Efficient Multivariate Quantile Regression Estimation

Sung Jae Jun*and Joris Pinkse†
The Pennsylvania State University

September 2006

Abstract

We propose an efficient semiparametric estimator for the multivariate linear quantile regression model in which the conditional joint distribution of errors given regressors is unknown. The procedure can be used to estimate multiple conditional quantiles of the same regression relationship. The proposed estimator is asymptotically as efficient as if the conditional distribution were known. Simulation results suggest that the estimation procedure works well in practice and dominates an equation–by–equation efficiency correction if the errors are dependent conditional on the regressors.

*Department of Economics, The Pennsylvania State University, 608 Kern Graduate Building, University Park PA 16802, sjun@psu.edu

†joris@psu.edu We thank participants at the Carnegie Mellon departmental seminar and Ari Kang for their useful suggestions.

1 Introduction

We propose an efficient semiparametric estimator for the multivariate linear quantile regression model in which the conditional joint distribution of errors given regressors is unknown. The procedure can be used to estimate multiple conditional quantiles of the same regression relationship. The proposed estimator is asymptotically as efficient as if the conditional distribution were known. Simulation results suggest that the estimation procedure works well in practice and dominates an equation–by–equation efficiency correction if the errors are dependent conditional on the regressors.

The proposed method entails the nonparametric estimation of optimal instruments for a set of moment conditions corresponding to the conditional quantiles of interest and subsequently using these estimated optimal instruments to obtain the efficient quantile estimates. Two–step efficiency corrections like this go back to at least Aitken (1935), but semiparametric corrections like ours have been around for a while, also. Carroll (1982), Delgado (1992) and Robinson (1987) achieve full GLS[1] asymptotic efficiency by estimating the conditional error variance function nonparametrically. Newey (1990, 1993) proposes methods for estimating optimal instruments nonparametrically, thereby allowing for multivariate regressions and ones with endogenous regressors. Pinkse (2006) introduces a method which addresses the *curse of dimensionality* associated with the nonparametric estimation of functions with many arguments. Finally, Zhao (2001), Whang (200?) and Komunjer and Vuong (2005) propose efficiency corrections for the univariate median regression model.

The multivariate quantile regression case is of interest for applied work for several reasons. First, even absent dependence between errors and regressors quantile regression estimators tend to have greater asymptotic variances than mean regression ones[2] and efficiency improvements are hence more valuable. Further, an optimal parametric correction in the mean regression model requires one to guess the correct parametric form of the conditional variance function (matrix–valued in the multivariate case), which is difficult since little reliable information may be available as to its shape. In the quantile case, one would need to know the marginal conditional error densities at zero plus, in the multivariate case, the probability for each pair of errors that both are negative, conditional on the regressors. It is even more unrealistic for an empirical researcher to possess that much information; incorrect guesses will lead to inefficient estimators, quite possibly to ones that have lesser asymptotic efficiency than uncorrected ones.[3] Unless the errors are independent conditional on the regressors *and* there are no cross–equation restrictions on the regression coefficients, multivariate efficiency corrections are moreover generally more efficient than univariate ones. Finally, with quantile estimation it is possible to estimate multiple quantiles of the same regression relationship (i.e. the same dependent variable and the same regressors) simultaneously, which would imply strong dependence between the corresponding errors and hence more scope for efficiency

[1]Generalized Least Squares

[2]The asymptotic relative efficiency for a median regression estimator versus a median regression estimator for a model with normally distributed errors is $2/\pi$. Please note that median and mean regression estimators typically estimate different coefficients.

[3]In the univariate case it can be reasonable to assume that errors factor as the product of a function of regressors and some error independent of the regressors, see e.g. Koenker (2005), section 5.3.2, and Koenker and Zhao (1994).

improvements.

Like all of the above semiparametric estimators ours relies on the availability of a $\sqrt{n}$–consistent first round estimator; a natural choice is the standard quantile regression estimator. A problem with such a two–step procedure is that the first round estimation error, while asymptotically absent, can be such that correction is not worthwhile in small samples. This is especially true when the number of regressors is large due to the fact that nonparametric estimators of high–dimensional functions are notoriously inaccurate. Please note however, that our correction does not require (nor do we establish) pointwise consistent estimation of the optimal instruments and since the uncorrected estimates are special cases of the correction procedure for particular values of the input parameters of the semiparametric procedure, the semiparametric procedure is in principle never worse irrespective of the sample size. Please note, however, that we offer no procedure for the optimal selection of the input parameters; our simulation results indicate that the performance is comparatively insensitive to their choice.

This paper contains several theoretical innovations. While Newey (1990, 1993) allows for multiple equations to be estimated jointly, his results do not cover the current case because of the nondifferentiability of the optimal instruments. Zhao (2001), Whang (200?) and Komunjer and Vuong (2005) propose estimators for the single equation case. In the single equation case the nuisance function is just conditional error density at zero instead of the product of a matrix and the inverse of another matrix, as is the case here. Whang (200?) and Komunjer and Vuong (200?) achieve the semiparametric efficiency bound (the latter for time series) by optimizing an objective function involving a series expansion of the nuisance function; the nondifferentiability problems we solve do not arise then.

Our paper is closer to Zhao (2001) in that we use a nonparametric plugin estimator. The nondifferentiability issue is only partly addressed by Zhao (2001); Zhao's results rely on *sample splitting*. He requires that the first step estimator used to estimate the weights for half the observations is computed using only the other half and vice versa. Although sample splitting does not affect the asymptotic efficiency it is likely to have an effect in samples of finite size and is more cumbersome. Our results obviate the need for sample splitting with Zhao's (2001) estimator, also, since his estimation problem is a special case of ours.

The new proof (contained in the last two lemmas of Appendix C and using L1 of Appendix A) entails ratcheting up of the established uniform convergence rate of the feasible estimator of the moment condition and the feasible estimator of the parameter vector of interest alternately. This method of proof has uses that go well beyond the particular problem at hand or indeed differentiability problems or ones involving nonparametric estimation.

To compute our estimates we use a procedure which involves a standard linear programming problem followed by one or more Newton steps. The procedure is guaranteed to yield estimates satisfying our constraints — we prove this — and does so fast; computing the nonparametric weights takes the most time. The reason that computation here is simple, in contrast to e.g. Chernozhukov and Hansen's (2006) estimator, is that we have an initial easily computable $\sqrt{n}$–consistent but inefficient estimator at our disposal, namely the standard least absolute deviations estimator. The Matlab code is available from the authors on request.

The outline of the paper is as follows. In section 2 we introduce the setup and define our estimator. Section 3 contains the theoretical results for our estimator, whose computation and performance are studied in sections 4 and 5, respectively. Section 6 concludes.

2 Model and Estimator

Let $\{y_i, X_i\}$ be an i.i.d. sequence for which

$$Q(y_i|X_i) = X_i'\theta_0 \text{ a.s.}, \quad i = 1, \ldots, n, \tag{1}$$

or equivalently,

$$y_i = X_i'\theta_0 + u_i, \quad Q(u_i|X_i) = 0 \text{ a.s.}, \quad i = 1, \ldots, n, \tag{2}$$

where $y_i \in \mathbb{R}^d$, $X_i \in \mathbb{R}^{K \times d}$ and Q denotes the vector of quantiles of interest.

The formulations in (1) and (2) allow for several possibilities. The restriction that the regression coefficients are the same in all regression equations is not restrictive because we can make the choices

$$X_i = \begin{bmatrix} x_{i1} & & \\ & \ddots & \\ & & x_{id} \end{bmatrix}, \quad \theta_0 = \begin{bmatrix} \theta_{01} \\ \vdots \\ \theta_{0d} \end{bmatrix},$$

resulting in

$$y_{ij} = x_{ij}'\theta_{0j} + u_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, d. \tag{3}$$

So (1) allows for arbitrary amounts of overlap between the vectors of regression coefficients across equations. An assumption implicit in (1) is that the regressors in equation ℓ do not enter the conditional quantile function in equation $j \neq \ell$ insofar the two regressor vectors do not overlap. This is where part of the efficiency gain originates; it is akin to an orthogonality condition between regressors and errors across equations in the mean regression case.[4] It is possible to choose $y_{ij} = y_{i\ell}$, $x_{ij} = x_{i\ell}$, $j \neq \ell$, for all i in (3) if different regression quantiles of the same regression relationship are desired. Assuming multiple quantiles of the same relationship to all be linear, however, imposes strong restrictions on the types of dependence between errors and regressors that can be accomodated and a procedure that exploits such restrictions will likely work better in practice than the more general procedure proposed here; a more fruitful avenue would be to estimate the median and mean jointly, a possibility not covered by our results.

