AN ALTERNATIVE WAY OF COMPUTING EFFICIENT INSTRUMENTAL VARIABLE ESTIMATORS*

Xiaohong Chen and Oliver Linton†
London School of Economics

September 7, 2001

Abstract

We propose a new way of constructing efficient semiparametric instrumental variable estimators. Our method is to combine a large number of possibly inefficient estimators rather than combining the instruments into an optimal instrument function. We establish the consistency and asymptotic normality for a class of estimators that are linear combinations of a set of root-n consistent estimators whose cardinality increases with sample size. We show that the semiparametrically efficient estimator lies in our class. We investigate the finite sample performance of our estimator and show that it does quite well.

Journal of Economic Literature Classification: C12, C13, C14

Keywords and phrases. Instrumental Variables; Minimum Distance; Semiparametric Efficiency; Two-Stage Least Squares

*We would like to thank Javier Hidalgo, Roger Koenker, Guido Kuersteiner, Benno Pötscher, and Tom Rothenberg for helpful comments, and Ayan Sen for research assistance. We thank STICERD, the NSF and the ESRC for financial support.

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail addresses: x.chen1@lse.ac.uk and lintono@lse.ac.uk.

1 Introduction

Instrumental variables estimation procedures are widely used and taught in econometrics courses. The approach is semiparametric in the sense that the joint distribution of the data is unspecified apart from a finite number of conditions. Frequently, this information is in the form of a finite number of conditional moments. These imply an infinite number of unconditional moment conditions, although a certain finite dimensional combination of them gives full efficiency – see e.g., Hansen (1985), Chamberlain (1987), Newey (1990, 1993) and Robinson (1991). The optimal instrument function involves an unknown conditional expectation. Therefore, to obtain full efficiency, it is necessary to use nonparametric regression techniques to estimate the optimal instruments. Newey (1990, 1993) established the asymptotic properties of a semiparametric instrumental variable estimator based on a nonparametric estimate (specifically, nearest neighbors and series estimates) of the optimal instrument function. Under regularity conditions, he showed that the standardized estimator is asymptotically normal with zero mean and finite variance and moreover has the same variance as the procedure based on the true unknown optimal instrument function. This method has been used in numerous applications (e.g., Olley and Pakes (1996)) and has been generalized in numerous ways (e.g., Ai and Chen (1999), Newey (1999)).

We propose an alternative approach to efficient estimation in models defined through conditional moment restrictions. Our method is based on combining many consistent but possibly inefficient GMM estimators in an optimal way [in a central case, the optimal weighting is inversely proportional to the asymptotic variance of the individual estimator and so estimation of the optimal weights is trivial]. This is in contrast to the usual method that proceeds by first combining the instruments into an optimal instrument function and then computing the estimator based on the corresponding criterion. Our approach has the advantage that one can see how much variation there is in the 'sample' of inefficient but $\sqrt{n}$-consistent estimators, where n is sample size. A common criticism of semiparametric estimation is that 'the results don't change much' as compared with 'parametric' methods. Our method allows the practitioner to see how much variation there is in the parameter estimates, and indeed one can see how much weight the optimal estimator puts on the different estimators in the pool. In cases where there truly is little variation then one can presumably make do with very simple inference rules. By contrast, there is not much interpretable information in the sample of different valid instruments. Nevertheless, the two methods achieve the same objective under similar conditions. In certain linear models, the two versions of the optimal estimator are actually identical, but in nonlinear models they are generically different. Our method is in effect a generalization of the classical method of minimum chi-squared or minimum distance discussed in Rothenberg (1973), which was conceived as a way of imposing equality restrictions in estimation via first estimating an unrestricted model and then finding the best combination of the unrestricted estimators that imposes the restrictions. In a number of cases this strategy is preferable to solving the constrained estimation problem directly. In our case, the best combination is linear with weights that sum to one. We first establish consistency and $\sqrt{n}$-asymptotic normality of a class of estimators that involve finite linear combinations of an infinite dimensional set of estimators, where the cardinality of the linear combinations increases with sample size. We allow the class of estimators considered to include those computed from discontinuous criterion functions that are nonlinear in the parameters and data. We also establish that a member of our class of estimators achieves the semiparametric efficiency bound for the conditional moment model. We discuss how to estimate the optimal combination of estimators. We also give a Monte Carlo experiment that shows how our procedure works in practice.

