DEPARTMENT OF ECONOMICS WORKING PAPER SERIES

Two Stage Semi Parametric Quantile Regression

J. M. Krief
Louisiana State University

Working Paper 2009-05
http://www.bus.lsu.edu/economics/papers/pap09_05.pdf

Department of Economics
Louisiana State University
Baton Rouge, LA 70803-6306
http://www.bus.lsu.edu/economics/

Two Stage Smooth Semi Parametric Quantile Regression

Jerome Krief

Louisiana State University*

March 30,2009

Abstract

We propose a root n consistent estimator for β_0 when the qth conditional quantile of Y given X=x and Z=z takes the semi linear form $g(x)+z\prime\beta_0$ where $g(.)$ is an unknown real valued function,β_0 a finite dimensional parameter and (X,Z)a couple of explanatory variables.Importantly, our estimator attains,under homoscedasticity,the semi parametric efficiency bound.This estimation is conducted in two steps. First, a Robinson's like demeaning of the original model is employed which provides a new quantile regression whose nuisance terms are estimated via a non parametric procedure.In the second stage, the quantile regression is conducted by smoothing the check function.We show that the previous estimator belongs to a class of estimators we propose to name "two stage smooth semi parametric quantile"

JEL-codes: C22, C51. **Key words:** M-Smoothing,Quantile Regression,Adaptive Estimation,Semi Parametric model.

*Department of Economics,2125 CEBA Bldg., Baton Rouge, LA 70803, phone: (225) 388-3806,e-mail:jkrief1@lsu.edu

1 Introduction

Quantile regression serves many important purposes in Econometrics. First, even under the Gauss Markov assumptions the LAD (least absolute deviation) estimator minimizing the ℓ_1 norm of the errors is well acknowledged as a non linear estimator asymptotically more efficient than the OLS (Koenker and Basset 1978) when the error distribution departs from Normality. Also, conducting a quantile regression permits researchers to obtain a more comprehensive picture of the stochastic relationship between the dependant and the explanatory variables by learning about the "marginal effects" of the covariates on the various quantiles such as in the field of Labor Economics(Buchinsky 1994).Finally, a valid conditional quantile restriction on the unobservable term of a structural equation permits to identify the parameters of interest due to the equivariance property of the conditionals quantile operator to monotonic transformations, which has proved valuable in the context of censored data (Chen and Khan 2001)and binary choice choice modeling when positing a parametric family for the latent error distribution is untenable[1].(Manski 1985,Horowitz 1992)

Similarly to a conditional mean regression, the risk of specification for the conditional quantile is present. Thus, non-parametric point wise estimators for estimating a conditional quantile function have been proposed, which essentially extend the kit of kernel based procedures for local mean regression (Watson 1964) to the realm of quantile regression(Fan et Al 1994). The local quantile regression (Chaudhuri 1994) is probably the most popular as the asymptotic using the Local Bahadur Representation has been well developed while other approaches seek to improve small sample performance such as spline smoothing (Koeneker 1994), double Kernel smoothing (Yu and Jones 1998) or tackle endogeneity (Horowitz and Lee 2007, Cherozhukov Gagliardini Scaillet 2007).

[1]The maximum score estimator permits identification of the parameters up to a positive scale, see Horowitz 1992 for identifying restrictions.

Even when asymptotically Gaussien distributed,the above mentioned estimators are not root n consistent with the speed of convergence in probability deteriorating exponentially as the number of explanatory variables increases (often called the "curse of dimensionality" in the non parametric jargon). As a reaction to this latter issue, models emerged imposing some form on the multivariate quantile function such as the Additive Quantile Model (Horowitz and Lee 2004) in which case a root n consistent point wise estimator can be constructed via sieves estimation provided the conditional quantile is infinitely smooth.

An interesting sub model is the semi-parametric model for quantile regression, which offers a compromise between efficiency and specification.Lee 2003 is a seminal paper for semi-parametric quantile. First the Average Quantile Estimator(AQR) for the linear part ,under homoscedasticity [2],is root n consistent while simultaneously efficient (Newey 1990). Also, under Heteroscedasticity another "one step" efficient estimator reaching the efficiency bound is proposed.To the best of my knowledge those two estimators are the sole procedures to reach efficiency in the context of a semi parametric quantile regression.

Yet, it is puzzling to notice that the nature of the efficient estimator under homoscedasticity,average derivative based(Chaudhuri,Doksum and Samarov 1997), marketably differs from the efficient one under heteroscedasticity, score approximation based(Stone 1975,Bickel 1982).As Econometricians we are familiar with dealing with a class of estimators for parametric models containing (in the sense of efficiency) a simpler class of estimators such as G.M.M. and Two Stage Least Squares to cite probably the most recognized.This inclusive property is not only theoretically interesting but also it brings guidelines as to what estimator to use in finite sample based upon our testing over the

[2]Throughout this paper we will employ the same terminology as Lee 2003 for the sake of consistency keeping in mind that the term homoscedastic error in this context is a substitute for $f(0|X,Z) = f_\varepsilon(0)$ a.s. where $f(0|X,Z)$ is the density of the error ε conditional on (X,Z) while f_ε denotes the marginal density of the error. .

statistical relationship between the unobservable term and the explanatory variables.This last point is all the more relevant in the context of a semi parametric quantile regression because of the complex manner in which the stochastic relationships affect the efficiency bound (see section 6).

In this paper we wish to define a general class of estimators for estimating efficiently the linear part, thus offering a unifying approach to efficient estimation in the context of a semi parametric quantile regression. We introduce this class of estimators we name "two stage smooth semi parametric quantile"(2SSSPQ)and show that in any stochastic contingency there exists an efficient estimator belonging to this class.The rest of the paper is organized in two parts. In section 2, we rapidly remind the reader about the semi parametric efficiency bound.In section 3-7, we are to focus on the homoscedastic case where the consistency and asymptotic efficiency of a two stage smooth estimator is derived.In section 8,we show that this previous estimator belongs to the 2SSSPQ family whose efficient property for heteroscedastic models are derived generalizing the approach from section 3-7.In section 9, a Monte Carlo experiment illustrates the finite sample properties of a 2SSSPQ.

2 The Semi Parametric Efficiency Bound

In this section we define the efficiency bound for the slope coefficient of a general semi-linear model satisfying for some given $\psi(.)$:

$$E[\psi(Y - Z'\beta_0 - g(X))|X, Z] = 0 \ \textit{almost surely} \ \text{(a.s.)}$$

where (Y, Z, X) is a $\mathbb{R} \otimes \mathbb{R}^K \otimes \mathbb{R}^d$ valued random variable such that $(K, d) \in {\mathbb{N}_*}^2$ and $(\beta_0, g(.)) \in \mathbb{R}^K \otimes L^2(\mathbb{R}^d, \mu)$ is an unknown parameter where μ will indicate the Lebesgue measure in the appropriate Euclidean space.Notice that this model is the semilinear conditional mean case(Robinson 1988)when $\psi(.)$ is the identity function while $\psi(.) = 1_{.<0} - q$ for some q $\in (0, 1)$ yields the semilinear conditional quantile model(Lee 2003).

We will note $\pi_0(.|x,z)^2$ the Lebesgue density of Y conditional on $X=x$ and $Z=z$ and $\Pi = \{\pi(.|x,z) \in L^2(\mathbb{R},\mu), \pi(.|x,z)^2 > 0 \text{ and } \int \pi(.|x,z)^2 d\mu = 1\}$.The interest is to find the minimum variance achievable by regular[3]estimators of β_0.The concept of an efficiency bound was introduced in Stein 1956 and its the computation for Econometrics models has been typically conducted via the projection method(Bickel 1983,Newey 1990) ,which was successfully employed in Lee 2003 to obtain the efficient bound for a semi parametric conditional quantile model. The next definition is largely adapted from Severini and Tripathi 2001 which provides an alternative to the projection formula in order to obtain the efficient bound. We believe that this approach is more closely linked to the maximum likelihood origin of this concept while additionally often more rapid at retrieving the bound of semi linear models as extensively illustrated in Severini and Tripathi 2001.

Definition

Let suppose that $\pi_0(.|x,z)$ *belongs to* Π.

Let $\{\pi_t\}_{t\in[0,b]} \subseteq \Pi$ *be an arbitrary curve passing through* π_0 *at* $t=0$ *for some* $b>0$ *and which is also compatible with the true semi linear conditional model.*

Let $\overline{\mathbb{T}(\Pi,\pi_0)} = \{\dot{\pi} \in L^2(\mathbb{R},\mu) a.e., \int \dot{\pi}\pi_0 d\mu = 0 a.e.\}$ *where* $\dot{\pi} = \frac{\partial \pi_t}{\partial t}|_{t=0}$.

Let $\mathcal{I} = 4E[\int \dot{\pi}^2 d\mu]$ *be the Maximum Likelihood information for a one parameter problem and* $<.,.>_F$ *the Fisher inner product on* $\overline{\mathbb{T}(\Pi,\pi_0)}$ *inducing the norm* $||\dot{\pi}||_F = \mathcal{I}^{1/2}$.

For any $c \in \mathbb{R}^K$ *let* $A_c : \Pi \to \mathbb{R}$ *such that* $A_c(\pi_t) = c'\beta_t$.*Suppose that there exists a linear functional* $\nabla A_c : (\overline{\mathbb{T}(\Pi,\pi_0)}, <.,.>_F) \to \mathbb{R}$ *such that:*

(i)$lim_{t\to 0+}|\frac{A_c(\pi_t)-A_c(\pi_0)}{t} - \nabla A_c(\dot{\pi})|$=0

(ii)*For any* $\{v_n\}_{n\in\mathbb{N}} \subseteq \overline{\mathbb{T}(\Pi,\pi_0)}$*such that* $lim_n||v_n||_F = 0$ *implies* $lim_n|\nabla A_c(v_n)| = 0$.

[3]See Newey 1990 and Hajek's Theorem (1970)for the definition of a regular estimator. This class of estimators rules out super efficient estimators that may for a family of parametric densities "beat" the Cramer bound.

Then $||\nabla A_c||^2 = c'\Omega_0 c$ *for some K by K matrix* Ω_0*, which is called the semi parametric efficiency bound for regular estimators of* β_0*.*

Comments:

The efficiency bound is the supremum of the Cramer bound for β_0 over all possible parametric conditional densities that agree with the true density for some one dimensional parameter.Sometimes,this bound is called the Cramer bound of the "least favorable" model (Stein 1956)because it corresponds to the least efficient Maximum likelihood estimator of β_0.As indicated in the definition,the existence of such bound needs assumptions.The notion of a curve is the generalization of the Taylor's representation in Hilbert spaces,which relates to the regularity conditions on the parametric density adopted for ML estimation. In our context $\{\pi_t\}_{t\in[0,b]} \subseteq \Pi$ is a curve passing through π_0 means that for all t $\in [0, b]$ we have for almost all $(x, z) \in \mathbb{R}^d \otimes \mathbb{R}^K$:

$$\pi_t = \pi_0 + t\dot{\pi} + r_t \text{ for some } r_t \in L^2(\mathbb{R}, \mu)\text{such } lim_{t\to 0+} \int |\frac{r_t}{t}|^2 d\mu = 0.$$

The curve is said compatible with the true semi linear conditional model when :

$$\int \psi(y - z'\beta_t - g_t(x))\pi_t^2 d\mu = 0 \text{ a.e.}$$

so that the functional $A_c : \Pi \to \mathbb{R}$ exits, which is needed for deriving the variance of the maximum ML of β_0 under a parametric submodel of density. Once, those later conditions hold,the most important requirement is the existence of the pathwise derivative at the true density i.e.π_0, which is ensured by (i) and (ii).The intuition behind (i) is that small changes in the parameter of the conditional density around the true value do not abruptly alter the value of the K dimensional parameter of interest. Notice that without(ii)the efficiency bound does not exists so the continuity of the linear functional ∇A_c relates to the existence of regular estimators (Chamberlain 1986),which is an identification problem.In the semi linear conditional quantile case,(i) and (ii) are satisfied under mild requirement on the density of the error term and the inability to predict Z

from some measurable functions of X.Finally, $\overline{\mathbb{T}(\Pi, \pi_0)}$ is the linear closure of the "Cone Tangent" (Severini and Tripathi 2001),which is our domain of reference for only Cramer bounds of parametric densities satisfying the maximum likelihood score condition are relevant i.e $E[\frac{\partial log \pi_t^2}{\partial t}|_{t=0}] = 0$.

We are to succinctly sketch how the efficiency bound is constructed summarizing Severini and Tripathi 2001.In the definition we normalize the true parameter to be 0 for any parametric family of densities for it does not change the problem. So,let π_t be a curve passing through π_0 when $t = t_0$.Since that curve is locally compatible with the semilinear model the functional F such that $F\pi_t = \beta_t$ is well defined. Consequently $\hat{\beta}_{ML}$,the maximum likelihood estimator of β_0, is given by $F\pi_{\hat{t}_{ML}}$ where $\hat{t}_{ML} = Argmax\hat{E}[log\pi(y|x, z, t)^2]$ because of the invariance principle of the ML and the assumption of pathwise differtiability of F at $\dot{\pi} = \frac{\partial \pi}{\partial t}|_{t=t_0}$. Hence, there exits a linear functional $\nabla F(\dot{\pi})$ such that:

$$F\pi_{\hat{t}_{ML}} - F\pi_{t_0} = (\hat{t}_{ML} - t_0)\nabla F(\dot{\pi}) + o_p(n^{-1/2})$$

and consequently:

$$asymvar\sqrt{n}(\hat{\beta}_{ML} - \beta_0) = \Omega(\dot{\pi})$$

where $\Omega(\dot{\pi}) = \frac{\nabla F(\dot{\pi})\nabla F(\dot{\pi})'}{||\dot{\pi}||_I^2}$ is just one Cramer bound.Consequentely,any regular estimator β_R will satisfy for all c in $\mathbb{R}^K$:

$$asymvar\sqrt{n}c'(\hat{\beta}_R - \beta_0) \geq c'\Omega(\dot{\pi})c$$

Since there are possibly an infinity of one parameter problems permitting to recoverer the true conditional density π_0^2,the supremum of $c'\Omega(\dot{\pi})c$ over $\overline{\mathbb{T}(\Pi, \pi_0)}$ provides a lower bound for regular estimators.[4].Subsequentely,the conclusion of the definition arises because:

[4]Notice that when any one dimensional parametric model of density returns the same bound, the search is over. This is the adaptive case which occurs when $|\nabla F|$ is constant on the unit ball of $\overline{\mathbb{T}(\Pi, \pi_0)}$.

$$c'\Omega(\dot{\pi})c = (\frac{|c'\nabla F(\dot{\pi})|}{||\dot{\pi}||_F})^2$$

where $c'\nabla F(\dot{\pi})$ is what we called $\nabla A_c(\dot{\pi})$ which is under the previous assumptions (i) and (ii) well defined as a bounded linear functional on the Hilbert space$(\overline{\mathbb{T}(\Pi, \pi_0)}), < ., . >_F)$ yielding the efficiency bound as the squared norm of the linear functional.[5]

3 Motivation for a Robinson's Like Estimator For Semi parametric Quantile Regression under homoscedasticity

In this section we introduce the semi linear quantile regression model and rapidly described the computational steps required from the AQR estimator in order to reach the semi parametric efficiency bound under Homoscedasticity.

The semi parametric quantile regression model posits:

$$Y = g(X) + Z'\beta_0 + \varepsilon \text{ (I)}$$

$$\text{P}[\varepsilon < 0|X, Z]=\text{q } a.s. \text{ for some given q} \in (0, 1).$$

where Y is an observable variable,(X, Z) a couple of observable explanatory variables such that $(DimX, DimZ) = (d, K)$ with $min(d, K) \geq 1$,$g(.)$ is an unknown function ,β_0 a parameter of interest while ε is the error term.There are essentially two ways to interpret this model. First, the researchers may be primarily interested in estimating the q^{th} conditional quantile function of $Y|X, Z$ positing P$[Y < g(X) + Z'\beta_0|X, Z]$=q(a.s.)in which case P$[\varepsilon < 0|X, Z]=q$ a.s. is merely tautological. A good illustration is the conditional value at risk(R.Engle and S.Manganelli 2005).The second more common interpretation is

[5]By the Riesz representation theorem there exists a unique $\pi_* \in \mathbb{T}(\Pi, \pi_0)$ such that $||\nabla A_c||^2 = ||\pi_*||_F^2$ so that $\delta = 2(\int \pi_*^2 d\mu)^{1/2}$ corresponds to the "efficient influence function"(Newey 1990).

that Y is a response variable explained according to some Economic theory where ε contains unobservable terms (and/or variables omitted from the underlying theory)and $g(.)$ is left unspecified motivated by the researcher suspicion on the high non-linearity of the relationship between Y and X.This later choice serves two purposes simultaneously.First, it reduces the risk of inconsistent estimation[6]on the parametric part caused by badly specified $g(.)$, which has been beneficial to testing the relative income hypothesis in Health Economics(A.Jones and J.Wildman 2008).Additionally, relaxing the assumption on $g(.)$ permits to learn more about the relationship between Y and X such as in the field of social learning where the nature of the peer effect can be better uncovered(G.Bobonis and F. Finan 2005).In this instance,P$[\varepsilon < 0|X, Z]$=q a.s. is a judiciously chosen assumption on the unobservable component to identify $(g(.), \beta_0)$ among many others constant conditional location restrictions of the form $E[\Psi(\varepsilon - \alpha)|X, Z] = 0$ a.s. for some constant α and function $\Psi(.)$ satisfying $\Psi(.)(.) \geq 0$(Powell 1994).

In a seminal 2003 paper S.Lee showed that the "Average Quantile Derivative" (AQR) estimator can, under iid sampling, estimate β_0 consistently and efficiently. We are thus to remind the reader briefly about the AQR estimator.For a positive integer k we note A_k=$\{u \in \mathbb{N}^d : \sum u_i \leq k\}$ and N_k its cardinality.Also,for any $v \in \mathbb{R}^d$ and $u \in A_k$ we use the condensed notation v^u for $\prod_{i=1}^{d} v_i^{u_i}$.Let assume that $g(.)$ is m times continuously differentiable with its m^{th} derivative also hölder continuous of exponent $\gamma \in (0, 1]$ where $s = m + \gamma$ meets $s > 3d/2$.Given an iid sequence of observations $\{Y_i, X_i, Z_i\}_{i=1}^{n}$, the efficient AQR estimator is obtained in two steps.In the first stage, this consists of minimizing in c $\in \mathbb{R}^{N_k}$ and β:

$$\sum_{i \in \mathrm{I}_{j,n}} \rho_q(Y_i - P_n(c, X_i, X_j) - Z_i'\beta) \text{ for } j = 1, \ldots, n$$

where$P_n(c, t, X_j) = \sum_{u \in A_k} c_u \delta_n^{-u}(t - X_j)^u$ is a modified version of the Taylor's expansion of $g(.)$at some order k around X_j, $I_{j,n} = \{i \neq j : |X_i - X_j| < \delta_n\}$, $\delta_n = O(n^{-\alpha})$ for some

[6]We employ the term reduce because the estimator is still inconsistent if the error term is endogenous.

$\alpha \in (1/2s, 1/3d)$ and $\rho_q(t) = (2q-1)t + |t|$ is the "check function" (Koenker and Basset 1978).This first stage provides a n-sequence $\{\hat{\beta}_j\}_{j=1}^n$, all of which converging in probability to β_0 at a non parametric rate(Chaudhuri 1991).Hence, the second stage consists of combining this sequence using a judicious weighting system to reach efficiency.Under some mild conditions,β_α,the efficient AQR under homoscedasticity is given by:[7]

$$\beta_\alpha \equiv [\tfrac{1}{n}\sum_{j=1}^n \hat{\Omega}(X_j)]^{-1}(\tfrac{1}{n}\sum_{j=1}^n \hat{\Omega}(X_j)\hat{\beta}_j)$$

where $\hat{\Omega}(x)$ is a non parametric estimator of $Var(Z|X=x)$.

As showed in Lee 2003, β_α has desirable statistical properties in that $\sqrt{n}(\beta_\alpha - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{H})$ where $\mathcal{H}$ is the efficiency bound, for regular estimators of β_0 under the condition that the model is homoscedastic.