We now formulate an infeasible efficient estimation procedure for θ_0. Let $s_i(\theta) = I(y_i \leq X_i'\theta) - \tau$, where τ is the vector indicating which quantiles are desired (a vector with values 0.5 in case of the median) and I is the *indicator function*, where for any $v \in \mathbb{R}^{d_v}$, $I(v) = [I(v_1), \ldots, I(v_{d_v})]'$. Then the conditional moment condition is ($s_i = s_i(\theta_0)$)

$$E(s_i|X_i) = 0 \text{ a.s.}.$$

The corresponding optimal unconditional moment conditions are

$$E(A_i s_i) = 0, \tag{4}$$

[4] It is possible to obtain efficiency improvements when the conditional quantiles do not depend on some but not all of the regressors in another equation; this possibility can be accomodated in our setup by a judicious choice of y and X.

where $A_i = S_i' T_i^{-1}$ with

$$S_i = F_i X_i', \quad F_i = \begin{bmatrix} f_{u_{i1}|X_i}(0) & & \\ & \ddots & \\ & & f_{u_{id}|X_i}(0) \end{bmatrix}, \quad T_i = E(s_i s_i' | X_i). \tag{5}$$

The asymptotic variance of an infeasible estimator $\hat{\theta}_I$ based on (4) will later be shown to be V^{-1} with

$$V = E(A_1 s_1 s_1' A_1') = E(S_1' T_1^{-1} S_1). \tag{6}$$

The proposed procedure yields a natural efficiency improvement over equation–by–equation estimation when there is overlap between the regression coefficients across equations. Absent such overlap, the asymptotic variance of $\hat{\theta}_{I1}$, the infeasible estimator of the first subvector θ_{01}, is for $d = 2$ equal to

$$V_{I1} = \Big(E[t_i^{11} f_{i1}^2 x_{i1} x_{i1'}] - E[t_i^{12} f_{i1} f_{i2} x_{i1} x_{i2}'] \big\{ E[t_i^{22} f_{i2}^2 x_{i2} x_{i2}'] \big\}^{-1} E[t_i^{12} f_{i1} f_{i2} x_{i2} x_{i1}'] \Big)^{-1},$$

where $f_{ij} = f_{u_{ij}|X_i}(0)$ and $t_i^{j\ell}$ is the (j, ℓ)–element of

$$T_i^{-1} = \frac{1}{t_{i11} t_{i22} - t_{i12}^2} \begin{bmatrix} t_{i22} & -t_{i12} \\ -t_{i12} & t_{i11} \end{bmatrix}; \quad T_i = \begin{bmatrix} t_{i11} & t_{i12} \\ t_{i12} & t_{i22} \end{bmatrix}.$$

The corresponding asymptotic variances for the inefficient and efficient single equation estimators are

$$V_{SI1} = \tau_1(1-\tau_1) \big(E[\tilde{f}_{i1} x_{i1} x_{i1}'] \big)^{-1} E(x_{i1} x_{i1}') \big(E(\tilde{f}_{i1} x_{i1} x_{i1}') \big)^{-1}, \quad V_{SE1} = \tau_1(1-\tau_1) \big(E[\tilde{f}_{i1}^2 x_{i1} x_{i1}'] \big)^{-1},$$

where $\tilde{f}_{i1} = f_{u_{i1}|x_{i1}}(0)$. It is necessarily true that $V_{I1} \le V_{SE1} \le V_{SI1}$; we now discuss when they are equal. When $\tilde{f}_{i1}$ does not depend on x_{i1}, $V_{SI1} = V_{SE1}$; otherwise equality only occurs in exceptional cases.

Using information from different equations is useful because one can exploit (i) the information that regressors in equation 2 do not impact the conditional quantile of equation 1 and (ii) the fact that u_{i1} and u_{i2} are not necessarily independent conditional on X_i. Consideration (i) can be accomodated in the single equation case (as in Zhao (2001)) by extending the conditioning set to regressors outside of the equation being estimated; in the multivariate case the conditioning vector can likewise be extended (and the efficiency thereby improved) by including variables from outside the system. But (ii) cannot be used in the single equation setup.

So even if the regressors in both equations are the same and $\tilde{f}_{i1} = f_{i1}$, there is still an efficiency gain from our method unless u_{i1}, u_{i2} are independent conditional on X_i,[5] in which case $t_i^{12} = t_{i12} = 0$, or if u_{i1}, u_{i2} do not depend on X_i. Conversely, even if u_{i1}, u_{i2} are independent of X_i there is still an efficiency gain unless $x_{i1} = x_{i2}$. All of this is similar to a SUR model with random regressors where no efficiency gain obtains from joint estimation if the errors are uncorrelated conditional on the regressors or if the regressors

[5] Or more precisely: if $I(u_{i1} \le 0)$ and $I(u_{i2} \le 0)$ are independent conditional on X_i.

are identical and independent of the errors.[6] Table 1 in the appendix contains the full details of when efficiency improvements obtain for the various estimators.

If the errors are known to be independent of the regressors, then no nonparametric correction is needed since only the joint distribution of $I(u_{ij} \leq 0)$ with $I(u_{i\ell} \leq 0)$ for all j, ℓ is needed, and this distribution entails only $d(d-1)/2$ unknowns. The types of dependence between errors and regressors that lead to efficiency improvements is different from the mean regression case. In the mean regression case efficiency improvements obtain only if $\Sigma(X_i) = V(u_i|X_i)$ varies with X_i whereas in the quantile regression case improvements obtain if the conditional error densities at zero vary with X_i or if $P(u_{ij} \leq 0, u_{i\ell} \leq 0|X_i)$ varies with X_i for some j, ℓ. Neither situation implies the other, except in special models like

$$u_i = \big(\Sigma(X_i)\big)^{1/2} e_i, \tag{7}$$

where the elements of e_i are independent with unit variances and $\Sigma_i = \Sigma(X_i)$ is some positive definite matrix. The problem with (7) is that quantiles are generally not invariant to linear transformations, e.g. $\mathrm{Med}(a+b) \neq \mathrm{Med}(a) + \mathrm{Med}(b)$. If the e_i's are mean zero normal, however, then so are the u_i's and their conditional median is zero.[7] With (7), $f_{u_i|X_i}(0) = f_{e_i}(0)/\sqrt{|\Sigma_i|}$ and hence varies with X_i unless Σ_i is constant.

We now proceed with the formulation of our estimators. We begin with the infeasible estimator $\hat\theta_I$ which is defined as any estimator satisfying

$$m_n(\hat\theta_I) = o_p(n^{-1/2}), \quad \text{where } m_n(\theta) = n^{-1}\sum_{i=1}^n A_i s_i(\theta). \tag{8}$$

We do not set m_n equal to zero in (8) because no value of θ may exist that satisfies $m_n(\theta) = 0$ since s_i involves an indicator function. m_n converges to m with

$$m(\theta) = E\big[A_1 s_1(\theta)\big].$$

$\hat\theta_I$ is infeasible since the A_i's in (8) are unknown. We will estimate them and using their estimates $\hat A_i$ we can define $\hat{\hat\theta}$ as any value satisfying

$$\hat m_n(\hat{\hat\theta}) = o_p(n^{-1/2}), \quad \text{where } \hat m_n(\theta) = n^{-1}\sum_{i=1}^n \hat A_i s_i(\theta). \tag{9}$$

The only remaining question is how to estimate A_i. Let $\hat\theta$ be any $\sqrt{n}$–consistent first stage estimator of θ_0, e.g. based on single equation quantile estimation. We estimate T_i, S_i separately using KNN estimators

$$\hat T_i = n^{-1}\sum_{j=1}^n w_{ij}\hat s_j \hat s_j', \quad \hat S_i = n^{-1}\sum_{j=1}^n w_{ij}\hat F_j X_i', \tag{10}$$

[6] In the classical SUR model errors are assumed independent of the regressors, in which case no efficiency gain arises when the regressors are identical or the errors are uncorrelated.

[7] This holds for any class of multivariate distributions that is closed to linear transformations and which are element–wise even.

where $\hat{s}_i = I(\hat{u}_i \le 0) - \tau$, $\hat{F}_i = \operatorname{diag}\big(I(|\hat{u}_i| \le \beta_n \iota)/(2\beta_n)\big)$ with ι a vector of ones, β_n a *bandwidth* parameter, $\hat{u}_i = y_i - X_i'\hat{\theta}$ and w_{ij} a KNN weight,[8] setting $\hat{A}_i = \hat{S}_i'\hat{T}_i^{-1}$.

The KNN weights are all nonnegative and w_{ij} is positive only if observation is among observation i k_n closest neighbors in terms of the distance between X_i and X_j; ties only occur when all regressors are discrete and can be resolved by randomizing among the tying observations. The only other constraints we impose are upper and lower bounds to their values and conditions on the rate at which the number of neighbors should increase.

