Efficient Semiparametric Seemingly Unrelated Quantile Regression Estimation

Sung Jae Jun* and Joris Pinkse†
The Pennsylvania State University
and
The Center for the Study of Auctions, Procurements and Competition Policy

June 2008

Abstract

We propose an efficient semiparametric estimator for the coefficients of a multivariate linear regression model — with a conditional quantile restriction for each equation — in which the conditional distributions of errors given regressors are 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 true optimal instruments 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.

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

† joris@psu.edu We thank the coeditor, two anonymous referees and participants at the Carnegie Mellon departmental seminar and Ari Kang for their useful suggestions. We thank the Human Capital Foundation for their support.

1 Introduction

We propose an efficient semiparametric estimator for the coefficients of a multivariate linear regression model — with a conditional quantile restriction for each equation — in which the conditional distributions of errors given regressors are 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 true optimal instruments 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), and semiparametric corrections like ours have been around for a while, also. Carroll (1982), Delgado (1992) and Robinson (1987) show that by estimating the conditional error variance function nonparametrically the same asymptotic variance can be achieved as with (infeasible) GLS. 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 (2006), and Komunjer and Vuong (2006) propose efficiency corrections for the univariate quantile regression model.

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. Further, the uncorrected estimates are special cases of the correction procedure for particular values of the input parameters of the semiparametric method, but we offer no procedure for the optimal selection of the input parameters. Earlier work (Pinkse, 2006) and some experimentation (not reported) suggest that our procedure is comparatively insensitive to the choice of input parameters.

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 nondifferentiability issues. Zhao (2001), Whang (2006) and Komunjer and Vuong (2006) 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 (2006) and Komunjer and Vuong (2006) 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, but they differ in several dimensions. Zhao's results only cover the single–equation case and are not readily generalized to the multivariate case. Further, Zhao uses a less primitive technical condition on the construction of the weights, while we have specifically opted for nearest neighbor estimation; we believe that nearest neighbor weights satisfy Zhao's condition. A final difference concerns the way in which the first step estimation error is addressed. For both methods (Zhao's and ours) first step estimates enter the second step via a nondifferentiable function. Zhao proposes two distinct procedures to address this problem. The first procedure entails *sample splitting*, i.e. using half the data to get the first step estimator to be used as a plug–in in the second step for the second half of the data and vice versa. This procedure does not make a difference asymptotically, but is less attractive due to its inherent (finite sample) inefficiency. His second procedure assumes that the first step estimator has a certain *Bahadur representation,* which we believe is likely to hold in practice.

We do not know whether Zhao's methods can be extended to the multivariate case. Instead, we follow a new line of proof which neither entails sample–splitting nor does it require any assumptions on the first step estimator beyond a convergence rate. The new proof (contained in the last two lemmas of appendix C and using lemma 2 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 propose procedures involving solving standard linear programming (LP) problems possibly combined with taking a Newton step. 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 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 section 4.

2 Model and Estimator

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

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

or equivalently,

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

where Q denotes the vector of quantiles of interest, $y_i \in \mathbb{R}^d$, and $X_i \in \mathbb{R}^{K\times d}$ with d the number of regression equations and K the total number of unknown regression coefficients.

The restriction that the same θ_0–vector occurs in all regression equations is not restrictive, because we can make the choices $X_i = \oplus_{j=1}^{d} x_{ij}$ and $\theta_0 = [\theta_{01}', \dots, \theta_{0d}']'$,[1] resulting in

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

Note that (1) allows for linear cross–equation restrictions on the parameters.

[1] $\oplus$ denotes 'matrix direct sum,' i.e. the result is like a block–diagonal matrix with nonsquare diagonal blocks x_{ij}.

An assumption implicit in (1) is that $Q(y_{ij}|x_{i\ell}, x_{ij}) = Q(y_{ij}|x_{ij})$ a.s.. 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. A more detailed discussion of this and related issues follows further below.

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 $\mathrm{E}(s_i|X_i) = 0$ a.s.. ($s_i = s_i(\theta_0)$). The corresponding optimal unconditional moment conditions are

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

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

$$S_i = F_i X_i', \quad F_i = \oplus_{j=1}^{d} f_{u_{ij}|X_i}(0), \quad T_i = \mathrm{E}(s_i s_i'|X_i).^{2} \tag{5}$$

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

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

The proposed procedure yields a natural efficiency improvement over equation–by–equation estimation when there are cross–equation restrictions on the regression coefficients.

Absent such restrictions, the intuition for the nature of the efficiency improvement can be understood by comparing four estimators. The first estimator is $\hat\theta_{\mathrm{SI}} = [\hat\theta_{\mathrm{SI}1}', \ldots, \hat\theta_{\mathrm{SI}d}']'$, where $\hat\theta_{\mathrm{SI}j}$ is the traditional

[2] These unconditional moments are optimal in the sense that estimators based on them achieve the semiparametric efficiency bounds. See e.g. Chamberlain (1987), Newey (1993).

single equation quantile regression estimator. The second and third estimators are $\hat{\theta}_{\mathrm{SE}}$ and $\hat{\theta}_{\mathrm{SE}^*}$ which are constructed similarly with $\hat{\theta}_{\mathrm{SE}j}, \hat{\theta}_{\mathrm{SE}^*j}$ infeasible versions of Zhao's (2001) single equation estimator where the conditioning variables used are x_{ij} and X_i, respectively. Finally, $\hat{\theta}_{\mathrm{I}}$ is the infeasible version of our estimator defined in (8). In the mean regression case, $\hat{\theta}_{\mathrm{SI}}$ would correspond to doing OLS, $\hat{\theta}_{\mathrm{SE}}$ to equation–by–equation heteroskedasticity–corrected GLS, $\hat{\theta}_{\mathrm{SE}^*}$ to ditto but using regressors in all equations (see equation (7)) and $\hat{\theta}_{\mathrm{I}}$ to full GMM estimation with optimal instruments.

All four estimators can be expressed in the form (4) if A_i is replaced with some function of X_i. Adding a suffix to indicate the corresponding estimator, they make use of

$$A_{\mathrm{SI};i} = \oplus_{j=1}^{d} x_{ij}, \quad A_{\mathrm{SE};i} = \oplus_{j=1}^{d} \phi_j(x_{ij}), \quad A_{\mathrm{SE}^*;i} = \oplus_{j=1}^{d} \phi_j^*(X_i), \quad A_{\mathrm{I};i} = A_i = A(X_i), \tag{7}$$

where $\phi_j(x) = f_{u_{ij}|x_{ij}}(0) x_i = \tilde{f}_{ij} x_i$, $\phi_j^*(X_i) = f_{u_{ij}|X_i}(0) x_i = f_{ij} x_i$ and A_i as in the discussion following (4). If $d = 2$, the asymptotic variances of the estimators of the vector of coefficients in the first regression equation are

$$V_{\mathrm{I}1} = \big(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\big)^{-1}, \qquad V_{\mathrm{SE}^*1} = \tau_1(1-\tau_1)\big(\mathrm{E}[f_{i1}^2 x_{i1} x_{i1}']\big)^{-1},$$
$$V_{\mathrm{SI}1} = \tau_1(1-\tau_1)\big(\mathrm{E}[\tilde{f}_{i1} x_{i1} x_{i1}']\big)^{-1} \mathrm{E}[x_{i1} x_{i1}']\big(\mathrm{E}[\tilde{f}_{i1} x_{i1} x_{i1}']\big)^{-1}, \qquad V_{\mathrm{SE}1} = \tau_1(1-\tau_1)\big(\mathrm{E}[\tilde{f}_{i1}^2 x_{i1} x_{i1}']\big)^{-1},$$

where $\Omega_{j\ell} = \mathrm{E}[t_i^{j\ell} f_{ij} f_{i\ell} x_{ij} x_{i\ell}']$ and $t_i^{j\ell}$ is the (j, ℓ) element of T_i^{-1}.

