator of $P_{Y|X=x_i}$, say $\hat{P}_{Y|X=x_i}$, with the support $\{z_{ik}\colon 1\le k\le n_i\}$. At the same time we obtain the local empirical log-likelihood $(1/n_i)\sum_{k=1}^{n_i}\log\{\hat{P}_{Y|X=x_i}(z_{ik})\}$. Furthermore, treating the augmented sample sets $\{z_{ik}\colon 1\le k\le n_i\}$, $1\le i\le n$ as if they are independent, we have a total conditional log-likelihood simply by adding up all these local empirical log-likelihoods.

An improved version of the above idea is to include a smoothing weight function, which weighs down the contribution to the ith 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 ith 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 $q_{ji}=Q_{Y|X=x_i}\{y_j\}$ denote the mass we put on y_j. For each i, we maximize

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

subject to

$$\sum_{j=1}^{n}q_{ji}G(y_j,\theta(x_j))=0,$$

$$\sum_{j=1}^{n}q_{ji}=1,\quad q_{ji}\ge 0,\quad j=1,\ldots,n.$$

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

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

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

$$\sum_{j=1}^{n}w_{ji}\frac{G(y_j,\theta(x_j))}{1+\alpha_n(x_i,\theta)^{\tau}G(Y_j,\theta(x_j))}=0. \tag{4}$$

Define

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

Step 2. Define the log-SEL by

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

Note that without the local constraints (1) the log-SEL becomes $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ji}\log w_{ji}$. Furthermore, if there is no knowledge of the continuity of $P_{Y|X=x}$ with respect to x, then we should set $w_{ii}=1$, $w_{ji}=0, j\neq i$. In this case, the log-SEL agrees with the profile conditional log-likelihood in (3).

The logarithm of the likelihood ratio between the approximate empirical likelihoods with and without local constraints 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 $N_i\subset\{1,\ldots,n\}$ and define

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

in step 1. This leads to a possible reduction of the degree of correlation among $l(i,\theta)$, $1\le i\le n$, but on the other hand also to a loss of some information due to dropping $\hat{\boldsymbol{q}}_{ji}, j\notin N_i, 1\le i\le n$. For simplicity, in the following sections, we consider only the cases with $N_i=\{1,\ldots,n\}, 1\le i\le n$ and $N_i=\{i\},\ 1\le i\le n$, respectively. Note that when $N_i=\{i\}, 1\le i\le n$, we recover LeBlanc and Crowley's log-likelihood l_c and likelihood ratio l_{cr} which are defined by

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

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

$$\hat{\boldsymbol{\beta}}=\arg\max l_{sr}(\beta).$$

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

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

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

Estimators of conditional distributions. Let $\hat{\theta}$ be the estimator defined above. Then, the estimate of $F_{Y|X=x}$, namely, $\hat{F}_{Y|X=x}$, is obtained by replacing θ and x_i in (4) by $\hat{\theta}$ and x respectively. Note that, it is supported by $\{y_j, 1\le j\le n\}$ with masses $\hat{\boldsymbol{p}}_j(x), 1\le j\le n$.

2.2. Quadratic approximations

To get a better insight into the above procedure, we first investigate the behaviour of the proposed likelihood. Suppose for each i, $\alpha_n(x_i,\theta)$ in (4) tends to zero. Then, we have the following informal approximations of $\alpha_n(x_i,\theta)$ and l_{sr},

$$\alpha_n(x_i,\theta)=\left\{\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j))\right\}^{-1}\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))(1+o_p(1)),\quad 1\le i\le n;$$

$$l_{sr}=-\frac{1}{2}\sum_{i=1}^n\left(\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))\right)^\tau\left\{\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j))\right\}^{-1}\times\left(\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))\right)(1+o_p(1)). \tag{5}$$

and we also have the following population version in the neighbourhood of the true value of parameter, say θ_o,

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

Applying the same argument, we have

$$\begin{aligned}l_{cr}=&-\sum_{i=1}^n\sum_{j=1}^n w_{ji}G^\tau(y_j,\theta(x_j))\left\{\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j))\right\}^{-1}G(y_j,\theta(x_j))\\&+\frac{1}{2}\sum_{i=1}^n\left(\sum_{j=1}^n w_{ji}G^\tau(y_j,\theta(x_j))\right)\left\{\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j))\right\}^{-1}\\&\times G(y_i,\theta(x_i))G^\tau(y_i,\theta(x_i))\left\{\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j))\right\}^{-1}\\&\times\left(\sum_{j=1}^n w_{ji}G(y_j,\theta(x_j))\right)(1+o_p(1)).\end{aligned} \tag{6}$$

Clearly, the approximation of l_{sr} is quadratic while that of l_{cr} is not. However, the population versions of l_{cr} and l_{sr} are the same.

Let $K(\cdot)$ be a bounded univariate kernel function. Set

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

for $1\le i,j\le n$, where the bandwidth $h=h(x)$ may depend on x. Note that the dependence of $w_j(x)$ on h is suppressed for notational convenience. Recall the definition of α_o in the condition **(P2)** in appendix A. We show that $\alpha_n(x_i,\theta)$ does tend to zero uniformly in x_i below.