3 Results

We now discuss our main result, formulated in T3, which shows that the feasible estimator $\hat{\hat{\theta}}$ has a limiting normal distribution with variance V^{-1}. For our main result, we need the following assumptions.[9]

A1 *θ_0 is an interior point of the compact parameter space Θ.*

A2 *For some $C_T > 0$, $P\big(\lambda_{\min}(T_1) \ge C_T\big) = 1$.*

A3 *$E(X_1X_1') > 0$.*

A4 *For some $0 < C_f < \infty$, and all $j = 1, \ldots, d$, $P\big(f_{u_{1j}|X_1}(0) \ge 1/C_f\big) > 0$, $P\big(f_{u_{1j}|X_1}(0) \le C_f\big) = 1$, $P\big(\sup_t \big|f'_{u_{1j}|X_1}(t)\big| \le C_f\big) = 1$ and $P\big(\sup_t \big|f''_{u_{1j}|X_1}(t)\big| \le C_f\big) = 1$.*

A5 *$\forall \theta \in \Theta : m(\theta) = 0 \Leftrightarrow \theta = \theta_0$.*

A6 *The weights w_{ij} are nonnegative and all k_n nonzero weights take values in the range $[1/(C_w k_n); C_w/k_n]$.*

A7 *Let for any $p > 0$, $\zeta_{npT} = n^{1/p_x - 1/2} + n^{1/p} k_n^{-1/2}$ and $\zeta_{npS} = n^{1/p_x} k_n^{-1/2} \beta_n^{1/p_x - 1} + n^{1/p_x} \beta_n^2 + n^{1/2} k_n^{-1} \beta_n$. Then for some $p < \infty$, $\sqrt{n}\zeta_{npT}^2 \to 0$, $\sqrt{n}\zeta_{npT}\zeta_{npS} \to 0$ and $k_n/n \to 0$, as $n \to \infty$.*

A1 and A3 are standard. A2 essentially says that $\operatorname{Corr}\big[I(u_{i1} \le 0), I(u_{i2} \le 0)|X_i\big]$ should be a.s. bounded away from ± 1; this is reasonable and similar to a condition used in Pinkse (2006). The assumption (A4) that the conditional error densities have two uniformly bounded derivatives excludes distributions like the Laplace distribution, but is otherwise reasonable within the context of nonparametric estimation.[10] The assumption that the conditional densities at zero are bounded away from zero with positive probability is needed for the invertibility of V. Further, A6 is not a restriction on the model, but rather on how to choose the nearest neighbor weights and is hence innocuous.

[8] See Newey and Powell (1990) for a similar use of $\hat{F}_i$.

[9] We have not separated the assumptions by theorem since we are mostly concerned with T3.

[10] The Laplace distribution could be accomodated since its density has bounded first left and right derivatives at zero, but this would come at the expense of longer proofs, stronger conditions on the value of p_x and more restrictive choices of $\{k_n\}$.

That leaves A5 and A7. A5 is not primitive. It is a necessary and sufficient condition to ensure identification. In the univariate case A5 is implied by A2, A3 and A4, but we have failed to find a natural and primitive sufficient condition in the multivariate case. Finally, A7 deals with the rate at which k_n increases. As long as a sequence exists that satisfies the restrictions, A7 is merely a prescription on how to choose k_n. A7 is for instance satisfied when $p_x = 6$, $\beta_n \sim k_n^{-3/17}$ and $k_n \sim n^{35/36}$. It can be shown that A7 can only be satisfied for values of p_x greater than $3+\sqrt{8}$. However, if an expansion taken in L21 and L22 in the appendix is taken beyond the second order the requirements would improve but would never be better than $\sqrt{n}\zeta_{npT}^{o} \to 0$, $\sqrt{n}\zeta_{npT}^{o-1}\zeta_{npS} \to 0$ where o denotes the order of the expansion. Since with cross–sectional data fat regressor tails are rarely an issue and the extension would merely involve a repetition of the same arguments, we have omitted it in the interest of brevity.

The assumptions above are stronger than those required for Zhao's (2001) estimator for several reasons. First, the model is more general; the above conditions would be weaker in the single–equation case or if the T_i–matrices are known. We need some further conditions to avoid his sample splitting procedure.

We now state our theorems.

T1 *For any estimator* $\hat{\theta}_I$ *satisfying (8),* $\hat{\theta}_I \xrightarrow{p} \theta_0$.

T2 *For any estimator* $\hat{\theta}_I$ *satisfying (8),* $\sqrt{n}(\hat{\theta}_I - \theta_0) \xrightarrow{d} N(0, V^{-1})$.

T3 *For any estimator* $\hat{\hat{\theta}}$ *satisfying (9),* $\sqrt{n}(\hat{\hat{\theta}} - \theta_0) \xrightarrow{d} N(0, V^{-1})$.

For the purpose of hypothesis testing the matrix V needs to be estimated. The assumptions made are amply sufficient to guarantee convergence of our estimator $\hat{V}$ of V.

T4 $\hat{V} = n^{-1}\sum_{i=1}^{n} \hat{A}_i \hat{S}_i \xrightarrow{p} V$.

4 Computation

The computation of estimates $\hat{\hat{\theta}}$ that satisfy (9) is straightforward. We are helped by the availability of a $\sqrt{n}$–consistent inefficient estimator $\hat{\theta} = \hat{\theta}_{(0)}$, which is absent in the much harder procedure for computing quantile instrumental variables estimates; see Chernozhukov and Hansen (2006). We use the well–known procedure of taking one or more Newton steps in the direction of the 'minimum,' where the objective function is given by $||\hat{m}_n||$ and the 'gradient' and 'Hessian' by $\hat{m}_n$ and $\hat{V}$. So

$$\hat{\theta}_{(1)} = \hat{\theta}_{(0)} - \hat{V}^{-1}(\hat{\theta}_{(0)})\hat{m}_n(\hat{\theta}_{(0)}),$$

satisfies (9), but we use a general Newtonian optimization procedure with starting value $\hat{\theta}_{(0)}$; doing so will necessarily give an $||\hat{m}_n||$ value no worse than $||\hat{m}_n(\hat{\theta}_{(1)})||$ and hence also satisfies (9).

The only complication is that $\hat{m}_n$ is nondifferentiable, but all fundamental results to deal with the nondifferentiability issue were established in the proofs to earlier theorems.

T5 $\hat{\theta}_{(1)}$ *solves (9).*

5 Simulations

6 Conclusions

Appendices

A Infeasible Estimator

Proof of T1: Consider the following class of functions:

$$\mathcal{F} \equiv \Big\{c'A_1 s_1(\theta) = \sum_{j=1}^{d} c'A_{1j}s_{1j}(\theta) : \theta \in \Theta \subset \mathbb{R}^D\Big\},$$

where $c = [c_1, c_2, ..., c_d]'$ is an arbitrary vector and A_{1j} is the j^{th} column vector of A_1. Since $\mathcal{G}_j \equiv \{1(y_{1j} \leq X'_{1j}\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$ is a Vapnik Červonenkis subgraph class (or simply VČ class),[11] it follows that $\mathcal{F}_j \equiv \{c'A_{1j}s_{1j}(\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$ is also a VČ class by lemma 2.6.18 of van der Vaart and Wellner (1996). Since a VČ class is Euclidean for every envelope function (Pakes and Pollard (PP), 1989, lemma 2.12), we know that $\mathcal{F}_j$ is Euclidean with envelope function $\mathcal{E}_j = c'A_{1j}$. Therefore, by lemma 2.14 of PP, $\mathcal{F}$ is Euclidean with envelope function $\mathcal{E} = \sum_{j=1}^{n} \mathcal{E}_j$. Since $E(\mathcal{E}) < \infty$ by A3 and A4, it follows from lemma 2.8 of PP that

$$\sup_{\theta\in\Theta}\big|c'm_n(\theta) - c'm(\theta)\big| = o_p(1).$$

Since c is arbitrary, we have $\sup_{\theta\in\Theta}\big|\big|m_n(\theta) - m(\theta)\big|\big| = o_p(1)$. Now, by the triangle inequality

$$||m(\hat\theta_I)|| \leq ||m_n(\hat\theta_I)|| + ||m(\theta) - m_n(\theta)|| = o_p(n^{-1/2}) + o_p(1) = o_p(1).$$

Hence, by assumptions A1, A4 and A5, $\hat\theta_I - \theta_0 = o_p(1)$. ■

L1 *For any positive sequence $\{r_n\}$ and a consistent estimator θ_n, $m_n(\theta_n) = o_p(r_n)$ implies $||\theta_n - \theta_0|| = O_p(n^{-1/2}) + o_p(r_n)$.*
Proof: Let $\{\delta_n\}$ be a sequence such that $P\big(||\theta_n - \theta_0|| > \delta_n\big) = o(1)$. Then, recalling that $A_i s_i(\theta)$ is VČ,

$$\begin{aligned}||m(\theta_n)|| &\overset{\text{triangle}}{\leq} ||m_n(\theta_n) - m(\theta_n)|| + ||m_n(\theta_n)|| \lesssim \sup_{||\theta-\theta_0||<\delta_n} ||m_n(\theta) - m(\theta)|| + o_p(r_n)\\ &\leq \sup_{||\theta-\theta_0||<\delta_n} ||m_n(\theta) - m(\theta) - m_n(\theta_0) + m(\theta_0)|| + ||m_n(\theta_0)|| + o_p(r_n)\\ &= o_p(n^{-1/2}) + O_p(n^{-1/2}) + o_p(r_n). \quad (11)\end{aligned}$$

A2, A3 and A4 imply that

$$m(\theta) = V(\theta - \theta_0) + o(||\theta - \theta_0||). \quad (12)$$

Hence

$$\lambda_{\min}(V)||\theta_n - \theta_0|| \leq ||V(\theta_n - \theta_0)|| \leq ||m(\theta_n)|| + o_p(||\theta_n - \theta_0)||,$$

[11] See problem 14 on page 152 of van der Vaart and Wellner (1996).