The restrictions imposed on A in (7) weaken left to right and hence efficiency improves left to right, also. Specifically, because $\hat{\theta}_{SEs}$ allows the ϕ–function to depend on regressors in all equations, it is no less efficient than $\hat{\theta}_{\mathrm{SE}}$, which in turn is no less efficient than $\hat{\theta}_{SI}$ which requires $\phi_j(x_{ij}) = x_{ij}$. Our estimator gains because it does not require the 'off–diagonal' vectors in A_i to be zero.

Note that equivalence of $\hat{\theta}_{\mathrm{SE}}$ and $\hat{\theta}_{\mathrm{SE}^*}$ occurs trivially if the regressors are the same in all equations. Our estimator yields an efficiency improvement over $\hat{\theta}_{\mathrm{SE}^*}$ if the u_{ij}'s are dependent conditional on X_i[3] unless *both* the errors are independent of the regressors *and* the regressors are the same in all equations. This is similar to the situation in a mean regression seemingly unrelated regessions (SUR) model with random regressors in which an efficiency improvement does not obtain if either the errors are uncorrelated

[3] More precisely, if for some $j \neq \ell$, $P[t_{ij\ell} \neq 0] > 0$, i.e. if $I(u_{ij} \leq 0)$ and $I(u_{i\ell} \leq 0)$ are dependent conditional on X_i.

conditional on the regressors or the regressors are identical and independent of the errors.[4] Table 1 contains full details of when efficiency improvements obtain in the quantile model for the various estimators.

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

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

$\hat{\theta}_I$ 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

$$\hat{T}_i = \sum_{j=1}^{n} w_{ij}\hat{s}_j\hat{s}_j', \quad \hat{S}_i = \sum_{j=1}^{n} w_{ij}\hat{F}_j X_i', \tag{10}$$

where $\hat{s}_i = I(\hat{u}_i \le 0) - \tau$, $\hat{F}_i = \mathrm{diag}\big(I(|\hat{u}_i| \le \beta_n \iota)/(2\beta_n)\big)$ 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,[5] 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 j is among observation i's 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.

[4] The classical SUR model assumes deterministic regressors and homoskedastic errors which corresponds to independence of errors and regressors when regressors are random.

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

3 Results

We now discuss our main result, formulated in theorem II, which shows that the feasible estimator $\hat{\hat{\theta}}$ has a limiting normal distribution with variance V. For our main result, we need the following assumptions.[6]

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

Assumption A2 *For some* $C_T > 0$, $P\big(\lambda_{\min}(T_i) \geq C_T\big) = 1$*; the smallest eigenvalue of* T_i *is bounded away from* 0 *with probability* 1.

Assumption A3 $\mathrm{E}(X_i X_i') > 0$.

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

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

Assumption A6 *The weights* w_{ij} *are nonnegative and all* k_n *nonzero weights take values in the range* $[1/(C_w k_n); C_w/k_n]$ *for some* $C_w > 0$.

Assumption A7 *Let* $p_x > 0$ *be such that* $\mathrm{E}(||X_i||^{p_x}) < \infty$ *and define for any* $p > 0$, $\zeta_{npT} = (n^{1/p_x - 1/2} + n^{1/p} k_n^{-1/2}) \log n$ *and* $\zeta_{npS} = (n^{1/p_x} k_n^{-1/2} \beta_n^{1/p_x - 1} + n^{1/p_x} \beta_n^2 + n^{1/2} k_n^{-1} \beta_n) \log n$. *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 $\mathrm{Corr}\big[I(u_{i1} \leq 0), I(u_{i2} \leq 0)|X_i\big]$ 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.[7] The assumption that the conditional densities at zero are bounded away from zero with positive probability

[6] We have not separated the assumptions by theorem since we are mostly concerned with theorem II.
[7] 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 $\{k_n\}$.

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.

That leaves A5 and A7. A5 is not primitive. It is a necessary and sufficient condition to ensure identification. In the single equation case A5 is implied by A2, A3, and A4 because the quantile regression function (e.g. least absolute deviations) is convex, but we have failed to find a natural and primitive sufficient condition in the multiple equations case; note that the convexity argument of the single equation case does not carry over. 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 lemmas 21 and 22 in the appendix is taken beyond the second order the requirements would improve but would never be better than $\sqrt{n}\zeta_{npT}^{\omega} \to 0, \sqrt{n}\zeta_{npT}^{\omega-1}\zeta_{npS} \to 0$ as $n \to \infty$ where ω 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. We now state our theorems.

Theorem I *Let assumptions A1–A7 hold. Then for any estimator $\hat{\theta}_I$ satisfying (8), $\sqrt{n}(\hat{\theta}_I - \theta_0) \xrightarrow{d} N(0, V)$.*

Theorem II *Let assumptions A1–A7 hold. Then for any estimator $\hat{\hat{\theta}}$ satisfying (9), $\sqrt{n}(\hat{\hat{\theta}} - \theta_0) \xrightarrow{d} N(0, V)$.*

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

Theorem III *Let assumptions A1–A7 hold. Then $\hat{V} \xrightarrow{p} V$.*

4 Computation and Simulations

In this section, we report the results of a small simulation study and we discuss issues of computation. We begin by outlining a simple method for the computation of estimates $\hat{\hat{\theta}}$ that satisfy (9). This procedure entails taking one Newton step from any $\sqrt{n}$–consistent starting value, e.g. $\hat{\theta}_{(0)} = \hat{\theta}$, i.e. computing

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

This is a familiar procedure, where only the nondifferentiability issues provide minor complications; complications which were largely addressed in the earlier theorems.

Theorem IV *Let assumptions A1–A7 hold. Then* $\hat{\theta}_{(1)}$ *solves (9).*

Experience based on our simulations suggests that the above–described procedure often leads to an increase of the value of $||\hat{\hat{m}}_n||$, and the resulting estimate $\hat{\theta}_{(1)}$ does not behave as well as the theory predicts. We therefore propose an alternative procedure to ensure that (9) remains satisfied, but it only works when there are no cross–equation restrictions on the coefficients.