Theorem 1

(*i*) (*Parametric regression setting*): *If for some constants* $d_o>0$, $d_1>1$ *and* $0<\eta<(\alpha_o-2)/(\alpha_o+2)$, $d_o\le h(x)^p n^\eta\le d_1$, $x\in\Omega$, *then under the conditions* **(K0)**, **(X0)**, **(P1)**–**(P6)** *in appendix A*, (5) *and* (6) *hold uniformly in* $\beta, ||\beta-\beta_o||\le r_n=O(n^{-1/\alpha_o})$. *The similar result holds for the case* $\theta(x)=(\beta^\tau x,\lambda)$.

(*ii*) (*Non-parametric regression setting*): *If for some positive constants* d_o, d_1 *and* η, $d_o\le h(x)^p n^\eta\le d_1$, *where for* w^* *in* **(N6)**, η *satisfies*

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

then under conditions **(K1)**, **(X0)**, **(N1)**–**(N6)** *in appendix A,* (5) *and* (6) *hold uniformly in* θ, $||\theta-\theta_o|| \le r_n = O(n^{-1/\alpha_o})$.

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

The above theorem implies that SEL is asymptotically equivalent to the following modifications of the classical weighted least squares criterion.

Population version

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

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

Sample versions

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

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

where $\hat{U}(X,\theta) = \hat{E}_{Y|X}G(Y,\theta(X))G^\tau(Y,\theta(X))$, and $\hat{E}$, $\hat{E}_{Y|X}$ 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 (LeBlanc & Crowley, 1995, p. 102). In fact, there is room for improving the EGLS estimator via the following equality: for any measurable function $H(Y,\theta(X))$,

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

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

$$R_M(\theta) = \frac{1}{2}E\{\{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}^\tau U^{-1}(X,\theta) \times \{E_{Y|X}(G(Y,\theta(X)) - H(Y,\theta(X))) + H(Y,\theta(X))\}\}.$$

The associated sample version:

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

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 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 $\hat{R}_c(\beta,\lambda)$.

3. Asymptotic properties

This section presents some general theory on the consistency, convergence rates and asymptotic normalities of the SEL estimators. As before we use the kernel weights. We assume that the true value of the parameter of interest is an inner point of the parameter space.

3.1. Parametric regression setting

For simplicity, we consider the case that $\theta(x) = x^\tau\beta$. Denote by β_o the true value of β. Recall the definition of α_o in the condition (**P2**). We first show the SEL estimator is weakly consistent in the following theorem.

Theorem 2
Suppose that the conditions (**X0**), (**K0**), (**P1**)–(**P7**) *in appendix A hold. If, for some positive constants* d_o, d_1 *and* η, $d_o \le h(x)^p n^\eta \le d_1$, *and* $0 < \eta < (\alpha_o - 4)/\alpha_o$, *then* $\hat{\beta} - \beta_o = o_p(n^{-1/\alpha_o})$.

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

Assume that there exists a positive definite matrix V such that

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

V can be derived by calculating the second derivative of $R(\beta)$. Then the next theorem states that $\hat{\beta}$ is asymptotically normal.

Theorem 3
We suppose that when G is bounded, the conditions **(K0)**, **(X0)**, **(P1)**–**(P8)** *in appendix A hold; and that when G is unbounded, the conditions* **(K0)**, **(X0)**, **(P1)**–**(P9)** *in appendix A hold. Suppose* $E_{Y|X=x}G(Y, X^\tau\beta)$ *has a continuous derivative with respect to x. Suppose that for some positive constants* d_o, d_1, $d_o \le h(x)n^\eta \le d_1$, $x \in \Omega$, *where*

$$0 < \eta < \begin{cases} 1/2, & G \text{ is bounded}, \\ \min\left\{\dfrac{\alpha_o - 4}{\alpha_o}, \dfrac{\alpha_o - 1}{\alpha_o + 1}\left(\dfrac{1}{2} + \dfrac{\zeta_1 - 1}{\alpha_o - 1}\right)\right\}, & G \text{ is unbounded}. \end{cases}$$

Then

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

Remark 2. In practice, G often satisfies the Lipschitz condition that $\zeta_1 = 1$ in the condition **(P9)**. For an unbounded G, the restriction on η in theorem 3 becomes $0 < \eta < 7/18$ if $\alpha_o = 8$ and $\zeta_1 = 1$; $0 < \eta < 1/3$ if $\alpha_o = 6$ and $\zeta_1 = 1$. This implies theorem 3 does cover the optimal bandwidth $h = n^{-1/5}$.

Theorems 1, 2 and 3 show that the SEL and EGLS estimators asymptotically have the same distributions as the GMM estimators based on the optimal instrumental variables. Furthermore, they are asymptotically efficient in the sense of Hansen (1982) and Bickel *et al.* (1993) even if G is discontinuous. When G is discontinuous, by the result of Hansen (1982), we can get the optimal instrumental variable, $\partial E_{Y|X}G(Y, X^\tau\beta)/\partial\beta U^{-1}(X)$. However, it is still unknown whether the GMM estimator is asymptotically efficient in this case.

3.2. Non-parametric regression setting

Let $||\theta||_\infty = \sup_{x\in\Omega} ||\theta(x)||$ and $||\theta||^2 = E_X||\theta(X)||^2$. Let θ_o be the true value of θ. The following theorem shows that the SEL estimator $\hat{\theta}$ is also weakly consistent.