which, together with the consistency of θ_n, implies that

$$\big(\lambda_{\min}(V) - o_p(1)\big)||\theta_n - \theta_0|| \leq ||m(\theta_n)|| = O_p(n^{-1/2}) + o_p(r_n).$$

Since V is positive definite, $||\theta_n - \theta_0|| = O_p(n^{-1/2}) + o_p(r_n)$. ■

Proof of T2: First, recall that $\mathcal{F}$ is a Euclidean class with envelope function $\mathcal{E} = \sum_{j=1}^d \mathcal{E}_j = \sum_{j=1}^d c' A_{1j}$. Note also that $E\big(\mathcal{E}^2\big) = c'\{\sum_{j=1}^d \sum_{t=1}^d E\big(A_{1j}A_{1t}'\big)\}c < \infty$. Therefore, it follows from lemma 2.17 of PP that

$$\sup_{||\theta-\theta_0||<\delta_n} \big|\sqrt{n}(c'm_n(\theta) - c'm(\theta)) - \sqrt{n}(c'm_n(\theta_0) - c'm(\theta_0))\big| = o_p(1)$$

for any sequence $\{\delta_n\}$ with $\delta_n = o(1)$. Since c is arbitrary, it implies that

$$\sup_{||\theta-\theta_0||<\delta_n} \big|\big|\sqrt{n}(m_n(\theta) - m(\theta)) - \sqrt{n}(m_n(\theta_0) - m(\theta_0))\big|\big| = o_p(1)$$

The asserted result now follows from theorem 3.3 of PP. Specifically, note that by lemma L1, $\hat\theta_I - \theta_0 = O_p(n^{-1/2})$. Using derivations similar to those in (11) and (12) we have

$$\begin{aligned} o_p(n^{-1/2}) = m_n(\hat\theta_I) &= \big(m_n(\hat\theta_I) - m(\hat\theta_I) - m_n(\theta_0) + m(\theta_0)\big) + m(\hat\theta_I) + m_n(\theta_0) \\ &= o_p(n^{-1/2}) + V(\hat\theta_I - \theta_0) + o_p(n^{-1/2}) + m_n(\theta_0) = m_n(\theta_0) + V(\hat\theta_I - \theta_0) + o_p(n^{-1/2}). \end{aligned}$$

Hence since $E(A_1 s_1 s_1' A_1') = E(A_1 T_1 A_1') = E(A_1 F_1 T_1^{-1} F_1 A_1') = V > 0$,

$$\sqrt{n}(\hat\theta_I - \theta_0) = -V^{-1}\sqrt{n} m_n(\theta_0) + o_p(1) \xrightarrow{d} N(0, V^{-1}).$$ ■

B Nonparametric Approximation

In addition to $\hat T_i, T_i, \hat S_i, S_i$ we define

$$\tilde T_i = \sum_{j=1}^n w_{ij} s_j s_j', \quad \bar T_i = \sum_{j=1}^n w_{ij} T_j, \quad \tilde S_i = \sum_{j=1}^n w_{ij} \tilde F_j X_j', \quad \bar S_i = \sum_{j=1}^n w_{ij} S_j,$$

where $\tilde F_j = \operatorname{diag}\big(I(|u_{jt}| \leq \beta_n)\big)/(2\beta_n)$.

B.1 Lemmas showing that $\max_i ||\hat{A}_i - \bar{A}_i|| = o_p(1)$.

Note that

$$\hat{A}_i - \bar{A}_i = (\hat{S}_i' - \bar{A}_i\hat{T}_i)\hat{T}_i^{-1} = \big((\hat{S}_i - \bar{S}_i)' - \bar{A}_i(\hat{T}_i - \bar{T}_i)\big)\big(\bar{T}_i^{-1} + (\hat{T}_i^{-1} - \bar{T}_i^{-1})\big)$$
$$= \Big((\hat{S}_i - \tilde{S}_i)' + (\tilde{S}_i - \bar{S}_i)' - \bar{A}_i\big((\hat{T}_i - \tilde{T}_i) + (\tilde{T}_i - \bar{T}_i)\big)\Big)\Big(\bar{T}_i^{-1} + (\hat{T}_i^{-1} - \bar{T}_i^{-1})\Big). \quad (13)$$

We deal with the uniform convergence of the differences in turn and then find a bound on the growth of $\bar{A}_i$.

B.1.1 $\tilde{T}_i - \bar{T}_i$

L2 $\exists \epsilon > 0 : \forall n : P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) = 0.$
Proof:

$$P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) \leq P\big(\min_i \lambda_{\min}(T_i) < \epsilon\big) = 0,$$

by A2. ■

L3 *For any $p > 2$ for which $E(R_{ni}|X_i) = 0$ a.s. and $\limsup E||R_{ni}||^p < \infty$, $E||\sum_{j=1}^n w_{ij}R_{nj}||^p = O(k_n^{-p/2})$.*
Proof: This is a special case of Pinkse (2006), L3, which was inspired by Robinson (1987), lemma ???. ■

L4 *For any $\{\xi_{ni}\}$ for which $E||\xi_{ni}||^p < \infty$ for all i, n and any $\epsilon > 0$,*

$$P(\max_i ||\xi_{ni}|| \geq \epsilon) \leq \epsilon^{-p}\sum_{i=1}^n E||\xi_{ni}||^p.$$

Proof: The LHS is bounded by $\sum_i P(||\xi_{ni}|| \geq \epsilon)$ which is bounded by the RHS by the Markov inequality. ■

L5 *For any $p > 2$ for which $E(R_{ni}|X_i) = 0$ a.s. and $\limsup E||R_{ni}||^p < \infty$, $\max_i ||\sum_{j=1}^n w_{ij}R_{nj}|| = O_p(n^{1/p}k_n^{-1/2})$.*
Proof: Take $\xi_{ni} = n^{-1/p}k_n^{1/2}\sum_j w_{ij}R_j$ in L4 to obtain

$$P\Big(\max_i \Big|\Big|n^{-1/p}k_n^{1/2}\sum_{j=1}^n w_{ij}R_j\Big|\Big| \geq \epsilon\Big) \leq n^{-1}k_n^{p/2}\epsilon^{-p}\sum_{i=1}^n E\Big|\Big|\sum_{j=1}^n w_{ij}R_j\Big|\Big|^p \overset{\text{L3}}{=} O(1)\epsilon^{-p} \to 0,$$

as $\epsilon \to \infty$. ■

L6 *For all values of $p > 2$, $\max_i ||\tilde{T}_i - \bar{T}_i|| = O_p(k_n^{-1/2}n^{1/p})$.*
Proof: Use L5 with $R_i = s_is_i' - T_i$. ■

B.1.2 $\hat{T}_i - \tilde{T}_i$

We will make frequent use of the inequality

$$||\hat{s}_j\hat{s}_j' - s_js_j'|| \le ||\hat{s}_j - s_j||^2 + ||s_j|| \cdot ||\hat{s}_j - s_j|| \le C_s||\hat{s}_j - s_j||, \tag{14}$$

which holds for some $0 < C_s < \infty$ since both s_j and $\hat{s}_j$ are vectors of zeroes and ones. We will also make multiple use of the inequality

$$\begin{aligned}||\hat{s}_j - s_j|| &= \big|\big|I\big(u_j \le X_j'(\hat{\theta} - \theta_0)\big) - I(u_j \le 0)\big|\big| \le \big|\big|I\big(|u_j| \le ||X_j|| \cdot ||\hat{\theta} - \theta_0||\iota\big)\big|\big| \\ &\le \big|\big|I\big(|u_j| \le ||X_j||r_n\iota\big)\big|\big| + I(||\hat{\theta} - \theta_0|| > r_n) = ||\alpha_{jr_n}|| + I(||\hat{\theta} - \theta_0|| > r_n),\end{aligned} \tag{15}$$

which holds for any sequence $\{r_n\}$.