Consider the case of two equations. Computing our estimator then entails solving

$$\hat{m}_n(\hat{\hat{\theta}}) = [\hat{m}'_{n1}(\hat{\hat{\theta}}), \hat{m}'_{n2}(\hat{\hat{\theta}})]' = o_p(n^{-1/2}), \quad \text{with } \hat{m}_{nj}(\hat{\hat{\theta}}) = n^{-1}\sum_{i=1}^{n} x_{ij}\sum_{\ell=1}^{2}\hat{\delta}_{ij\ell}\big\{I(y_{i\ell} \le x'_{i\ell}\hat{\hat{\theta}}_\ell) - \tau_\ell\big\}, \tag{11}$$

where $\hat{\delta}_{ij\ell}$ is the (j,ℓ)–element of $\hat{\Delta}_i = (\sum_{j=1}^n w_{ij}\hat{F}_j)\hat{T}_i^{-1}$. Starting from an initial point $\hat{\hat{\theta}}_{(1)}$ define

$$\hat{\theta}_{(t+1)j} = \operatorname*{argmin}_{\theta_j} n^{-1}\sum_{i=1}^{n}\big[\hat{\delta}_{ijj}\varrho_{\tau_j}(y_{ij} - x'_{ij}\theta_j) + 2\hat{\delta}_{ijj'}\big\{I(y_{ij'} \le x'_{ij'}\hat{\theta}_{(t)j'}) - \tau_{j'}\big\}\big], \quad j = 1,2;\ j' = 3-j, \tag{12}$$

where $\varrho_\tau(s) = |s| + (2\tau - 1)s$ is Koenker's check function. Note that the linear programming (LP) problems in (12) have the asymptotic first order conditions $\big|\big|\hat{m}_{n1}\big(\hat{\theta}_{(t+1)1}, \hat{\theta}_{(t)2}\big)\big|\big| = \eta_{n1t} = o_p(n^{-1/2})$ and $\big|\big|\hat{m}_{n2}\big(\hat{\theta}_{(t)1}, \hat{\theta}_{(t+1)2}\big)\big|\big| = \eta_{n2t} = o_p(n^{-1/2})$. We can therefore choose $\hat{\theta}_{(t)}$ as a solution to (11) if $\big|\big|\hat{m}_{nj}(\hat{\theta}_{(t+1)1}, \hat{\theta}_{(t+1)2})\big|\big| \le C\eta_{njs_j}$ for $j = 1,2$, some $s_1, s_2 \le t$ and some prespecified constant C.

In our experiments, we use $\hat{\theta}_{\mathrm{SI}}$ as the starting value and $\hat{\theta}_{(2)}$ as the estimates. This computational strategy outlined above generalizes naturally to the case with more than two equations.

The design of our experiment follows Zhao (2001), i.e.

$$y_{ij} = \theta_{0j0} + x_{ij1}\theta_{0j1} + x_{ij2}\theta_{0j2} + u_{ij}, \quad j = 1,2,$$

where $\theta_{0j} = [\theta_{0j0}, \theta_{0j1}, \theta_{0j2}]' = [10, -4, 2]'$. Like Zhao (2001), we generate the regressors and errors of equation j by $x_{ij1} = N_{ij} + 0.2U_{ij}$, $x_{ij2} = 0.2N_{ij} + U_{ij}$, and $u_{ij} = h_j(X_i)\epsilon_{ij}$, where $N_{ij} \sim N(5,9)$, $U_{ij} \sim U(0,4)$, and $(\epsilon_{i1}, \epsilon_{i2})$ are jointly normally distributed with mean zero and variance one. In the homoskedastic experiments we set $h_1 = h_2 = 1$ and the heteroskedastic form used is

$$h_1(X_i) = \exp(|x_{i1}'\theta_{01} + x_{i2}'\theta_{02}|/10), \quad h_2(X_i) = 1 + 3\exp(-(x_{i1}'\theta_{01} + x_{i2}'\theta_{02} + 10)^2/100),$$

where $x_{ij} = [1, x_{ij1}, x_{ij2}]'$. This is the same design as Zhao (2001) except that we allow the regressors in one equation to enter the conditional variance function of the other equation. We allow two different values of the error correlation parameter ρ, namely zero and 0.7. The values for the input parameters used are $\beta_n \propto n^{-1/6}$ and $k_n \propto n^{4/5}$. The number of replications for each experiment is 1,000 and the results are summarized in Table 2. For each scenario, we compute regular median regression estimates (Unweighted LAD), within–equation efficient median estimates (Weighted LAD 1 and 2; à la Zhao (2001)), and our own estimates (SUR Estimation).

The experiments are designed to investigate the effects of (i) dependence of errors on regressors and (ii) dependence of errors across equations. In the absence of error correlation with homoskedasticity (top set of four rows) all estimators have the same limiting distribution, but unweighted LAD does somewhat better than the others because it does not have the overhead of nonparametric first step estimation. The effect of such overhead appears, as one would expect, to be less when the number of observations is larger. With heteroskedasticity, the benefits of the nonparametric correction methods become apparent. Zhao's estimator using all regressors (Weighted LAD 2) appears to dominate the competition when there is no correlation between the errors (third set of rows). It does better than the proposed method because it has less overhead and beats the other two estimators because of its greater asymptotic efficiency. Our estimator again narrows the gap when the number of observations increases.

When the errors are correlated, however, the proposed estimator does better than the others. The degree of such improvement likely depends on the amount of correlation and the sample size. In small samples with a small amount of error correlation, we recommend Zhao's procedure using all regressors. If there is substantial correlation or the data set is sufficiently large then our procedure appears preferable.

A Infeasible Estimator

Lemma 1 $\hat{\theta}_I \xrightarrow{p} \theta_0$.

Proof: Consider the class of functions $\mathcal{F} \equiv \left\{c'A_1 s_1(\theta) = \sum_{j=1}^{d} c'A_{1j}s_{1j}(\theta) : \theta \in \Theta \subset \mathbb{R}^D\right\}$, where $c = [c_1, c_2, ..., c_d]'$ is arbitrary and A_{1j} is the j^{th} column vector of A_1. Since $\mathcal{G}_j = \{1(y_{1j} \leq X'_{1j}\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$ is a Vapnik Červonenkis subgraph class (VČ class),[8] so is $\mathcal{F}_j \equiv \{c'A_{1j}s_{1j}(\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$.[9] Hence $\mathcal{F}$ is Euclidean with envelope function $\mathcal{E} = \sum_{j=1}^{n} \mathcal{E}_j = \sum_{j=1}^{n} c'A_{1j}$ (Pakes and Pollard (PP), 1989, lemmas 2.12 and 2.14). Because $\mathrm{E}(\mathcal{E}) < \infty$ by A3 and A4, it follows from lemma 2.8 of PP that $\sup_{\theta\in\Theta}\left|c'm_n(\theta) - c'm(\theta)\right| = o_p(1)$, and since c is arbitrary, we have $\sup_{\theta\in\Theta}\left|\left|m_n(\theta) - m(\theta)\right|\right| = o_p(1)$. 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, $\hat{\theta}_I - \theta_0 = o_p(1)$. ■