Theorem 4
Suppose the conditions **(X0)**, **(K1)**, **(N1)**–**(N7)** *in appendix A hold. If, for some positive constants* d_o, d_1 *and* η, $d_o \le h(x)^p n^\eta \le d_1$, *and* $0 < \eta < 2(\alpha_o - 4)/(\alpha_o(2 + w^*))$ *with* w^* *in* **(N6)**, *then* $||\hat{\theta} - \theta_o|| = o_p(n^{-1/\alpha_o})$. *Similar to remark* 1, *we obtain the weak consistency of the EGLS estimator.*

To obtain a better convergence rate of $\hat{\theta}$, we consider the following parameter space of θ,

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

where $r > 0$, and $L_j, j = 0, \ldots, q$ are fixed constants and $||\theta||_H = \sup_{x\neq y} |\theta(x) - \theta(y)|/|x - y|^r$. Set $\Omega = [0, 1]$, $w^* = 1/(q + r) < 1$, and $q \ge 1$. Let $b_1 = 2/(2 + w^*)$, $b_2 = 2(1 - w^*)/(2 - w^*)$, $1 \le a_1 \le 2$ and

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

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

$$\eta_1(\pi_o) = \min\left\{\pi_o(a_1 b_1 - b_2), \frac{2\pi_o}{1-w^*} - \pi_o(b_2+1)\right\},$$

$$\eta_2(\pi_o) = \min\left\{\frac{1}{3}(1 - 2b_2\pi_o + a_1 b_1 \pi_o), \frac{2}{6+w^*}(1 - 2b_2\pi_o + w^*\pi_o + \frac{1}{2}a_1 b_1(2-w^*)\pi_o)\right\}.$$

Theorem 5

Under the conditions **(X0)**, **(K1)**, **(N1)**, **(N2)**, **(N3)**, **(N6)**, **(N7)**, *and* **(N8)** *in appendix A, for any* $0 \le \pi_o < \pi^*$, $\eta_1(\pi_o) < \eta < \eta_2(\pi_o)$, *we have*

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

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

Remark 3. Obviously, theorem 5 can be directly extended to the p-dimension. Also the exponential order moment condition **N8(i)** can be relaxed via the technique used in theorem 1.

It follows from theorem 5 that when $w^* \le 1/4$,

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

which are the optimal convergence rates in the ordinary non-parametric regression case. See example 6 in the next section and Stone (1982). When $w^* = 1/2, 1/3$, $\pi^*/(1-w^*) = 0.276$ and 0.411, both of which have not attained the corresponding optimal values 0.4 and 4/9 in the ordinary non-parametric 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 non-parametric 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. 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), $i = 1, \ldots, n$. Let $F_{e|X}$ and $f_{e|X}$ denote the conditional distribution function and density of e given X. Let β_o and λ_o denote the true value of β and λ, respectively. We demonstrate that the proposed methods are more generally applicable than both the generalized least squares and GMM through the following examples.

We begin with the following two prototype examples, which are the generalizations of the well-known symmetric location model (e.g. Bickel *et al.* 1993, pp. 75, 400–405).

Example 1 (Linear regression with a symmetric error distribution). Consider the linear regression model $Y = X^\tau\beta + e$. Suppose that given X, e is symmetrically distributed. To use the information about e, we choose $0 = s_o < s_1 < \cdots < s_{k_o}$ and $S_k = [s_{k-1}, s_k)$, $1 \le k \le k_o$. Set $G_k(Y, X^\tau\beta) = I(Y - X^\tau\beta \in S_k) - I(Y - X^\tau\beta \in -S_k)$, $1 \le k \le k_o$. Then, G satisfies the conditional equations in (1). Furthermore, we have

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

We assume $f_{e|X=x}(z)$ 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 1, 2 and 3 that

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

with

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

This directly implies that if $f_{e|X=x}(z)$ can be written in the form $g_o(z/s(x))$ with known g_o and unknown $s(x)$, then $\hat{\boldsymbol{\beta}}$ 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

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

Thus, when the underlying probability distribution for observations is symmetric, an approximately efficient and adaptive estimator can be obtained by refining the sequence of $\{s_k\}$. See Bickel *et al*. (1993).

Example 2 (Non-parametric regression with a symmetric error distribution). Consider the regression model $Y = \theta(X) + e$ with $E_{e|X}e^2 = \theta(X)^2 + 1$. Suppose that given X, e is symmetric distributed. 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).

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

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

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 $v_{ij}(X) = E_{e|X}\exp(2(i+j-3))$, $2 \le i, j \le k_o$. Then, as $||\theta - \theta_o|| \to 0$,

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

This shows that SEL and EGLS can weight the observations according to the conditional moments. Note that, the local quasi-likelihood in Fan & Gijbels (1996) allows for only the second moment information and in general fails to produce an efficient estimator when the other moment information is available.

Assume that $U(x) = E_{Y|X=x} G(Y, \theta_o(X)) G^\tau(Y, \theta_o(X))$ is uniformly positive definite with respect to $x \in \Omega = [0, 1]$, and that $\sup_{x\in\Omega} ||U(x+t) - U(x)|| \le c|t|$ for some positive constant c. Assume that $E_{e|X}\exp(t_o|e|) < \infty$ for some positive constant t_o. Then, theorem 5 holds.