L7 *For some $C > 0$ and any $r \ge 0$, $E(||\alpha_{ir}||\,|X_i) \le C||X_i||r$ a.s.*
Proof: Note that

$$0 \le E(\alpha_{irj}|X_i) = P(|u_{ij}| \le r||X_i||\,|X_i) = F_{u_{ij}|X_i}(r||X_i||) - F_{u_{ij}|X_i}(-r||X_i||) \overset{\text{A4}}{\le} 2C_f||X_i||r. \quad \blacksquare$$

L8 *For any $p > 0$, $\max_i ||\hat{T}_i - \tilde{T}_i|| = O_p(\zeta_{npT})$.*
Proof: First,

$$\begin{aligned}C_s^{-1}||\hat{T}_i - \tilde{T}_i|| &= C_s^{-1}\Big|\Big|\sum_{j=1}^n w_{ij}(\hat{s}_j\hat{s}_j' - s_js_j')\Big|\Big| \overset{(14)}{\le} \sum_{j=1}^n w_{ij}||\hat{s}_j - s_j|| \\ &\overset{(15)}{\le} \sum_{j=1}^n w_{ij}\big(||\alpha_{jr_n}|| - E(||\alpha_{jr_n}||\,|X_j)\big) + \sum_{j=1}^n w_{ij}E(||\alpha_{jr_n}||\,|X_j) + I(||\hat{\theta} - \theta_0|| > r_n).\end{aligned} \tag{16}$$

Take $r_n = 1/(\sqrt{n} - \log n)$. Since $e^{-1/t}$ is an increasing function of t and for arbitrary positive a, b $I(a > b) \le g(a)/g(b)$ for any increasing function g,

$$I(||\hat{\theta} - \theta_0|| > r_n) \le e^{1/r_n}e^{-1/||\hat{\theta} - \theta_0||} = O_p(e^{1/r_n - \sqrt{n}}) = O_p(e^{-\log n}) = O_p(n^{-1}). \tag{17}$$

For the second RHS term in (16), note that

$$\max_i \sum_{j=1}^n w_{ij}E(||\alpha_{jr_n}||\,|X_j) \overset{\text{L7}}{\le} C_\alpha r_n \max_i \sum_{j=1}^n w_{ij}||X_j|| \le C_\alpha r_n \max_i ||X_i|| \overset{\text{L4}}{=} O_p(r_n n^{1/p_x}) = O_p(n^{1/p_x - 1/2}).$$

Finally, noting that the $||\alpha_{jr_n}||$'s are uniformly bounded, L5 implies that for any $p > 0$,

$$\max_i \Big|\Big|\sum_{j=1}^n w_{ij}\big(||\alpha_{jr_n}|| - E(||\alpha_{jr_n}||\,|X_j)\big)\Big|\Big| = O_p(n^{1/p}k_n^{-1/2}),$$

which takes care of the first RHS term in (16). $\blacksquare$

L9 *For any $p > 0$, $\max_i ||\hat{T}_i^{-1} - \bar{T}_i^{-1}|| = O_p(\zeta_{npT})$.*
Proof: Since $\hat{T}_i^{-1} = \bar{T}_i^{-1}\big(I + (\hat{T}_i - \bar{T}_i)\bar{T}_i^{-1}\big)^{-1}$, the result follows from lemmas L2, L6 and L8. $\blacksquare$

B.1.3 $\tilde{S}_i - \bar{S}_i$

L10 $\max_i ||\bar{S}_i|| = O_p(n^{1/p_x})$ *and* $\max_i ||\bar{A}_i|| = O_p(n^{1/p_x})$.
Proof: Note that for some $0 < C < \infty$,

$$\max_i ||\bar{A}_i|| \leq \max_i ||\bar{S}_i|| \max_i ||\bar{T}_i^{-1}|| \overset{\text{L2}}{\leq} C \max_i ||\bar{S}_i|| \leq C \max_i ||S_i|| \overset{\text{A4}}{\leq} CC_f \max_i ||X_i|| \overset{\text{L4}}{=} O_p(n^{1/p_x}). \quad ■$$

L11 $\max_i ||\tilde{S}_i - \bar{S}_i|| = O_p\big(n^{1/p_x}(k_n^{-1/2}\beta_n^{1/p_x-1} + \beta_n^2)\big)$.
Proof: Note that

$$\tilde{S}_i - \bar{S}_i = \sum_{j=1}^{n} w_{ij}\big(\tilde{F}_j - E(\tilde{F}_j|X_j)\big)X_j' + \sum_{j=1}^{n} w_{ij}\big(E(\tilde{F}_j|X_j) - F_j\big)X_j'. \tag{18}$$

Take $R_{nj} = \beta_n^{1-1/p_x}\big(\tilde{F}_j - E(\tilde{F}_j|X_j)\big)X_j'$ in L5 to obtain the rate $O_p(n^{1/p_x}k_n^{-1/2}\beta_n^{1/p_x-1})$ for the first RHS term in (18). For the second RHS term note that by the mean value theorem for all $t = 1, \ldots, d$,

$$\big|\big|E(\tilde{F}_{jt}|X_j) - F_{jt}\big|\big| = \big|\big|6^{-1}\beta_n^2 f''_{u_{jt}|X_j}(\cdot)\big|\big| \overset{\text{A4}}{\leq} 6^{-1}C_f\beta_n^2. \tag{19}$$

Hence the second RHS term in (18) is bounded by

$$6^{-1}C_f\beta_n^2 \max_i \sum_{j=1}^{n} w_{ij}||X_j|| \leq 6^{-1}C_f\beta_n^2 \max_i ||X_i|| = O_p(n^{1/p_x}\beta_n^2). \quad ■$$

B.1.4 $\hat{S}_i - \tilde{S}_i$

L12 $\max_i ||\hat{S}_i - \tilde{S}_i|| = O_p(n^{1/2}k_n^{-1}\beta_n^{-1})$.
Proof: Let $r_n = 1/(\sqrt{n} - \log n)$. Now,

$$\max_i ||\hat{S}_i - \tilde{S}_i|| = \max_i ||\hat{S}_i - \tilde{S}_i||I(||\hat{\theta} - \theta_0|| \leq r_n) + \max_i ||\hat{S}_i - \tilde{S}_i||I(||\hat{\theta} - \theta_0|| > r_n). \tag{20}$$

By (17) $I(||\hat{\theta} - \theta_0|| > r_n) = O_p(n^{-1})$, such that the second RHS term in (20) converges faster than the first. Now the first RHS term in (20). Using the inequality (for generic a, b, t)

$$\big|I(|a| \leq t) - I(|b| \leq t)\big| \leq I(|b| \leq t + |a - b|) - I(|b| \leq t - |a - b|),$$

it follows that

$$||\hat{F}_j - \tilde{F}_j||I(||\hat{\theta} - \theta_0|| \leq r_n) \leq \Big|\Big|I\big(|u_j| \leq (\beta_n + ||X_j||r_n)\iota\big) - I\big(|u_j| \leq (\beta_n - r_n||X_j||)\iota\big)\Big|\Big|, \tag{21}$$

and hence

$$\max_i ||\hat{S}_i - \tilde{S}_i|| I(||\hat{\theta} - \theta_0|| \le r_n) \le \max_i \sum_{j=1}^n w_{ij} ||X_j|| \cdot ||\hat{F}_j - \tilde{F}_j||$$
$$\le \beta_n^{-1} \max_i ||X_i|| \max_i \sum_{j=1}^n w_{ij} \Big|\Big| I\big(|u_j| \le (\beta_n + ||X_j|| r_n)\iota\big) - I\big(|u_j| \le (\beta_n - r_n ||X_j||)\iota\big) \Big|\Big|$$
$$\stackrel{A6}{\le} C_w (k_n \beta_n)^{-1} \sum_{j=1}^n \Big|\Big| I\big(|u_j| \le (\beta_n + r_n ||X_j||)\iota\big) - I\big(|u_j| \le (\beta_n - r_n ||X_j||)\iota\big) \Big|\Big|. \quad (22)$$

Since for all $t = 1, \dots, d$,

$$E\Big(I\big(|u_{jt}| \le (\beta_n + r_n ||X_j||)\big) - I\big(|u_{jt}| \le (\beta_n - r_n ||X_j||)\big) |X_j \Big) = \mathcal{F}_{u_{jt}|X_j}(\beta_n + r_n ||X_j||) - \mathcal{F}_{u_{jt}|X_j}(\beta_n - r_n ||X_j||)$$
$$= f_{u_{jt}|X_j}(\cdot) ||X_j|| r_n \le C_f r_n ||X_j||, \quad (23)$$

the unconditional expectation of (22) is bounded by

$$d C_w C_f r_n (k_n \beta_n)^{-1} \sum_{j=1}^n E||X_j||^2 = O\big(n r_n (k_n \beta_n)^{-1}\big) = O(n^{1/2} k_n^{-1} \beta_n^{-1}). \quad ■$$

L13 $\max_i ||\hat{A}_i - \bar{A}_i|| = o_p(1)$.
Proof: Using L2, L6, L8, L9, L10, L11 and L12 in (13) yields

$$\hat{A}_i - \bar{A}_i = O_p\big((n^{1/p_x} \zeta_{npT} + \zeta_{npS})(1 + \zeta_{npT}\big) = o_p(1),$$

by A7. ■

B.2 $\sqrt{n}\big(\hat{m}_n(\theta_0) - m_n(\theta_0)\big)$

Observe that

$$\sqrt{n}\big(\hat{m}_n(\theta_0) - m_n(\theta_0)\big) = n^{-1/2} \sum_{i=1}^n (\hat{A}_i - A_i) s_i = n^{-1/2} \sum_{i=1}^n (\hat{A}_i - \bar{A}_i) s_i + n^{-1/2} \sum_{i=1}^n (\bar{A}_i - A_i) s_i. \quad (24)$$

We use the expansion in (13) to deal with the first RHS term and show the following results.