Lemma 2 *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, since $A_i s_i(\theta)$ is VČ,

$$\begin{aligned} ||m(\theta_n)|| &\overset{\text{triangle}}{\leq} ||m_n(\theta_n) - m(\theta_n)|| + ||m_n(\theta_n)|| \lesssim^{10} \sup_{||\theta-\theta_0||<\delta_n} ||m_n(\theta) - m(\theta)|| + o_p(r_n) \\ &\leq \sup_{||\theta-\theta_0||<\delta_n} ||m_n(\theta) - m(\theta) - m_n(\theta_0) + m(\theta_0)|| + ||m_n(\theta_0)|| + o_p(r_n) \\ &= o_p(n^{-1/2}) + O_p(n^{-1/2}) + o_p(r_n). \end{aligned} \tag{13}$$

A2, A3 and A4 imply that

$$m(\theta) = \Psi(\theta - \theta_0) + o(||\theta - \theta_0||). \tag{14}$$

Hence $\lambda_{\min}(\Psi)||\theta_n - \theta_0|| \leq ||\Psi(\theta_n - \theta_0)|| \leq ||m(\theta_n)|| + o_p(||\theta_n - \theta_0)||$, which, together with the consistency of θ_n, implies that $\big(\lambda_{\min}(\Psi) - o_p(1)\big)||\theta_n - \theta_0|| \leq ||m(\theta_n)|| = O_p(n^{-1/2}) + o_p(r_n)$. Since Ψ is positive definite, $||\theta_n - \theta_0|| = O_p(n^{-1/2}) + o_p(r_n)$. ■

[8] van der Vaart and Wellner (1996), p.52, problem 14.
[9] van der Vaart and Wellner (1996), lemma 2.6.18.
[10] $\lesssim$ means that the inequality holds with probability approaching one.

Proof of theorem I: Recalling that $\mathcal{F}$ is a Euclidean class with envelope function $\mathcal{E}$ and noting that $\mathrm{E}\big(\mathcal{E}^2\big) < \infty$ and that c is arbitrary, it follows from lemma 2.17 of PP that for any sequence $\{\delta_n\}$ with $\delta_n = o(1)$,

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

Noting that by lemmas 1 and 2, $\hat{\theta}_I - \theta_0 = O_p(n^{-1/2})$, using derivations similar to those in (13) and (14) we have

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

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

$$\sqrt{n}(\hat{\theta}_I - \theta_0) = -V\sqrt{n} m_n(\theta_0) + o_p(1) \xrightarrow{d} N(0, V). \quad \blacksquare$$

B Nonparametric Approximation

In addition to $\hat{T}_i, T_i, \hat{S}_i, S_i$ we define $\tilde{F}_j = \mathrm{diag}\big(I(|u_{jt}| \le \beta_n)\big)/(2\beta_n)$ and

$$\tilde{T}_i = \sum_{j=1}^n w_{ij} s_j s_j', \quad \bar{T}_i = \sum_{j=1}^n w_{ij} T_j, \quad \tilde{S}_i = \sum_{j=1}^n w_{ij} \tilde{F}_j X_j', \quad \bar{S}_i = \sum_{j=1}^n w_{ij} S_j.$$

Note that

$$\begin{aligned} \hat{A}_i - \bar{A}_i = (\hat{S}_i' - \bar{A}_i \hat{T}_i)\hat{T}_i^{-1} &= \big((\hat{S}_i - \bar{S}_i)' - \bar{A}_i(\hat{T}_i - \bar{T}_i)\big)\big(\bar{T}_i^{-1} + (\hat{T}_i^{-1} - \bar{T}_i^{-1})\big) \\ &= \Big((\hat{S}_i - \tilde{S}_i)' + (\tilde{S}_i - \bar{S}_i)' - \bar{A}_i\big((\hat{T}_i - \tilde{T}_i) + (\tilde{T}_i - \bar{T}_i)\big)\Big)\Big(\bar{T}_i^{-1} + (\hat{T}_i^{-1} - \bar{T}_i^{-1})\Big). \end{aligned} \quad (15)$$

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

Lemma 3 $\exists \epsilon > 0 : \forall n : P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) = 0.$

Proof: $P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) \le P\big(\min_i \lambda_{\min}(T_i) < \epsilon\big) = 0$, by A2. $\blacksquare$

Lemma 4 *For any $\{\xi_{ni}\}$ for which $\mathrm{E}\,||\xi_{ni}||^p < \infty$ for all i, n and any $\epsilon > 0$, $P(\max_i ||\xi_{ni}|| \ge \epsilon) \le \epsilon^{-p} \sum_{i=1}^n \mathrm{E}\,||\xi_{ni}||^p$.*

Proof: We have $P(\max_i ||\xi_{ni}|| \ge \epsilon) \le \sum_{i=1}^n P(||\xi_{ni}|| \ge \epsilon)$; use the Markov inequality. $\blacksquare$

Lemma 5 *For any $p > 2$ for which $\mathrm{E}(R_{ni}|X_i) = 0$ a.s. and $\limsup \mathrm{E}\,||R_{ni}||^p < \infty$,*[11] $\max_i ||\sum_{j=1}^n w_{ij} R_{nj}|| = O_p(n^{1/p} k_n^{-1/2})$.

Proof: Take $\xi_{ni} = n^{-1/p} k_n^{1/2} \sum_j w_{ij} R_j$ in lemma 4 to obtain

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

by lemma L3 of Pinkse (2006). Letting $\epsilon \to \infty$ completes the proof. ■

Lemma 6 *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 lemma 5 with $R_i = s_i s_i' - T_i$. ■

We will make frequent use of the inequality

$$||\hat{s}_j \hat{s}_j' - s_j s_j'|| \leq ||\hat{s}_j - s_j||^2 + ||s_j|| \cdot ||\hat{s}_j - s_j|| \leq C_s ||\hat{s}_j - s_j||, \tag{16}$$

which holds for some $0 < C_s < \infty$ since both s_j and $\hat{s}_j$ are vectors of zeroes and ones. Let $\alpha_{jr} = \big|\big| I\big(|u_j| \leq ||X_j|| r\iota\big)\big|\big|$. We will also use the fact that for any sequence $\{r_n\}$,

$$\begin{aligned} ||\hat{s}_j - s_j|| &= \big|\big| I\big(u_j \leq X_j'(\hat{\theta} - \theta_0)\big) - I(u_j \leq 0)\big|\big| \leq \big|\big| I\big(|u_j| \leq ||X_j|| \cdot ||\hat{\theta} - \theta_0|| \iota\big)\big|\big| \\ &\leq \big|\big| I\big(|u_j| \leq ||X_j|| r_n \iota\big)\big|\big| + I(||\hat{\theta} - \theta_0|| > r_n) = ||\alpha_{jr_n}|| + I(||\hat{\theta} - \theta_0|| > r_n). \end{aligned} \tag{17}$$

Lemma 7 *For some $C > 0$ and any $r \geq 0$, $\mathrm{E}(||\alpha_{ir}||\,|X_i) \leq C||X_i|| r$ a.s.*

Proof: Note that

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

Lemma 8 *For any $p > 0$, $\max_i ||\hat{T}_i - \tilde{T}_i|| = O_p(\zeta_{npT})$.*

Proof: First,

$$\begin{aligned} C_s^{-1} ||\hat{T}_i - \tilde{T}_i|| = C_s^{-1} \Big|\Big| \sum_{j=1}^n w_{ij} (\hat{s}_j \hat{s}_j' - s_j s_j') \Big|\Big| &\overset{(16)}{\leq} \sum_{j=1}^n w_{ij} ||\hat{s}_j - s_j|| \\ &\overset{(17)}{\leq} \sum_{j=1}^n w_{ij} \big(||\alpha_{jr_n}|| - \mathrm{E}(||\alpha_{jr_n}||\,|X_j)\big) + \sum_{j=1}^n w_{ij}\, \mathrm{E}(||\alpha_{jr_n}||\,|X_j) + I(||\hat{\theta} - \theta_0|| > r_n). \end{aligned} \tag{18}$$

[11] $||R_{ni}||$ means the square root of the maximum eigenvalue of $R_{ni}' R_{ni}$.