There is a vast literature on estimating models defined through conditional moment restrictions. We just mention one recent paper that is particularly relevant to our study, Koenker and Machado (1999). They considered a similar problem albeit restricted to certain linear models and to a rather specific estimator. They proved that a sufficient condition for the usual asymptotics for GMM to be valid when the number of moment conditions τ increases with n is that $\tau^3/n \to 0$. Their results can be interpreted as a warning not to include too many moment conditions in GMM: that the consequences of so doing are not just that no improvement results, but that the distributional approximation can break down. Our objective is quite different and we deal with nonlinear models.[1]

Our focus is on the variance of our estimators, but there may be some case for combining estimators to improve higher order bias as well. For example, Sawa (1973) considered combining k-class estimators in simultaneous equations systems, for the reason of improved bias. In the nonparametric literature, Gray and Schucany (1972) and Bierens (1987) have proposed 'jacknife' estimators that combine different kernel smoothers in order to reduce bias. Powell and Honoré (2001) have used essentially this device to reduce bias in a class of semiparametric estimation problems.

There are a number of other recent papers that have exploited the idea of combining estimators and forecasts, and Granger (2000) provides a nice discussion. Breiman (1996,1999) introduced the idea of bagging, which is based on using bootstrap resamples to compute a largeish sample of subsample estimators and then combining them. Watson (2000) and Stock and Watson (1999) propose various methods for combining large numbers of predictors to improve forecasting performance.

We use $\|A\| = (\mathrm{tr}(A'A))^{1/2}$ for any matrix A. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of a real symmetric matrix A.

[1] We do not search for the largest value of τ consistent with our asymptotics, although of course the Koenker and Machado *op. cit.* results provide an upper bound.

2 The Standard Approach

We observe an independent and identically distributed sample $\{Z_i\}_{i=1}^n$, where $Z_i = (Y_i, X_i)$. We suppose that there is a unique $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ satisfying the conditional moment conditions

$$E[\rho(Z_i, \theta_0)\,|X_i] = 0$$

with probability one, where $\rho(z, \theta)$ is a scalar residual function.[2] This implies the unconditional moment conditions

$$E[A(X_i)\rho(Z_i, \theta_0)] = 0, \tag{1}$$

for any $p \times 1$ vector $A(X_i)$ [for which the expectation exists]. The sample version of (1) is the basis of estimation as described in many previous papers, including Amemiya (1974) and Hansen (1982).

Suppose that $E[\rho(Z_i, \theta_0)^2\,|X_i] = \sigma_0^2(X_i)$ is positive with probability one, and that

$$D_0(X_i) = \left(E\left[\frac{\partial \rho}{\partial \theta}(Z_i, \theta)|X_i\right]\right)_{\theta=\theta_0}$$

exists with probability one. In this case, the optimal (instrumental variables) matrix is proportional to $A_{oiv}(X_i) = D_0(X_i)\sigma_0^{-2}(X_i)$, and the resulting optimal instrumental variables (or optimal GMM) estimator $\widetilde{\theta}_{oiv}$ has asymptotic variance $\Sigma_{oiv} = \{E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)']\}^{-1}$ – see for example Hansen (1985), Chamberlain (1987) and Newey (1990, 1993).[3]

Suppose that the optimal matrix $A_{oiv}(\cdot)$ is of unknown form, but can be represented, in an L_2 sense, by the following series expansion

$$A_{oiv}(x) = D_0(x)\sigma_0^{-2}(x) = \sum_{j=1}^{\infty} \beta_{j0}\phi_j(x),$$

where $\phi_j(\cdot)$ are known basis functions chosen by the practitioner, while β_{j0} are unknown coefficients determined uniquely by the basis.[4] For notational convenience we shall allow ϕ_j to be $p \times 1$ vectors;

[2]The generalization to a system setting is conceptually straightforward; in order to keep the notation simple we shall concentrate on the scalar single equation case.