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\hat{T}_i-\tilde{T}_i)\bar{T}_i^{-1}s_i=o_p(1), \tag{25}$$

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\tilde{T}_i-\bar{T}_i)\bar{T}_i^{-1}s_i=o_p(1), \tag{26}$$

$$n^{-1/2}\sum_{i=1}^{n}(\hat{S}_i-\tilde{S}_i)'\bar{T}_i^{-1}s_i=o_p(1), \tag{27}$$

$$n^{-1/2}\sum_{i=1}^{n}(\tilde{S}_i-\bar{S}_i)'\bar{T}_i^{-1}s_i=o_p(1), \tag{28}$$

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\hat{T}_i-\bar{T}_i)(\hat{T}_i^{-1}-\bar{T}_i^{-1})=o_p(1), \tag{29}$$

$$n^{-1/2}\sum_{i=1}^{n}(\hat{S}_i-\bar{S}_i)'(\hat{T}_i^{-1}-\bar{T}_i^{-1})=o_p(1), \tag{30}$$

$$n^{-1/2}\sum_{i=1}^{n}(\bar{A}_i-A_i)s_i=o_p(1). \tag{31}$$

B.2.1 (25)

L14 *Let* $\{\xi_i\}$ *be an i.i.d. random sequence for which* $E(\xi_i|X)=0$ *a.s. and* $\operatorname{ess\,sup}(||\xi_i||)\leq 1$. *Then*

$$\max_j\Big|\Big|\sum_{i=1}^{n}w_{ij}\xi_i\Big|\Big|=o_p(\sqrt{n\log n}/k_n).$$

Proof: Let $\epsilon_n=C_w\sqrt{3n\log n}/k_n$. Then

$$P\Big(\max_j\Big|\Big|\sum_i w_{ij}\xi_i\Big|\Big|\geq 2\epsilon_n\Big)\leq P\Big(\max_j\Big|\Big|\sum_{i\neq j} w_{ij}\xi_i\Big|\Big|\geq \epsilon_n\Big)+P\Big(\max_j||w_{jj}\xi_j||\geq\epsilon_n\Big). \tag{32}$$

The second RHS term in (32) is bounded by $I(C_w/k_n\geq\epsilon_n)$, which equals zero for sufficiently large n. We now deal with the first RHS term in (32). By the *Hoeffding inequality*,[12] noting that $||w_{ij}\xi_i||\leq C_w/k_n$ for

[12] The Hoeffding inequality says that if $\{\mu_i\}$ is an independent sequence of mean zero random variables taking values on $[a_i,b_i]$, then $P(||\sum_i\mu_i||>\epsilon_n)\leq\exp\big[-2\epsilon_n^2/\sum_{i=1}^{n}(b_i-a_i)^2\big]$.

all i, j,

$$P\Big(\max_j \Big|\Big|\sum_{i\neq j} w_{ij}\xi_i\Big|\Big| \geq \epsilon_n | X_j, \xi_j\Big) \leq \sum_{j=1}^n P\Big(\Big|\Big|\sum_{i\neq j} w_{ij}\xi_i\Big|\Big| \geq \epsilon_n | X_j, \xi_j\Big)$$
$$\leq \sum_{j=1}^n \exp\Big(-\frac{\epsilon_n^2 k_n^2}{2nC_w^2}\Big) = n\exp\big(-(3/2)\log n\big) = n^{-1/2} = o(1). \quad \blacksquare$$

L15 *Let $\{\xi_i\}$ be as in L14 and let $\xi_{ni} = \Xi_{ni}(X)\xi_i$, where for some $p_\Xi > 0$, $\limsup E||\Xi_{ni}(X)||^{p_\Xi} < \infty$. Then*

$$\max_j \Big|\Big|\sum_{i=1}^n w_{ij}\xi_{ni}\Big|\Big| = o_p\big(n^{1/p_\Xi + 1/2} k_n^{-1}\log n\big).$$

Proof: Let $\epsilon_n^* = n^{1/p_\Xi}\sqrt{\log n}$, $\epsilon_n = \sqrt{3}C_w n^{1/p_\Xi+1/2}\log n/k_n$ and $\xi_{ni}^* = \xi_{ni}I\big(||\Xi_{ni}(X)|| \leq \epsilon_n^*\big)/\epsilon_n^*$. Then

$$P\Big(\max_j \Big|\Big|\sum_{i=1}^n w_{ij}\xi_{ni}\Big|\Big| \geq 2\epsilon_n\Big) = P\Big(\max_j \Big|\Big|\sum_{i=1}^n w_{ij}\big(\epsilon_n^*\xi_{ni}^* + \xi_{ni}I(||\Xi_{ni}(X)|| > \epsilon_n^*)\big)\Big|\Big| \geq 2\epsilon_n\Big)$$
$$\leq P\Big(\max_j \Big|\Big|\sum_{i=1}^n w_{ij}\xi_{ni}^*\Big|\Big| \geq 2\frac{\epsilon_n}{\epsilon_n^*}\Big) + P\Big(\max_i ||\Xi_{ni}(X)|| \geq \epsilon_n^*\Big). \quad (33)$$

The second RHS term in (33) is by L4 bounded by

$$(\epsilon_n^*)^{-p_\Xi}\sum_{i=1}^n E||\Xi_{ni}||^{p_\Xi} = O\big((\log n)^{-p_\Xi/2}\big) = o(1).$$

The first RHS term in (33) is also $o(1)$ because ess sup $||\xi_{ni}^*|| \leq 1$ by construction and since

$$\frac{\epsilon_n}{\epsilon_n^*} = \frac{\sqrt{3}C_w n^{\frac{p_\Xi+2}{2p_\Xi}}\log n/k_n}{n^{1/p_\Xi}\sqrt{\log n}} = \frac{C_w\sqrt{3n\log n}}{k_n},$$

L14 can be applied. $\blacksquare$

L16 $n^{-1/2}\sum_i \bar{A}_i(\hat{T}_i - \tilde{T}_i)\bar{T}_i^{-1}s_i = o_p(1)$.
Proof: The LHS is

$$\Big|\Big|n^{-1/2}\sum_{j=1}^n\sum_{i=1}^n w_{ij}\bar{A}_i(\hat{s}_j\hat{s}_j' - s_js_j')\bar{T}_i^{-1}s_i\Big|\Big| \overset{\text{L15}}{\leq} \sum_{j=1}^n ||\hat{s}_j\hat{s}_j' - s_js_j'|| \times o_p(n^{1/p_x}k_n^{-1}\log n)$$
$$\overset{(14)}{\leq} C_s\sum_{j=1}^n ||\hat{s}_j - s_j|| \times o_p(n^{1/p_x}k_n^{-1}\log n). \quad (34)$$

Set $r_n = 1/(\sqrt{n} - \log n)$. Now,

$$\sum_{j=1}^{n} ||\hat{s}_j - s_j|| \overset{(15)}{\leq} \sum_{j=1}^{n} \big(||\alpha_{jr_n}|| - E(||\alpha_{jr_n}||\,|X)\big) + \sum_{j=1}^{n} E(||\alpha_{jr_n}||\,|X) + nI(||\hat{\theta} - \theta_0|| > r_n). \tag{35}$$

The third RHS term is $O_p(1)$ by (17) and the second RHS term is by L7 bounded by $C_\alpha r_n \sum_{j=1}^{n} ||X_j|| = O_p(nr_n) = O_p(n^{1/2})$. Squaring the first RHS term and taking its expectation yields

$$\sum_{j=1}^{n} E\big(||\alpha_{jr_n}|| - E(||\alpha_{jr_n}||\,|X)\big)^2 \overset{\text{L7}}{\leq} Cnr_n = O(nr_n).$$

Hence the RHS in (35) is $O_p(\sqrt{nr_n}) + O_p(\sqrt{n}) + O_p(1) = O_p(\sqrt{n})$, which implies that the RHS in (34) is $o_p(n^{1/p_x+1/2}k_n^{-1}\log n) = o_p(1)$ by A7. ■

B.2.2 (26)

L17 *Let $\xi_{nij} = \xi_n(u_i, u_j; X)$ be such that $E(\xi_{nij}|u_i, X) = E(\xi_{nij}|u_j, X) = 0$ a.s. for all i, j and $\max_{i,j} E||\xi_{nij}||^2 = O(1)$. Then $n^{-1}\sum_{i,j=1}^{n} w_{ij}\xi_{nij} = O_p(k_n^{-1})$.*
Proof: Square the LHS and take the expectation to obtain

$$n^{-2}\sum_{i,j=1}^{n}\Big(E(w_{ij}^2||\xi_{nij}||^2) + E(w_{ij}w_{ji}\xi_{nij}'\xi_{nji})\Big) \overset{\text{A6}}{\leq} 2C_w^2 k_n^{-2}\max_{i,j} E||\xi_{nij}||^2 = O(k_n^{-2}). \quad ■$$

L18 $n^{-1/2}\sum_{i=1}^{n} \bar{A}_i(\tilde{T}_i - \bar{T}_i)\bar{T}_i^{-1}s_i = o_p(1)$.
Proof: In L17, take $\xi_{nij} = \bar{A}_i(s_js_j' - T_j)\bar{T}_i^{-1}s_i$ to obtain a convergence rate of $O_p(n^{1/2}k_n^{-1}) = o_p(1)$. ■

B.2.3 (27) and (28)