Take $r_n = \log n/\sqrt{n}$. The third RHS term in (18) is $o_p(\kappa_n)$ for any positive sequence $\{\kappa_n\}$ since

$$P\big[I(||\hat{\theta}-\theta_0|| > r_n) > \kappa_n\big] = P\big[||\hat{\theta}-\theta_0|| > r_n\big] = o(1). \tag{19}$$

For the second RHS term in (18), note that

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

Finally, noting that the $||\alpha_{jr_n}||$'s are uniformly bounded and independent conditional on the regressors, Hoeffding's theorem implies that $\max_i \Big|\Big|\sum_{j=1}^n w_{ij}\big(||\alpha_{jr_n}|| - \mathrm{E}(||\alpha_{jr_n}||\,|X_j)\big)\Big|\Big| = o_p(k_n^{-1/2}\log n)$, which takes care of the first RHS term in (18). ■

Lemma 9 *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 3, 6 and 8. ■

Lemma 10 *$\max_i ||\bar{S}_i|| = O_p(n^{1/p_x})$ and $\max_i ||\bar{A}_i|| = O_p(n^{1/p_x})$.*

Proof: Note that for some $0 < C < \infty$,

$$\max_i ||\bar{A}_i|| \leq \max_i ||\bar{S}_i|| \max_i ||\bar{T}_i^{-1}|| \overset{\text{L3}}{\leq} C \max_i ||\bar{S}_i|| \leq C \max_i ||S_i|| \overset{\text{A4}}{\leq} CC_f \max_i ||X_i|| \overset{\text{L4}}{=} O_p(n^{1/p_x}). \quad ■$$

Lemma 11 *$\max_i ||\tilde{S}_i - \bar{S}_i|| = O_p\big(n^{1/p_x}(k_n^{-1/2}\beta_n^{1/p_x-1} + \beta_n^2)\big)$.*

Proof: Note that

$$\tilde{S}_i - \bar{S}_i = \sum_{j=1}^n w_{ij}\big(\tilde{F}_j - \mathrm{E}(\tilde{F}_j|X_j)\big)X_j' + \sum_{j=1}^n w_{ij}\big(\mathrm{E}(\tilde{F}_j|X_j) - F_j\big)X_j'. \tag{20}$$

Take $R_{nj} = \beta_n^{1-1/p_x}\big(\tilde{F}_j - \mathrm{E}(\tilde{F}_j|X_j)\big)X_j'$ in lemma 5 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 (20). For the second RHS term note that by the mean value theorem for all $t=1,\ldots,d$,

$$\big|\big|\mathrm{E}(\tilde{F}_{jt}|X_j) - F_{jt}\big|\big| = \big|\big|6^{-1}\beta_n^2 f''_{u_{jt}|X_j}(\cdot)\big|\big| \overset{\text{A4}}{\leq} 6^{-1}C_f\beta_n^2. \tag{21}$$

Hence the second RHS term in (20) is bounded by

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

Lemma 12 $\max_i ||\hat{S}_i - \tilde{S}_i|| = O_p(n^{1/2}k_n^{-1}\beta_n^{-1}\log n)$.

Proof: Let $r_n = \log n/\sqrt{n}$. Now,

$$\max_i ||\hat{S}_i - \tilde{S}_i|| = \max_i ||\hat{S}_i - \tilde{S}_i||I(||\hat{\theta} - \theta_0|| \leq r_n) + \max_i ||\hat{S}_i - \tilde{S}_i||I(||\hat{\theta} - \theta_0|| > r_n). \tag{22}$$

By (19), the second RHS term in (22) is negligble. For the first RHS term in (22) 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

$$2\beta_n||\hat{F}_j - \tilde{F}_j||I(||\hat{\theta} - \theta_0|| \leq r_n) \leq \Big|\Big|I\big(|u_j| \leq (\beta_n + ||X_j||r_n)\iota\big) - I\big(|u_j| \leq (\beta_n - r_n||X_j||)\iota\big)\Big|\Big|, \tag{23}$$

and hence

$$\begin{aligned}2\max_i ||\hat{S}_i - \tilde{S}_i||I(||\hat{\theta} - \theta_0|| \leq r_n) &\leq 2\max_i \sum_{j=1}^{n} w_{ij}||X_j|| \cdot ||\hat{F}_j - \tilde{F}_j|| \\ &\leq \beta_n^{-1}\max_i \sum_{j=1}^{n} w_{ij}||X_j||\Big|\Big|I\big(|u_j| \leq (\beta_n + ||X_j||r_n)\iota\big) - I\big(|u_j| \leq (\beta_n - r_n||X_j||)\iota\big)\Big|\Big| \\ &\overset{A6}{\leq} C_w(k_n\beta_n)^{-1}\sum_{j=1}^{n} ||X_j||\Big|\Big|I\big(|u_j| \leq (\beta_n + r_n||X_j||)\iota\big) - I\big(|u_j| \leq (\beta_n - r_n||X_j||)\iota\big)\Big|\Big|. \end{aligned} \tag{24}$$

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

$$\begin{aligned}\mathrm{E}\Big(I\big(|u_{jt}| \leq (\beta_n + r_n||X_j||)\big) - I\big(|u_{jt}| \leq (\beta_n - r_n||X_j||)\big)|X_j\Big) &= F_{u_{jt}|X_j}(\beta_n + r_n||X_j||) - F_{u_{jt}|X_j}(\beta_n - r_n||X_j||) \\ &= f_{u_{jt}|X_j}(\cdot)||X_j||r_n \leq C_f r_n||X_j||, \end{aligned} \tag{25}$$

the unconditional expectation of (24) is bounded by

$$dC_wC_fr_n(k_n\beta_n)^{-1}\sum_{j=1}^{n} E||X_j||^2 = O\big(nr_n(k_n\beta_n)^{-1}\big) = O(n^{1/2}k_n^{-1}\beta_n^{-1}\log n). \quad ■$$

Lemma 13 $\max_i ||\hat{A}_i - \bar{A}_i|| = o_p(1)$.