In the next section, we offer a root n consistent estimator circumventing the AQR first stage while retaining efficiency.The main conditions we introduce deal with the stochastic relationship between X and the error term and the nature of the random variable X. To be more precise, we impose (i) E[$\varepsilon|X$]=E[ε] a.s. and (ii)X contains either discrete or continuous bounded variables.Even though assuming statistical independence between X and the error term would suffice for (i) it is generally too strong a condition with economic data.Hence, (i) requiring at least that ε and X be uncorrelated,relaxes the stringency on the degree of stochastic proximity.Finally, it is important to bear in mind that unlike the discrete case ,the bounded support imposes by (ii)does facilitate the derivation of our results but is not necessary when X is a continuously distributed random vector.[8]

[7]We removed the X measurable trimming function used in the computation of the AQR,which filters the sequence of estimators making up the AQR depending on their non parametric part origin,so strictly speaking the efficiency bound is only "almost" attained because it aims at offering satisfactory finite sample properties.However,as explained in Lee 2003 this has no practical implication asymptotically.

[8]The intermediate case where X contains a mixture of discrete and continuous variables is possible.The structure of the proofs being almost identical apart for the more tedious notations we decided not to cover it.

4 The Estimation Strategy

Under (i) we can use the operator $E[.|X]$ (Robinson 1988) on both sides of (I) which yields a new equation and subtracting this latter from (I) results in:

$$T = w'\theta_0 + \varepsilon \text{ (II)}$$

where $\theta_0' = (-E[\varepsilon], \beta_0')$,$T = Y - E[Y|X]$ and $w' = [1, (Z - E[Z|X])']$.

We notice that part of the efficient score (Lee 2003)for a semi parametric quantile model emerges in (II),which suggests estimating efficiently θ_0 by minimizing $\sum_{i=1}^{n} \rho_q(\hat{T}_i - \hat{w}_i\prime\theta)$ where the hat stands for non parametric estimates of the nuisance functions. The main issue to overcome pertains to the inevitable first stage estimation of $\{T_i, w_i\}_{i=1}^{n}$, which are known up to X measurable nuisance functions we note τ.Let $\hat{\tau}$ be some non parametric estimator such that $plim\ \mathfrak{d}(\hat{\tau}, \tau)$=0 for a pseudo metric $\mathfrak{d}$ defined on some infinite dimensional functional space[9]containing our functions of interest(Andrews 94)[10].In general, showing that the asymptotic will be preserved using preliminary nonparametric estimates demands assumptions.Robinson 1988 succeeded in the context of a semi parametric model for conditional mean, assuming a particular smoothness for $g(.)$ and statistical independence between the error term and (X,Z). Subsequently, Andrews 1994 offered a general sufficient condition with the concept of Stochastic Equicontinuity, which holds under some regularity conditions [11].

Yet, the score for a quantile regression is not differentiable.This prevents using Stochastic Equicontinuity as an argument relying on the standard asymptotic theory with the Taylor representation of the score.One solution is provided in the seminal work of Chen et Al

[9]The term "pseudo" refers to the fact that $\mathfrak{d}(f_1, f_2) = 0 \Leftrightarrow f_1 = f_2$ almost everywhere. For instance,$\mathfrak{d}$ may be induced from a norm $N(f) = (\int_{\mathfrak{X}} |f|^r d\mu)^{1/r}$ where r is a positive integer and $(\mathfrak{X}, \mathfrak{B}, \mu)$ some measure space because $\mathfrak{d}(f_1, f_2) = N(f_1 - f_2)$ satisfies this condition.

[10]When the support of X is countably finite,the functional space is finite dimensional. However, our results extend to the case where X is continuous so we adopt a general treatment of the problem.

[11]See Andrews 94,Handbook of Econometrics,Volume4.

2003 which,under some regularity conditions, would permit us to derive the asymptotic of the unsmooth feasible estimator relying on the empirical process since this later is path wise differentiable.Even though this last approach could be employed,we rely instead on the smoothing of the objective function because we believe this approach allows for simpler proofs for our specific problem using classic non parametric results for Kernel density estimation.Additionnaly, our approach does not impose to find a pseudo metric satisfying $\mathfrak{d}(\hat{\tau}, \tau) = o_p(n^{-1/4})$(see Chen et Al 2003,Theorem 2-2.4) which is in general demanding where the dimension of X exceeds the smoothness order of the nuisance functions.

In this paper, we propose to estimate β_0 by smoothing the Check function(Amemiya 82, Horowitz 98) minimizing instead:

$$\sum_{i=1}^{n} \rho_n(\hat{T}_i - \hat{w}_i\prime\theta) \text{ (III)}$$

where $\{\rho_n\}_{n\in\mathbb{N}}$ is a sequence of twice differentiable real valued functions,converging uniformly to ρ_q.Those functions are build from integrating kernel functions as to approximate the absolute value function.The uniform rate of convergence to the check function i.e.sup $|\rho_n - \rho_q|$ will be given by the underlying bandwidth h of the Kernel employed.

The root n consistency and efficiency(under homoscedasticity) of the estimator of β_0 based upon (III), which we note $\beta(\hat{\tau})$, is derived using the following argument.First,using an appropriate smoothing scheme(Horowitz 1998)for the check function will establish that $\sqrt{n}(\beta(\tau) - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{H})$ where $\beta(\tau)$ corresponds to the estimator of (III) when the nuisance parameter is known. Then $\sqrt{n}(\beta(\tau) - \beta(\hat{\tau})) = o_p(1)$ will follow principally by letting h vanish as n approaches infinity at a sufficient slow rate, which is decided by the rate of convergence on the nuisance parameters.In other words, our feasible estimator from (III)is root n consistent while simultaneously efficiency in the class of regular estimators of β_0.The logic behind our admissible bandwidth spectrum is intuitive if one

thinks of h as inversely related to the smoothness of the score derived from (III): we need the smoothness of the score to deteriorate slowly enough as to let the estimation mistakes on the nuisance terms have no impact asymptotically.

As explained above, the choice of the bandwidth for smoothing the check function is critical: we must choose $h = O(1/n^p)$ for some $p \in (1/2r, c)$ where r corresponds to the(uniform)order of smoothness of the density of the error conditional on the explanatory variables and $0 < c < 1$ depends on the nature of X.When X contains discrete random variables $c = 1/4$ while a model where X is continuously distributed imposes $c = m/4(m+d)$ with $m > 1$ indicating the minimum order of smoothness between the density of X and the nuisance functions.

It is important to stress that the uniform rate of convergence on the nuisance terms plays a pivotal role in deciding the smoothness required on the conditional density for our estimation to be successful. When X comprises discrete random variables,the uniform rate of convergence in probability on the nuisance terms is parametric i.e.$\sqrt{n}$ imposing $r > 2$ for the density.In the instance where X contains continuous random variables,the (optimal)uniform rate on the nuisance functions is $n^{m/2m+2d}$ dictating $r > 2(m+d)/m$.

We thus observe two important distinctive features when X contains continuous regressors. First, the existence of a trade-off between the smoothness assumption of the nuisance functions and the error density. Secondly,the presence of a "linear curse of dimensionality in the smoothness" in that the minimal degree of smoothness on the density of the error is increasing in the number of explanatory variables entering $g(.)$[12]. In this paper we opted for $m = 2$ for we wish to be conservative on the class of nuisance functions and we believe the cost in terms of r to be very reasonable owing to the small dimensionality of X frequently encountered in semi parametric applications.

[12]Interestingly, this dimensionality problem attenuates as m becomes large so that the choice for h becomes identical to the discrete case when the density of X and the nuisance functions are infinitely smooth.

Before providing the Model and its full conditions we need to introduce some notations used throughout the paper:

(1)For $r > 0$ and $z \in \mathbb{R}$ we note $B(z,r) = \{x \in \mathbb{R} || x - z| < r\}$

(2) $1_A(x) = 1$, if $x \in A$,where A is some real Borel set.

(3) (Ω, σ, P) refers to a probability space where Ω is the space of states of nature,σ is the sigma field of measurable events and P indicates the probability measure.

(4) $\mathfrak{B}$= space of real valued Borel measurable functions

(5) For any real valued random variable X and positive integer k we note:

$Ł_X^k = \{f(X), f \in \mathfrak{B}, \int_\Omega |f(X)|^k dP < \infty\}$, $Ł_X^\infty$ the space of X measurable random variables bounded almost surely and $\hat{E}(f(X))$ the plug in estimator of $E(f(X))$.

(6) For f:$\mathbb{R}^d \longrightarrow \mathbb{R}$ we note $f^{(j)}(X)$ its j^{th} derivative at X whenever $\frac{\partial^{|j|} f(X)}{\partial x_1^{u_1} ... \partial x_d^{u_d}}$ exists for all $u \in \mathbb{N}^d$ such that$\sum u_i = j$.

(7) we note $||X||$ the Euclidean Norm of a vector $X = (x_1, ..., x_d)$ and $|X|_\infty$=$Max_{i=1...d}|x_i|$ where $d \in \mathbb{N}$.

(8) we note $|||M||| = \sqrt{trM'M}$ where M is a finite dimensional real valued Matrix and M' its transpose.

(9) we note $X_n \rightsquigarrow X$ for X_n converging in distribution to X.

(10) For a joint couple of real valued random variables (A, B) we use $f_b(a)$ as the Lebesgue density of A conditional on $B = b$.

(11) we use $||L||$ for the norm of a linear operator L whenever the context precludes confusion with the Euclidean norm.

(12)for$(d, m) \in \mathbb{N}*^2$we note:

$\Phi_{d,m}$=$\{f : \mathbb{R}^d \longrightarrow \mathbb{R}$,$f \in \mathfrak{B}$; $||f||_{sup} < \infty$(i)$\int f(X)dX = 1$(ii) $\int X^u f(X)dX = 0$ for $[u] = 1, ..., m-1$(iii)$\int |X^u f(X)|dX < \infty$ for $[u] = 0, m\}$.

with the standard notations X^u=$\prod_{i=1}^{d} x_i^{u_i}$ for $u \in \mathbb{N}^d$ and $[u] = \sum u_i$.

(13) For $m \in \mathbb{N}$ and $\mathcal{X} \subset \mathbb{R}^d$ open convex, we note $\mathcal{C}^m(\mathcal{X})$=$\{f : \mathcal{X} \longrightarrow \mathbb{R}, f^{(j)}$ exists and is continuous for $j = 0, ..., m$ uniformly over $\mathcal{X}$ $\}$.

(14)we use $||f||_{\infty,\mathcal{X}*}$ for the essential supremum of $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ on $\mathcal{X}* \subset \mathbb{R}^n$ where $n \in \mathbb{N}*$.

For any even integer r greater than 4 we note:

$\mathfrak{K}_\mathfrak{r}$=$\{K : \mathbb{R} \longrightarrow \mathbb{R}$, $K(t) = Q(t)1_{[-1,1]}(t)$ where Q is a symmetric polynomial of degree r satisfying (i)Q has (r-2)/2 distinct roots in (0,1)(ii)$Q(1) = 0; Q(0) > 0$ $Q^{(1)}(1) = Q^{(1)}(-1) = 0(iii)\int K(t)dt = 1$ and $\int t^j K(t)dt = 0$ for $j = 1, ..., r-1$ $\}$.

$\mathfrak{F}_\mathfrak{r}$=$\{\varphi : \mathbb{R} \longrightarrow \mathbb{R}, \varphi(u) = \int_u^\infty \int_x^\infty K(t)dtdx$, $K \in \mathfrak{K}_\mathfrak{r}\}$.

Finally, for $\alpha > 0$ and $s > 1$ we note $\mathfrak{H}_{\mathfrak{s},\alpha} = \{f : \mathbb{R} \longrightarrow \mathbb{R}$(i)f is r times continously differentiable where $r-1 < s \leq r$ (ii) $\int |f^{(r)}(t)|dt < \infty\}(iii)sup_{y \in B(x,\varrho)} \frac{|f(y)-T_{r-1}(y-x)|}{|y-x|^s} \leq \psi(x)$ for all x and some $\varrho > 0$ where $T_{r-1}(y-x)$ is the Polynomial in the Taylor's expansion at order r-1 of f(y) around x (iv)$\int f^\alpha(x)dx < \infty$ and $\int \psi^\alpha(x)dx < \infty$ $\}$.

5 The Model

$$Y = g(X) + Z'\beta_0 + \varepsilon$$

Assumption 1:

P[$\varepsilon < 0|X, Z$]=q a.s.

Assumption 2:

(a) E$|\varepsilon|^2 < \infty$ and (b) E[$\varepsilon|X$]=E[ε] a.s.

Assumption 3:

The support of Z, noted $\mathcal{Z}$, is a compact subset of $\mathbb{R}^K$ where K ≥ 1.

Assumption 4:

$\theta_0' = (-E[\varepsilon], \beta_0')$ is an interior point of Θ, which is a compact subset of $\mathbb{R}^{K+1}$.

Assumption 5:

The support of X, noted $\mathcal{X}$, is a countable subset of $\mathbb{R}^d$ where $d \geq 1$ satisfying (i) $inf_{\mathcal{X}} P[X = x] > 0(ii)\sum_{x\in\mathcal{X}} P[X = x] = 1$(iii) $inf_{t\in\mathcal{X}\setminus x}|t - x| > 0$ uniformly over $\mathcal{X}$.

Assumption 6:

g is a Borel measurable real valued function satisfying $sup_{x\in\mathcal{X}}|g(x)| < \infty$

Assumption 7:

For almost all (x,z) $\in \mathcal{X} \times \mathcal{Z}$ there exists $r(x, z) > 0$ such that $f_{x,z}(e) > 0$ on B(0,r(x,z))

where $f_{x,z}(.)$ is the density of ε conditional on X=x and Z=z.

Assumption 8:

$E[ww']$ is positive definite where $w' = [1, (Z - E[Z|X])']$.

Assumption 9:

$f_{x,z}(.)$ is in $\mathfrak{H}_{\mathfrak{r},\mathfrak{1}}$ almost everywhere on $\mathcal{X} \times \mathcal{Z}$.

Assumption 10:

The probability distribution measure of ε is absolutely continuous with respect to the Lebesgue measure.

Comments:

The assumptions follow mostly the literature for linear quantile regression(Koenker and Basset 1978, Amemiya 1982) because our transformed model is in effect linear.We will subsequently elaborate on the assumptions to highlight their relevance in obtaining the results.Nevetheless,it is worth discussing assumptions 2a,4, 5,6 and 9 at this point.Assumption 2a is stronger than usually required where the existence of the first moment of the error suffices. This extra condition originates from the presence of nuisance terms whose root n convergence holds provided the central limit theorem applies.Assumption 4 is introduced for simplicity but our results remain valid when Θ is simply assumed totally bounded(Andrews 1992),which permits models where strict inequality constraints are imposed on the parameters.Assumption 5,directly taken from Bierens 87,is the definition of a well behaved discrete random variable with (iii) excluding degenerated cases. Assumption 6 is technical but permits along with assumption 2a the convergence of our nuisance terms at the parametric rate(i.e $\sqrt{n}$).Finally, assumption 9 is a stronger requirement on the conditional density than proposed in the semi parametric quantile literature(Lee 2003,Chen and Kahn 2001).This type of smoothness requirement on the density is common in the literature of semi parametric estimation based upon prior nuisance terms(Robinson 1988, Florens et al 2006).That is, just like a classic non parametric density estimation,using a Kernel of order $r > 2$ demands the density of interest to be r times differentiable. Yet, in this paper we also assume the r^{th} derivative of the conditional

density is locally Lipschitz.[13].This last modification plays a major role in eliminating the asymptotic bias on the limiting distribution of the smooth quantile estimator.In the reminding part of the paper we will remove the q subscript for the check function being well understood that the quantile of interest has been chosen i.e. $\rho(.) = (2q-1)(.) + |.|$

6 Results

Proposition 1 (identification)

Under assumptions 1 through 8

θ_0 *is the global minimum of* $\mathrm{E}[\rho(T - w'\theta)]$ *on* Θ *where* $T = Y - E[Y|X]$

Comments:

Assumptions 8 and 7 are the most crucial for our parameter of interest to be identified.Assumption 8 requires that Z cannot be perfectly predicted via its minimum MSE predictor on $Ł^2_X$. Thus, this last condition discards models where Z contains a constant or X measurable functions(power of X for instance)[14].This last condition appeared identically in Robinson 1988. Finally, assumption 7 is a classic condition for quantile regression, relevant for unlike mean regression one does not have a globally convex population moment function which prevents the first order condition to suffice.In an Econometrics Model,this condition may be interpreted in terms of the purity of the unobservable component, which must have some strictly positive probability of getting arbitrary small in absolute term.Using this assumption permits to guarantee that the q^{th} quantile of the error conditional on $X = x$ and $Z = z$ is unique which, combined to assumption

[13]This type of condition is useful for dealing with the integrated bias of a kernel density estimator for a random variable whose support is not compact and can be loosely interpreted as a stability condition on the $L1$ norm of the r^{th} derivative to small perturbation

[14]This is not an issue in our model since assumption 3 and 5 together exclude this case.

8, translates into θ_0 being the sole local minimum of our population moment and consequently the global minimum. It is worth mentioning that the empirical counterpart of $\mathrm{E}[\rho(T - w'\theta)]$is not the minimization of interest but the consistency of our smooth estimator(feasible or not) originates from proposition 1.

Proposition 2

let $\{T_i, w_i'\}_{i=1\ldots n}$ *be an iid sequence from a joint couple* $\{T, w'\}$ *defined on* (Ω, σ, P).*For any* $q \in (0,1)$ *let* ρ_n *be a real valued function such that* $\rho_n(u) = 2(qu + \varphi_n(u))$ *where* $\varphi_n(u) = h\varphi(u/h)$ *for some* $\varphi \in \mathfrak{F}_{\mathfrak{r}}$ *and some* $h = O(1/n^p)$ *with* $p \in (1/2r, 1/4)$.*Then under assumption* 1 *through* 10 *the followings hold:*

(i) $\theta_* \equiv \mathbf{Argmin}_{\boldsymbol{\Theta}} \sum_{i=1}^n \rho_n(T_i - w_i'\theta)$ *is consistent for* θ_0.

(ii)$\sqrt{n}(\theta_* - \theta_0) \rightsquigarrow \mathcal{N}(0, q(1-q)E(f_{x,z}(0)ww')^{-1}E(ww')E(f_{x,z}(0)ww')^{-1})$.

Comments:

Consequentially, under homoscedasticity this smooth estimator reaches the efficiency bound for the linear part, which is $\frac{q(1-q)}{f(0)^2}E[V(Z|X)]^{-1}$(Lee 2003).The idea of smoothing M-estimators is not new (Huber 1964)but in the context of a linear quantile regression this consists of mimicking the empirical counterpart for the gradient and Hessian of the population function i.e. $\mathrm{E}[\rho(T - w'\theta)]$,which permits a more rapid derivation of the smooth estimator's asymptotic because it avoids having to work from a Taylor's representation of the empirical process (Koenker and Basset 1978).It is important to stress than even though our bandwidth constraint precludes the root n equivalency between this smooth quantile estimator and the minimizer of $\sum_{i=1}^n \rho(T_i - w_i'\theta)$, both estimators'asymptotic are identical.

The smoothing technique employed in the context of the 2SLAD(Two Stage Least Absolute Deviation)(Amemiya 1982)is simple and analytically tractable since build from

the logistic kernel which is of order 2. Unfortunately, in the context of estimation with nuisance functions one needs a kernel of higher order $r > 2$ capable of handling bandwidth h such that h^r=$o(1/\sqrt{n})$ to obtain a smooth estimator asymptotically Gaussien and simultaneously $h^{-4} = o(n)$ for the nuisance terms to have no impact. Thus, we rely on a variant of Horowitz's uniform kernel approach(Horowitz 1998) as employed in the context of Bootstrapping. The integration of such kernels of order r is easy to compute and yield polynomials of degree $r + 2$ on a compact support which after tuning with a bandwidth approximate the "check function". A good example when $r = 4$ (i.e. for constructing a function in $\mathfrak{F}_4$) would be the Epanechnikov Kernel given by $K(t) = 15/32(7t^4-10t^2+3)1_{(|t|\leq 1)}$ resulting in $\varphi(u) = 15/32(7/30u^6-5/6u^4+3/2u^2-16/15u+1/6)1_{(|u|\leq 1)} - u1_{(u<-1)}$.