[3]Note that if ρ is not differentiable but the matrix

$$D_0(X_i) = \left(\frac{\partial}{\partial \theta} E\left[\rho(Z_i, \theta)|X_i\right]\right)_{\theta=\theta_0}$$

exists, then one might still obtain efficiency by extending the proof in Newey and Powell (1990). We thank Whitney Newey for suggesting this.

[4]In order for the sum to converge, the coefficients β_j must decline as $j \to \infty$, at least when the basis functions are of fixed magnitude in j.

in general, β_{j0} depends on θ_0 and is a $p \times p$ matrix. A common approach here is to estimate the coefficients β_{j0} and then to let

$$\widehat{A}_\theta(x) = \sum_{j=1}^{\tau(n)} \widehat{\beta}_j(\theta)\phi_j(x),$$

where $\tau(n)$ is some truncation sequence that goes to infinity with sample size but at a slow rate.[5] Then let $\widetilde{\theta}$ be any sequence that satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{A}_{\widetilde{\theta}}(X_i)\rho(Z_i, \widetilde{\theta}) = o_p(n^{-1/2}).$$

In current parlance this would be called a continuously updated optimal IV estimator. An alternative method is to use some preliminary estimator of θ to first construct an estimator of A, and then to solve a similar first order condition with the estimated instrument. Newey (1990, 1993) showed that such an estimator is asymptotically equivalent to the instrumental variable procedure based on knowing the optimal instrument function A_{oiv} and computing solutions $\widetilde{\theta}$ to

$$\frac{1}{n}\sum_{i=1}^{n} A_{oiv}(X_i)\rho(Z_i, \widetilde{\theta}) = o_p(n^{-1/2}).$$

See Newey and McFadden (1994) for discussion. There have been a number of alternative suggestions made more recently with a view to improving small sample performance, Newey and Smith (2001) contains an excellent review of this literature.

3 Our Estimation Idea

We take a different approach. Instead of estimating the optimal instrument function we will estimate the optimal way to combine all the available estimators. We consider a sequence of pre-specified basis ($p \times 1$ vector-valued) functions $\{A_j(\cdot)\}$ such that $E[||A_j(X_i)||^2] < \infty$; for instance, we may take a uniformly bounded basis such as the B-spline basis. We define the estimators $\widehat{\theta}_j$ $[j = 1, 2, \ldots]$ as any sequence that satisfies

$$G_{nj}(\widehat{\theta}_j) = \frac{1}{n}\sum_{i=1}^{n} A_j(X_i)\rho(Z_i, \widehat{\theta}_j) = o_p(n^{-1/2}). \quad (2)$$

For each j, this problem is completely parametric and will result in a root-n consistent and asymptotically normal estimator $\widehat{\theta}_j$ (under standard conditions).[6] We combine these estimators in a linear

[5] One can also directly estimate the conditional expectations inside A by nearest neighbor, kernels, or series methods.

[6] It is easy to allow the data in (2) to depend on j, but we have suppressed this notationally. For example, we could have $Z_i \in \mathbb{R}^\infty$ but with only a finite number of variables in each estimating equation.

fashion to produce a new estimator

$$\widehat{\theta} = \sum_{j=1}^{\tau(n)} W_{nj}\widehat{\theta}_j, \tag{3}$$

where $\tau = \tau(n)$ is a truncation parameter and W_{nj} are some given matrix weights that sum to the identity. This defines a class of estimators $\mathcal{E}$ indexed by the weighting matrices $\{W_{nj}, j = 1, \ldots, \tau(n)\}$; as we show below, by an appropriate choice of weights one can achieve the semiparametric efficiency bound for this problem, i.e., the semiparametrically efficient estimator is a member of $\mathcal{E}$. Before we get to that point we discuss the estimator with given weights.