Proof: Using lemmas 3, 6, 8, 9, 10, 11 and 12 in (15) yields

$$\hat{A}_i - \bar{A}_i = O_p\big((n^{1/p_x}\zeta_{npT} + \zeta_{npS})(1 + \zeta_{npT}\big) \overset{A7}{=} o_p(1). \quad ■$$

Observe that

$$\sqrt{n}\big(\hat{m}_n(\theta_0) - m_n(\theta_0)\big) = n^{-1/2}\sum_{i=1}^{n}(\hat{A}_i - A_i)s_i = n^{-1/2}\sum_{i=1}^{n}(\hat{A}_i - \bar{A}_i)s_i + n^{-1/2}\sum_{i=1}^{n}(\bar{A}_i - A_i)s_i. \quad (26)$$

We use the expansion in (15) to deal with the first RHS term and show the following results.

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\hat{T}_i - \tilde{T}_i)\bar{T}_i^{-1}s_i = o_p(1), \quad (27)$$

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\tilde{T}_i - \bar{T}_i)\bar{T}_i^{-1}s_i = o_p(1), \quad (28)$$

$$n^{-1/2}\sum_{i=1}^{n}(\hat{S}_i - \tilde{S}_i)'\bar{T}_i^{-1}s_i = o_p(1), \quad (29)$$

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

$$n^{-1/2}\sum_{i=1}^{n}\bar{A}_i(\hat{T}_i - \bar{T}_i)(\hat{T}_i^{-1} - \bar{T}_i^{-1}) = o_p(1), \quad (31)$$

$$n^{-1/2}\sum_{i=1}^{n}(\hat{S}_i - \bar{S}_i)'(\hat{T}_i^{-1} - \bar{T}_i^{-1}) = o_p(1), \quad (32)$$

$$n^{-1/2}\sum_{i=1}^{n}(\bar{A}_i - A_i)s_i = o_p(1). \quad (33)$$

Condition (27) is dealt with in lemmas 14–16, (28) in lemmas 17–18, (29) and (30) in lemmas 19–20, (31) and (32) in lemmas 21–22 and 33 in lemmas 23–25.

Lemma 14 *Let* $\{\xi_i\}$ *be a sequence of random variables for which* $\mathrm{E}(\xi_i|X, \xi_{i-1}, \ldots, \xi_1) = 0$ *and for which* $\operatorname{ess\,sup}(||\xi_i||) \le 1$. *Then* $\max_j ||\sum_{i=1}^n w_{ij}\xi_i|| = o_p(\sqrt{n\log n}/k_n)$.

Proof: Let $\epsilon_n = C_w(\sqrt{3n\log n} + 2)/k_n$. Then

$$P\Big(\max_j\Big|\Big|\sum_i w_{ij}\xi_i\Big|\Big| \ge 2\epsilon_n\Big) \le P\Big(\max_j\Big|\Big|\sum_{i\neq j}^{n} w_{ij}\xi_i\Big|\Big| \ge \epsilon_n\Big) + P\Big(\max_j w_{jj}||\xi_j|| \ge \epsilon_n\Big)$$

$$\stackrel{\text{Bonferroni}}{\le} \sum_{j=1}^{n} P\Big(\Big|\Big|\sum_{i\neq j}^{n} w_{ij}\xi_i\Big|\Big| \ge \epsilon_n\Big) + I(C_w/k_n \ge \epsilon_n) = \sum_{j=1}^{n}\mathrm{E}\,P\Big(\Big|\Big|\sum_{i\neq j}^{n} w_{ij}\xi_i\Big|\Big| \ge \epsilon_n|X\Big),$$

where we used the fact that $\max_j w_{jj}||\xi_j|| \le C_w/k_n$ with probability one. We then used the law of iterated expectations and the fact that $I(\epsilon_n \le C_w/k_n) = 0$. Since $\tilde{\xi}_i = w_{ij}\xi_i$ forms a martingale difference sequence for each j conditional on X and $||\tilde{\xi}_i|| \le w_{ij}$ a.s., we apply Azuma's inequality (e.g. Davidson (1994, p245)) to the RHS to obtain an upper bound of

$$2n\exp\Big(-\frac{\epsilon_n^2}{2\sum_i w_{ij}^2}\Big) \le 2n\exp\Big(-\frac{3C_w^2 nk_n^{-2}\log n}{2nC_w^2k_n^{-2}}\Big) = 2n^{-1/2} = o(1). \quad \blacksquare$$

Lemma 15 *Let* $\{\xi_i\}$ *be as in lemma 14 and let* $\xi_{ni} = \Xi_{ni}(X)\xi_i$*, where for some* $p_\Xi > 0$, $\limsup \mathrm{E}\,||\Xi_{ni}(X)||^{p_\Xi} < \infty$*. 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 $\xi_{ni}^* = \xi_{ni}I\big(||\Xi_{ni}(X)|| \le \epsilon_n^*\big)/\epsilon_n^*$. Then

$$P\Big(\max_j\Big|\Big|\sum_{i=1}^n w_{ij}\xi_{ni}\Big|\Big| \ge 2\epsilon_n\Big) = P\Big(\max_j\Big|\Big|\sum_{i=1}^n w_{ij}\big(\epsilon_n^*\xi_{ni}^* + \xi_{ni}I(||\Xi_{ni}(X)|| > \epsilon_n^*)\big)\Big|\Big| \ge 2\epsilon_n\Big)$$
$$\le P\Big(\max_j\Big|\Big|\sum_{i=1}^n w_{ij}\xi_{ni}^*\Big|\Big| \ge 2\frac{\epsilon_n}{\epsilon_n^*}\Big) + P\Big(\max_i ||\Xi_{ni}(X)|| \ge \epsilon_n^*\Big). \quad (34)$$

The second RHS term in (34) is by lemma 4 bounded by $(\epsilon_n^*)^{-p_\Xi}\sum_{i=1}^n \mathrm{E}\,||\Xi_{ni}||^{p_\Xi} = O\big((\log n)^{-p_\Xi/2}\big) = o(1)$. The first RHS term in (34) is also $o(1)$ because ess sup $||\xi_{ni}^*|| \le 1$ by construction and lemma 14 can be applied since

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

Lemma 16 $n^{-1/2}\sum_i \bar{A}_i(\hat{T}_i - \tilde{T}_i)\bar{T}_i^{-1}s_i = o_p(1)$.