Example 3 (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

$$R(\beta) = \frac{1}{2} E_X \left\{ (\beta_o - \beta)^\tau \frac{XX^\tau}{\sigma^2(X)} (\beta_o - \beta) \frac{1}{1 + (\beta_o - \beta)^\tau XX^\tau (\beta_o - \beta)/\sigma^2(X)} \right\}$$

$$R_M(\beta) = \frac{1}{2} E \left\{ \frac{1}{\sigma(X)^2} (Y - X^\tau\beta)^2 \frac{1}{1 + (\beta_o - \beta)^\tau XX^\tau (\beta_o - \beta)/\sigma^2(X)} \right\}.$$

Assume that the error e satisfies the following conditions: $E_{e|X}e = 0$, $\sup_{x\in\Omega} E_{Y|X=x}|e|^{\alpha_o} < \infty$, $\alpha_o \ge 6$. Then, under the conditions **(K0)**, **(X0)**, $d_o \le h^p n^\eta \le d_1$ and $0 < \eta < 1/3$, it follows from theorems 1, 2 and 3 that

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

with $V^{-1} = E_X(XX^\tau / E_{e|X}e^2)$. Thus, $\hat{\boldsymbol{\beta}}$ is asymptotically as efficient as Carroll's estimator (Carroll, 1982; Robinson, 1987).

Example 4 (An endogenous dummy variable model) (Newey, 1993). Consider the non-linear 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 SEL. First, we define $Y^* = (Y, S)^\tau$, $G(Y^*, X^\tau\beta, \lambda) = Y - \lambda S - \phi(X, \beta)$. Then, we have

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

Under the conditions similar to those in example 3, using theorem 1, 2 and 3, we have

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

with

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

Example 5 (Regression quantiles) (Koenker & 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 the 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 $\partial E_{Y|X=x}G(Y, X^\tau\beta)/\partial\beta$. 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 $\underline{h}^* = (h^*, \ldots, h^*)$ 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 $E_{Y|X=x}G(Y, x^\tau\beta)$ is involved.

Assume that $f_{e|X=x}(z)$ is continuous and bounded with respect to (z, x). Note that $f_{e|X=x}(0)$ 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 1, 2 and 3 that

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

with $V^{-1} = [1/q(1-q)]E_X((f_{e|X}(0))^2XX^\tau)$. Thus, $\hat{\boldsymbol{\beta}}$ 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 behaviour of conditional error densities given different values of the predictor.

The similar phenomenon is also found in the multiple quantile regression (Welsh *et al.*, 1994; Zhang & Gijbels, 1999).

Note that in example 5, 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 (Powell, 1994). However, it seems difficult to find such kind of characterizations in example 1.

Example 6. Consider the ordinary non-parametric regression model $Y = \theta(X) + e$ with unknown function $\theta(X)$ and $E_{e|X}e = 0$. Let (y_i, x_i), $i = 1, \ldots, n$ be i.i.d. observations. Write $G(Y, \theta(X)) = Y - \theta(X)$. Then

$$l_{sr}(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ji}(y_j - \theta(x_j))\right)^2\left[\sum_{j=1}^{n} w_{ji}(y_j - \theta(x_j))^2\right]^{-1}(1 + o_p(1))$$

$$\hat{R}_M(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta(x_i))^2\left[\sum_{j=1}^{n} w_{ji}(y_j - \theta(x_j))^2\right]^{-1}.$$

$l_{sr}(\theta)$ 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.

Let $\Omega = [0, 1]$. Assume that $E_{e|X}\exp(t_o|e|) < \infty$ 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 2 holds. In particular, when $w^* \le 1/4$, the SEL estimator attains the optimal global convergence rate in the L_2 norm.

5. Discussions

5.1. Bandwidth

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

(i) 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;
(ii) the kernel weights with a nearest neighbour 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 $K(||t||) = (1 - ||t||^3)_+^3$ (Cleveland, 1979).

We extend a theorem due to Fan & Gijbels (1996) that the nearest neighbour bandwidth is adaptive for the design points.

Proposition 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 $x \in \Omega$.

We note that under condition (**X0**), by the above proposition, there exist two positive constants d_o and d_1 independent of x such that when n is large, for $h = h(x, m_n)$, we have $h_{nl}^p = d_o n^{-\eta} \leq h^p \leq d_1 n^{-\eta} = h_{nu}^p$. Some issues on how to select the bandwidth for a finite sample in practice and how to compute the associated estimators have been addressed in LeBlanc & Crowley (1995). There is some further progress in this direction (Zhang & Liu, 2000).

5.2. Conclusions

The empirical likelihood cannot be directly applied to a problem with certain infinite dimensional parameters of interest involved. To overcome this difficulty, we proposed the sieve empirical likelihood (SEL) approach. The large sample study shows such a method is promising. Like the parametric likelihood based maximum estimators (Lehmann, 1983; Wong & Severini, 1991), the SEL based maximum estimators are asymptotically optimal in the parametric regression setting and can achieve the optimal global convergence rate in the non-parametric regression setting. 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 (LeBlanc & Crowley, 1995). Recently this theorem has been extended to the case of infinite dimensional parameter (Fan *et al*., 2001; Fan & Zhang, 2000). Like LeBlanc & Crowley (1995), the SEL can be constructed for censored survival data and for a random effect regression model. However, their properties are still unknown.

Acknowledgements

The first author has greatly benefitted from several long discussions with Professor Wing Hung Wong, Department of Biostatistics, Harvard University. Example 1 is borrowed from him. The helpful comments of the Editor and the referee are gratefully acknowledged. The work was partially supported by Department of Statistics, Chinese University of Hong Kong, by "Project d'Actions de Recherche Concertées" (No. 93/98-164) from the Belgian government, and by the National Natural Science Foundation of China.

References

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

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

Carroll R. J. & Ruppert, D. (1988). *Transformation and weighting in regression*. Chapman & Hall, New York.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc*. **74**, 829–836.