This estimator (3) is a form of minimum distance where the number of restrictions could increase with sample size.[7] Even though each criterion function G_{nj} is a nonlinear function of θ, the computational costs of this procedure may not be so great, since one can use the first estimator as starting values in the computation of the second estimator etc. Additional computational issues arise in connection with the weights W_{nj} but these are discussed below.

Example 1. Classical two stage least squares in simultaneous equations. Suppose that

$$y_{1i} = \theta y_{2i} + \varepsilon_i \quad ; \quad y_{2i} = \pi_2' X_i + u_i,$$

where (ε_i, u_i) are i.i.d. error terms, $E[\varepsilon_i|X_i] = 0$, $E[u_i|X_i] = 0$ and $X_i \in \mathbb{R}^k$. The two stage least squares estimator is

$$\widetilde{\theta} = \frac{\sum_{i=1}^n \widehat{y}_{2i} y_{1i}}{\sum_{i=1}^n [\widehat{y}_{2i}]^2} = \frac{\sum_{i=1}^n \widehat{y}_{2i} y_{1i}}{\sum_{i=1}^n \widehat{y}_{2i} y_{2i}}, \tag{4}$$

where $\widehat{y}_{2i} = \widehat{\pi}_2' X_i$ and $\widehat{\pi}_2$ is the vector of least squares estimates obtained from the reduced from regression of y_{2i} on all the instruments $X_i = (X_{1i}, \ldots, X_{ki})'$. Our estimator is

$$\widehat{\theta} = \sum_{j=1}^{k} W_{nj}\widehat{\theta}_j,$$

where

$$\widehat{\theta}_j = \frac{\sum_{i=1}^n \widehat{y}_{2i}^j y_{1i}}{\sum_{i=1}^n [\widehat{y}_{2i}^j]^2} = \frac{\sum_{i=1}^n \widehat{y}_{2i}^j y_{1i}}{\sum_{i=1}^n \widehat{y}_{2i}^j y_{2i}^j}, \tag{5}$$

where $\widehat{y}_{2i}^j = \widehat{\pi}_{2j} X_{ji}$, and $\widehat{\pi}_{2j}$ is the least squares estimates obtained from the reduced from regression of y_{2i} on the single instrument X_{ji} for $j = 1, \ldots, k$. Here, W_{nj} are scalar weights that satisfy $\sum_{j=1}^k W_{nj} = 1$. There is a choice of W_{nj} that makes $\widehat{\theta}$ asymptotically equivalent to the two stage least squares estimator $\widetilde{\theta}$, see below. The classical minimum distance estimator (generalized indirect least squares) exploits the relationship between the reduced form coefficients and the structural parameter, i.e., $\pi_{1j}/\pi_{2j} = \theta$, where $\pi_{\ell j} = E[y_{\ell i} X_{ji}]/E[X_{ji}^2]$ are the parameters of the reduced form of $y_{\ell i}$ on X_{ji}

[7] See Rothenberg (1973) and Newey and McFadden (1994) for finite fixed τ.

for $\ell = 1, 2$ and $j = 1, \ldots, k$ [the estimator is a linear combination of $\widehat{\pi}_{1j}/\widehat{\pi}_{2j}$, where $\widehat{\pi}_{\ell j}$ are the corresponding reduced from estimators], see Rothenberg (1973).