Proof: The LHS is

$$\Big|\Big|n^{-1/2}\sum_{j=1}^n\sum_{i=1}^n w_{ij}\bar{A}_i(\hat{s}_j\hat{s}_j' - s_js_j')\bar{T}_i^{-1}s_i\Big|\Big| \overset{\text{L15}}{\le} \sum_{j=1}^n ||\hat{s}_j\hat{s}_j' - s_js_j'|| \times o_p(n^{1/p_x}k_n^{-1}\log n)$$
$$\overset{(16)}{\le} C_s\sum_{j=1}^n ||\hat{s}_j - s_j|| \times o_p(n^{1/p_x}k_n^{-1}\log n). \quad (35)$$

Set $r_n = \log n/\sqrt{n}$. Now,

$$\sum_{j=1}^n ||\hat{s}_j - s_j|| \overset{(17)}{\le} \sum_{j=1}^n\big(||\alpha_{jr_n}|| - \mathrm{E}(||\alpha_{jr_n}||\,|X)\big) + \sum_{j=1}^n \mathrm{E}(||\alpha_{jr_n}||\,|X) + nI(||\hat{\theta} - \theta_0|| > r_n). \quad (36)$$

The third RHS term is $o_p(1)$ by (19) and the second RHS term is by lemma 7 bounded by $C_\alpha r_n\sum_{j=1}^n ||X_j|| = O_p(nr_n) = O_p(n^{1/2}\log n)$. Squaring the first RHS term and taking its expectation yields

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

Hence the RHS in (36) is $O_p(\sqrt{nr_n}) + O_p(\sqrt{n}\log n) + O_p(1) = O_p(\sqrt{n}\log n)$, which implies that the RHS in (35) is $o_p\big(n^{1/p_x+1/2}k_n^{-1}(\log n)^2\big) = o_p(1)$ by A7. ■

Lemma 17 *Let $\xi_{nij} = \xi_n(u_i, u_j; X)$ be such that $\mathrm{E}(\xi_{nij}|u_i, X) = \mathrm{E}(\xi_{nij}|u_j, X) = 0$ a.s. for all i, j and $\max_{i,j} \mathrm{E}\,||\xi_{nij}||^2 = O(1)$. 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

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

Lemma 18 $n^{-1/2}\sum_{i=1}^n \bar{A}_i(\tilde{T}_i - \bar{T}_i)\bar{T}_i^{-1}s_i = o_p(1)$.

Proof: In lemma 17, take $\xi_{nij} = \bar{A}_i(s_j s_j' - T_j)\bar{T}_i^{-1}s_i$ to obtain a convergence rate of $O_p(n^{1/2}k_n^{-1}) = o_p(1)$. $\blacksquare$

Lemma 19 $n^{-1/2}\sum_i(\hat{S}_i - \tilde{S}_i)'\bar{T}_i^{-1}s_i = o_p(1)$.

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 lemma 12) $r_n = \log n/\sqrt{n}$. Then

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

$$\overset{(19)}{=} \sum_{j=1}^n||\hat{F}_j - \tilde{F}_j|| \times ||X_j||I(||\hat{\theta} - \theta_0|| \leq r_n) + o_p(1).$$

Finally,

$$\frac{\sqrt{\log n}}{k_n}\sum_{j=1}^n||\hat{F}_j - \tilde{F}_j|| \times ||X_j||I(||\hat{\theta} - \theta_0|| \leq r_n) \overset{(23),(25)}{\leq} \frac{C_f d r_n\sqrt{\log n}}{k_n\beta_n}\sum_{j=1}^n||X_j||^2 = O_p\Big(\frac{\sqrt{n}(\log n)^2}{k_n\beta_n}\Big) \overset{\text{A7}}{=} o_p(1). \quad \blacksquare$$

Lemma 20 $n^{-1/2}\sum_{i=1}^n(\tilde{S}_i - \bar{S}_i)'\bar{T}_i^{-1}s_i = o_p(1)$.

Proof: The LHS is

$$n^{-1/2}\sum_{i,j=1}^n w_{ij}\big(\tilde{F}_j - \mathrm{E}(\tilde{F}_j|X_j)\big)X_j'\bar{T}_i^{-1}s_i + n^{-1/2}\sum_{i,j=1}^n w_{ij}\big(\mathrm{E}(\tilde{F}_j|X_j) - F_j\big)X_j'\bar{T}_i^{-1}s_i. \tag{37}$$

The first RHS term is $O_p(n^{1/2}\beta_n^{-1/2}k_n^{-1}) = o_p(1)$ by lemma 17. The norm of the second RHS term is bounded by

$$n^{-1/2}\max_j\Big|\Big|\sum_{i=1}^n w_{ij}\bar{T}_i^{-1}s_i\Big|\Big|\sum_{j=1}^n w_{ij}\big|\big|\mathrm{E}(\tilde{F}_j|X_j)-F_j)\big|\big|\times||X_j||$$
$$\overset{\text{L14,(21)}}{\leq} O_p(k_n^{-1}\sqrt{\log n})6^{-1}C_f\beta_n^2\sum_j||X_j|| = O_p(nk_n^{-1}\beta_n^2\sqrt{\log n}) \overset{\text{A7}}{=} o_p(1). \quad ■$$

Lemma 21 $n^{-1/2}\sum_{i=1}^n \bar{A}_i(\hat{T}_i-\bar{T}_i)(\hat{T}_i^{-1}-\bar{T}_i^{-1})s_i = o_p(1)$.

Proof: Note that by lemmas 8, 9 and assumption A7,

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

Lemma 22 $n^{-1/2}\sum_{i=1}^n(\hat{S}_i-\bar{S}_i)'(\hat{T}_i^{-1}-\bar{T}_i^{-1})s_i = o_p(1)$.

Proof: Use a similar inequality to the one used in lemma 21 to obtain a rate of $n^{1/2}\zeta_{npS}\zeta_{npT} = o(1)$ by A7. ■

Lemma 23 $\mathrm{E}\,||\bar{A}_i - A_i||^2 = o(1)$.

Proof: The square of the LHS is bounded by $C\Big(\mathrm{E}\,||A_i||^4\,\mathrm{E}\,||\bar{T}_i-T_i||^4 + (\mathrm{E}\,||\bar{S}_i-S_i||^2)^2\Big) = o(1)$, by theorem 1 of Stone (1977). ■

Lemma 24 $n^{-1/2}\sum_{i=1}^n(\bar{A}_i-A_i)s_i = o_p(1)$.

Proof: $\mathrm{E}\Big|\Big|n^{-1/2}\sum_{i=1}^n(\bar{A}_i-A_i)s_i\Big|\Big|^2 \leq \mathrm{E}\,||\bar{A}_i-A_i||^2 = o(1)$, by lemma 23. ■

Lemma 25 $\hat{m}_n(\theta_0) - m_n(\theta_0) = o_p(n^{-1/2})$.

Proof: Using the expansion in (26) and (27)–(33), the stated result follows from lemmas 16, 18, 19, 20, 21, 22, and 24. ■

C Feasible Estimator

Lemma 26 *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 $\mathrm{E}\,||\bar{A}_i - A_i||^2 = o(\mu_{1n}^2)$; such μ_{1n} exist by lemma 23. Now, $\mathrm{E}\big(||\bar{A}_i - A_i||\,||\alpha_{ir_n}||\big) \overset{\text{L7}}{\leq} Cr_n\,\mathrm{E}\big(||\bar{A}_i - A_i||\,||X_i||\big) \overset{\text{Schwarz}}{\leq} Cr_n\sqrt{\mathrm{E}\big(||\bar{A}_i - A_i||^2\big)}\sqrt{\mathrm{E}\,||X_i||^2} = o(r_n\mu_{1n})$. ■

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

Lemma 27 *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_{1n} be as in lemma 26 and 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 lemma 13. 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 $\mu_n = \mu_{1n} + \mu_{2n}$. ■

Lemma 28 $m_n(\hat{\hat{\theta}}) = o_p(n^{-1/2})$.