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. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

Fan, J. & Zhang, J. (2000). Sieve empirical likelihood ratio tests for non-parametric functions. EURANDOM report 2000-46, Eindhoven, The Netherlands.

Fan, J., Zhang, C. & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist*. **26**, 153–193.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Jung, S. (1996). Quasi-likelihood for median regression models. *J. Amer. Statist. Assoc.* **91**, 251–257.
Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25**, 2084–2102.
Kitamura, Y. & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**, 861–874.
Koenker, R. & Bassett, G. (1978). Regression quantitles. *Econometrica* **46**, 33–50.
LeBlanc, M. & Crowley, J. (1995). Semiparametric regression functionals. *J. Amer. Statist. Assoc.* **90**, 95–105.
Lehmann, E. L. (1983). *Theory of point estimation*. Wiley, New York.
Mammen, E. & van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25**, 1014–1035.
Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Handbook of statistics* (eds G. S. Maddala, C. R. Rao & H. D. Vinod), **11**, 419–453. Elsevier Science, Amsterdam.
Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.
Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York.
Powell, J. L. (1994). Estimation of semiparametric models. In *Handbook of econometrics* (eds R. F. Engle & D. L. McFadden), **IV**, 2443–2521 Elsevier Science, Amsterdam.
Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity. *Econometrica* **55**, 875–891.
Shen, X. (1997). On methods of sieves and penalization. *Ann. Statist.* **25**, 2555–2591.
Shen, X. & Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580–615.
Shen, X., Shi, J. & Wong, W. H. (1999). Random sieve likelihood and general regression models. *J. Amer. Statist. Assoc.* **94**, 835–846.
Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussions). *J. Roy. Statist. Soc. Ser. B* **47**, 1–52.
Stone, C. (1982). Optimal global rates of convergence for non-parametric regression. *Ann. Statist.* **10**, 1040–1053.
Tibshirani, R. & Hastie, T. J. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559–567.
Welsh, A. H., Carroll, R. J. & Ruppert, D. (1994). Fitting heteroscedastic regression models. *J. Amer. Statist. Assoc.* **89**, 100–116.
Wong, W. H. & Severini, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter space. *Ann. Statist.* **19**, 603–632.
Wong, W. H. & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLES. *Ann. Statist.* **23**, 339–362.
Zhang, J. & Gijbels, I. (1999). Sieve empirical likelihood and extensions of the generalized least squares. Discussion Paper 9911, Institut de Statistique, Université Catholique de Louvain (http://www.stat.ucl.ac.be/).
Zhang, J. & Liu, A. (2000). Local polynomial fitting based on empirical likelihood. Report 2000-025, EURANDOM, The Netherlands.

Received February 2000, in final form November 2001

Jian Zhang, EURANDOM, Den Dolech 2, 5612 AZ, Eindhoven, The Netherlands.
E-mail: jzhang@euridice.tue.nl

Appendix A. Technical conditions

In this appendix, we collect the conditions used in the previous sections. We begin with some notations. Let $N(\epsilon, \rho, \mathscr{F})$ be the covering number, the smallest number of ϵ-balls in metric ρ needed to cover $\mathscr{F}$. Let $N^B(\epsilon, \rho, \mathscr{F})$ be the bracketing covering number, the smallest number m for which there exist $f_k^l \le f_k^u$, $k = 1, \ldots, m$ with $\max_{1\le k\le m} \rho(f_k^u - f_k^l) \le \epsilon$ and $\mathscr{F} \subset \bigcup_{k=1}^m [f_k^l, f_k^u]$. 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 $H(\epsilon, \rho, \mathscr{F}) = \log N(\epsilon, \rho, \mathscr{F})$ and $H^B(\epsilon, \rho, \mathscr{F}) = \log N^B(\epsilon, \rho, \mathscr{F})$. Let $K(\cdot)$ be a univariate

density function. Let r_o denote a predetermined positive constant. We first introduce the condition for X.

(X0)

X has a compact and convex support Ω. And the density f of X is continuous over Ω and $0 < \sup_{x\in\Omega} f(x) < \infty$.

The conditions for the parametric regression setting are as follows.

As before, consider the case that $\theta(x) = x^\tau\beta$. Denote by β_o the true value of β. For $1 \le k, m \le k_o$, set

$$\mathscr{F}_{G_k} = \{G_k(\cdot, \cdot^\tau\beta) : ||\beta - \beta_o|| \le r_o\}, \quad \mathscr{F}_{G_kG_m} = \{G_k(\cdot, \cdot^\tau\beta)G_m(\cdot, \cdot^\tau\beta) : ||\beta - \beta_o|| \le r_o\}.$$

In addition to (**X0**), we assume:

(K0)

$K(s)$ is non-increasing as s goes from 0 to ∞. For any $\delta_o > 0$,

$$\int K(||t||)dt = 1, \quad \int K(||t||)t\,dt = 0, \quad \int K(||t||)^{(1+\delta_o)/\delta_o}dt < \infty. \tag{7}$$

(P1)

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

(P2)

There exists $\alpha_o > 4$ such that

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

(P3)

For $1 \le k \le k_o$,

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

(P4)

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

(P5)

For $1 \le k \le k_o$,

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

(P6) (Covering number)

For $1 \le k \le k_o$,

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

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

(P7)

As $z \to 0$,

$$\sup_{x\in\Omega, ||\beta-\beta_o||\le r_o} |E_{Y|X=x+z}G(Y, X^\tau\beta) - E_{Y|X=x}G(Y, X^\tau\beta)| \to 0,$$

$$\sup_{x\in\Omega, ||\beta-\beta_o||\le r_o} |E_{Y|X=x+z}G_m(Y, X^\tau\beta)G_k(Y, X^\tau\beta) - E_{Y|X=x}G_m(Y, X^\tau\beta)G_k(Y, X^\tau\beta)| \to 0,$$

for $1 \le m, k \le k_o$.