Example 2. Now consider the infinite order regression model

$$Y_i = \sum_{k=1}^{\infty} X_{ki}\beta_k(\theta) + \varepsilon_i, \tag{6}$$

where θ is some finite dimensional parameter and ε_i is an error term satisfying $E(\varepsilon_i X_{ji}) = 0$, $j = 1, 2, \ldots$. Consider the special case that $\beta_k(\theta) = \theta$ for all k. Then, we need at least that $E[(\sum_{k=1}^{\infty} X_{ki})^2] < \infty$ in order for the summation in (6) to be well defined; this would be satisfied if $\sigma_k^2 = E(X_{ki})^2$ goes to zero at a rate faster than k^{-1} as $k \to \infty$. The optimal estimator under homoskedasticity is the OLS estimator of Y_i on $\sum_{k=1}^{\infty} X_{ki}$. If also the regressors are mutually orthogonal, i.e., $E(X_{ji}X_{ki}) = 0$ for all $j \neq k$, the OLS estimators of Y_i on X_{ki} are consistent, and so will any linear combination thereof, and so we can construct estimators of θ by taking linear combinations of these marginal OLS regressions.[8]

There are two tasks we now pursue. The first is to prove that such an estimator (3) is consistent and root-n asymptotically normal under general conditions on the truncation parameter and weighting sequence. The second task is to determine the optimal choice of weights.

4 Large Sample Properties

We begin by defining the sample and population first order conditions. For $j = 1, 2, \ldots,$ let

$$G_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A_j(X_i)\rho(Z_i, \theta) \text{ and } G_j(\theta) = EG_{nj}(\theta)$$

We do not assume that the function $G_{nj}(\theta)$ is differentiable or even continuous, although smoothness conditions are imposed on the expectation $G_j(\theta)$. In this way, we allow also quantile regression estimators (e.g., Koenker and Bassett, 1978), Huber's (1967) M-estimators, and simulation-based estimators (e.g., McFadden (1989) and Pakes and Pollard (1989)). For some of the arguments we only require high level conditions on the sample and population first order conditions, and so our results can apply more generally to any linear combination of estimators that have appropriate expansions.

[8] By changing variables to X_{ki}/σ_k the parameters become $\theta \cdot \sigma_k$ in which case the problem is more like the instrumental variables regression because the regressors have the same variance but the parameters decline in importance.

4.1 Consistency

In this subsection we give our consistency result for the estimator (3). We make the following assumptions.

Assumption A

1. The triangular array $\{W_{nj}\}_{j=1}^{\tau(n)}$, $n = 1, \ldots,$ satisfies

$$\sum_{j=1}^{\tau(n)} W_{nj} = I_p \text{ and } \sup_n \sum_{j=1}^{\tau(n)} \|W_{nj}\| < \infty \tag{7}$$

   with probability tending to one. Here, $\tau(n)$ satisfies $\tau(n) \to \infty$ as $n \to \infty$.

2. For each j, $\|G_j(\theta_0)\| = 0$.

3. For all $\delta > 0$ and $n \geq 1$, there is an $\epsilon_n(\delta) > 0$ (with $\epsilon_n(\delta) \to 0$) such that

$$\min_{1 \leq j \leq \tau(n)} \inf_{\|\theta - \theta_0\| > \delta} \|G_j(\theta)\| \geq \epsilon_n(\delta) > 0.$$

4. For the sequences $\epsilon_n(\delta), \tau(n)$ defined above, there exists a positive sequence α_{1n} with $\sup_n(\alpha_{1n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{1 \leq j \leq \tau(n)} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\theta \in \Theta} \|G_{nj}(\theta)\| \right) = o_p(\alpha_{1n}),$$

5. For the sequences $\epsilon_n(\delta), \tau(n)$ defined above, there exists a positive sequence α_{2n} with $\sup_n(\alpha_{2n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{1 \leq j \leq \tau(n)} \sup_{\theta \in \Theta} \|G_{nj}(\theta) - G_j(\theta)\| = o_p(\alpha_{2n}).$$

The assumptions on the weights are quite weak and are satisfied by many suitable weighting sequences both random and non-random. For example, equal weighting $W_{nj} = 1/\tau(n)$ satisfies the assumption A1. There are no explicit conditions on the truncation sequence $\tau(n)$ here, but the assumptions A3-A5 may require some restrictions on the rate at which $\tau(n)$ increases with n. Assumption A4 is just a definition of the estimator and is a little bit stronger than usual due to the uniformity over j requirement.