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 lemma 27. 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 lemma 2, $||\hat{\hat{\theta}} - \theta_0|| = o_p(\psi_n) + O_p(n^{-1/2})$. Hence $\psi_n \sim n^{-1/2}$. ■

Proof of theorem II: By lemma 28, $\hat{\hat{\theta}}$ satisfies (8). ■

D Covariance Matrix Estimation

Proof of theorem III: Let $\bar{\Psi} = n^{-1}\sum_{i=1}^{n} \bar{A}_i \bar{S}_i$. Then

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

such that $\hat{\Psi} - \bar{\Psi} = o_p(1)$ by lemmas 11, 12 and 13. To show $\bar{\Psi} - \Psi = o_p(1)$, which is sufficient since $\Psi > 0$ by assumption, we can use an expansion similar to the one above, which leads to

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

Apply lemma 23, theorem 1 of Stone (1977) and the fact that $\mathrm{E}\,||A_i||^2, \mathrm{E}\,||S_i||^2 < \infty$ by assumption. ■

E Computation

Proof of theorem IV: By lemma 27 and theorem III it follows that $\hat{\theta}_{(1)} = O_p(n^{-1/2})$. Hence by lemma 2, $\hat{m}_n(\hat{\theta}_{(j)}) - m_n(\hat{\theta}_{(j)}) = o_p(n^{-1/2})$ for $j = 0, 1$. Because $\{A_i s_i\}$ is a VČ class (see (13)), 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 $m(\hat{\theta}_{(1)}) - m(\hat{\theta}_{(0)}) = \Psi(\hat{\theta}_{(1)} - \hat{\theta}_{(0)}) + o_p(n^{-1/2})$ (see (14)), it follows that

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

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

References

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

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

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–334.

Davidson, J. (1994) *Stochastic limit theory.* New York: Oxford University Press.

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

Koenker, R. (2005) *Quantile regression.* New York: Cambridge University Press.

Koenker, R. & Q. Zhao (1994) L–estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics* **3**, 223–235.

Komunjer, I. & Q. Vuong (2006) Efficient Conditional Quantile Estimation: The Time Series Case. UCSD working paper.

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

Newey, W. K. (1993), Efficient estimation of models with conditional moment restrictions. In G.S. Maddala, C.R. Rao and H.D. Vinod (ed.) *Handbook of Statistics* **11**. Amsterdam: North Holland, pp 419–454.

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

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica*

57, 1027–1057.

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

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

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

van der Vaart, A. & J. A. Wellner (1996) *Weak convergence and empirical processes: with applications to statistics*. New York: Springer.

Whang, Y. (2006) Smoothed Empirical Likelihood Methods for Quantile Regression Models. *Econometric Theory* **22**, 173–205.

Zhao, Q. (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≽S*,S,O	J≽S*,S,O
	$u_{i1} \not\perp\!\!\!\perp u_{i2}$	all same	J≽S*,S,O	J≽S*,S,O	J≽S*,S,O
$x_{ij} \perp\!\!\!\perp u_{ij^*}; j \neq j^*$	$u_{i1} \perp\!\!\!\perp u_{i2} \| x_{i1}, x_{i2}$	all same	J,S*,S≽O	J≽S*,S,O	J≽S*,S≽O
	$u_{i1} \not\perp\!\!\!\perp u_{i2} \| x_{i1}, x_{i2}$	all same	J≽S*,S≽O	J≽S*,S,O	J≽S*,S≽O
$x_{ij} \not\perp\!\!\!\perp u_{ij^*}$	$u_{i1} \perp\!\!\!\perp u_{i2} \| x_{i1}, x_{i2}$	J,S*,S≽O	J,S*≽S≽O	J≽S*,S≽O	J≽S*≽S≽O
	$u_{i1} \not\perp\!\!\!\perp u_{i2} \| x_{i1}, x_{i2}$	J≽S*,S≽O	J≽S*≽S≽O	J≽S*,S≽O	J≽S*≽S≽O

The entries indicate which methods are preferable to others in terms of asymptotic efficiency in various situations. '$\perp\!\!\!\perp$' denotes independence and '≽' 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* there are no cross–equation restrictions on the coefficient vectors, 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

		Slope Coefficients of Equation 1				Slope Coefficients of Equation 2			
		θ_{011}		θ_{012}		θ_{021}		θ_{022}	
	$n =$	100	500	100	500	100	500	100	500
	Unweighted LAD	0.0025	0.0004	0.0122	0.0025	0.0025	0.0005	0.0136	0.0026
Homo	Weighted LAD 1	0.0026	0.0004	0.0125	0.0026	0.0026	0.0005	0.0139	0.0027
$\rho = 0$	Weighted LAD 2	0.0026	0.0004	0.0125	0.0025	0.0025	0.0005	0.0136	0.0027
	SUR Estimation	0.0026	0.0004	0.0125	0.0026	0.0026	0.0005	0.0140	0.0027
	Unweighted LAD	0.0025	0.0005	0.0136	0.0027	0.0024	0.0005	0.0132	0.0027
Homo	Weighted LAD 1	0.0025	0.0005	0.0141	0.0027	0.0025	0.0005	0.0137	0.0027
$\rho = 0.7$	Weighted LAD 2	0.0025	0.0005	0.0137	0.0027	0.0025	0.0005	0.0139	0.0027
	SUR Estimation	0.0020	0.0004	0.0112	0.0020	0.0021	0.0004	0.0110	0.0021
	Unweighted LAD	0.0543	0.0097	0.1426	0.0255	0.0052	0.0009	0.0377	0.0078
Hetero	Weighted LAD 1	0.0496	0.0087	0.1361	0.0239	0.0054	0.0009	0.0378	0.0077
$\rho = 0$	Weighted LAD 2	0.0475	0.0083	0.1325	0.0230	0.0053	0.0009	0.0378	0.0076
	SUR Estimation	0.0487	0.0084	0.1383	0.0233	0.0054	0.0009	0.0389	0.0076
	Unweighted LAD	0.0567	0.0091	0.1384	0.0237	0.0053	0.0011	0.0423	0.0086
Hotero	Weighted LAD 1	0.0516	0.0081	0.1305	0.0222	0.0054	0.0011	0.0429	0.0085
$\rho = 0.7$	Weighted LAD 2	0.0498	0.0076	0.1258	0.0212	0.0053	0.0011	0.0418	0.0083
	SUR Estimation	0.0429	0.0062	0.1091	0.0166	0.0047	0.0009	0.0340	0.0061

Table 2: MSE of slope coefficients by Monte Carlo. Weighted LAD 1 used $\hat{f}_{u_{ij}|x_{ij}}(0)$ while Weighted LAD 2 used $\hat{f}_{u_{ij}|X_i}(0)$ with $j = 1, 2$. Typical convex and inverse U-shaped forms of heteroskedasticity were used for equations 1 and 2, respectively.