For any $\rho > 0$,

$$\sup_{||\beta-\beta_o||\geq\rho} ||EG(Y, X^\tau\beta)|| \neq 0;$$

and for any sequences of constants, $0 < \rho_{n1} < \rho_{n2} \to 0$, there exists a constant c such that

$$\sup_{\rho_{n2}\geq||\beta-\beta_o||\geq\rho_{n1}} ||EG(Y, X^\tau\beta)|| \geq c\rho_{n1}.$$

(P8)

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

$$E_{Y|X=x}|G_k(Y, X^\tau\beta) - G_k(Y, X^\tau\beta_o)|^2 \leq b_1||\beta - \beta_o||,$$
$$|E_{Y|X=x+u}G_k(Y, X^\tau\beta) - E_{Y|X=x}G_k(Y, X^\tau\beta)| = o(||\beta - \beta_o||), \quad \text{as } u \to 0.$$

(P9)

There exist $\alpha_o > 4$ and $0 < \zeta_1 \leq 1$ such that for $1 \leq k \leq k_o$ and any $||\beta - \beta_o|| \leq r_o$,

$$|G_k(y, x^\tau\beta) - G_k(y, x^\tau\beta_o)| \leq M(y, x)||\beta - \beta_o||^{\zeta_1}$$

with

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

The conditions for the non-parametric regression setting are as follows.

For $1 \leq k, m \leq k_o$, and for any positive sequence $\{r_n\}$ with $r_n \to 0$, write

$$\mathscr{F}^*_{G_k} = \{G_k(\cdot, \theta(\cdot)) : ||\theta - \theta_o|| \leq r_n\}, \quad \mathscr{F}^*_{G_kG_m} = \{G_k(\cdot, \theta(\cdot))G_m(\cdot, \theta(\cdot)) : ||\theta - \theta_o|| \leq r_n\}.$$

In addition to the condition **(X0)**, we assume

(K1)

For some positive constant c_o, and for any $s_1, s_2 \in R^1$, $|K(s_1) - K(s_2)| \leq c_o|s_1 - s_2|$. And (7) holds.

We define the conditions **(N1)**, **(N2)**, **(N3)** and **(N5)** by replacing $x^\tau\beta$, $X^\tau\beta$ and $||\beta - \beta_o|| \leq r_o$ in the conditions **(P1)**, **(P2)**, **(P3)** and **(P5)** by $\theta(x)$, $\theta(X)$ and $||\theta - \theta_o|| \leq r_o$, $\theta \in \Theta$, respectively.

(N4)

For $1 \leq k, m \leq k_o$, $D_{km} = G_kG_m$, as $u \to 0$.

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

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

(N6) (Entropy)

For $1 \leq k, m \leq k_o$,

$$H^B(\epsilon, L_2(P), \mathscr{F}^*_{G_k}) \leq A^*(P)(\epsilon/r_n)^{-w^*}, \quad H^B(\epsilon, L_2(P), \mathscr{F}^*_{G_kG_m}) \leq B^*(P)(\epsilon/r_n)^{-w^*},$$

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

(N7)

As $z \to 0$,

$$\sup_{x\in\Omega,||\theta-\theta_o||\le r_o} ||E_{Y|X=x+z}G(Y,\theta(X)) - E_{Y|X=x}G(Y,\theta(X))|| \to 0,$$

$$\sup_{x\in\Omega,||\theta-\theta_o||\le r_o} |E_{Y|X=x+z}G_m(Y,\theta(X))G_k(Y,\theta(X)) - E_{Y|X=x}G_m(Y,\theta(X))G_k(Y,\theta(X))| \to 0,$$

for $1 \le m,k \le k_o$.

For any $\rho > 0$,

$$\sup_{||\theta-\theta_o||\ge\rho} ||EG(Y,\theta(X))|| \ne 0.$$

Moreover, for any $0 < \rho_{n1} < \rho_{n2} \to 0$, there some constant c such that

$$\sup_{\rho_{n2}\ge||\theta-\theta_o||\ge\rho_{n1}} ||EG(Y,\theta(X))|| \ge c\rho_{n1}.$$

(N8)

(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 $k = 1,\ldots,k_o$,

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

and

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

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

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

where c is a constant.

(iii) Let $U(x) = E_{Y|X=x}G(Y,\theta_o(X))G^\tau(Y,\theta_o(X))$. $U(x)$ 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 $x \in \Omega$.

Remark 4. These conditions are not necessarily independent. For example, the condition (**N8**) used to prove theorem 5 implies the condition (**N6**).