The identification Assumption A3 takes account of the fact that each additional moment condition is adding less and less information. The rate at which $\epsilon_n(\delta)$ declines is determined by the sequence

$\tau(n)$ and by the sequence A_j, in particular the rate at which $||E[A_j(X)]||$ decreases. By choosing $\tau(n)$ to grow very slowly we can compensate for a rapid decline in the moments of the instruments.

The uniform convergence Assumption A5 is easy to verify, although it is slightly stronger than usual due to the $\max_{1\leq j\leq\tau(n)}$ factor. This factor costs little extra, as can be verified from the Bonferroni and exponential inequalities, see below. Since we must have $\epsilon_n(\delta)$ of larger order than $n^{-1/2}$ in the case of i.i.d. data this puts an upper limit on the rate at which $\tau(n)$ can grow, but no lower limit. If $\tau(n)$ only increases very slowly, say like $\log n$, the stated rate is easy to achieve.

Theorem 1 (i). *Suppose that Assumptions A1-A5 hold. Then* $\widehat{\theta}-\theta_0=o_p(1)$.

For the purpose of obtaining $\sqrt{n}-$asymptotic normality of $\widehat{\theta}$ in the next subsection, we need to first establish that $\widehat{\theta}-\theta_0=o_p(n^{-1/4})$ under the following stronger version of Assumption A:

Assumption A'

1. A1 holds.

2. A2 holds.

3. For all $\delta_n=o(1)$ and $n\geq 1$, there is a positive c_n which could slowly increase to $+\infty$ such that

$$\min_{1\leq j\leq\tau(n)}\inf_{\|\theta-\theta_0\|>\delta_n}\|G_j(\theta)\|\geq\delta_n c_n>0.$$

4. For all $\delta_n=o(1)$ and $n\geq 1$,

$$\max_{1\leq j\leq\tau(n)}\left(\|G_{nj}(\widehat{\theta}_j)\|-\inf_{\|\theta-\theta_0\|\leq\delta_n}\|G_{nj}(\theta)\|\right)=o_p(n^{-1/4}).$$

5. For all $\delta_n=o(1)$ and $n\geq 1$,

$$\max_{1\leq j\leq\tau(n)}\sup_{\|\theta-\theta_0\|\leq\delta_n}\|G_{nj}(\theta)-G_j(\theta)\|=o_p(n^{-1/4}).$$

Assumption A4' is just a definition of the estimator and is a little bit stronger than usual due to the uniformity over j requirement. Assumption A5' is stronger than usual, in that we are taking a maximum over an increasing number of first order conditions and requiring a rate at which the resulting random variable goes to zero. However, it is likely to be satisfied in most problems. The uniformity across θ is usually satisfied, indeed we can expect in many cases that $\sup_{\theta\in\Theta}\|G_{nj}(\theta)-G_j(\theta)\|=O_p(1/\sqrt{n})$ for any compact parameter set Θ. Below we provide a lemma that can be used to verify the uniformity across j condition and may be useful elsewhere.

Theorem 1 (ii). *Suppose that Assumptions A1-A2, A3'-A5' hold. Then* $\widehat{\theta} - \theta_0 = o_p(n^{-1/4})$.

Of course there are many alternative ways to impose sufficient conditions which lead to convergence rate. We conclude this subsection with a result that is needed in verifying Assumption A5' etc. above.