Our next step will be to use a modified version of our smooth estimator using non parametric estimates for $M(X) = E[Y|X]$ and $\vartheta(X)$=$E[Z|X]$,which we propose to estimate(pointwise) using the following estimators:

$$\hat{M}(x) = \sum_{i=1}^{n} k_i(x)Y_i \text{ and } \hat{\vartheta}(x)=\sum_{i=1}^{n} k_i(x)Z_i$$

where $k_i(x) = \phi((X_i - x)/\zeta)/\sum_{j=1}^{n} \phi((X_j - x)/\zeta)$ with ϕ a symmetric Kernel while ζ is a sequence of bandwidth. We will briefly remind the reader about the conditions upon ϕ leading to our unusual uniform rate of convergence for $\hat{M}$ and $\hat{\vartheta}$ by stating the conditions directly taken from Bierens 87.

Proposition 3(Bierens 1987)

let $\{Y_i, X_i, Z_i\}_{i=1...n}$ be an iid sequence from (Y, X, Z),a triplet defined on (Ω, σ, P) where X meets assumption 5. Let ϕ :$\mathbb{R}^d \longrightarrow \mathbb{R}$ be a real valued function satisfying

(i) ϕ is symmetric

(ii) $\phi(0) = 1$

(iii) $\limsup \sqrt{n}|\phi(t)1_{|t|>1/\zeta}| = 0$ for $\zeta = O(n^{-\alpha})$ where $\alpha > 0$.

For all $x \in \mathcal{X}$ *let define* $\hat{M}(x) = \sum_{i=1}^{n} k_i(x)Y_i$ *and* $\hat{\vartheta}(x)=\sum_{i=1}^{n} k_i(x)Z_i$ *where* $k_i(x) = \phi((X_i - x)/\zeta)/\sum_{j=1}^{n} \phi((X_j - x)/\zeta)$.

Then $sup_{x\in\mathcal{X}}\{\sqrt{n}(\hat{M} - M)\}$=$O_p(1)$ and$sup_{x\in\mathcal{X}}\{\sqrt{n}||\hat{\vartheta} - \vartheta||\}$=$O_p(1)$.

Comments:

This results originates from the fact that the kernel based estimators converges in probability to the empirical counterpart of the conditional mean which is root n consistent at any point of the conditioning. The condition $\phi(0) = 1$,not typically met by Kernel functions, is at core origin of this convergence success. The intuition is that realizations of X happening to "hit" the very point of the chosen conditioning $x \in X$ must ensure $\phi((X - x)/\zeta)$=1 for mimicking the empirical estimator in question. Finally, $\limsup \sqrt{n}|\phi(t)1_{|t|>1/\zeta}| = 0$ for ζ=$o(1)$ is met by Kernels belonging to the exponential family.In practice,ϕ (.) can be constructed in a simple manner as shown in Bierens 1987 as a linear combination of two normal Kernels:

$$\phi = \alpha_1\phi_1 + \alpha_2\phi_2$$

where $\phi_j(x)$=$\sigma_j^{-d}(2\pi)^{-d/2}e^{-x'x/2\sigma_j^2}$ for $j = 1, 2$

$\alpha_j = (\sigma_i^{-d} - (2\pi)^{d/2})\frac{\sigma_j^d\sigma_i^d}{\sigma_j^d - \sigma_i^d}$ for $j \neq i$

while $\{\sigma_j\}_{j=1,2}$ are arbitrary chosen strictly positive real numbers.

Proposition 4

Under assumption 1 through 10

$$\tilde{\theta}_* \equiv \mathbf{Argmin}_{\mathbf{\Theta}} \sum_{i=1}^{n} \rho_n(\hat{T}_i - \hat{w}_i\prime\theta) \text{ is consistent for } \theta_0$$

Assumption 11

For any $\delta > 0$ there exists $\xi > 0$ and N_0 such that $sup_{n\geq N_0}P[sup_{\Theta_n(\xi)}|\upsilon_n(\Delta)| > \delta] < \delta$ where $\upsilon_n(\Delta) = \sum \frac{1}{\sqrt{n}}\frac{w_i}{h}K(\varepsilon_i/h)\Delta_i$ and $\Theta_n(t) = \{\{\Delta\}_{i=1..n} : |\Delta|_\infty \leq t\}$ for any $t > 0$.

Comments:

This assumption ensures $\upsilon_n(\hat{\Delta}) = o_p(1)$,which we found to be sufficient to show that the feasible estimator is asymptotically equivalent. The structure of the condition is inspired from the notion of stochastic equicontinuity (SEC)(Andrews 94).The viability of this assumption may be judged by observing that under our previous assumptions $|\upsilon_n(\hat{\Delta})| \leq \sqrt{n}|\hat{\Delta}|_\infty$ and is thus bounded in probability[15].

Proposition 5

Under assumption 1 through 11

$$plim|\sqrt{n}(\tilde{\theta}_* - \theta_0) - \sqrt{n}(\theta_* - \theta_0)| = 0$$

Comments:

Proposition 5 establishes therefore that our feasible estimator reaches the efficiency bound under homoscedasticity.There are two practical concerns. First, the estimator will be computed minimizing the non linear function $\sum_{i=1}^{n} \rho_n(\hat{T}_i - \hat{w}_i\prime\theta)$ using an iterated procedure(i.e.Newton's and its variants)or a direct search method such as simulated annealing(Kirkpatrick et Al 1983).Secondly, to conduct inferences the covariance matrix needs consistent estimators of $H_0 = E[f_{x,z}(0)ww']$ and $M_0 = E[ww']$ which are given respectively by $\hat{H}_0 = \frac{1}{nh}\sum \hat{w}_i\hat{w}_i' K(\frac{\hat{T}_i - \hat{w}_i'\tilde{\theta}_*}{h})$ and $\hat{M}_0 = \frac{1}{n}\sum \hat{w}_i\hat{w}_i'$.[16]Finally a point wise estimator of g is given by $\hat{g} = \hat{M} + \hat{u} - \hat{\vartheta}'\hat{\beta}_0$ where $\hat{u}$ is the estimator of the intercept in $\tilde{\theta}_*$ while $\hat{\beta}_0$ its reminding sub vector. Then $\sqrt{n}(\hat{g} - g) = O_p(1)$ follows immediately from propositions 3 and 5.Next we are to provide the conditions for our results to hold in the case where X contains continuously distributed random variables.

[15]This is no longer true when X has a compact support.

[16]See Lemma 3 for proof of $\hat{H}_0$ and $\hat{M}_0$ consistency.

Corollary

Let the previous assumptions of our model hold except assumptions 2a,5,9. Also, let the followings hold:

(a)X*is a* $\mathcal{X}$ *valued random variable where* $\mathcal{X} \subset \mathbb{R}^d$ *open convex bounded.*

(b)$\mathcal{X}* \subset \mathcal{X}$ *compact non empty such that* $\{x \in \mathcal{X}*|Var(Z|X = x)$*positive definite*$\}$ *has a strictly positive Lebesgue measure.*

(c) *The distribution function of* X *is absolutely continuous with respect to the Lebesgue measure and the density of* X*,noted* π*,is strictly positive on* $\mathcal{X}$*.*

(d)π*,g, and* ϑ *belong to* $\mathcal{C}^2(\mathcal{X})$*.*

(e)$E|\varepsilon|^{2+a} < \infty$ for some $a > 0$.

(f) *There exists constants* C_1*,* C_2 *and* C_3 *such that:*

$$|||E(ZZ'|X = x_1) - E(ZZ'|X = x_2)||| \leq C_1||x_1 - x_2|| \text{ for all } (x_1, x_2) \in \mathcal{X} \times \mathcal{X}.$$

$$|E(\varepsilon^2|X = x_1) - E(\varepsilon^2|X = x_2)| \leq C_2||x_1 - x_2|| \text{ for all } (x_1, x_2) \in \mathcal{X} \times \mathcal{X}.$$

$$||E(Z\varepsilon|X = x_1) - E(Z\varepsilon|X = x_2)|| \leq C_3||x_1 - x_2|| \text{for all } (x_1, x_2) \in \mathcal{X} \times \mathcal{X}.$$

(g)$f_{x,z}$ *belongs to* $\mathfrak{H}_{\mathfrak{r},\mathfrak{1}}$ *for almost all* $(x, z) \in \mathcal{X} \times \mathcal{Z}$ *with* $r > 2 + d$*.*

The nuisance functions are estimated pointwise with:

$\hat{M}(x)$=$\sum_{i=1}^n$ $k_i(x)Y_i$ and $\hat{\vartheta}(x)$=$\sum_{i=1}^n$ $k_i(x)Z_i$

where $k_i(x) = \phi((X_i - x)/\zeta)/\sum_{j=1}^n \phi((X_j - x)/\zeta)$

(h)$\phi \in \Phi_{d,2}$ *and* $\zeta = O(n^{-1/4+2d})$*.*

(i)$\int |\int e^{it'X}\phi(X)dX|dt < \infty$ *where* $i = \sqrt{-1}$*.*

(j)ρ_n *from proposition 2 is such that* $h = O(1/n^p)$ *with* $p \in (1/2r, 1/4 + 2d)$*.*

(k) *For any* $\varepsilon > 0$ *and* $\eta > 0$ *there exists* $\delta > 0$ *such that* $\overline{lim} P^*[sup_{\mathcal{B}(\tau_0,\delta)} \sqrt{n} || \nabla S_*(\tau, \theta_0) - \nabla S_*(\tau_0, \theta_0)|| > \eta] < \varepsilon$ *where* $\nabla S_*(\tau_0, \theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\theta} \rho_n(T_i - w_i'\theta) \mid_{\theta=\theta_0}$ *is the true empirical gradient while* $\nabla S_*(\tau, \theta_0)$ *is that using some other nuisance functions* $\tau \in \mathcal{F}$ *with* $\mathcal{F} = \{(f,g) : ||f||_{\infty,\mathcal{X}*} + ||g||_{\infty,\mathcal{X}*} < \infty\}$,$\mathcal{B}(\tau_0, \delta) = \{\tau \in \mathcal{F} | \urcorner_{\mathcal{F}}(\tau, \tau_0) < \delta\}$ *and* $\urcorner_{\mathcal{F}}(\tau_1, \tau_2) = ||f_1 - f_2||_{\infty,\mathcal{X}*} + ||g_1 - g_2||_{\infty,\mathcal{X}*}$ *for any* $(\tau_1, \tau_2) = \{(f_1, g_1), (f_2, g_2)\} \in \mathcal{F} \otimes \mathcal{F}$.

Then $\tilde{\theta}_* \equiv \mathbf{Argmin}_{\mathbf{\Theta}} \sum_{i=1}^n \lambda(X_i) \rho_n(\hat{T}_i - \hat{w}_i\prime\theta)$,*where* $\lambda(X) = 1_{X \in \mathcal{X}*}$, *satisfies the followings:*

(I) $\tilde{\theta}_*$ *is consistent for* θ_0.

(II) $\sqrt{n}(\tilde{\theta}_* - \theta_0) \rightsquigarrow \mathcal{N}(0, q(1-q) E(\lambda f_{x,z}(0) ww')^{-1} E(\lambda ww') E(\lambda f_{x,z}(0) ww')^{-1})$.

Comments:

Hence,the semi parametric efficiency bound will be attained under homoscedasticity apart for the presence of the trimming function.This "almost" efficiency is also a characteristics of the efficient AQR when the unobservable term is homoscedastic.It is interesting to notice that while the AQR trimming function has a practical origin(Lee 2003, page7),our trimming criteria is introduced for theoretical reason which are to be explained shortly. In practice one can render the trimming effect inconsequential in large samples by gauging the support of X.Assumption(a)is standard for continuously distributed random variables entering nuisance functions, ensuring a support of "minimal smoothness"(Andrews 1994).The extra condition we impose is that the support is also bounded,which simplifies many of the proofs but needs not to hold.Assumption (b)ensures that $E[\lambda ww']$ is positive definite which plays the same role as assumption 8 in the context of our trimmed estimator.This condition is weaker than $V(Z|X = x)$ positive definite a.e. because it allows Z to be perfectly predicted by X on some strict subsets of $\mathcal{X}*$,which may be relevant in applications when (Z, X) share a perfect relationship around some

level of X [17].Assumptionc(c),(d) and (e) ensures the classic conditions to obtain a uniform rate of convergence in probability on the nuisance functions over compact subsets of $\mathcal{X}$.Notice that Assumption (d) is conservative on the nuisance functions which comes at a cost in terms of the smoothness required on the conditional density in (g) where $r > 2 + d$ is assumed.However,we feel this later condition on the density of the error to be mild as the dimension of the variable entering the non parametric part is small in most economic applications.It is interesting compare the smoothness tradeoff between the AQR and our suggested estimation procedure.Unlike the AQR,it is not the smoothness of $g(.)$ that must grow with the dimension of X but that of the conditional density of the error term.The bandwidth in (h) for the Kernel employed to estimate the nuisance functions is the optimal one under the smoothness conditions previously enumerated using Kernels of bounded variations(Silverman 1978,Bierens 1987).Assumption (i),required to obtain a uniform rate of convergence in probability on our nuisance functions,demands a Kernel whose Fourier transform is absolutely integrable,which will hold for instance when $\phi(x)=(2\pi)^{-d/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}x'\Sigma^{-1}x}$ for some positive definite matrix Σ.Finally,the trimming componentλ is introduced because the uniform rate of convergence on conditional mean functions is guaranteed only on compact subsets.This filtering of observations has thus been widely used in estimation based upon nuisance functions(Andews 1994,Robinson 1988).Even though trimming imposes a sacrifice in large sample in terms of the efficiency of our estimator,it offers more robust finite sample properties by discarding observations close to the cluster points of the support of X.Finally, assumption(k) imposes stochastic equicontinuity (Andrews 94) on the smooth score because assumption(d)is not strong enough to ensure the analogue of the discrete case to show directly $\sqrt{n}\{\nabla S_*(\hat{\tau},\theta_0) - \nabla S_*(\tau_0,\theta_0)\} = o_p(1)$. The *Caratheodory* measure P^* is introduced in

[17]For instance,in a simple wage equation using $X = age$ and $Z = schooling$ will have the variable age in a low range as a perfect predictor of schooling as compulsory enrollment prevent any variation to occur for Z.More generally, this type of lack of variation arises with Economic data when some ranges of the variables X are constrained by law to a unique choice for Z.

order to handle instances where $sup_{\mathcal{B}(\tau_0,\delta)}\sqrt{n}||\nabla S_*(\tau,\theta_0) - \nabla S_*(\tau_0,\theta_0)||$is not a σ measurable sequence of maps[18].Because of our choice for the pseudo metric,this condition can be interpretation as follows: for any $\tau \in \mathcal{F}$ the measure(outer)of the discrepancy i.e.$\sqrt{n}||\nabla S_*(\tau,\theta_0) - \nabla S_*(\tau_0,\theta_0)||$ exceeding an arbitrary level can be rendered arbitrary small provided that the worst absolute difference over $\mathcal{X}*$ between τ and τ_0 is kept under control.Even though (k) is a demanding assumption,it is important to keep in mind that this condition is not necessary to achieve $\sqrt{n}\{\nabla S_*(\hat{\tau},\theta_0) - \nabla S_*(\tau_0,\theta_0)\} = o_p(1)$.

The testing of hypothesis on the slope coefficient will be conducted in practice plugging consistent estimators of $H_0 = E(\lambda f_{x,z}(0)ww')$ and $M_0 = E(\lambda ww')$ which are given by $\hat{H}_0 = \frac{1}{nh}\sum \lambda_i \hat{w}_i\hat{w}_i' K(\frac{\hat{T}_i - \hat{w}_i'\tilde{\theta}_*}{h})$ and $\hat{M}_0 = \frac{1}{n}\sum \lambda_i \hat{w}_i\hat{w}_i'$.Similarly to the discrete case, $g(.)$ will be estimated pointwise as explained on page 22 but with a slower convergence rate imposed by that achieved on the nuisance functions i.e. $n^{\frac{1}{2+d}}(\hat{g} - g) = O_p(1)$.Furthermore, in applications the testing of a null hypothesis of the form $H_o : R\nabla g(x) = r$ where $\nabla g(x)$ is the gradient of g evaluated at some x $\in \mathcal{X}$,R is a d by d matric of rank L and r $\in \mathbb{R}$ may be an object of interest.Under mild regularity conditions provided in Pagan and Ullah 1999 one can use the fact that that $\sqrt{n\gamma^{d+2}}(\nabla\hat{g}(x) - \nabla g(x)) \equiv \sqrt{n\gamma^{d+2}}(\nabla\hat{E}[y - z'\beta_0 \mid X = x] - \nabla E[y - z'\beta_0 \mid X = x]) + o_p(1)$ to derive under the Null:

$$\sqrt{n\gamma^{d+2}}R\nabla\hat{g}(x) - r \rightsquigarrow \mathcal{N}(0, R\Xi(x)R')$$

for some $\Xi(x)$ which can estimated consistently non parametrically by $n\gamma^{d+2}\hat{\Xi}(x)$, thus providing a practical testing from $(R\nabla\hat{g}(x) - r)'(R\hat{\Xi}(x)R')^{-1}(R\nabla\hat{g}(x) - r) \rightsquigarrow \chi^2(L)$.

[18] P^* "measures" a non measurable event A by using measurable coverings of A i.e.$P^*(A) = inf\{\sum P(A_i)|A \subseteq \cup A_i, \{A_i\} \subseteq \sigma\}$.In our context,$P^*$ is useful to show that $E \subseteq B$ with E measurable and B non measurable still allows for $P(E) \leq P^*(B)$ because P^* coincides with P on σ while P^* is monotonic by construction.See Hopf's extension Theorem.

7 Bandwidth selection

We have not addressed so far the selection of the bandwidth for smoothing the check function.Our Monte Carlo experiments suggest that the size of the t-test is highly sensitive to the choice of the bandwidth.Even though selection procedures for mean squared error loss have been developed for some M-estimators based upon a smoothing of the density(Horowitz and Hall 1990)the body of research is scant when nuisance functions are present and limited for testing purposes(Gao and Gijbels 2008). Thus, we are to offer a simple rule of thumb based upon the fact that under our assumptions $\hat{H}_0 = HS_*(\theta_0,\tau)+o_p(1)$ where $HS_*(\theta_0,\tau) = \frac{1}{nh}\sum \lambda_i w_i w_i' K(\frac{T_i - w_i'\theta_0}{h})$ has an asymptotic mean squared error(componentwise) easy to establish.The following proposition offers an expression for this optimal bandwidth.

proposition 5 bis

Let $\mathfrak{L}$ *be the* $K+1$ *by* $K+1$ *matrix such that* $\mathfrak{L}_{ij} = E|HS_*(\theta_0,\tau)_{ij} - E(\lambda f_{x,z}(0) ww')_{ij}|^2$ *for* $(i,j) \in \{1,...,K+1\}^{\times 2}$. *Then under the assumptions of the corollary:*

$$\zeta_0^{\frac{1}{2r+1}} n^{-\frac{1}{2r+1}} = Argmin_h |||\mathfrak{L}|||^2 \text{ as } n \text{ approaches infinity}$$

where:

$\zeta_0 = \frac{b(1-2r)+\sqrt{(QB)}}{4ar}$

$a = trace(M_1^2); b = trace(M_1 M_2); c = trace(M_2^2)$

$M_1 = (\frac{\mu_r}{r!})^2 E[\lambda ww' f_{x,z}^{(r)}(0)]_{22}$

$M_2 = \int K^2(t)dt E[\lambda w_{22} w_{22}' f_{x,z}(0)]$

$\mu_r = \int t^r K(t)dt$

$f_{x,z}^{(r)}(0)$ *indicating the* $r^t h$ *derivative of the density of* ε *conditional on* x, z *evaluated at* 0 *and* $A_{22} = \{a_{ij}^2\}$ *for any matrix* $A = \{a_{ij}\}$.