Appendix B. Proofs

Throughout the remains of the paper, we denote $f_{\max} = \max_x f(x)$, $f_{\min} = \min_x f(x)$, $h_{nl}^p = h_{nl}^p(\eta) = d_o n^{-\eta}$ and $h_{nu}^p(\eta) = d_1 n^{-\eta}$ for some positive constants d_o and d_1. Write

$$A_n(x,\theta) = \sum_{j=1}^n w_j(x)G(y_j,\theta(x_j)),$$

$$S_n(x,\theta) = \sum_{j=1}^n w_j(x)G(y_j,\theta(x_j))G^\tau(y_j,\theta(x_j)),$$

$$Z_n(\theta) = \max_{1\le j\le n}||G(y_j,\theta(x_j))||.$$

Let $e_n(x,\theta)$ be the minimum eigenvalue of $S_n(x,\theta)$. For convenience, we use $A_n(x,\beta)$, $Z_n(\beta)$, $S_n(x,\beta)$ and $e_n(x,\beta)$ to denote $A_n(x,\theta)$, $Z_n(\theta)$, $S_n(x,\theta)$ and $e_n(x,\theta)$ in lemmas 1 and 3–6 below.

The proofs of the lemmas in this appendix are omitted but available from Zhang & Gijbels (1999).

Lemma 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 $o_p(1)$ *in* (5) *and* (6) *tends to zero uniformly in* $\theta\in\Theta, ||\theta-\theta_o||\le r_n$, *and* $x\in\Omega$.

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

Lemma 2
Let $\mathscr{F}_n$ *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 $g\in\mathscr{F}_n$. *Suppose that for some constants* v, w *and* d_n,

$$\sup_{g\in\mathscr{F}_n}\mathrm{var}(g)\le v,\quad N(\epsilon,L_1(P_n),\mathscr{F}_n)\le A(P_n)(d_n\epsilon)^{-w}$$

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

$$P(\sup_{\mathscr{F}_n}|\upsilon_n(g)|>M)\le c_1(P)(\sqrt{n}(Md_n)^{-1})^w\exp(-M^2/(2\times 64^2\times v))+c_2(P)v^{-w}\exp(-nv)$$

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

Lemmas 3–6 below will be used to prove theorem 1.

Lemma 3
Suppose $\theta(X)=X^\tau\beta$. *Then, under the conditions* (**K0**), (**X0**), (**P1**)–(**P3**), (**P5**) *and* (**P6**), $0<\eta<(\alpha_o-2)/\alpha_o$ *and* $r_n=O(n^{-1/\alpha_o})$, *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 4
Suppose $\theta(X)=X^\tau\beta$. *Then, under the conditions* (**K0**), (**X0**), (**P1**)–(**P4**), $0<\eta<(\alpha_o-2)/(\alpha_o+2)$ *and* $r_n\to 0$, *we have* $S_n(x,\beta)=E_{Y|X=x}G(Y,X^\tau\beta)G^\tau(Y,X^\tau\beta)+o_p(1)$, *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 5
Suppose θ_0 *is an inner point of* Θ. *Under the conditions* (**K1**), (**X0**), (**N1**)–(**N3**), (**N5**) *and* (**N6**), $r_n=O(n^{-1/\alpha_o})$, *and*

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

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 6
Suppose that θ_0 *is an inner point of* Θ. *Then, under the conditions* (**K1**), (**X0**), (**N1**)–(**N4**), (**N6**), $r_n\to 0$, *and*

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

we have $S_n(x,\theta) = E_{Y|X=x}G(Y,\theta(X))G^\tau(Y,\theta(X)) + o_p(1)$ *where* $o_p(1)$ *is uniform in* $x \in \Omega$, $||\theta - \theta_o|| \le r_n$, $\theta \in \Theta$, *and* $h_{nl}^p \le h^p \le h_{nu}^p$, *and* θ_o *is the true value of* θ.

Proof of theorem 1. It follows from lemmas 1, 3, 4, 5 and 6.

Proof of theorem 2. It is similar to the proof of theorem 1 in Zhang & Liu (2000) and thus omitted.

Recall that $U(x) = E_{Y|X=x}G(Y,X^\tau\beta_o)G^\tau(Y,X^\tau\beta_o)$. To facilitate the proof of theorem 3, we establish lemmas 7–10 below.

Lemma 7

We suppose that when G is bounded, the conditions (**K0**), (**X0**), (**P1**), (**P4**), (**P5**), (**P6**) *and* (**P8**) *hold; and that when G is unbounded, the conditions* (**K0**), (**X0**), (**P1**), (**P2**) *and* (**P6**)–(**P9**) *hold. If* $r_n \to 0$ *and*

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

then

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

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

Lemma 8

Under the same conditions in lemma 7, *we have*

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

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

Lemma 9

Suppose that when G is bounded, (**P1**), (**P3**)–(**P8**) *hold*; *when G is unbounded,* (**P1**) *to* (**P9**) *hold. Then,* $EG(Y,X^\tau\beta) = O(||\beta - \beta_o||)$ *and*

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

Lemma 10

Suppose that when G is bounded, (**P1**), (**P3**)–(**P6**), (**P8**) *hold*; *when G is unbounded,* (**P1**)–(**P9**) *hold. Then*

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

Proof of theorem 3. We take for example $\theta(x) = x^\tau\beta$. It suffices to establish the asymptotic normality of the estimator defined by minimizing

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

where the minimization is performed over $||\beta - \beta_o|| \le r_n$, and $\{r_n\}$ is a sequence of constants satisfying $r_n \ge cn^{-1/\alpha_o}$ for some constant $c > 0$.