Lemma 1. *Let* U_{ji} *be a triangular array of random variables,* $i = 1, \ldots, n$, $j = 1, \ldots, \tau(n)$, *i.i.d. across* i *for each* j *with* $E(U_{ji}) = 0$ *and* $E[|U_{ji}|^{\kappa}] = c_j < \infty$ *for some* $\kappa \geq 2$. *Let* $s_{nj}^2 = \sum_{i=1}^n \text{var}(U_{ji}) = n\sigma_j^2$, *where* $\sigma_j^2 \to \infty$ *as* $j \to \infty$, *and let*

$$a_n = (\max_{1 \leq j \leq \tau(n)} \sigma_j^2) \log \tau(n) + \left(\sum_{j=1}^{\tau(n)} \frac{c_j^2}{\sigma_j^{2\kappa}} \right)^{1/\kappa}. \tag{8}$$

Then we have for $\delta_n = a_n \varrho_n$ *for any increasing sequence* ϱ_n *that*

$$\max_{1 \leq j \leq \tau(n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n U_{ji} \right| = o_p(\delta_n).$$

For example if we take $\kappa = 2$, then $a_n = (\max_{1 \leq j \leq \tau(n)} \sigma_j^2) \log \tau(n) + \sqrt{\tau(n)}$. One application of this lemma is when $n^{-1/2} \sum_{i=1}^n U_{ji}$ is the leading term of the estimator $\widehat{\theta}_j$, in which case, σ_j^2 would be Γ_j^{-1} (under homoskedasticity.). Therefore, the corresponding a_n is of order $\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)}$. Provided $\tau(n)$ does not increase too rapidly, this is less than $n^{1/4}$ as would be required by assumption A5'. Furthermore, it implies that $\max_{1 \leq j \leq \tau} ||\widehat{\theta}_j - \theta_0||$ goes to zero no slower in probability ability than $(\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)})/\sqrt{n}$.

4.2 Asymptotic Normality

In this subsection we derive the asymptotic distribution of our estimator $\widehat{\theta}$, under additional conditions. We strengthen the conditions of Pakes and Pollard (1989) and Newey and McFadden (1994) to accommodate our more general set-up, but again we do not require smoothness conditions on the residual function $\rho(Z_i, \theta)$. Let $g_j(Z_i, \theta) = A_j(X_i)\rho(Z_i, \theta)$ for each j. Then $G_{nj}(\theta) = n^{-1} \sum_{i=1}^n g_j(Z_i, \theta)$ and $G_j(\theta) = E[g_j(Z_i, \theta)]$. We denote

$$\Gamma_j = \frac{\partial}{\partial \theta'} G_j(\theta_0) = \frac{\partial}{\partial \theta'} E[A_j(X_i)\rho(Z_i, \theta)] \mid_{\theta = \theta_0}.$$

If $D_0(X_i) = \{\partial E[\rho(Z_i, \theta)|X_i]/\partial \theta\}|_{\theta = \theta_0}$ exists with probability one, then we have

$$\Gamma_j = E[A_j(X_i) D_0(X_i)'].$$

Assumption B

1. $\max_{1\le j\le\tau(n)}\left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\|\theta-\theta_0\|\le\delta_n}\|G_{nj}(\theta)\|\right) = o_p(1/\sqrt{n})$ for any $\delta_n = o(n^{-1/4})$;

2. There exists a finite C such that for any θ within a shrinking $(n^{-1/4}-)$ neighborhood of θ_0

$$\max_{1\le j\le\tau(n)}\|G_j(\theta) - \Gamma_j(\theta-\theta_0)\| \le C\|\theta-\theta_0\|^2,$$

where Γ_j is of full (column) rank for each j.

3. (a) $\max_{1\le j\le\tau(n)}\|\sqrt{n}[G_{nj}(\theta_0) - G_j(\theta_0)]\| = O_p(1)$;

   (b) For any $\delta_n = o(n^{-1/4})$,

$$\max_{1\le j\le\tau(n)}\sup_{\|\theta-\theta_0\|\le\delta_n}\|[G_{nj}(\theta) - G_j(\theta)] - [G_{nj}(\theta_0) - G_j(\theta_0)]\| = o_p(1/\sqrt{n}).$$

4. There exists a deterministic sequence of matrices W^0_{nj} satisfying: (a) $\sum_{j=1}^{\tau(n)}||(W_{nj}-W^0_{nj})\Gamma_j^{-1}|| = o_p(1)$; (b) $\limsup_n \sum_{j=1}^{\tau(n)}\left\|W^0_{nj}\Gamma_j^{-1}\right\| < \infty$.