Comments:

Our optimal rate is similar to that minimizing the mean squared integrated error of a Kernel density estimator. However, under assumption (j) of the corollary this optimal bandwidth is not attainable.Yet,this suggests using $h* = \zeta_0^{\frac{1}{2r+1}} n^{-p}$ for some p meeting assumption (j). In practice, M_1 and M_2 need some consistent estimators for the functions $f_{x,z}(0)$ and $f_{x,z}^{(r)}(0)$ which requires using the feasible version with our residuals to retrieve $\hat{f}_{x,z}(0)$and $\hat{f}_{x,z}^{(r)}(0)$ as explained in section 6.Hence, a natural way to proceed in order to estimate the proposed optimal bandwidth consists of using:

$$\hat{E}[\lambda w w' f_{x,z}^{(r)}(0)]_{22} = [\tfrac{1}{n} \textstyle\sum \lambda_i \hat{w}_i \hat{w}_i' \hat{f}_{x_i,z_i}^{(r)}(0)]_{22}$$

and

$$\hat{E}[\lambda w_{22} w_{22}' f_{x,z}(0)] = \tfrac{1}{n} \textstyle\sum \lambda_i \hat{w}_{22,i} \hat{w}_{22,i}' \hat{f}_{x_i,z_i}(0).$$

It is yet not clear whether this will provides consistent estimators(under the assumptions of the corollary)for h* when nuisance functions are present because the theory of asymptotic interchangeability between consistent residuals and error terms applies for root-n consistent residuals(Hall and Horowitz 1990) which does not hold under the continuous model exposed in the coralary.Thus, one may have to impose assumptions similar to those of section 6 (assumption $H4$).Finally, it is important to stress that this optimal criteria is merely suggestive because our approximation on the Hessian holds in probability only [19] and the asymptotic optimal choice may not be relevant in finite sample.However, we believe that this rule of thumb offers a starting point in applications for choosing a range of values for the bandwidth,which is useful should one adopts bootstrapping driven bandwidth selection(Horowitz 1998)or plug in methods(Hall,Sheater,Jones,Marron 1991).

[19] $\hat{H}_0 - HS_*(\theta_0, \tau) = o_p(1)$ is not sufficient to conclude that the asymptotic mean squared error of $\hat{H}_0$ will be equal to that of $HS_*(\theta_0, \tau)$ because the moments need not to converge unless strong uniform integrability assumptions are imposed i.e.$sup_n E|||\hat{H}_0 - HS_*(\theta_0, \tau)|||^s$ for some $s > 2$,see Chung page 100-101.

8 Generalization and discussion

Our paper has introduced an approach to semi parametric quantile regression for estimating efficiently β_0, which is generalizable to various stochastic relationships of the triplet (ε, X, Z). In this section the continuous case is treated. For the sake of clarity, it will be convenient to introduce the operator A from $L^1(\Omega)$ to $L^\infty_X(\Omega)$ satisfying $Ar = E[r|X]$ and T from $L^\infty_{X,Z}(\Omega)$ to $L^\infty_X(\Omega)$ satisfying $T\psi = E[\psi|X]$(Carrasco,Florens,Renault 2007) where $L^\infty_{X,Z}(\Omega)$=$\{\Psi(X,Z)$,Ψ $\mathbb{R}$ valued Borel: $||\Psi(X,Z)||_\infty < \infty$ $\}$ and $L^\infty_X(\Omega)$=$\{\Lambda(X)$,Λ :$\mathbb{R}$ valued Borel:$||\Lambda||_\infty < \infty$ $\}$.Given $Y = g(X) + Z'\beta_0 + \varepsilon$ as the model and using the linearity of A one can show (Newey and Powell 1990)that:

$$\beta_0 = Argmin_\beta E[f\rho(V - w'\beta)] \text{ (IV)}$$

with

$V = Y - \frac{A(f^2Y)}{A(f^2)} + \frac{A(f^2\varepsilon)}{A(f^2)}$

$w = Z - \frac{A(f^2Z)}{A(f^2)} = Z - \Gamma$

where $\Gamma = \frac{A(f^2Z)}{A(f^2)}$ and f indicates $f_{x,z}(0)$ [20].

The demeaning employed in our previous section is therefore transposable to the general case for(IV).Given an iid sequence of observations, one can easily derive that $\hat{\beta}_0 = Argmin_\beta \hat{E}[f\rho(V - w'\beta)]$ (where $\hat{E}$ denotes the empirical counterpart of (IV)) satisfies:

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{VB}) \text{ (V)}$$

where $\mathcal{VB} = q(1-q)E[f^2ww']^{-1}$ is the semi parametric efficiency bound(Lee 2003).This suggests that the parametric part can always be estimated efficiently via a smooth linear quantile regression adjusting for the presence of nuisance terms.

[20]Our previous model is a special case under what we called homoscedasticity which furnished $A(f^2Y) = A(Y)$ and $A(f^2) = f^2$ while assumption 2.*b* yielded $\frac{A(f^2\varepsilon)}{A(f^2)}$ as a constant.

Henceforth, we define $\tilde{\beta}_* = Argmin_\beta \hat{E}[\hat{f}\rho_n(\hat{V} - \hat{w}'\beta)]$,the smoothed version of $\hat{\beta}_0$, as the Adaptive Semi Parametric Quantile Estimator (ASPQ) where $\tau = (f, g, \Gamma, \beta_0)$ is the nuisance parameter, which must be estimated from a first stage. The consistency of $\tilde{\beta}_*$ can be derived from that of $\tilde{\theta}_*$ established in section 5 imposing uniform consistency conditions on $\hat{\tau} = (\hat{f}, \hat{g}, \hat{\Gamma}, \hat{\beta})$. In practice, both $\hat{g}$ and $\hat{\beta}$ can be conveniently estimated from the AQR first stage.Also,Γ may be estimated by:

$$\hat{\Gamma}(X) = \frac{\hat{T}(\hat{f}^2 Z)}{\hat{T}(\hat{f}^2)} \simeq \frac{\sum \kappa(\frac{X_i - X}{c_n})\hat{f}_{X_i,Z_i}(0)^2 Z_i}{\sum \kappa(\frac{X_i - X}{c_n})\hat{f}_{X_i,Z_i}(0)^2} \quad (V)$$

for some strictly positive $\kappa \in \Phi_{d,2}$ and $c_n = o(1/n)$ a sequence of bandwidth.

using

$$\hat{f}_{X,Z}(0) = h_{2,n}^{d+K} h_{1,n}^{-(d+K+1)} \frac{\sum \kappa_{exz}(\frac{e_i}{h_{1,n}}, \frac{X_i - X}{h_{1,n}}, \frac{Z_i - Z}{h_{1,n}})}{\sum \kappa_{xz}(\frac{X_i - X}{h_{2,n}}, \frac{Z_i - Z}{h_{2,n}})} \quad (VI)$$

where $\{e_i\}_{i=1...n}$ are the consistent residuals retrieved from $(\hat{g}, \hat{\beta})$, $(\kappa_{exz}, \kappa_{xz}) \in \Phi_{d+K+1,2} \otimes \Phi_{d+K,2}$ while $(\{h_{1,n}\}, \{h_{2,n}\})$ are two sequences of bandwidth meeting the same condition as c_n[21].

However,the analogy with the Homoscedastic case in terms of the efficiency requires more caution. This arises because the estimator of $T(\Psi)$, where Ψ are the relevant projected elements in Γ, relies on $\hat{T}(\hat{\Psi}) = \int \hat{\Psi}\hat{f}_x(z)dz$ where $\hat{f}_x(z)$ is the non parametric estimator of $f_x(z)$ while $\hat{\Psi}$ that of Ψ retrieved from consistent residuals i.e. $Y - \hat{g} - Z'\hat{\beta}$. Thus,even though $||\hat{T}(\Psi) - T(\Psi)||_\infty$ may converges in probability at an acceptable rate ,the same may no longer apply to $||\hat{T}(\hat{\Psi}) - T(\Psi)||_\infty$. We are to give next some generic conditions to ensure consistency and efficiency in this more general setting.

[21] These suggested feasible versions of non parametric estimators for the nuisance function are the same as proposed in Lee 2003 to compute the efficient one step estimator under Heteroscedasticity.It is a very natural way to proceed when no parametrization of f and Γ is assumed

Assumption H1:

(i)Assumptions 1,3,6,7,10 and (a),(c)of the Corollary hold.

(ii)β_0 is an interior point of $\mathbf{B} \subset \mathbb{R}^K$ compact.

Assumption H2:

(i)$\mathcal{X}* \subset \mathcal{X}$ compact non empty such that $\{x \in \mathcal{X}*|E[f^2(Z-\Gamma)(Z-\Gamma)'|X=x]$positive definite$\}$ has a strictly positive Lebesgue measure.

(ii)$E|\varepsilon| < \infty$

Assumption H3:

There exists $(\hat{g}, \hat{\beta})$ satisfying:

(i)$||\hat{g} - g||_\infty = O_p(n^{-\gamma})$ for some $\gamma > 0$

(ii)$\hat{\beta} - \beta_0 = O_p(n^{-1/2})$.

Assumption H4:

There exists $a > 0$ and $b > 0$ such that:

(i) $sup_{\mathcal{X}\times\mathcal{Z}}|\hat{f} - f| = O_p(\frac{1}{n^a})$

(ii)$||\hat{T} - T|| = O_p(\frac{1}{n^b})$

Assumption H5:

$f_{x,z}(.)$ belongs to $\mathfrak{H}_{\mathfrak{r},\mathfrak{1}}$ for almost all $(x,z) \in \mathcal{X} \times \mathcal{Z}$

with $r > \frac{1}{m}$ where $m = min\{a, b, \gamma\}$.

Assumption H6:

For any $\varepsilon > 0$ and $\eta > 0$ there exists $\delta > 0$ such that $\overline{lim}P^*[sup_{\mathcal{B}(\tau_0,\delta)}\sqrt{n}||\nabla S_*(\tau,\beta_0) - \nabla S_*(\tau_0,\beta_0)|| > \eta] < \varepsilon$ where $\nabla S_*(\tau_0,\beta_0) = \frac{\partial}{\partial\beta}\hat{E}[f\rho_n(Y - g - \Gamma'\beta_0 - (Z-\Gamma)'\beta]\ |_{\beta=\beta_0}$ and $\nabla S_*(\tau,\beta_0) = \frac{\partial}{\partial\bar{\beta}}\hat{E}[\bar{f}\rho_n(Y - \bar{g} - \bar{\Gamma}'\bar{\beta} - (Z-\bar{\Gamma})'\beta]\ |_{\beta=\beta_0}$ for any $\tau = (\bar{f}, \bar{g}, \bar{\Gamma}, \bar{\beta}) \in \mathcal{F}$ where $\mathcal{F} = \{(f,g,t,b) : f \in L^\infty_{\mathcal{X}*\otimes\mathcal{Z}}, g \in L^\infty_{\mathcal{X}*}, t \in \otimes_{k=1}^K L^\infty_{\mathcal{X}*}, b \in \mathbb{R}^K\}$,$\mathcal{B}(\tau_0,\delta) = \{\tau \in \mathcal{F}|\urcorner_{\mathcal{F}}(\tau,\tau_0) < \delta\}$ and $\urcorner_{\mathcal{F}}(\tau_1,\tau_2) = ||f_1 - f_2||_{\infty,\mathcal{X}*\otimes\mathcal{Z}} + ||g_1 - g_2||_{\infty,\mathcal{X}*} + sup_{\mathcal{X}*}||t_1 - t_2|| + ||b_1 - b_2||$ for any $(\tau_1,\tau_2) = \{(f_1,g_1,t_1,b_1),(f_2,g_2,t_2,b_2)\} \in \mathcal{F}\otimes\mathcal{F}$.

Assumption H7:

ρ_n from proposition 2 is such that $h = O(1/n^p)$ with $p \in (1/2r, m/2)$.

Assumption H8:

$\{Y_i, X_i, Z_i\}_{i=1...n}$ is an i.i.d. sequence from (Y, X, Z).

Proposition 6

Under assumption H1 through H8

$\tilde{\beta}_* \equiv Argmin_{\mathbf{B}} \sum_{i=1}^n \lambda_i \hat{f}_i \rho_n(Y_i - \hat{g}_i - \hat{\Gamma}_i'\hat{\beta} - (Z_i - \hat{\Gamma}_i)'\beta)$ *,with* $\lambda_i = 1_{X_i \in \mathcal{X}_*}$*, is consistent for* β_0 *and* $\sqrt{n}(\tilde{\beta}_* - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{VB}_\lambda)$ *where* $\mathcal{VB}_\lambda = q(1-q)E[\lambda f^2 (Z-\Gamma)(Z-\Gamma)']^{-1}$.

Comments:

Identically to the homoscedastic case, the efficiency bound is almost reached because of the trimming term, which can be eliminated if the support of X is assumed to be compact. However, in small sample,it is preferable to retain this filtering of observations as explained on page 25. Assumptions $H1$ and $H2$ permit identification of β_0. Assumptions $H3$ requires to find some prior estimator of $g(.)$ converging uniformly over $\mathcal{X}$ and a root-n estimator of β_0. This will be satisfied under the conditions of the AQR which are provided in Lee 2003 in which case $\gamma = 1/3$. Alternatively,there may be other estimators meeting $H3$ for a semi parametric model if the error term satisfies other scale location invariance restrictions (Robinson 1988,Powel 1994) or specific Heteroscedasticity(section 4).Assumptions $H4$, whose sufficient conditions are provided in Bierens 1983 and Horowitz Hall 1990,delivers consistency by ensuring a uniform rate of convergence in probability on the nuisance functions $\hat{\tau} = (\hat{f}, \hat{g}, \hat{\Gamma}, \hat{\beta})$.Conditions $H4(ii)$refers to$||\hat{T} - T|| = sup_{||\psi|| \neq 0} \frac{||\hat{T}\psi - T\psi||}{||\psi||}$ whose rate of convergence depends on the "general quality" of the Kernel employed in dealing with the estimation of the projection of (X, Z) measurable elements[22].

[22]see proposition 6 proof $H(ii)$for the almost sure existence of $||\hat{T} - T||$.

Assumption $H5$ is the familiar smoothness condition on the density of the error term imposed by the uniform of convergence rate in probability achieved on the nuisance functions.Assumption $H6$ is the stochastic equicontinuity condition, which suffices to ensures the root-n equivalence of the empirical gradients. As in (j) of the Corollary, this seemingly strong assumption it not necessary to obtain this equivalence.

remarks:

(1)It is interesting to compare the one step efficient estimator β_{OS} proposed in Lee 2003 with the ASPQ.The ASPQ minimizing $\mathcal{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i \rho_n(Y_i - \hat{g}_i - \hat{\Gamma}_i'\hat{\beta} - (Z_i - \hat{\Gamma}_i)'\beta)$ has the asymptotic representation:

$$\tilde{\beta}_* = \beta_0 - HS_*(\ddot{\beta}, \hat{\tau})^{-1}\nabla S_*(\beta_0, \hat{\tau}) \text{ for some } \ddot{\beta} \text{ , } wpa.1.$$

$$\nabla S_*(\beta_0, \hat{\tau}) = \frac{\partial \mathcal{S}}{\partial \beta} \mid_{\beta=\beta_0}$$

$$HS_*(\ddot{\beta}, \hat{\tau}) = \frac{\partial^2 \mathcal{S}}{\partial \beta \beta'} \mid_{\beta=\ddot{\beta}}.$$

Conversely the one step efficient estimator suggested in Lee 2003 is computed by:

$$\beta_{OS} = b_n - \{\frac{\partial \nabla S_n(\beta,\hat{\tau})}{\partial \beta'} \mid_{\beta=b_n}\}^{-1}\nabla S_n(b_n, \hat{\tau}),\ wpa.1.$$

$$-\nabla S_n(\beta, \hat{\tau}) \propto \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i \hat{w}_i\ [q - D_n(Y_i - \hat{g}_i - Z_i'\beta)]$$

where b_n is some available estimator such that $\sqrt{n}(b_n - \beta_0) = O_p(1)$ and $D_n(.)$is a smooth function whose derivative is a Kernel.Thus,even though the empirical gradient∇S_n and ∇S_* differ,both estimators are based upon the same principle of approximating the function $d(.) = 1_{.<0}$ with the integral of a Kernel function. That is, both rely on some differentiable $M_n(\beta, \hat{\tau})$ satisfying :

$$plim \frac{\partial M_n(\beta,\hat{\tau})}{\partial \beta'} \mid_{\beta=\beta_n} = q(1-q)\mathcal{V}\mathcal{B}^{-1} \text{for any consistent } \beta_n.$$

and

$$-\sqrt{n}M_n(\beta_0,\hat{\tau})\rightsquigarrow\mathcal{N}(0,[q(1-q)]^2\mathcal{V}\mathcal{B}^{-1}).$$

which yields β_{eff}, the solver of $M_n(\beta,\hat{\tau})=0$ as efficient[23].

Hence, unlike the ASPQ,the nature of the one step estimator is approximative because it estimates in finite sample the true representation of its corresponding β_{eff} with the aid of some root-n consistent estimator.Alternatively, the distinctive nature of the one step may be understood from the perspective of numerical optimization where β_{eff} is approximated by the Newton'a algorithm using only one iteration and some consistent estimator as the starting value while the ASPQ uses as many iterations as necessary furnishing β_{eff}.

(2)Both estimator can be interpreted as GMM estimators minimizing $M_n(\beta,\hat{\tau})'M_n(\beta,\hat{\tau})$ but the A.S.P.Q. has also(smooth)quantile regression interpreting as the regression $\hat{f}(Y-\hat{g}-\hat{\Gamma}'\hat{\beta})$ on $\hat{f}(Z-\hat{\Gamma})$.

(3) In applications,a more expedition way to compute the ASPQ is to notice that $\tilde{\beta}_*=\hat{\beta}+\hat{\delta}$ where $\hat{\delta}=Argmin_{\mathbf{B}}\Sigma\hat{f}_i\rho_n(e_i-\hat{w}_i'\delta)$ where $\{e_i\}_{i=1...n}$ are the residuals from a first stage.Hence,$\hat{\delta}$ can be interpreted as the efficiency adder.

Identically to the one step estimator,the ASPQ is adaptive in the sense that the semi-parametric efficient bound is always reached.However, there are stochastic relationships for (X,Z,ε) (which can be tested using the AQR first stage residuals)simplifying efficient estimation.To clarify this point let $\mathbf{F}$ be the space of all joint density for(X,Z,ε) meeting the assumptions of our model (2.b excluded) and let $\mathbf{H}$ be what we shall name the set of conditions, which is a subset of $\mathbb{R}^7$.We define the "condition mapping" C as follows:

[23]For the sake of simplicity the trimming term is removed for both estimators.

C:$\mathbf{F} \rightarrow \mathbf{H}$

F $\mapsto (||f-f(0)||_{L^1}, ||f-f(0|X)||_{L^1}, ||f-f(0|Z)||_{L^1}, ||\mu(X)-\mu||_{L^1}, ||\mu(Z)-\mu||_{L^1}, ||\mu(X,Z)-\mu||_{L^1}, ||\pi(X,Z)-\chi(X)\zeta(Z)||_{L^1})$

where $\mu(J)$=$E[\varepsilon|J]$ for some random variable J, μ=$E[\varepsilon]$, χ is the marginal density of X and ζ that of Z.Furthermore,we note {f(h),V(h),W(h)}the random vector from (IV) when the condition h $\in \mathbf{H}$ holds and $\{\hat{f}(h), \hat{V}(h), \hat{w}(h)\}$its corresponding nonparametric estimator.For instance, if h contains $||f - f(0|X)||_{L^1} = 0$ and $||\mu(X) - \mu||_{L^1} > 0$ we get $f(h) = f(0|X)$,$V(h) = Y - E[Y|X] + E[\varepsilon|X]$ and$w(h) = Z - E[Z|X]$.

Given F_0 as the true joint distribution of (X, Z, ε), $C(F_0)$ is true set of conditions which we naturally note h_0 with its associated efficient bound $B(h_0)$.It follows that $\hat{\beta}(h_0) = Argmin_\beta \hat{E}[\hat{f}(h_0)\rho_n\{\hat{V}(h_0) - \hat{w}(h_0)'\beta\}]$ satisfies $\sqrt{n}(\hat{\beta}(h_0) - \beta_0) \rightsquigarrow \mathcal{N}(0, B(h_0))$. For instance using our previous example about h_0 a simpler estimator than the ASPQ is the minimizer of $\hat{E}[\hat{f}_x(0)\rho_n\{Y - \hat{E}[Y|X] + \hat{E}[\varepsilon|X] - (Z - \hat{E}[Z|X])'\beta\}]$which resembles the estimator covered in section 4-5 apart from the fact that the assumption 2.b no longer holds and that the density weighting approach is employed to reach efficiency (otherwise the estimator would be solely C.A.N.).We define $\{\hat{\beta}(h)|h \in \mathbf{H}\}$ as a class of estimators we name "two stage smooth semiparametric quantile" given a semi parametric model where $\hat{\beta}(h)$ is efficient for $h = h_0$.