For this purpose, we first easily observe from lemmas 7, 8 and 9 that under the conditions (P1) to (P9), and $0 < \eta < (\alpha_o - 2)/(\alpha_o + 2)$,

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

where, without loss of generality, we drop the factor $1 + o_p(1)$. Here and hereafter $o_p(1)$ is always uniform in (x, β, h), $x \in \Omega$, $||\beta - \beta_o|| \le r_n$ and $h_{nl}^p \le h^p \le h_{nu}^p$. Then via the same argument and symbols of th. 5 of Pollard (1984, p. 141), we have

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

and $E\Delta = 0$,

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

Therefore, let $V = E\Delta\Delta^\tau$, we have

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

The proof is completed.

Proof of theorem 4. It is similar to th. 1 of Zhang & Liu (2000) and thus omitted. Lemmas 11–14 below are employed to prove theorem 5.

Lemma 11

Under the conditions **(X0)**, **(K1)**, **(N1)**, **N8(i)** *and* **N8(ii)**, *if* $\pi_1 \ge 0$, $q \ge 1$, *and*

$$\eta \ge \pi_1(1 - b_2), \quad \eta > \pi_1(a_1b_1 - b_2), \quad \eta < \frac{1}{3}(1 - 2b_2\pi_1 + a_1b_1\pi_1),$$
$$\eta < \frac{2}{6 + w^*}\left(1 - 2b_2\pi_1 + w^*\pi_1 + a_1b_1\left(1 - \frac{w^*}{2}\right)\right),$$

then

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

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 12
In addition to the assumptions in lemma 11, *we assume that* **N8(iii)** *holds. Then, we have*

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

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 13
Under the conditions (**X0**), (**N1**), **N8(i)** *and* **N8(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 14
Under the conditions (**X0**), (**N1**) *and* (**N8**), *if* $0 < \pi_1 < 1/(2+w^*)$ *and* $0 < w^* < 2$. *Then uniformly in* θ *with* $||\theta-\theta_o|| \le c_1 n^{-\pi_1}$ *and* $\theta \in \Theta$

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

Proof of theorem 5. First of all, we claim O_p and o_p below are uniform in $\theta \in \Theta$. By theorems 1 and 4, we need only to consider the EGLS estimator defined on the parameter set $\{\theta \in \Theta : ||\theta-\theta_o|| \le r_n\}$ with the constant $r_n \to 0$. Note that, as $||\theta-\theta_o|| \le r_n \to 0$, applying the similar arguments used to prove theorem 1, we obtain

$$\hat{R}_s(\theta) - \hat{R}_s(\theta_o) = -\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ji}G(y_j,\theta(x_j))\right)^\tau U^{-1}(x)\sum_{j=1}^{n} w_{ji}G(y_j,\theta(x_j))(1+o_p(1)). \quad (8)$$

The strategy adopted in the remainder of the proof, similar to Shen & 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 neighbourhoods.

Iteration. 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

$$\eta \ge \pi_1(1-b_2), \quad \eta > \pi_1(a_1b_1 - b_2), \quad \eta < \frac{1}{3}(1-2b_2\pi_1 + a_1b_1\pi_1),$$
$$\eta < \frac{2}{6+w^*}\left(1-2b_2\pi_1 + w^*\pi_1 + a_1b_1\left(1-\frac{w^*}{2}\right)\right),$$

then in the next step we can show

$$||\hat{\theta}-\theta_o|| = O_p(n^{-(\pi_1+\pi_2)/2}) + O_p(n^{-(\pi_1/2+1/(2+w^*))}), \quad (9)$$
$$||\hat{\theta}^{(1)} - \theta_o^{(1)}|| = O_p(n^{-(\pi_1+\pi_2)(1-w^*)/2}) + O_p(n^{-(\pi_1/2+1/(2+w^*))(1-w^*)}), \quad (10)$$

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

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

Write $\pi_2 = \eta + b_2\pi_1$ and $b_2 = 2(1-w^*)/(2-w^*)$. Then, the combination of (8) and lemmas 11 to 14 gives

$$\hat{R}_s(\theta) - \hat{R}_s(\theta_o) = -\frac{1}{n}\sum_{j=1}^{n} a^{\tau}(x_j,\theta(x_j))W_j + \frac{1}{2n}\sum_{j=1}^{n}||a(x_j,\theta(x_j))||^2 + O_p(n^{-(\pi_2+\pi_1)}) + O_p(n^{-(\pi_1+2/(2+w^*))}) \tag{11}$$

provided

$$1/(2+w^*) \geq \pi_1 \geq 0, \quad w^* \leq 1, \tag{12}$$

$$\eta > \pi_1(a_1b_1 - b_2), \quad \eta < \frac{1}{3}(1-2b_2\pi_1 + a_1b_1\pi_1), \tag{13}$$

$$\eta < \frac{2}{6+w^*}\left(1-2b_2\pi_1 + w^*\pi_1 + a_1b_1\left(1-\frac{w^*}{2}\right)\right). \tag{14}$$

Let $||a||^2 = E|a(X,\theta(X))|^2$ and O_p be uniform with respect to $\theta \in \Theta$. By (11) and the definition of $\hat{\theta}$, and by th. 2.2 and 2.3 of Mammen & van de Geer (1997), we find

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

which together with the assumptions implies (9). 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 (10).

The above iteration continues if (12)–(14) hold, $\pi_1 < (1-w^*)/(2-w^*)$, $\eta \geq \pi_1(1-b_2)$, and $(\eta + 2b_2\pi_1 + \pi_1)(1-w^*) > 2\pi_1$. Now the proof is completed by some simple calculations.

Proof of proposition 1. See Zhang & Gijbels (1999).