L19 $n^{-1/2}\sum_i(\hat{S}_i - \tilde{S}_i)'\bar{T}_i^{-1}s_i = o_p(1)$.
Proof: The norm of the LHS is

$$\Big|\Big|n^{-1/2}\sum_{j=1}^{n}(\hat{F}_j - \tilde{F}_j)X_j'\sum_{i=1}^{n} w_{ij}\bar{T}_i^{-1}s_i\Big|\Big| \leq \max_j\Big|\Big|n^{-1/2}\sum_{i=1}^{n} w_{ij}\bar{T}_i^{-1}s_i\Big|\Big|\sum_{j=1}^{n}||\hat{F}_j - \tilde{F}_j|| \times ||X_j||$$
$$\overset{\text{L14}}{=} O_p(k_n^{-1}\sqrt{\log n})\sum_{j=1}^{n}||\hat{F}_j - \tilde{F}_j|| \times ||X_j||.$$

Let (as in L12) $r_n = 1/(\sqrt{n} - \log n)$. Then

$$\sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| = \sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| I(||\hat{\theta} - \theta_0|| \le r_n) + \sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| I(||\hat{\theta} - \theta_0|| > r_n)$$
$$\overset{(17)}{=} \sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| I(||\hat{\theta} - \theta_0|| \le r_n) + \sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| \times O_p(n^{-1}).$$

Finally,

$$\frac{\sqrt{\log n}}{k_n} \sum_{j=1}^{n} ||\hat{F}_j - \tilde{F}_j|| \times ||X_j|| I(||\hat{\theta} - \theta_0|| \le r_n) \overset{(21),(23)}{\le} \frac{C_f d r_n \sqrt{\log n}}{k_n \beta_n} \sum_{j=1}^{n} ||X_j||^2 = O_p\Big(\frac{\sqrt{n \log n}}{k_n \beta_n}\Big) = o_p(1),$$

by A7. ■

L20 $n^{-1/2} \sum_{i=1}^{n} (\tilde{S}_i - \bar{S}_i)' \bar{T}_i^{-1} s_i = o_p(1)$.
Proof: The LHS is

$$n^{-1/2} \sum_{i,j=1}^{n} w_{ij} \big(\tilde{F}_j - E(\tilde{F}_j | X_j)\big) X_j' \bar{T}_i^{-1} s_i + n^{-1/2} \sum_{i,j=1}^{n} w_{ij} \big(E(\tilde{F}_j | X_j) - F_j\big) X_j' \bar{T}_i^{-1} s_i. \tag{36}$$

The first RHS term is $O_p(n^{1/2} \beta_n^{-1/2} k_n^{-1}) = o_p(1)$ by L17. The norm of the second RHS term is bounded by

$$n^{-1/2} \max_j \Big|\Big| \sum_{i=1}^{n} w_{ij} \bar{T}_i^{-1} s_i \Big|\Big| \sum_{j=1}^{n} w_{ij} \big|\big| E(\tilde{F}_j | X_j) - F_j \big) \big|\big| \times ||X_j||$$
$$\overset{L14,(19)}{\le} O_p(k_n^{-1} \sqrt{\log n}) 6^{-1} C_f \beta_n^2 \sum_j ||X_j|| = O_p(n k_n^{-1} \beta_n^2 \sqrt{\log n}) = o_p(1),$$

by A7.

B.2.4 (29) and (30)

L21 $n^{-1/2} \sum_{i=1}^{n} \bar{A}_i (\hat{T}_i - \bar{T}_i)(\hat{T}_i^{-1} - \bar{T}_i^{-1}) s_i = o_p(1)$.
Proof: Note that

$$\Big|\Big| n^{-1/2} \sum_{i=1}^{n} \bar{A}_i (\hat{T}_i - \bar{T}_i)(\hat{T}_i^{-1} - \bar{T}_i^{-1}) s_i \Big|\Big|$$
$$\le \max_i \big|\big| \hat{T}_i - \tilde{T}_i \big|\big| \times \big|\big| \hat{T}_i^{-1} - \bar{T}_i^{-1} \big|\big| \times n^{-1/2} \sum_{i=1}^{n} ||\bar{A}_i|| \times ||s_i|| = O_p(\sqrt{n} \zeta_{npT}^2) = o_p(1),$$

by L8, L9 and A7. ■

L22 $n^{-1/2}\sum_{i=1}^n(\hat{S}_i-\bar{S}_i)'(\hat{T}_i^{-1}-\bar{T}_i^{-1})s_i=o_p(1)$.
Proof: Use a similar inequality to the one used in L21 to obtain a rate of $n^{1/2}\zeta_{npS}\zeta_{npT}=o(1)$ by A7. ■

B.2.5 (31)

L23 $E||\bar{A}_i-A_i||^2=o(1)$.
Proof: The square of the LHS is bounded by

$$C\Big(E||A_i||^4E||\bar{T}_i-T_i||^4+(E||\bar{S}_i-S_i||^2)^2\Big)=o(1),$$

by theorem 1 of Stone (1977). ■

L24 $n^{-1/2}\sum_{i=1}^n(\bar{A}_i-A_i)s_i=o_p(1)$.
Proof:

$$E\Big|\Big|n^{-1/2}\sum_{i=1}^n(\bar{A}_i-A_i)s_i\Big|\Big|^2\le E||\bar{A}_i-A_i||^2=o(1),$$

by L23. ■

L25 $\hat{m}_n(\theta_0)-m_n(\theta_0)=o_p(n^{-1/2})$.
Proof: Using the expansion in (24) and (25)–(31), the stated result follows from lemmas L16, L18, L19, L20, L21, L22, and L24. ■

C Feasible Estimator

L26 *There exists a positive sequence* $\{\mu_{1n}\}$ *with* $\mu_{1n}=o(1)$ *such that for any positive sequence* $\{r_n\}$, $n^{-1}\sum_{i=1}^n||\bar{A}_i-A_i||\,||\alpha_{ir_n}||=o_p(r_n\mu_{1n})$.
Proof: Let μ_{1n} be such that $\mu_{1n}=o(1)$ and $E||\bar{A}_i-A_i||^2=o(\mu_{1n}^2)$; such μ_{1n} exist by lemma L23. Now,

$$E\big(||\bar{A}_i-A_i||\,||\alpha_{ir_n}||\big)\overset{\text{L7}}{\le}Cr_nE\big(||\bar{A}_i-A_i||\,||X_i||\big)\overset{\text{Schwarz}}{\le}Cr_n\sqrt{E\big(||\bar{A}_i-A_i||^2\big)}\sqrt{E||X_i||^2}=o(r_n\mu_{1n}).$$ ■

Let $\Theta_r=\{\theta\in\Theta:||\theta-\theta_0||<r\}$.

L27 *There exists a positive sequence* $\{\mu_n\}$ *with* $\mu_n=o(1)$ *such that for any positive sequence* $\{r_n\}$,

$$\sup_{\theta\in\Theta_{r_n}}||\hat{m}_n(\theta)-m_n(\theta)||=o_p(r_n\mu_n+n^{-1/2}).$$

Proof: First note that

$$\sup_{\theta\in\Theta_{r_n}} ||\hat{m}_n(\theta)-m_n(\theta)|| \overset{\text{triangle}}{\leq} \sup_{\theta\in\Theta_{r_n}} ||\hat{m}_n(\theta)-m_n(\theta)-\hat{m}_n(\theta_0)+m_n(\theta_0)|| + ||\hat{m}_n(\theta_0)-m_n(\theta_0)||$$

$$\overset{\text{L25}}{\leq} \sup_{\theta\in\Theta_{r_n}} n^{-1}\sum_{i=1}^{n} ||\hat{A}_i - A_i||\,||s_i(\theta)-s_i(\theta_0)|| + o_p(n^{-1/2})$$

$$\leq n^{-1}\sum_{i=1}^{n} ||\hat{A}_i - A_i||\,||\alpha_{ir_n}|| + o_p(n^{-1/2}).$$

Now, let μ_{2n} be such that $\max_i ||\hat{A}_i - A_i|| = o_p(\mu_{2n})$ and $\mu_{2n} = o(1)$; such μ_{2n} exist by L13. Then by the triangle inequality,

$$n^{-1}\sum_{i=1}^{n} ||\hat{A}_i - A_i||\,||\alpha_{ir_n}|| \leq n^{-1}\sum_{i=1}^{n} ||\hat{A}_i - \bar{A}_i||\,||\alpha_{ir_n}|| + n^{-1}\sum_{i=1}^{n} ||\bar{A}_i - A_i||\,||\alpha_{ir_n}||$$

$$\leq \max_i ||\hat{A}_i-\bar{A}_i|| n^{-1}\sum_{i=1}^{n}||\alpha_{ir_n}|| + n^{-1}\sum_{i=1}^{n}||\bar{A}_i - A_i||\,||\alpha_{ir_n}|| \overset{\text{L7,L13,L26}}{=} o_p(\mu_{2n})O_p(r_n)+o_p(\mu_{1n}r_n) = o_p\big((\mu_{1n}+\mu_{2n})r_n\big),$$

Take $\mu_n = \mu_{1n}+\mu_{2n}$. ■

L28 $m_n(\hat{\hat{\theta}}) = o_p(n^{-1/2})$.