9 Monte Carlo Simulation

In this section we examine the finite sample properties of the suggested estimator described in section 6 for the median case(i.e.q=1/2).This estimation strategy is used to estimate the parameter $\beta = 1$ and the function $g(x) = x + \frac{4}{\sqrt{2\pi}} e^{-2x^2}$(pointwise)when the data generating process obeys:

$$Y = g(X) + \beta Z + U$$

where (X, Z) is a standard bivariate Normal couple of correlation coefficient 0.5.This design was examined in Lee 2003 where $g(.)$ is a bell curve around the origin with the 45 degree line as asymptote.Even though the support of (X, Z) violates assumption 3 and (a) of the corollary the results are not affected The disturbance has the form U=$\sigma\epsilon$ where ϵ ,independent of (X, Z),is either drawn from a standard normal distribution or from a Student distribution with 4 degrees of freedom (normalized to have a unit variance).We used $\sigma = 1$ for the homoscedastic case while $\sigma = e^{\nu(X+Z)}$ for the heteroscedastic model with ν chosen as to normalize the variance of U.We thus examine four designs,the Normal homoscedastic(NHO),the Normal heteroscedastic(NHE), the Student homoscedastic(SHO)and the Student heteroscedastic(SHE).It is rapid to verify that our designs meet assumptions (c) (d) (e) and (f) of the corollary. A simulation of the estimator for a sample size of $n = 50, 200$ and 800 consists of 1000 replications.The simulations are conducted in Gauss.

The smoothing of the check functions follows proposition 2 and the corollary. We used $\rho_n(u) = u + 2h\varphi(u/h)$ where φ is (as described on page 15)derived from the Epanechnikov Kernel of order $r = 4$, which meets assumption(g)of the corollary owing to the fact that(under the type of distributions adopted for ϵ)the smoothness of the density of $U|X, Z$ is infinite a.s. A sequence of bandwidth $h = O(1/n^p)$ with $1/8 < p < 1/6$ satisfies assumption (j). Our preliminary simulations showed that the value of p is immaterial in affecting the results so we decided to use $p = 1/7$.Hence, our simulations are performed

employing $h = cn^{-1/7}$with $c \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$.This last range of values for the bandwidth constant is chosen as to contain $c*$, the optimal values from the perspective of proposition 5 bis which permits to judge whether,at least locally,the optimal choice put forth in this paper is desirable for inferential purposes. In the model with a normal error we found $c* = 3.086$ for the homoscedastic case and $c* = 2.50$ under heteroscedasticity while the model with a Student error yielded $c* = 2.62$ under homoscedasticity and $c* = 2.17$ for the heteroscedastic case. The estimation of the nuisance functions follows (h) of the Corollary where the order 2 kernel $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ is employed along with the bandwidth sequence $\zeta = n^{-1/6}$.

Finally, the estimator is computed minimizing by quadratic hill climbing(Goldfeld, Quandt and Trotter 1966) $\mathcal{S}(\theta) = \sum_{i=1}^{n} \lambda(X_i)\rho_n(\hat{T}_i - \hat{w}_i\prime\theta)$ where $\lambda(X) = 1_{|X|<2}$ is used for the trimming criteria which satisfies assumptions (b) because of the joint normality of (X, Z).Given a n-sample, a search for the global minimum consists of selecting out of 10 iterative searches, the local minimum minimizing $\mathcal{S}$ [24]as there is no guaranty in finite sample that the local minimum is unique because the class of Kernel required for smoothing the check function is negative on some intervals.For instance,in our simulation the Kernel of order 4 utilized is strictly negative on $(-1, -\sqrt{3/7}) \cup (\sqrt{3/7}, 1)$.

A useful check on whether a local minimum is the global minimum consists of obtaining a lower bound B for $\mathcal{S}$ on the complement of $\mathcal{P}$=$\{\ \theta : \mathcal{S}$:strictly convex$\}$(Demindenko 2000).Let $J_n = \{i \in \{1, .., n\} : \lambda_i = 1\}$ and $\hat{W}_{J_n}$ the $\#J_n$ by $K + 1$ matrix of regressors excluding observations not in J_n. Let further suppose that the sample at hand is such that $\hat{W}_{J_n}$has full rank.Since $\frac{\partial^2}{\partial^2\theta}\mathcal{S} \propto \sum_{J_n} \hat{w}_i\hat{w}_i' K_n(\hat{T}_i - \hat{w}_i'\theta)$ where $K_n(t) = \frac{1}{h}K(\frac{t}{h})$ we have $\mathcal{P} = \{\theta : K_n(\hat{T}_i - \hat{w}_i'\theta) > 0\, \forall i \in J_n\} = \{\theta : \hat{T}_i - \hat{w}_i'\theta \in K_n^{-1}(0, \infty) \forall i \in J_n\}$ where $K_n^{-1}(0, \infty) = \cup_{k=1}^{\frac{r}{2}-1} O_{k,n}$ and $\{O_{k,n}\} \subseteq [0, 1]$ are open disjoints intervals which can be found analytically from the roots of the Kernel on $(0, 1)$.Hence,$\theta \in \mathcal{P}^{\complement}$ implies

[24]The different starting values are drawn from a joint $\mathcal{N}(\theta_0, 25Id)$ distribution where Id refers the identity matrix.

$\hat{T}_j - \hat{w}_j'\theta \in K_n^{-1}(0,\infty)^{\complement}$ for some $j \in J_n$ so that a simple lower bound for $\mathcal{S}$ on $\mathcal{P}^{\complement}$ is given by $B = (\#J_n - 1)min\rho_n + min_{K_n^{-1}(0,\infty)^{\complement}}\rho_n$.It follows that a sufficient condition for a local minimum θ_{iter} to be the smooth quantile estimator is $\mathcal{S}(\theta_{iter}) < B$.Even though this check suffices it is not a necessary condition and having too low a bound may not be informative.

For a given sample size ,a table contains four measures enabling to assess, the quality of the estimator $\hat{\theta}$ of θ_0'=(0, 1).The bias column refers to absolute value of the bias i.e.$|E(\hat{\beta}) - \beta_0|$ where $\hat{\beta}$ is the slope coefficient estimator in $\hat{\theta}$. The $RMSE$ columns refer to the root mean squared error for the slope estimator i.e.$(E|\hat{\beta} - \beta_0|^2)^{1/2}$.The third column measures the accuracy of the estimator of $g(.)$ (retrieved as explained on page 22) by the expected value of the empirical $RMSE$ achieved on the nonparametric part i.e. $E[(\int |\hat{g} - g|^2 d\hat{F}_X)^{1/2}]$.Finally, the last column provides the size of the t-test for β_0 using the asymptotic critical values for a 5 percent type I error.For a sample size of $n \leq 5000$ observations, we found it takes approximately $n/100$ seconds to compute the estimator, which of course may vary with the iterative procedure adopted, the number of explanatory variables and the software employed. The global search methods such as SAN are likely to increase this computational time.

Overall,the qualitative behavior of the estimator agrees with the asymptotic theory developed in this paper.First,the RMSE for the slope parameter decreases at the $\sqrt{n}$ rate while the expected empirical RMSE on the non parametric function declines approximately at the $n^{1/3}$ rate.This last discrepancy may arise due to our biased plug in estimator of E$[(\int |\hat{g} - g|^2 dF_x)^{1/2}]$.Also,the disparities of the sizes across bandwidth constants shrinks as the sample size increases which agrees with the convergence in distribution of our t-statistics uniformly in c.Another interesting results from our Monte Carlo experiment pertains to the the absolute value of the bias which is (on average across bandwidth constants)3 percent(2.7 percent)of the parameter value under the Nor-

mal model(respectively the Student model) for a sample size of 50 observations and declines consistently across bandwidth as the sample size increases.As shown on tables 3-6-9-12,when n=800 observations the absolute bias is less than 0.5 percent for all designs.

Even though our theoretical section did not establish finite sample unbiasedness, we believe this finding to be encouraging as far as the ability of our estimation procedure to be on average correct at estimating the truth.

The figures 1 through 12 depict ,for a given sample size,the non parametric function(solid line) along with $E[\hat{g}]$(dashed line)both of which evaluated at a fixed design for $x = \{-2, -1.9, ..., 5.9, 6\}$. Those graphics illustrate an important fact about the estimator of the non parametric function obtained as explained on page 22. The bias i.e. $E[\hat{g}] - g$ declines as the sample size augments but the improvement is not uniform over the design with the right tail values of x above 2 being still inaccurately estimated on average even with a sample of 800 observations.This is a known finite sample problem for a Kernel regression estimator whose bias is inversely related to the density of the conditioning variable.Hence,in our designs low mass point of $X \sim \mathcal{N}(0, 1)$ will provide more pronounced biased estimator for our nuisance functions.It is worth pointing out that a local Kernel regression(Ruppert and Wand 1994)for estimating our nuisance functions would not have this bias issue. Consequently,once the finite dimensional parameter is estimated one may consider in applications using the approach described on page 22 with a local Kernel estimator for making point wise predictions.

Notice that for given sample size,the loss measures are relatively steady across bandwidth but our tables indicate that the size of the test is sensitive to the bandwidth constant adopted. As illustrated on table 3 and table 6,the optimal bandwidth selection criteria proposed in this paper does perform well under the normal model in that the type I error for a sample size of 800 observations is 5 percent for c somewhere between $c = 2.5$ and $c = 3$ under homoscedasticity and 5 percent for $c = 3$ under heteroscedas-

ticity. However,the Student model does not seem to bolster our bandwidth criteria .As illustrated on tables 9 and 12, the size returned with the calculated c* is below 5 percent for a sample size of 800 observations and this regardless of the scedasticity. This result hints that the Student designs require a larger sample size in order for the asymptotic critical values to achieve accurate probability coverage.This last difference between the Normal and Student model is not surprising as our covariance matrix is estimated with a Kernel density estimator which estimates the density of the error evaluated at 0. This last procedure is known to be inaccurate (i.e. have a large variance) when the mass of the distribution is more spread out around the origin.

Overall, our Monte Carlo simulations hint that conducting inferences using the estimated std errors may entail some risk in finite sample because the asymptotic critical values provide acceptable coverage for only specific bandwidth constants.The rule of thumb from proposition 5 bis is simple and did perform well for only some designs. Hence one may seek out alternative ways to conduct inferences. The results from Horowitz smooth LAD estimator suggests that Bootstrapping offers asymptotic improvement for Student and Chi square testing (for any q $\in (0,1)$)if one is willing to impose $r > \frac{7+4d}{2}$ in assumption (g)and use a Kernel 3 times differentiable instead.However,we did not attempt to bootstrapped our estimator.

Table 1: NHO model,n=50

c	Bias	RMSE slope	E[RMSE g]	size
1	0.031588	0.219078	0.432867	0.109
1.5	0.030976	0.209337	0.427383	0.100
2	0.031235	0.210668	0.429861	0.064
2.5	0.029199	0.201296	0.423307	0.054
3	0.031609	0.202413	0.427055	0.036
3.5	0.027463	0.193801	0.419970	0.028
4	0.031441	0.195683	0.425117	0.020

Table 2: NHO model, n = 200

c	bias	RMSE slope	E[RMSE g]	size
1	0.010642	0.106908	0.292745	0.069
1.5	0.008336	0.099368	0.292425	0.049
2	0.011327	0.103047	0.291585	0.052
2.5	0.009526	0.094850	0.291626	0.030
3	0.011819	0.099157	0.290635	0.037
3.5	0.009837	0.091550	0.290847	0.021
4	0.009840	0.090383	0.290499	0.020

Table 3: NHO model, n = 800

c	bias	RMSE slope	E[RMSE g]	size
1	0.002668	0.053722	0.196283	0.061
1.5	0.003066	0.053146	0.196242	0.060
2	0.003293	0.052351	0.196122	0.057
2.5	0.003367	0.051460	0.195925	0.053
3	0.003397	0.050567	0.195731	0.046
3.5	0.003451	0.049722	0.195562	0.042
4	0.003519	0.048945	0.195426	0.036

Table 4: NHE model, n=50

c	bias	RMSE slope	E[RMSE g]	size
1	0.033639	0.221999	0.448427	0.115
1.5	0.032424	0.214278	0.444255	0.083
2	0.033415	0.214371	0.445112	0.065
2.5	0.030145	0.206940	0.440343	0.057
3	0.029100	0.203535	0.438455	0.040
3.5	0.028271	0.200438	0.433846	0.028
4	0.027782	0.197920	0.435556	0.019

Table 5: NHE model, n =200

c	bias	RMSE slope	E[RMSE g]	size
1	0.010165	0.101104	0.304240	0.061
1.5	0.009967	0.098960	0.303665	0.046
2	0.009878	0.096847	0.303063	0.038
2.5	0.009842	0.095193	0.302493	0.032
3	0.009711	0.093821	0.301963	0.028
3.5	0.009508	0.092805	0.301470	0.023
4	0.009254	0.092026	0.301014	0.023

Table 6: NHE model, n = 800

c	bias	RMSE slope	E[RMSE g]	size
1	0.003604	0.052892	0.204410	0.066
1.5	0.003938	0.052172	0.204320	0.058
2	0.004052	0.051369	0.204170	0.056
2.5	0.003945	0.050612	0.203943	0.055
3	0.003738	0.049844	0.203687	0.050
3.5	0.003504	0.049082	0.203425	0.043
4	0.003278	0.048401	0.203180	0.033

Table 7: SHO model, n=50

c	bias	RMSE slope	E[RMSE g]	size
1	0.028374	0.176221	0.413889	0.079
1.5	0.028270	0.172627	0.412082	0.056
2	0.028389	0.168166	0.410688	0.034
2.5	0.028311	0.164013	0.409599	0.024
3	0.028148	0.161091	0.408962	0.015
3.5	0.028264	0.159465	0.408710	0.011
4	0.028446	0.158797	0.408657	0.008

Table 8: SHO model, n =200

c	bias	RMSE slope	E[RMSE g]	size
1	0.016048	0.0819181	0.287667	0.054
1.5	0.015215	0.0797703	0.287026	0.035
2	0.014580	0.0777787	0.286451	0.028
2.5	0.014253	0.0760414	0.285995	0.023
3	0.014047	0.0745962	0.285672	0.019
3.5	0.013940	0.0734891	0.285460	0.013
4	0.013912	0.0727618	0.285340	0.009

Table 9: SHO model, n =800

c	bias	RMSE slope	E[RMSE g]	size
1	0.005048	0.038464	0.195491	0.044
1.5	0.005033	0.037771	0.195373	0.036
2	0.005119	0.037194	0.195275	0.034
2.5	0.005421	0.036643	0.195246	0.026
3	0.005714	0.036220	0.195231	0.022
3.5	0.005870	0.035856	0.195199	0.020
4	0.005932	0.035582	0.195162	0.017

Table 10: SHE model, n=50

c	bias	RMSE slope	E[RMSE g]	size
1	0.027514	0.181987	0.426903	0.076
1.5	0.027752	0.178062	0.425157	0.056
2	0.027918	0.172998	0.423682	0.034
2.5	0.027709	0.168774	0.421838	0.022
3	0.027701	0.166064	0.421838	0.016
3.5	0.027716	0.164363	0.421401	0.014
4	0.027715	0.163341	0.421179	0.009

Table 11: SHE model, n =200

c	bias	RMSE slope	E[RMSE g]	size
1	0.017671	0.082785	0.298325	0.068
1.5	0.016614	0.080515	0.297658	0.044
2	0.015691	0.078640	0.297032	0.033
2.5	0.014870	0.076668	0.296444	0.026
3	0.014214	0.075626	0.295943	0.020
3.5	0.013743	0.074636	0.295548	0.012
4	0.013457	0.074070	0.295274	0.009

Table 12: SHE model, n =800

c	bias	RMSE slope	E[RMSE g]	size
1	0.005695	0.038509	0.203125	0.044
1.5	0.005750	0.037736	0.203012	0.034
2	0.005647	0.037150	0.202857	0.030
2.5	0.005482	0.036766	0.202677	0.024
3	0.005372	0.036473	0.202517	0.020
3.5	0.005326	0.036266	0.202383	0.017
4	0.005279	0.036139	0.202267	0.016

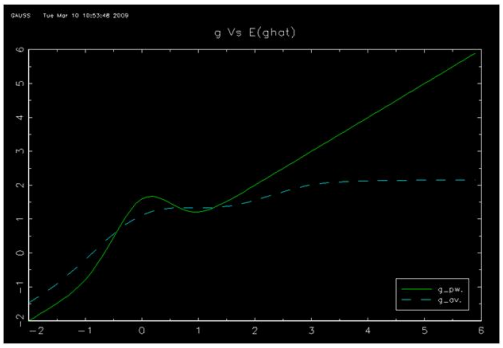

Figure 1: NHO,n=50

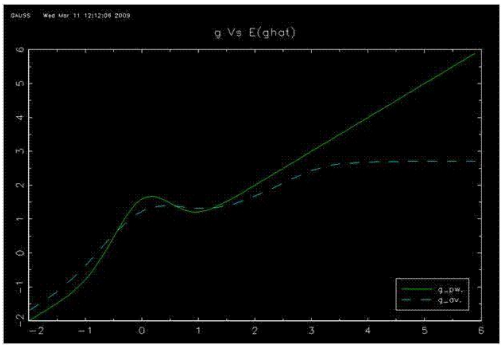

Figure 2: NHO,n=200

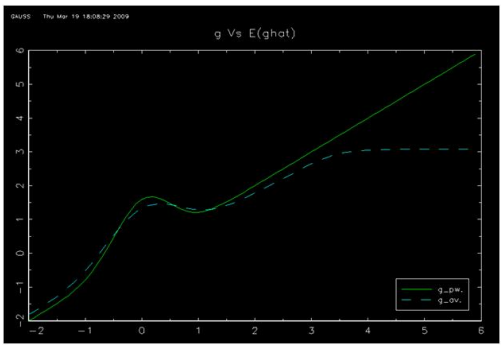

Figure 3: NHO,n=800

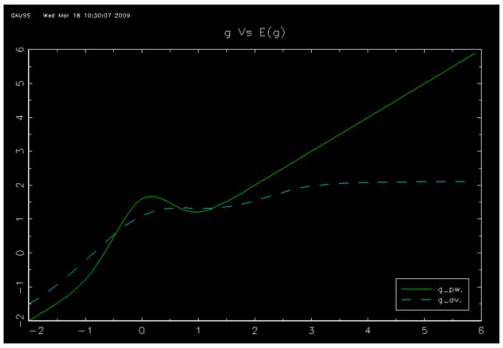

Figure 4: NHE,n=50

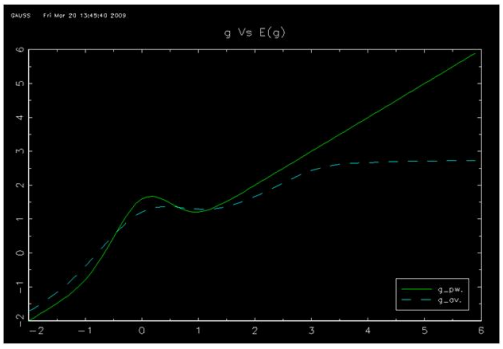

Figure 5: NHE,n=200

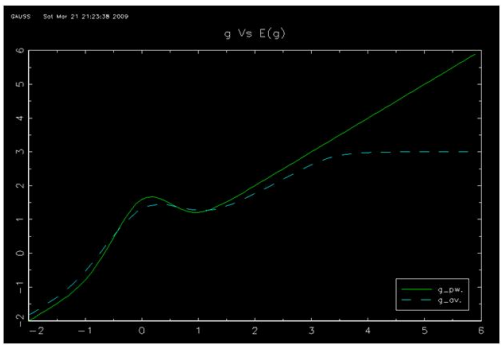

Figure 6: NHE,n=800

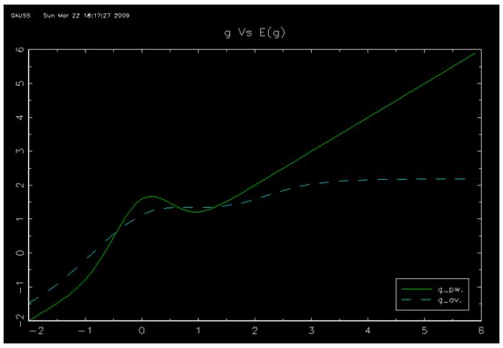

Figure 7: SHO,n=50

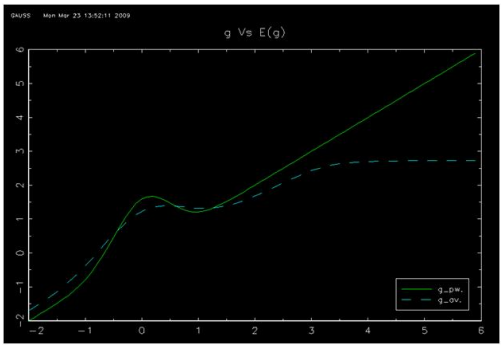

Figure 8: SHO,n=200

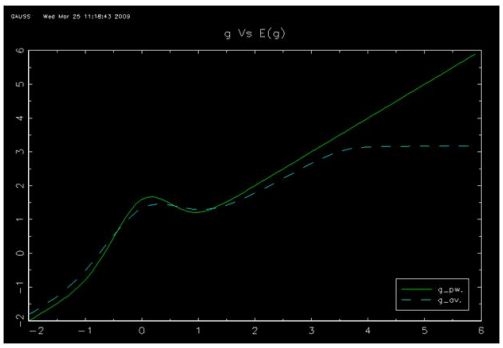

Figure 9: SHO,n=800

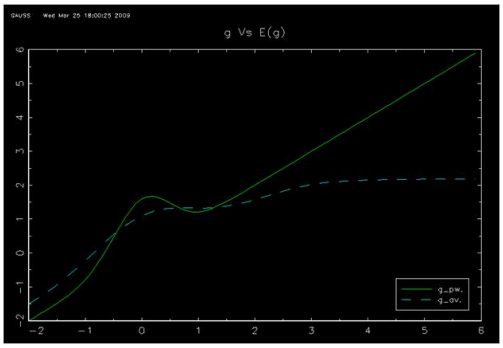

Figure 10: SHE,n=50

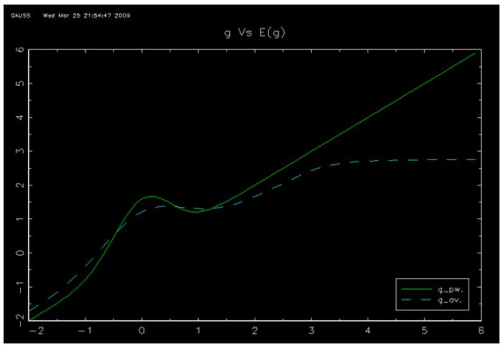

Figure 11: SHE,n=200

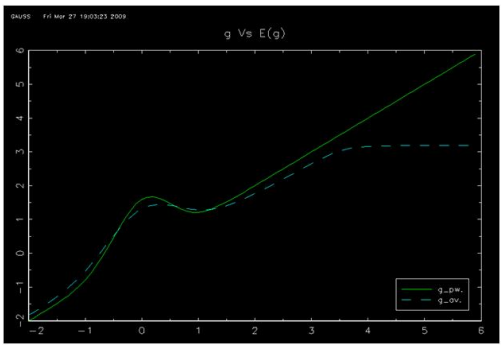

Figure 12: SHE,n=800

10 Conclusion

In this paper we have presented a root n consistent estimator for the slope parameter in a semi parametric quantile model which offers,under homoscedasticity,an efficient alternative to the AQR estimator.Our simulations show that this two stage smooth procedure behaves well in finite sample in terms of the bias and MSE but that a large sample size is needed for inferential purposes. Also, we discussed the generalization of this approach to any measurability of $f(0|X,Z)$ in order to reach the efficiency bound and the corresponding class of 2SSPQ estimators offering a systematic way to estimate the linear part efficiently via smooth quantile regression.We foresee four topics for future research related to the simple estimator suggested in section 3-7.First,the optimal bandwidth selection for testing purposes..Secondly,the testing of the homoscedastic assumption extending the slope invariance principle(Koenker and Basset 1982)for a smooth quantile estimator or using a direct non parametric approach from consistent residuals(Ullah 1996). Thirdly, the testing of assumptions 2b with the aid of another less efficient estimator "under the null" (non weighted AQR for instance)is an important question to explore as our estimator is no longer consistent should this condition be violated.Finally, the possible extension of this estimator when a subset of (Z,X) is endogenous as the conditional quantile may not the prime object of interest for policy making purposes. In that case we speculate that which one of the three existing approaches ,instrument variables (Honore and Hu 2004),"fitted value" (Amemiya 1982)and "Control function"(Lee 2004)is suitable will depend on which of X and Z is endogenous.

References

D.Andrews.1994. Asymptotic For Semi Parametric Econometrics Models Via Stochastic Equicontinuity. Econometrica.

D.Andrews.1992. Generic Uniform convergence. Econometric Theory.

S.Lee. 2003. Efficient Semi parametric Estimation of a Partially Linear Quantile Regression Model. Econometric Theory.

Bo Honoré and Luojia Hu 2004. On the Performance of Some Robust Instrumental Variables Estimators. Journal of Business and Economic Statistics.

Bickel,P.J.1982. On adaptive estimation. Annals of Statistics.

Manski,C.1975. Semi parametric Analysis of discrete Response,Asymptotic Properties of Maximum Score Estimator. Journal of Econometrics.

Stone,C.J.1975. Adaptive maximum likelihood estimators of a location parameter. Annals of Statistics.

Huber P.J.1964. Robust Estimation of a Location Parameter. Annals of Statistics,Volume 35,Number 1,73-101.

A.M.Jones,J.Wildman2008. Health, income and relative deprivation: Evidence from the BHPS. Journal of Health Economics.

G.Bobonis,F.Finan2005. Endogenous Peer Effect in School Participation. Working paper,University of Toronto.

J. Gao and I.Gijbels2008. Bandwidth Selection in Nonparametric Kernel Testing. Journal of The American Statistical Association.

P.Hall, S.J. Sheather,M.C.Jones and J.S.Marron1991. On optimal data-based bandwidth selection in kernel density estimation. Biometrika.

X.Chen ,O. Linton and I. Van Keilegom.2003. Estimation of Semiparametric Models when the Criterion Function is not Smooth. Econometrica.

Stein.C.1956. Efficient Nonparametric Testing and Estimation. Proc.Third Berkeley Symp.Math Statist.Prob

G. Tripathi and T.Severini.2001. A Simplified Approach to Computing Efficiency Bounds in Semi parametric Models. Journal of Econometrics.

K.YU and M.Jones.1998. Local Linear Quantile Regression. The Journal of the American Statistical Association.

J.Horowitz. 1998. Bootstrap Methods for Median Regression Models. Econometrica.

J.Horowitz. 1992. A Smooth Maximum Score Estimator For the Binary Response Model. Econometrica.

J.Horowitz and P.Hall 2005. Nonparametric Methods For Inference in The Presence of Instrumental Variables. Annals of Statistics.

J.Horowitz and P.Hall 1990. Bandwidth Selection in Semiparametric Estimation of Censored Linear Regression Models. Econometric Theory.

P.M. Robinson.1988. Root N Consistent Semi parametric Regression. Econometrica.

Ruppert.D.and Wand.M.P..1994. Multivariate Locally Weighted Least Squares Regression. Annals of Statistics 22(3): 1346-1370.

J.Hahn.1995. Bootstrapping Quantile Regression Estimators. Econometric Theory.

R.Koenker,P. Ng, S.Portnoy1994. Quantile Smoothing Spline. Biometrica.

R.Koenker and G.Basset.1978. Regression Quantiles. Econometrica.

R.Koenker and G.Basset.1982. Robust Tests for Heteroskedasticity Based on Regression Quantiles. Econometrica.

Kirkpatrick S., Gerlatt C. D. and Vecchi M. P.1983. Optimization by Simulated Annealing. Science 220,671-680.

P.Chaudhuri.1991 Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation. Annals of Statistics.

P.Chaudhuri,K.Doksum,A.Samarov.1997 On Average Derivatives Quantile regression. Annals of Statistics,Vol.25, No. 2,715.

E.Demidenko.2000 Is This the Least squares Estimate?. Biometrika.

J.Horowitz and S.Lee.2007. Non parametric Instrument Variables Estimation of A Quantile Regression Model. Econometrica.

V. Cherozhukov P.Gagliardini, O. Scaillet.2008. Nonparametric Instrument Variable Estimators of Quantile Structural Effects. Swiss Finance Institute Research Paper Series N08-03.

J.Horowitz and S.Lee2005 Nonparametric Estimation Of An Additive Quantile Regression Model. Journal Of The American Statistical Association.

W.Newey.1994. The Asymptotic Variance of Semi parametric Estimators Econometrica.

W.Newey. J.Powell.1990. Efficient Estimation of Linear and Type I Censored Regression Under Conditional Quantile Restrictions. Econometric Theory.

J. Fan, T.-C. Hu and Y.G. Truong.1994. Robust Non-Parametric Function Estimation. Scandinavian J. Statist.

H J.Bierens.1987. Kernel Estimators Of Regression Functions. Advances in Econometrics:Fifth World Congress,Vol I.

M. Buchinsky.1994. Change in the U.S. Wage structure 1962-1987: Application of Quantile Regression. Econometrica.

S.Chen.,S.Kahn.2001. Semi parametric Estimation of Partially Linear Censored Regression Model. Econometric Theory.

T.Amemiya.1982. Two Stage Least Absolute Deviation. Econometrica.

M. Carrasco,J.P. Florens, E.Renault.2007. Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. Handbook of Econometrics,Volume 6B.

J.Powell.1994. Estimation of Semiparametric Models. Handbook of Econometrics ,Volume 4,chap 41.

J.A.Hausman.1978. Specification Tests in Econometrics. Econometrica.

A.Pagan and A. Ullah.1999. Non parametric Econometrics. Cambridge University Press.

S.M.Goldfeld,R.E.Quandt and H.F.Trotte.1966. Maximization by Quadratic Hill-Climbing. Econometrica,Vol.34,No.3.

11 Appendix

In this section we provide the proofs to our propositions.

Proposition 1:

This is inspired from Amemiya(1982) approach in the context of a Median regression.

Existence of E[$\rho(T - w'\theta)$] uniformly over Θ

Writing $\varepsilon(\theta) = T - w'\theta$ for an arbitrary $\theta \in \Theta$ and ε as the true error we obtain:

$$\rho(\varepsilon(\theta)) = 2(\varepsilon - w'\Delta)(q - 1_{\varepsilon < w'\Delta}) \text{ where } \Delta = \theta - \theta_0$$

It follows that $|\rho(\varepsilon(\theta)| \leq 2max(q, 1-q)(|\varepsilon| + ||w||.||\Delta||)$.Using the compactness of Θ , $w \in L_\infty$ (from 3 and 5 together) along with assumption $(2a)$ ensures that $E|\rho(\varepsilon(\theta)|$ exists uniformly over Θ. We will subsequently note $S(\theta)$= E$[\rho(T - w'\theta)]$ for any $\theta \in \Theta$.

θ_0 as the global minimum of S

Because $|\varepsilon(\theta)| - \varepsilon(\theta) = 2\varepsilon(\theta)^-$ where $\varepsilon(\theta)^-$ is the negative part of $\varepsilon(\theta)$ we derive:

$$S(\theta) = 2\{qE[\varepsilon] - qE[w'\Delta] - E[\varepsilon 1_{\varepsilon < w'\Delta}] + E[w'\Delta 1_{\varepsilon < w'\Delta}]\}$$

where Δ is defined as before.Using iterated expectation and noting $F_{x,z}(.)$ the distribution function of ε conditional on $X = x$ and $Z = z$ furnishes the arranged expression:

$$S(\theta) = 2\{E[(w'\Delta)(F_{x,z}(w'\Delta) - q)] - E[V(w'\Delta)]\} + C$$

Where C is a constant and $V(w'\Delta)) = \int 1_{e < w'\Delta} e f_{x,z}(e)de$. By assumption $2a$, the function $G(t) = \int_{-\infty}^{t} e f_{x,z}(e)de$ will be differentiable almost everywhere with $G'(t) = t f_{x,z}(t)$. Thus, the Leibniz' rule provides the following expression for the gradient of $S(.)$:

$$\nabla S(\theta) = 2E[w\{F_{x,z}(w'\Delta) - q\}].$$

Clearly by assumption (1) θ_0 meets the first order condition for extremum. Furthermore, the Hessian of S is given by $HS(\theta) = 2E[ww' f_{x,z}(w'\Delta)]$. Using $f_{x,z}(0) > 0$ a.s. by assumption (7)(take the infinimum of all $r(x, z)$ we note r to construct a ball of center 0 and radius r where $f_{x,z}(.) > 0$ a.s.) and assumption (8) we conclude that $HS(\theta_0)$is definite positive and θ_0 is consequently a local minimum of S.Finally, let's show that it is indeed the global minimum.For all Δ of Euclidian norm strictly positive

we note $Z(\Delta) = w'\Delta[F_{x,z}(w'\Delta) - q]$. We have $P[|w'\Delta| > 0] > 0$ (from assumption 8) and $f_{x,z}(.) > 0$ on $B(0,r)$. It follows that $E[Z(\Delta)] > 0$ for all Δ such that $||\Delta|| > 0$. This implies (by the Cauchy-Schwartz's inequality) that $||\nabla S(\Delta)|| > 0$ holds whenever $||\Delta|| > 0$. In other words, θ_0 is the unique local minimum for S.

Proposition 2:

for all $\theta \in \Theta$ let $S_*(\theta) = n^{-1}\sum_{i=1}^{n} \rho_n(T_i - w_i'\theta)$and $\hat{S}(\theta) = n^{-1}\sum_{i=1}^{n} \rho(T_i - w_i'\theta)$.

(i)θ_* is consistent for θ_0

By the triangular inequality we get:

$$||S_* - S||_{sup\Theta} \leq ||S_* - \hat{S}||_{sup\Theta} + ||\hat{S} - S||_{sup\Theta}$$

By the uniform weak law of large numbers (UWLLN) we have $||\hat{S} - S||_{sup\Theta} = o_p(1)$ while lemma 1 yields $||S_* - \hat{S}||_{sup\Theta} = O(h)$.Consequently, $plim||S_* - S||_{sup\Theta} = 0$,which ensures θ_* weak consistency.Actually, one can show that θ_* is strongly consistent(Lemma 6).

(ii)asymptotic normality

Step1: S_* twice differentiability permits the following score representation:

$$\nabla S_*(\theta_*) = \nabla S_*(\theta_0) + HS_*(\bar{\theta})(\theta_* - \theta_0)$$

for some $\bar{\theta}$ in the line segment joining θ_* and θ_0.

Let's proof the claim that $plim\ HS_*(\bar{\theta}) = HS(\theta_0) = 2E[ww'f_{x,z}(0)]$.For that purpose we are to show first that $plim HS_*(\theta) = HS(\theta)$ uniformly over Θ.Apart from some minor differences ,the proof follows the non parametric literature for showing that the Kernel density estimator converges almost surely, examining the limiting behavior of the discrepancy between the average of random variables and the average of their means(Pagan and Ullah page 35-36).

we have $|HS_*(\theta) - HS(\theta)| \leq |HS_*(\theta) - EHS_*(\theta)| + |EHS_*(\theta) - HS(\theta)|$ (|.|for a matrix is to be understood componentwise)

where

$HS_*(\theta) - EHS_*(\theta) = n^{-1}\sum Z_{i,n}(\theta) - \mu_{i,n}(\theta)$

$Z_{i,n}(\theta) = h^{-1}2w_i w_i' K(\frac{\varepsilon_i - w_i'\Delta}{h})$

$\mu_{i,n}(\theta) = EZ_{i,n}(\theta)$.

Owing to the fact that $w_i w_i'$ is bounded componentwise and that $K(.)$ is a bounded function we obtain:

$$|Z_{i,n}(\theta) - \mu_{i,n}(\theta)| = O(1/h)$$

and

$$VZ_{i,n}(\theta) \leq EZ_{i,n}(\theta)Z_{i,n}(\theta)' \leq \frac{A}{h} sup_{x,z}\|f_{x,z}\|_{sup\mathbb{R}} \int K^2$$

where A is a constant due to $w'w \in L^\infty$ and $sup_{x,z}\|f_{x,z}\|_{sup\mathbb{R}}$ is the supremum over the compact set $\mathcal{X} \times \mathcal{Z}$ of the sup of the conditional density of the error.Because of our assumption this is also a constant.It follows that $VZ_i, n(\theta) = O(1/h)$.The Bennett's inequality hence yields that for an arbitrary $\delta > 0$ we have $P[|HS_*(\theta) - EHS_*(\theta)| > \delta] = O(e^{-l(\delta)nh})$ where $l(.) > 0$ on $\mathbb{R}_{++}$.Consequently $HS_*(\theta) - EHS_*(\theta) \longrightarrow 0$ a.s. follows by simply invoking the Borel-Cantelli lemma.

Finally, notice that:

$$EHS_*(\theta) - HS(\theta) = E[ww'E_{x,z}(\tfrac{1}{h}K(\tfrac{\varepsilon - w'\Delta}{h}) - f_{x,z}(w'\Delta))] = E[ww'b_n(w'\Delta)]$$

where$|b_n(w'\Delta)| \leq \frac{h^r}{r!}\|f_{x,z}^{(r)}\|_{sup\mathbb{R}} \int |t^r K(t)|dt = O(h^r)$ because of assumption 9, the compactness of $\mathcal{X} \times \mathcal{Z}$ and our Kernel choice. Clearly, we also have $EHS_*(\theta) - HS(\theta) = O(h^r)$ and subsequently $HS_*(\theta) \longrightarrow HS(\theta)$ a.s. uniformly over Θ. Henceforth plim $HS_*(\bar{\theta}) = HS(\theta_0)$ follows from lemma 4 of Amemiya 1973 using $\bar{\theta}$ weak consistency along with $f_{x,z}$continuity a.s.(i.e. assumption 9).

Step 2: from step 1 we have:

$n^{1/2}(\theta_* - \theta_0) \equiv HS_*(\bar{\theta})^{-1}\{-n^{1/2}\nabla S_*(\theta_0)\}$ wpa.1

where $HS_*(\theta) = \frac{2}{nh}\sum_{i=1}^n w_i w_i' K(\frac{\varepsilon_i - w_i'\Delta}{h})$

and

$-n^{1/2}\nabla S_*(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n 2w_i[q - d(\varepsilon_i/h)]$.

Noting $g_{i,n} = 2w_i[q - d(\varepsilon_i/h)]$ for $i = 1...n$ we have:

$$-n^{1/2}\nabla S_*(\theta_0) = U_{1,n} + U_{2,n}$$

where $U_{1,n} = n^{-1/2}\sum g_{i,n} - Eg_{i,n}$ and $U_{2,n} = n^{-1/2}\sum Eg_{i,n}$.