Proof: Let $\{\psi_n\}$ be such that $||\hat{\hat{\theta}}-\theta_0|| = O_p(\psi_n)$ but $||\hat{\hat{\theta}}-\theta_0|| \neq o_p(\psi_n)$. Let μ_n be as in L27. Then for $r_n = \psi_n/\sqrt{\mu_n}$ we have

$$||m_n(\hat{\hat{\theta}})|| \overset{\text{triangle}}{\leq} ||m_n(\hat{\hat{\theta}})-\hat{m}_n(\hat{\hat{\theta}})||+||\hat{m}_n(\hat{\hat{\theta}})|| \lesssim \sup_{\theta\in\Theta_{r_n}} ||m_n(\theta)-\hat{m}_n(\theta)||+o_p(n^{-1/2}) \overset{\text{L27}}{=} o_p(\psi_n\sqrt{\mu_n})+o_p(n^{-1/2}).$$

So by L1, $||\hat{\hat{\theta}}-\theta_0|| = o_p(\psi_n)+O_p(n^{-1/2})$. Hence $\psi_n \sim n^{-1/2}$. Apply L27 with $r_n = n^{-1/2}$. ■

Proof of T3: By L28, $\hat{\hat{\theta}}$ satisfies (8). ■

D Covariance Matrix Estimation

Let $\bar{V} = n^{-1}\sum_{i=1}^{n}\bar{A}_i\bar{S}_i$.

L29 $\hat{V}-\bar{V} = o_p(1)$.

Proof: Using the expansion

$$\hat{V}-\bar{V} = n^{-1}\sum_{i=1}^{n}(\hat{A}_i-\bar{A}_i)(\hat{S}_i-\bar{S}_i) + n^{-1}\sum_{i=1}^{n}(\hat{A}_i-\bar{A}_i)\bar{S}_i + n^{-1}\sum_{i=1}^{n}\bar{A}_i(\hat{S}_i-\bar{S}_i),$$

the stated result follows from L11, L12 and L13. ■

L30 $\bar{V} - V = o_p(1)$.
Proof: Using a similar expansion to the one in L29, we have

$$\begin{aligned} E||\bar{V} - V|| &= E\Big|\Big|n^{-1}\sum_{i=1}^{n}(\bar{A}_i\bar{S}_i - A_iS_i)\Big|\Big| \\ &\leq E\Big(||\bar{A}_i - A_i|| \times ||\bar{S}_i - S_i||\Big) + E\Big(||A_i|| \times ||\bar{S}_i - S_i||\Big) + E\Big(||\bar{A}_i - A_i|| \times ||S_i||\Big) \\ &\overset{\text{Schwarz}}{\leq} \sqrt{E||\bar{A}_i - A_i||^2}\sqrt{E||\bar{S}_i - S_i||^2} + \sqrt{E||A_i||^2}\sqrt{E||\bar{S}_i - S_i||^2} + \sqrt{E||\bar{A}_i - A_i||^2}\sqrt{E||S_i||^2}. \end{aligned}$$

Apply L23, theorem 1 of Stone (1977) and the fact that $E||A_i||^2, E||S_i||^2 < \infty$ by assumption. ■

Proof of T4: Combine the previous two lemmas. ■

E Computation

Proof of T5: By L27 and T4 it follows that $\hat{\theta}_{(1)} = O_p(n^{-1/2})$. Hence by L1, $\hat{m}_n(\hat{\theta}_{(j)}) - m_n(\hat{\theta}_{(j)}) = o_p(n^{-1/2})$ for $j = 0, 1$. Because $\{A_i s_i\}$ is a VČ class (see (11)), it follows that

$$\big|\big|m_n(\hat{\theta}_{(1)}) - m_n(\hat{\theta}_{(0)}) - m(\hat{\theta}_{(1)}) + m(\hat{\theta}_{(0)})\big|\big| = o_p(n^{-1/2}).$$

Since $m(\hat{\theta}_{(1)}) - m(\hat{\theta}_{(0)}) = V(\hat{\theta}_{(1)} - \hat{\theta}_{(0)}) + o_p(n^{-1/2})$ (see (12)), it follows that

$$\begin{aligned} \hat{m}_n(\hat{\theta}_{(1)}) - \hat{m}_n(\hat{\theta}_{(0)}) &= m_n(\hat{\theta}_{(1)}) - m_n(\hat{\theta}_{(0)}) + o_p(n^{-1/2}) = m(\hat{\theta}_{(1)}) - m(\hat{\theta}_{(0)}) + o_p(n^{-1/2}) \\ &= V(\hat{\theta}_{(1)} - \hat{\theta}_{(0)}) + o_p(n^{-1/2}) = -V\hat{V}^{-1}(\hat{\theta}_{(0)})\hat{m}_n(\hat{\theta}_{(0)}) + o_p(n^{-1/2}) \overset{\text{T4}}{=} -\hat{m}_n(\hat{\theta}_{(0)}) + o_p(n^{-1/2}). \end{aligned}$$

So $\hat{m}_n(\hat{\theta}_{(1)}) = o_p(n^{-1/2})$ and (9) is satisfied. ■

References Cited

Aitken, Alexander C. (1935) "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh* 55, 42–48.

Carroll, Raymond J. (1982) "Adapting for heteroscedasticity in linear models," *Annals of Statistics* 10, 1224–1233.

Chernozhukov, Victor and Christian Hansen (2006) "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics* 132, 491–525.

Delgado, Miguel (1992) "Semiparametric generalized least squares estimation in the multivariate nonlinear regression model," *Econometric Theory* 8, 203–222.

Koenker, Roger (2005) "Quantile regression," Cambridge University Press (New York).

Koenker, Roger and Quanshui Zhao (1994) "L–estimation for linear heteroscedastic models," *Journal of Nonparametric Statistics* 3, 223–235.

Newey, Whitney K. (1990) "Efficient instrumental variables estimation of nonlinear models," *Econometrica* 58, 809–837.

Newey, Whitney K. (1993) "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics* 11, G.S. Maddala, C.R. Rao and H.D. Vinod, eds., North Holland, Amsterdam.

Newey, Whitney K. and James L. Powell (1990) "Efficient estimation of linear and type I censored regression models under conditional quantile restrictions," *Econometric Theory* 6, 295–317.

Pakes, Ariel and David Pollard (1989) "Simulation and the asymptotics of optimization estimators," *Econometrica* 57, 1027–1057.

Pinkse, Joris (2006) "Heteroskedasticity correction and dimension reduction," Pennsylvania State University working paper.

Robinson, Peter M. (1987) "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica* 55, 875–891.

Stone, Charles J. (1977) "Consistent nonparametric regression," *Annals of Statistics* 5, 595–645.

van der Vaart, Aad, and Jon A. Wellner (1996) "Weak convergence and empirical processes: with applications to statistics," Springer (New York).

Zhao, Quanshui (2001) "Asymptotically efficient median regression in the presence of heteroskedasticity of unknown form," *Econometric Theory* 17, 765–784.

		No Overlap		**Overlap in θ_{01}, θ_{02}**	
		$x_{i1} = x_{i2}$	$x_{i1} \neq x_{i2}$	$x_{i1} = x_{i2}$	$x_{i1} \neq x_{i2}$
$x_{i1}, x_{i2} \perp\!\!\!\perp u_{i1}, u_{i2}$	$u_{i1} \perp\!\!\!\perp u_{i2}$	all same	all same	J$\succsim$S*SO	J$\succsim$S*SO
	$u_{i1} \not\perp\!\!\!\perp u_{i2}$	all same	J$\succsim$S*SO	J$\succsim$S*SO	J$\succsim$S*SO
$x_{ij} \perp\!\!\!\perp u_{ij^*}; j \neq j^*$	$u_{i1} \perp\!\!\!\perp u_{i2} \mid x_{i1}, x_{i2}$	all same	JS*S$\succsim$O	J$\succsim$S*SO	J$\succsim$S*S$\succsim$O
	$u_{i1} \not\perp\!\!\!\perp u_{i2} \mid x_{i1}, x_{i2}$	all same	J$\succsim$S*S$\succsim$O	J$\succsim$S*SO	J$\succsim$S*S$\succsim$O
$x_{ij} \not\perp\!\!\!\perp u_{ij^*}$	$u_{i1} \perp\!\!\!\perp u_{i2} \mid x_{i1}, x_{i2}$	JS*S$\succsim$O	JS*$\succsim$S$\succsim$O	J$\succsim$S*S$\succsim$O	J$\succsim$S*$\succsim$S$\succsim$O
	$u_{i1} \not\perp\!\!\!\perp u_{i2} \mid x_{i1}, x_{i2}$	J$\succsim$S*S$\succsim$O	J$\succsim$S*$\succsim$S$\succsim$O	J$\succsim$S*S$\succsim$O	J$\succsim$S*$\succsim$S$\succsim$O

The entries indicate which methods are preferable to others in terms of asymptotic efficiency in various situations. '$\perp\!\!\!\perp$' denotes independence and '$\succsim$' means "is typically more efficient but never less efficient than." J=joint estimation (new methodology), S=separate estimation (Zhao's method), S*=separate estimation using the regressors from both equations (Zhao's results can be used for this) and O=no efficiency correction.

Please note: when errors are independent of each other and of the regressors *and* the coefficient vectors do not overlap, then equation by equation adaptive (to error distribution) estimation dominates all of the other estimation methods mentioned here.

This comparison applies equally to mean and quantile regressions.

Table 1: Asymptotic Efficiency Comparison of Semiparametric Methods