Given the *iid* sequence $\{w_i, \varepsilon_i\}_{i=1\ldots n}$ we easily get:

$$\sum E|\tfrac{g_{i,n}-Eg_{i,n}}{\sqrt{n}}|^{2+\delta} = n^{-\delta/2}E|g_{i,n} - Eg_{i,n}|^{2+\delta} \text{ for any } \delta > 0.$$

Furthermore, w'w $\in L^{\infty}$ and $d(.)$ being a bounded function further give:

$$E|g_{i,n} - Eg_{i,n}|^{2+\delta} = O(1)$$

establishing $\sum E|\tfrac{g_{i,n}-Eg_{i,n}}{\sqrt{n}}|^{2+\delta} = O(n^{-\delta/2})$ and subsequently $lim \sum E|\tfrac{g_{i,n}-Eg_{i,n}}{\sqrt{n}}|^{2+\delta} = 0$ for some $\delta > 0$.Consequently, we can apply the *Liapounov's Central Limit Theorem* to our double array $\{g_{i,n} - Eg_{i,n}\}_{i=1\ldots n}$ which yields:

$$U_{1,n} \rightsquigarrow \mathcal{N}(0, limE(g_{i,n} - Eg_{i,n})(g_{i,n} - Eg_{i,n})')$$

Next, we must show that lim $E(g_{i,n} - Eg_{i,n})(g_{i,n} - Eg_{i,n})' = 4E[q(1-q)w_i w_i']$.From Lemma1 we know that lim $q - d(t/h) = q - 1_{t<0}$ a.e. which combined to assumption 10 ensures plim $w_i w_i'[q - d(\varepsilon_i/h)] = w_i w_i'[q - 1_{\varepsilon_i<0}]$. Furthermore $E|w_i w_i'| < \infty$ so can invoke *The Dominated Convergence Theorem* to conclude that:

$$\text{lim } E[4w_i w_i'\{q - d(\varepsilon_i/h)\}^2] = 4E[q(1-q)w_i w_i'].$$

Using a similar reasoning yields :

$$limE[2w_i\{q - d(\varepsilon_i/h)\}]E[2w_i'\{q - d(\varepsilon_i/h)\}] = 0.$$

Hence,$U_{1,n} \rightsquigarrow \mathcal{N}(0, 4E[q(1-q)w_i w_i'])$ is established.Finally,$U_{2,n} = O(h^r\sqrt{n}) = o(1)$ by lemma2 which yields:

$$-n^{1/2}\{\nabla S_*(\theta_0)\} \rightsquigarrow \mathcal{N}(0, 4q(1-q)E[w_i w_i'])$$

and proposition 3 directly follows from step 1 and step 2.

Proposition 3:

See Bierens (1987) page 115-116-117.

Proposition 4:

let $\hat{T}_i = Y_i - \hat{M}(X_i)$ and $\hat{w}_i = Z_i - \hat{\vartheta}(X_i)$ for $i = 1...n$.Also,let $\tilde{S}(\theta) = n^{-1}\sum_{i=1}^{n}\rho(\hat{T}_i - \hat{w}_i{}'\theta)$ be the counterpart to $\hat{S}(\theta)$ when nuisance parameters are estimated and $\tilde{S}_*(\theta) = n^{-1}\sum_{i=1}^{n}\rho_n(\hat{T}_i - \hat{w}_i{}'\theta)$.

It suffices to show that $plim||\hat{S} - \tilde{S}||_{sup\Theta} = 0$.Using Basic inequalities we obtain:

$$|\hat{S}(\theta) - \tilde{S}(\theta)| \leq \frac{1}{n}\sum |2q-1||U_i(\theta) - \hat{U}_i(\theta)| + |U_i(\theta) - \hat{U}_i(\theta)|$$

where $U_i(\theta) = T_i - w_i'\theta$ and $\hat{U}_i(\theta) = \hat{T}_i - \hat{w}_i{}'\theta$. simplifying further using our uniform rate of convergence on the non parametric terms easily yields:

$$||\hat{S} - \tilde{S}||_{sup\Theta} \leq O\{|\hat{M} - M|_{sup_{x\in\mathcal{X}}} + ||\hat{\vartheta} - \vartheta||_{sup_{x\in\mathcal{X}}}\} = o_p(1)$$

and $plim||\hat{S} - \tilde{S}||_{sup\Theta} = 0$ is proven.This suffices for proposition 4 because by the triangular inequalities we have:

$$||\tilde{S}_* - S||_{sup\Theta} \leq ||\tilde{S}_* - \tilde{S}||_{sup\Theta} + ||\tilde{S} - \hat{S}||_{sup\Theta} + ||\hat{S} - S||_{sup\Theta}$$

with plim $||\hat{S} - S||_{sup\Theta} = 0$ by the UWLLN while plim $||\tilde{S}_* - \tilde{S}||_{sup\Theta} = 0$ by lemma1. As a result plim $||\tilde{S}_* - S||_{sup\Theta} = 0$ which shows that $\tilde{\theta}_*$ is consistent.Similarly to proposition 1 one can show that $\tilde{\theta}_*$ is strongly consistent because both $|\hat{M} - M|_{sup_{x\in\mathcal{X}}}$ and $||\hat{\vartheta} - \vartheta||_{sup_{x\in\mathcal{X}}}$ convergence are almost sure(Bierens 1987).

Proposition 5:

we need to introduce some notations to ease the length of the proof.

Let $e'_{-1} = [0, I_K]$ the K by K+1 matrix (where I_K is the identity matrix of dimension K) and $e'_1 = (1, 0, ..., 0)$of dimension K+1.Let$\tau = \{T_i, w_i\}_{i=1}^n$ and $\hat{\tau} = \{\hat{T}_i, \hat{w}_i\}_{i=1}^n$.Also,for $k = 1, 2$ we note $D^k$the k^{th} derivative operator of a multivariate function defined on $\mathbb{R}^{K+1}$ where $k = 1$ corresponds to the gradient noted ∇ while $k = 2$ returns the Hessian noted H. Also, $D^k S_*(\eta, \tau)$ refers to the k^{th} derivatives of S_* with respect to θ evaluated at the finite dimensional parameter η and using τ. Similarly we write $D^k S_*(\eta, \hat{\tau})$as the k^{th} derivatives of S_* with respect to θ evaluated at the finite dimensional parameter η but using $\hat{\tau}$.For $i = 1...n$ we further employ the condensed notations $\Delta\vartheta_i = \vartheta(X_i) - \hat{\vartheta}(X_i)$ and $\Delta M_i = M(X_i) - \hat{M}(X_i)$. Finally we use $d_n(t) = q - d(t/h)$ and $K_n(t) = \frac{1}{h}K(t/h)$ as sequences of real valued functions.

Using a Taylor's expansion for the score around θ_0 taking the nuisance parameters as a constant(Andrews 94)yields:

$$\nabla S_*(\tilde{\theta}_*, \hat{\tau}) = \nabla S_*(\theta_0, \hat{\tau}) + HS_*(\ddot{\theta}, \hat{\tau})(\tilde{\theta}_* - \theta_0)$$

for some $\ddot{\theta}$ somewhere in the line segment joining $\tilde{\theta}_*$ and θ_0. By Lemma3 plim $HS_*(\ddot{\theta}, \hat{\tau}) - HS_*(\ddot{\theta}, \tau) = 0$ and consequently (by the same token as step1 of proposition 3 proof) plim $HS_*(\ddot{\theta}, \hat{\tau}) = HS(\theta_0)$. Thus, we have:

$$n^{1/2}(\tilde{\theta}_* - \theta_0) = HS_*(\ddot{\theta}, \hat{\tau})^{-1}\{-n^{1/2}\nabla S_*(\theta_0, \hat{\tau})\} \; wpa.1$$

where $-n^{1/2}\nabla S_*(\theta_0, \hat{\tau}) = \frac{1}{\sqrt{n}} \sum 2\hat{\omega}_i[q - d(\hat{\varepsilon}_i/h)] = \frac{1}{\sqrt{n}} \sum 2(w_i + \hat{w}_i - w_i)d_n(\hat{\varepsilon}_i)$.Using $\hat{\varepsilon}_i = \varepsilon_i + \Delta_i$ where $\Delta_i = \Delta M_i + \beta_0' \Delta\vartheta_i$ for $i = 1...n$ and $d_n(.)$ twice differentiability furnishes:

$$-n^{1/2}\nabla S_*(\theta_0, \hat{\tau}) = \frac{1}{\sqrt{n}} \sum 2(w_i + \hat{w}_i - w_i)[d_n(\varepsilon_i) + K_n(\varepsilon_i)\Delta_i + K_n^{(1)}(\xi_i)\Delta_i^2]$$

for some $\{\xi_i\}_{i=1}^n \in \otimes_{i=1}^n(\varepsilon_i, \varepsilon_i + \Delta_i)$.Hence,distributing breaks down the analysis of the limiting distribution in 4 blocks:

$$-n^{1/2}\nabla S_*(\theta_0, \hat{\tau}) = -n^{1/2}\nabla S_*(\theta_0, \tau) + R_{1,n} + R_{2,n} + E_n$$

where

$R_{1,n} = \frac{1}{\sqrt{n}} \sum 2w_i K_n(\varepsilon_i)\Delta_i$;

$R_{2,n} = \frac{1}{\sqrt{n}} \sum 2w_i K_n^{(1)}(\xi_i)\Delta_i^2$;

$e_1' E_n = 0; e_{-1}' E_n = R_{3,n} + R_{4,n} + R_{5,n}$;

$R_{3,n} = \frac{1}{\sqrt{n}} \sum 2\Delta\vartheta_i d_n(\varepsilon_i)$;

$R_{4,n} = \frac{1}{\sqrt{n}} \sum 2K_n(\varepsilon_i)\Delta\vartheta_i\Delta_i$;

$R_{5,n} = \frac{1}{\sqrt{n}} \sum 2K_n^{(1)}(\xi_i)\Delta_i^2\Delta\vartheta_i$

By lemma 4 and 5 we know that $R_{1,n} + R_{2,n} + E_n = o_p(1)$ which yields:

$$\sqrt{n}(\tilde{\theta}_* - \theta_0) = HS_*(\ddot{\theta}, \hat{\tau})^{-1}\{-\sqrt{n}\nabla S_*(\theta_0, \tau)\} + HS_*(\ddot{\theta}, \hat{\tau})^{-1}o_p(1).$$

exploiting plim $HS_*(\ddot{\theta}, \hat{\tau}) = HS(\theta_0)$ directly provides $plim|\sqrt{n}(\tilde{\theta}_* - \theta_0) - \sqrt{n}(\theta_* - \theta_0)| = 0$.

Corollary

The proofs of proposition $1, 2$ are identical apart from the trimming function and the uniform rate of convergence in probability achieved for the nuisance functions.Under assumptions (b),(c), and (d) we can easily show that $E[|Y - M|^{2+a}|X = x]\pi(x)$,$E[|Y - \vartheta|^{2+a}|X = x]\pi(x)$ are bounded functions and that both $v_1(x) = V(Y|X = x)$ and $v_2(x) = V(Z|X = x)$ belong to $\mathcal{C}(\mathcal{X})$.Hence, $M^2\pi$,$\vartheta^2\pi$, $v_1\pi$ and $v_2\pi$ are bounded continuous functions. It follows from Bierens 1987 that $sup_{x\in\mathcal{X}*}|\hat{M} - M|$=$O_p(1/a_n)$ and $sup_{x\in\mathcal{X}*}||\hat{\vartheta} - \vartheta||$=$O_p(1/a_n)$ where $a_n = n^{\frac{1}{2+d}}$ so that a similar reasoning as in proposition 4 is straightforward to show that the estimator is weakly consistent. Lastly,the analogue of proposition 5 can be conducted using a_n instead of root n, noticing that under (j) lim $a_n h^2$=∞ permits to show that plim $HS_*(\ddot{\theta}, \hat{\tau}) = HS(\theta_0)$ using the same approach as in Lemma 3.

Thus,showing $Z_n = \sqrt{n}||\nabla S_*(\hat{\tau}, \theta_0) - \nabla S_*(\tau_0, \theta_0)|| = o_p(1)$ would suffice to conclude $plim|\sqrt{n}(\tilde{\theta}_* - \theta_0) - \sqrt{n}(\theta_* - \theta_0)| = 0$. But this last condition on Z_n holds under assumption(k)because, as in Andrews 1994, one can use the fact that for any $\eta > 0$ we have:

$$P[Z_n > \eta] = P[Z_n > \eta \cap \{\urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) < \eta \cap \hat{\tau} \in \mathcal{F}\}] + P[Z_n > \eta \cap \{\urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) \geq \delta \cup \hat{\tau} \notin \mathcal{F}\}]$$

$$\leq P^*[sup_{\mathcal{B}(\tau_0,\delta)}\sqrt{n}||\nabla S_*(\tau, \theta_0) - \nabla S_*(\tau_0, \theta_0)|| > \eta] + P[\urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) \geq \delta] + P[\hat{\tau} \notin \mathcal{F}]$$

Noticing $\urcorner_{\mathcal{F}}(\hat{\tau}, 0) \leq \urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) + \urcorner_{\mathcal{F}}(\tau_0, 0)$ and $\urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) = o_p(1)$yields:

lim P$[\urcorner_{\mathcal{F}}(\hat{\tau}, 0) < A$ for some $A > \urcorner_{\mathcal{F}}(\tau_0, 0)]$=1 so that lim P$[\hat{\tau} \in \mathcal{F}]$=1 holds.

Finally using $\varepsilon = \eta$ in assumption(k) along with $\urcorner_{\mathcal{F}}(\hat{\tau}, \tau_0) = o_p(1)$ and $lim\ P[\hat{\tau} \in \mathcal{F}] = 1$ directly provides $\overline{lim}P[Z_n > \eta] < \eta$, completing the proof.

proposition 5 bis

Using proposition 2 step 1 and the same approach as in Pagan and Ullah 1999(page 29) yields (componentwise):

$Bias[HS_*(\theta_0, \tau)] = \frac{\mu_r}{r!}h^r E[\lambda w w' f^{(r)}_{x,z}(0)] + O(h^r)\iota_{K+1}\iota'_{K+1}$

and

$Var[HS_*(\theta_0, \tau)] = \frac{1}{nh}\int K^2(t)dt E[\lambda w_{22} w'_{22} f_{x,z}(0)] + O(\frac{1}{nh})\iota_{K+1}\iota'_{K+1}$

where ι_{K+1} is the $K+1$ by 1 vector where all entries are equal to 1. It follows that $\mathfrak{L}$,the asymptotic mean squared error of $HS_*(\theta_0, \tau)$ is given by:

$$\mathfrak{L} = h^{2r}M_1 + (nh)^{-1}M_2$$

where M_1 and M_2 are as defined in proposition 5 bis. Hence we obtain :

$$|||\mathfrak{L}|||^2 = h^{4r}|||M_1|||^2 + (nh)^{-2}|||M_2|||^2 + \frac{2h^{2r}}{nh} < M_1, M_2 >$$

with $< M_1, M_2 >= tr(M_1 M_2)$. Since $r > 3$ and both a and b are positive the first order condition suffices to minimize our loss and is given by :

$$\frac{\partial |||\mathfrak{L}|||^2}{\partial h} = 0 \text{ if and only if } L(h^{2r+1}) = 0$$

where L is a degree 2 polynomial such that $L(X) = 2arn^2X^2 + bn(2r-1)X - c$ where a,b and c are as defined in proposition 5 bis. Hence, the optimal bandwidth is $X_*^{\frac{1}{2r+1}}$ where $X_* = \frac{b(1-2r)+\sqrt{(}QB)}{n}$ is the positive root of $L(.)$ which is elementary to derive.Simplifying immediately yields the optimal bandwidth.

Proposition 6

we note $V = Y - g - \Gamma'\beta_0$,$w = Z - \Gamma$and$||T\Psi|| = sup_{k=1..K}||T\Psi_k||_\infty$whenever $\Psi' = (\Psi_1(X,Z), ..., \Psi_K(X,Z))$.

Finally f refers to $f_{x,z}(0)$ and the sequences of functions $K_n(t) = \frac{1}{h}K(t/h)$ and $d_n(t) = d(t/h)$ are used.

The consistency of $\beta_* = Argmin_{\mathbf{B}} \sum_{i=1}^n \lambda_i f_i \rho_n(Y_i - g_i - \Gamma_i'\beta_0 - (Z_i - \Gamma_i)'\beta)$ can be established as in proposition $2(i)$. First,$\beta_0 = Argmin_{\mathbf{B}} S(\beta)$ derives from proposition 1 using instead $S(\beta) = E[\lambda f \rho(V - w'\beta)]$,$\nabla S(\beta) = 2E[\lambda w f(F_{x,z}(w'\Delta) - q)]$ and $HS(\beta) = 2E[\lambda f w w' f_{x,z}(w'\Delta)]$ where $\Delta = \beta - \beta_0$ yielding β_0 as the sole local minimum because $E[\lambda E_x f^2 ww']$ is positive definite by $H2(i)$ and $P[f|w'\Delta| > 0] = P[|w'\Delta| > 0]$. Using proposition $2(i)$ we have $sup_{\mathbf{B}}|\hat{E}[\lambda f \rho_n(V - w'\beta)] - S(\beta)| = o_p(1)$ establishing the consistency of β_*.The asymptotic normality of β_* follows from proposition $2(ii)$ using instead $HS_*(\beta) = \frac{2}{n}\sum \lambda_i f_i w_i w_i' K_n(w_i'\Delta)$ whose almost sure convergence to $2E[\lambda f w w' f_{x,z}(w'\Delta)]$ (uniformly over $\mathbf{B}$)is direct from proposition $2(ii)step1$ and $-\sqrt{n}\nabla S_*(\beta_0) = \frac{2}{\sqrt{n}}\Sigma\lambda_i f_i w_i[q - d_n(\varepsilon_i)] \rightsquigarrow \mathcal{N}(0, 4q(1-q)E[\lambda f^2 ww'])$ can be established using the same approach as in $2(ii)step2$ by a double application of the *Dominated Convergence Theorem* and the fact that we choose $h = O(n^{-p})$ for some $p > 1/2r$.Hence,$\sqrt{n}(\beta_* - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{V}\mathcal{B}_\lambda)$ follows.

The proof of $\tilde{\beta}_*$ consistency needs further effort than proposition 4.we have:

$||\hat{S} - \tilde{S}||_{sup\mathbf{B}} \leq ||\hat{f}||_\infty\{||\hat{g} - g||_\infty + ||\beta_0||.sup_{\mathcal{X}*}||\hat{\Gamma} - \Gamma|| + ||\hat{\beta} - \beta_0||sup_{\mathcal{X}*}||\hat{\Gamma} - \Gamma|| + ||\beta_0||sup_{\mathcal{X}*}||\hat{\Gamma} - \Gamma|| + sup_{\mathbf{B}}||\beta||sup_{\mathcal{X}*}||\hat{\Gamma} - \Gamma||\} + sup_{\mathbf{B}}\frac{1}{n}\sum \rho(V_i - w_i'\beta)||\hat{f} - f||_\infty.(*)$

where $\tilde{S}(\beta) = n^{-1}\sum_{i=1}^{n}\lambda_i\hat{f}_i\rho(\hat{V}_i - \hat{w}_i'\beta)$ is the counterpart to $\hat{S}(\beta)$ when the nuisance functions are estimated.To show $plim||\hat{S} - \tilde{S}||_{sup\mathbf{B}} = 0$ we invoke the fact that $||T|| \leq 1$ and assumptions $H4(ii)$ [25]which imply that for any $(\Psi_1, \Psi_2) \in Ł^{\infty}_{X,Z}(\Omega)^2$ we have:

$$||T\Psi_1 - \hat{T}\Psi_2||_{\infty} \leq ||\Psi_1 - \Psi_2||_{\infty} + ||\hat{T} - T||.||\Psi_2||_{\infty}.$$

Applying this last inequality yields:

$$sup_{\mathcal{X}*}||\hat{\Gamma} - \Gamma|| \leq O||\hat{f} - f||_{\infty} + O_p(1)||\hat{T} - T||(**)$$

Using (**) and rearranging (*) provides:

$$||\hat{S} - \tilde{S}||_{sup\mathbf{B}} \leq O_p(1)||\hat{f} - f||_{\infty} + O_p(1)||\hat{g} - g||_{\infty} + O_p(1)||\hat{T} - T|| + o_p(1)$$

Hence,$||\hat{S} - \tilde{S}||_{sup\mathbf{B}} = O_p(n^{-min(a,b,\gamma)})$ establishing $plim||\hat{S} - \tilde{S}||_{sup\mathbf{B}} = 0$ and the consistency of $\tilde{\beta}_*$,the minimizer of $\tilde{S}_*$,follows using the analogue of proposition 4 with the aid of two triangular inequalities showing that $||\tilde{S}_* - S||_{sup\mathbf{B}}$ is dominated by three random variables,all of which $o_p(1)$.Finally, the asymptotic efficiency of $\tilde{\beta}_*$ is derived using $n^{1/2}(\tilde{\beta}_* - \beta_0) = HS_*(\ddot{\beta}, \hat{\tau})^{-1}\{-n^{1/2}\nabla S_*(\beta_0, \hat{\tau})\}$ $wpa.1.$ for some $\ddot{\beta}$ and lemma 3's approach, which yields :

$$sup_{\mathbf{B}}|||HS_*(\beta, \hat{\tau}) - HS_*(\beta, \tau)||| = O_p(h^{-2}n^{-min(a,b,\gamma)}) = o_p(1)$$

due to assumptions $H3$,$H4$ and $H7$ and subsequently $plimHS_*(\ddot{\beta}, \hat{\tau}) = HS(\beta_0)$for $plim\ddot{\beta} = \beta_0$.It then follows by assumptions $H6$ that$\sqrt{n}(\tilde{\beta}_* - \beta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{V}\mathcal{B}_{\lambda})$.

[25]Notice that $||\hat{T} - T||$exists a.s. because the trimming restrict the operator to have as range only $\mathcal{X}*$ supported functions i.e. $T_{res}\varphi = T\varphi 1_{\mathcal{X}*}$ and $||\hat{T}\Psi - T\Psi||_{\infty,\mathcal{X}*} \leq sup_{\mathcal{X}*}sup_{\mathcal{Z}}|\hat{f}_x(z) - f_x(z)|\ell(\mathcal{Z})||\Psi||_{\infty}$ where $\ell(\mathcal{Z})$ is the Lebesgue measure(in $\mathbb{R}^K$)of $\mathcal{Z}$and κ in (V) is strictly positive so the supremum in question exists.

Lemmas

Lemma 1:

let $\rho_n(u) = 2(qu + \varphi_n(u))$ *where* $\varphi_n(u) = h\varphi(u/h)$ *for some* $\varphi \in \mathfrak{F}_\mathfrak{r}$ *and some* $h = o(1)$. *Then (i)* $|\rho_n - \rho|_{sup\mathbb{R}} = O(h)$ *and (ii) lim* $d(u/h) = 1_{u<0}$ *a.e.*

proof: without loss of generality we are to show the case where r=4. The only difference deals with the number of roots of the polynomials Q on (-1,1). So let $K = Q1_{[-1,1]}$ where Q of degree 4, symmetric with one root in (0,1) we note ζ. Thus, Q will be decreasing on $(0, \zeta)$ and increasing on $(\zeta, 1)$.

let $d(x) = \int K(t)1_{t>x}dt$. By construction d is equal to 0 on $[1, \infty)$ and 1 on $(-\infty, -1]$ while the monotonicity on (-1,1) is given by the Fundamental Theorems of Calculus (FTC) as $d' = -Q$. Using the previous properties of Q yields that d is increasing on $(-1, -\zeta) \cup (\zeta, 1)$ while decreasing on $(-\zeta, \zeta)$.Notice that $d(0) = 1/2$.

Let $\varphi(u) = \int d(x)1_{x>u}dx$. By construction φ is 0 on $[1, \infty)$ while the monotonicity on $(-\infty, 1)$ is derived from the FTC as $\varphi' = -d$. It follows that φ will be increasing on $(\mu, 1)$ for some $\mu > 0$ and decreasing on $(-\infty, \mu)$. In other words,φ behaves almost like the negative part function which is the idea behind the approximation of the "Check function".

Finally, let $\varphi_n(u) = h\varphi(u/h)$ and $H(u) = -u1_{u<0}$.

We are to show that φ_n converges uniformly to H.Let $u \geq 0$. Because $u/h \geq 0$ will always hold and φ is bounded on $[0, \infty)$ we have $|\varphi_n(u)| \leq h|\varphi|_{Sup\mathbb{R}+}$ and thus $|\varphi_n - H| = O(h)$ uniformly when $u \geq 0$.Let $u < 0$. Examining $\varphi_n - H$ when $u < 0$ (using the properties of d) yields $|\varphi_n - H| \leq max\{(\varphi_n(0); |\varphi_n(\lambda h) - H(\lambda h)|\} = O(h)$ for λ somewhere in (-1,0) meeting $d(\lambda) = 1$.

Consequently, $sup_{u\in\mathbb{R}}|\varphi_n(u) - H(u)| = O(h)$ and the Lemma follows directly.

(ii) lim $d(u/h) = 1_{u<0}$ a.e.

for $u > 0$ we have lim $u/h = \infty$ yielding lim $d(u/h) = 0$. For $u < 0$ we get lim $u/h = -\infty$ and hence lim $d(u/h) = 1$. The almost everywhere convergence arises due to $d(0) = 1/2$.

Lemma 2:

Under the assumptions of the model we have $E_{x,z}d(\varepsilon/h) - q = O(h^r)$ *a.s. for sufficiently large n.*

proof:let $\hat{f}(e) = \frac{1}{nh}\sum K(\frac{\varepsilon_i - e}{h})$ be the non parametric estimator of $f(e)$ where f is the density of the error term. Using our iid assumptions for $\{\varepsilon_i\}_{i=1}^n$ we get:

$E_{x,z}d(\varepsilon/h) - q = E_{x,z}n^{-1}\sum d(\varepsilon_i/h) - q = E_{x,z}\int(\hat{f}(e) - f_{x,z}(e))1_{e<0}de$

where we used P[$\varepsilon < 0|X, Z$]=q a.s. along with the properties of $d(.)$.

Let us note $b_n(e, x, z) = E_{x,z}\hat{f}(e) - f_{x,z}(e)$. Notice that:

$E_{x,z}\hat{f}(e) = \int h^{-1}K(\frac{\varepsilon - e}{h})f_{x,z}(\varepsilon)d\varepsilon = \int K(t)f_{x,z}(e + th)dt.$

But by assumptions 9 we find:

$f_{x,z}(e + th) = f_{x,z}(e) + P_{x,z}(e - th) + R_{x,z}(e, e + th).$

where $P_{x,z}(e-th)$ is the Taylor's approximation of $f_{x,z}(e+th)$ around e at order r-1 and $R_{x,z}(e, e+th)$ its reminder. Hence, our Kernel of order r results in:

$b_n(e, x, z) = \int R_{x,z}(e, e + th)K(t)dt$

Finally, using the compact support of our Kernel ensures that for almost all (x,z) there exists a strictly positive constant $c_{x,z}$ and natural number $n(x, z)$ such that:

$|b_n(e, x, z)| \leq \int |R_{x,z}(e, e + th)||K(t)|1_{|th|<c_{x,z}}dt$ for $n > n(x, z)$

and consequently :

$|b_n(e, x, z)| \leq h^r\psi_{x,z}(e)\int |t^rK(t)|dt$ holds almost everywhere on $\mathcal{X} \times \mathcal{Z}$ for n large enough and some integrable function $\psi_{x,z}(.)$ due to assumption 9. It follows that $\int |b_n|de = O(h^r)$ a.s. for large n and this establishes Lemma 2.

Lemma 3:

Under assumptions 1-10 $sup_\Theta|||HS_*(\theta, \hat{\tau}) - HS_*(\theta, \tau)||| = o_p(1)$

proof:let $K_n(t) = h^{-1}K(t/h)$.Also for for $i = 1...n$ the followings will improve the clarity of the proof: $A_i = 2w_iw_i'$; $\hat{A}_i = 2\hat{w}_i\hat{w}_i'$;$U_i(\theta) = T_i - w_i'\theta$;$\hat{U}_i(\theta) = \hat{T}_i - \hat{w}_i'\theta$.

we have $HS_*(\theta,\hat{\tau}) - HS_*(\theta,\tau) = \frac{1}{n}\sum A_i K_n(U_i(\theta)) - \frac{1}{n}\sum \hat{A}_i K_n(\hat{U}_i(\theta))$. The triangular inequality further provides :

$sup_\Theta |||HS_*(\theta,\hat{\tau}) - HS_*(\theta,\tau)|||$

$\leq sup_\Theta |||\frac{1}{n}\sum A_i K_n(U_i(\theta)) - \frac{1}{n}\sum \hat{A}_i K_n(U_i(\theta))||| + sup_\Theta |||\frac{1}{n}\sum \hat{A}_i K_n(U_i(\theta)) - \frac{1}{n}\sum \hat{A}_i K_n(\hat{U}_i(\theta))|||$

Hence, $sup_\Theta |||HS_*(\theta,\hat{\tau}) - HS_*(\theta,\tau)||| \leq H_{1,n} + H_{2,n}$

where $H_{1,n} = \frac{1}{n}\sum |||A_i - \hat{A}_i|||sup|K_n| \leq O(1/h)|||A_i - \hat{A}_i|||_{supi=1..n} \leq O(1/h)O_p(1/\sqrt{n})$.

and

$H_{2,n} \leq \frac{1}{n}\sum |||\hat{A}_i|||sup|K_n^{(1)}|sup_\Theta|U_i(\theta) - \hat{U}_i(\theta)|$

$\leq O(h^{-2})\frac{1}{n}\sum |||\hat{A}_i|||\{sup_{x\in\mathcal{X}}|\hat{M} - M)| + Bsup_{x\in\mathcal{X}}||\hat{\vartheta} - \vartheta||\}$ where B is simply a constant due to the compactness of Θ.

Consequently, $H_{2,n} \leq O_p(\frac{1}{h^2 n^{1/2}})\frac{1}{n}\sum |||\hat{A}_i||| \leq O_p(\frac{1}{h^2 n^{1/2}})$.

Because $h = O(1/n^p)$ for some $p < 1/4$ we conclude that plim $H_{1,n} + H_{2,n} = 0$ which establishes Lemma 3.

Lemma 4:

$R_{1,n} + R_{2,n} + E_n = o_p(1)$

proof:

1. $R_{1,n} = o_p(1)$ by Lemma 5
2. $R_{2,n} = o_p(1)$

   We show $e'_{-1}R_{2,n} = o_p(1)$ since the proof of $e'_1 R_{2,n} = o_p(1)$ is similar.let a_n be the vector of dimension n where the i^{th} entry is $K_n^{(1)}(\xi_i)\Delta_i^2$. Since $||R_{2,n}||^2 = \frac{1}{n}a'_n(Z-\vartheta)(Z-\vartheta)'a_n$ (where $Z - \vartheta$ is the n by K Matrix of residuals from the projection of Z on X) we obtain :

   $||R_{2,n}||^2 \leq \lambda_{max}[\frac{1}{n}(Z-\vartheta)(Z-\vartheta)']||a_n||^2$

   where $\lambda_{max}(A)$ indicates the largest eigenvalues of a symmetric Matrix A.

   Furthermore, $||a_n||^2 \leq \sum \frac{1}{h^4}\Delta_i^4 = \frac{1}{nh^4}O_p(1)$ and $\frac{1}{n}(Z-\vartheta)'(Z-\vartheta) \longrightarrow M$ a.s. (by Kolgomorov's strong law of large numbers) where M is definite positive by assumption 8. Thus, $\lambda_{max}[\frac{1}{n}(Z - \vartheta)(Z-\vartheta)'] = O_p(1)$ and subsequently $||R_{2,n}||^2 \leq \frac{1}{nh^4}O_p(1)$ so that $||R_{2,n}||^2 = o_p(1)$ will hold due to our choice for h.

3. $R_{3,n} = o_p(1)$

   $||R_{3,n}||^2 = \frac{1}{n} d' \Delta\vartheta \Delta\vartheta' d$where d is a n by 1 vector whose i^{th} entry is $d_n(\varepsilon_i)$ while $\Delta\vartheta$ is n by K matrix whose k^{th} columns records the first stage "mistakes" on the conditional mean of Z_k on X.

   Because $\Delta\vartheta$ is measurable in $\{X_i, Z_i\}_{i=1...n}$ we have:

   $E_{X_i,Z_i}||R_{3,n}||^2 = \frac{1}{n} tr\{\Delta\vartheta\Delta\vartheta' E_{X_i,Z_i} dd'\}$.

   Additionally, the iid property of our errors combined to Lemma 2 gives $E_{X_i,Z_i} dd' = O(1)I_n + O(h^2)C$ where I_n is the identity matrix of dimension n while C is the n by n matrix whose diagonal is 0 and 1 elsewhere. Finally, using $sup_{x\in\mathcal{X}}\{\sqrt{n}(\hat{M} - M)\} = O_p(1)$ and $sup_{x\in\mathcal{X}}\{\sqrt{n}||\hat{\vartheta} - \vartheta||\} = O_p(1)$ yield:

   $\Delta\vartheta\Delta\vartheta' = O_p(\frac{1}{n})\Xi$

   where Ξ is the n by n matrix where all entries are equal to 1. Hence, $E_{X_i,Z_i}||R_{3,n}||^2 = \frac{1}{n} tr\{O_p(\frac{1}{n})\Xi[O(1)I_n + O(h^2)C]\}$.

   Noticing $\Xi C = (n-1)\Xi$ and keeping the largest order supplies $E_{X_i,Z_i}||R_{3,n}||^2 = \frac{1}{n} tr\{O_p(\frac{1}{n})\Xi\} = O_p(\frac{1}{n})$ and this achieves our objective by Dominated Convergence because $E_{X_i,Z_i}||R_{3,n}||^2 = O(1)$.

4. $R_{4,n} = o_p(1)$

   $||R_{4,n}|| \leq n^{1/2} sup_{x\in\mathcal{X}}||\hat{\vartheta} - \vartheta|| sup_{x\in\mathcal{X}}|\Delta_x| \frac{1}{n}\sum |K_n(\varepsilon_i)|$.

   Simplifying provides:

   $||R_{4,n}|| \leq sup_{x\in\mathcal{X}}|\Delta_x| O_p(1) = o_p(1)$

5. $R_{5,n} = o_p(1)$

   $||R_{5,n}|| \leq \frac{1}{\sqrt{n}} \sum sup|K_n^{(1)}| sup_{x\in\mathcal{X}}|\Delta_x|^2 sup_{x\in\mathcal{X}}||\hat{\vartheta} - \vartheta||$

   $\leq O(h^{-2}) O_p(1/n)$

   hence, $||R_{5,n}|| = O_p(\frac{1}{nh^2}) = o_p(1)$.

Lemma 5: *Under Assumption 11 we have plim* $\frac{1}{\sqrt{n}}\sum w_i K_n(\varepsilon_i)\hat{\Delta}_i = 0$

proof: let $\delta > 0$ be arbitrary. By assumption 11 there exists $\varepsilon > 0$ such that $P[sup_{\Theta_n(\varepsilon)}|\upsilon_n(\Delta)| > \delta] < \delta$ for n sufficiently large. Using basic probabilities inequalities we must also have:

$P[|\upsilon_n(\hat{\Delta})| > \delta] = P[|\upsilon_n(\hat{\Delta})| > \delta \cap |\hat{\Delta}|_\infty \leq \varepsilon] + P[|\upsilon_n(\hat{\Delta})| > \delta \cap |\hat{\Delta}|_\infty > \varepsilon]$

$\leq P[|\upsilon_n(\hat{\Delta})| > \delta \cap |\hat{\Delta}|_\infty \leq \varepsilon] + P[|\hat{\Delta}|_\infty > \varepsilon]$

$\leq P[sup_{\Theta_n(\varepsilon)}|\upsilon_n(\Delta)| > \delta] + P[|\hat{\Delta}|_\infty > \varepsilon]$

Using $sup_{x\in\mathcal{X}}\{\sqrt{n}(\hat{M} - M)\} = O_p(1)$,$sup_{x\in\mathcal{X}}\{\sqrt{n}||\hat{\vartheta} - \vartheta||\} = O_p(1)$ and $|\hat{\Delta}|_\infty = o_p(1)$ implies therefore the existence of a sample size $n*$ such that $\sup_{n\geq n*} P[|\upsilon_n(\hat{\Delta})| > \delta] < \delta$ Q.E.D.

Notice that we have not used the outer probability because for each sample size n the map $sup_{|\Theta_n(\varepsilon)}|\upsilon_n(\Delta)|$ is measurable due to our maximizing of a continuous function over $\Theta_n(\varepsilon)$ compact(Jenrich 1969,Lemma 2.1).

Lemma 6:

Under the assumptions of proposition 1

θ_* *converges almost surely to* θ_0

proof:Nothing original is presented here. We provide a proof restating in the context of our model known results for M estimators where the centered empirical moment is Liptschiz in the parameter(Mc Fadden 1991, Mc Fadden Newey 1994,Andrews 1992)and the space of parameters compact.It is understood that that the classic measurability conditions are met for $||\hat{S} - S||_{sup\Theta}$ (Jenrich 1969,lemma 2.1.).

θ_0 being the global minimum of S on Θ implies that for an arbitrary $\epsilon > 0$ there exits $\delta > 0$ such that $inf_{\{\theta:||\theta-\theta_0||>\epsilon\}}S(\theta) - S(\theta_0) \geq \delta$ and consequently $P[||\theta_* - \theta_0|| > \epsilon] \leq P[|S(\theta_*) - S(\theta_0)| \geq \delta]$. Additionally, one can show with two triangular inequalities that $|S(\theta_*) - S(\theta_0)| \leq 2||S_* - S||_{sup\Theta}$ due to the fact that $S_*(\theta_*) - S_*(\theta_0) \leq 0$ .Hence,$||S_* - S||_{sup\Theta}$ convergence almost surely to 0 would establish the lemma. Since $||S_* - \hat{S}||_{sup\Theta} = O(h)$ we need only to show to show $P[\omega \in \Omega : lim_n||\hat{S} - S||_{sup\Theta}(\omega) = 0] = 1$.

Under the assumptions of proposition 1,one can use $sup_\Theta||\nabla S(\theta)|| < \infty$ and the basic inequality $||a| - |b|| \leq |a - b|$ for any a, b real numbers to establish that for any $(\theta_1, \theta_2) \in \mathbb{R}^{2(K+1)}$ we have $|Q_n(\theta_1) - Q_n(\theta_2)| < C||\theta_1 - \theta_2||$ for some positive constant C, where $Q_n(\theta) = \hat{S}(\theta) - S(\theta)$ is the

centered empirical moment.Let $\varepsilon > 0$ arbitrary.Using the fact that Θ is compact[26] permits to invoke the Heine Borel Theorem to affirm the existence of a finite open covering of Θ. That is,$\Theta \subseteq \bigcup_k B(\theta_k, \varepsilon/2C)$ for some $\{\theta_k\}_{k=1..K}$ where $B(\theta_k, \varepsilon/2C) = \{\theta : ||\theta - \theta_k|| < \varepsilon/2C\}$.Since for any $\theta \in \bigcup_k B(\theta_k, \varepsilon/2C)$ implies $\theta \in B(\theta_k, \varepsilon/2C)$ for some θ_k, we further obtain:

$$|Q_n(\theta)| \leq |Q_n(\theta) - Q_n(\theta_k)| + |Q_n(\theta_k)| < \varepsilon/2 + |Q_n(\theta_k)|$$

Hence, we have :

$$sup_{\bigcup B(\theta_k, \varepsilon/2C)} |Q_n(\theta)| < \varepsilon/2 + sup_{k=1..K} |Q_n(\theta_k)|.$$

But, the iid sampling assumptions and $E|\rho(\varepsilon(\theta)| < \infty$ uniformly over Θ provides $|Q_n(\theta_k)| \longrightarrow 0$ almost surely by Kolmogorov strong law of large numbers. Hence, for any k $\in \{1...K\}$ there is a null set A_k such that $limQ_n(\theta_k)(\omega) = 0$ for all $\omega \in \Omega \backslash A_k$. It follows that for all $\omega \in \Omega \backslash \cup A_k$ there exists a sample size $n_k(\omega)$ such that $n \geq n_k(\omega)$ implies $|Q_n(\theta_k)(\omega)| < \varepsilon/2$ so that $n \geq N(\omega) = max_{k=1...K} n_k(\omega)$ exists to ensure $sup_{\bigcup B(\theta_k, \varepsilon/2C)} |Q_n(\theta)(\omega)| < \varepsilon$.

Since ε was arbitrary chosen we get $\Omega \backslash \cup A_k \subseteq \{\omega \in \Omega : lim||\hat{S} - S||_{sup\Theta}(\omega) = 0\}$which combined to $P[\omega \in \Omega \backslash \cup A_k] = 1$ yields $P[\omega \in \Omega : lim||\hat{S} - S||_{sup\Theta}(\omega) = 0] = 1$ Q.E.D.

[26]A totaly Bounded parameter space suffices to invoke the finite covering property,which is why the closeness of Θ imposed in assumption 4 can be relaxed(Andrews 1992).