5. (a) The matrix $\Sigma_n = \sum_{j=1}^{\tau(n)}\sum_{l=1}^{\tau(n)} W^0_{nj}V_{jl}W^{0\prime}_{nl}$ has a finite positive definite limit Σ, where for all $j, l = 1,\ldots,\tau(n)$,

$$V_{jl} = \Gamma_j^{-1}E[g_j(Z_i,\theta_0)g_l(Z_i,\theta_0)']\Gamma_l^{-1\prime} = \Gamma_j^{-1}E[A_j(X_i)\sigma_0^2(X_i)A_l(X_i)']\Gamma_l^{-1\prime};$$

   (b) the triangular array of random variables $f_n(Z_i) = n^{-1/2}\sum_{j=1}^{\tau(n)} c'W^0_{nj}\Gamma_j^{-1}g_j(Z_i,\theta_0)$ satisfies $nE|f_n(Z_i)|^{2+\kappa} \to 0$ for all vectors c and some $\kappa > 0$.

6. θ_0 is in the interior of Θ.

7. $\max_{1\le j\le\tau(n)}||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4})$.

Assumption B1 is just the definition of the estimator and is a little bit stronger than usual. Assumption B2 requires essentially two uniformly continuous derivatives for the population moment function at $\theta = \theta_0$ and that the first derivative matrix be of full rank.

For Assumption B3(b), the empirical distribution function satisfies

$$\sup_{|x-x_0|\le a/n^\alpha}\left|\sqrt{n}[F_n(x) - F(x)] - \sqrt{n}[F_n(x_0) - F(x_0)]\right| = O_p(n^{-\alpha/2})$$

for any $\alpha < 1$.[9] The cost of the additional max is typically no more than an additional factor of order $\sqrt{\tau(n)}$ as is evidenced in the Lemma 1.

[9]We are grateful to Benedikt Pötscher for pointing this out to us. This is due to the Hölder continuity of the limiting Brownian bridge process $B(\cdot)$ of $\sqrt{n}[F_n(\cdot) - F(\cdot)]$, i.e., $|B(x) - B(x_0)| \le c\cdot|x-x_0|^{1/2}$ for some random variable c with bounded moment. The local uniformity [across i] comes at very little extra cost.

In B4, we require that if the weights are random that they can be well approximated by some nonrandom sequence with certain summability properties. This condition entails some restrictions on the rate of growth of τ, and these restrictions can be as much as requiring that $\tau^3/n \to 0$, see Koenker and Machado (1999). The restrictions are not so stringent in special cases and really arise out of the nonlinearity of the estimating equation rather combined with the large number of parameters.

Assumption B5 allows us to apply the Liapounov's central limit theorem for triangular arrays to the leading term. This condition is satisfied for a variety of problems, and it implicitly imposes restrictions on how fast $\tau(n)$ could grow with sample size n. Notice that Assumption B5(b) is simply: for some $\kappa > 0$ and for all c,

$$E\left(\left|\sum_{j=1}^{\tau(n)} c' W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)\right|^{2+\kappa}\right) = o(n^{\kappa/2}).$$

For example, suppose we only require that $g_j(Z_i, \theta_0)$ have uniformly bounded fourth moments. Then, by the Cauchy-Schwarz inequality

$$nE[f_n(Z_i)^4] = \frac{1}{n} \sum_{j,k,l,m=1}^{\tau(n)} E[\varphi_{ji}\varphi_{ki}\varphi_{li}\varphi_{mi}] \leq \frac{1}{n\epsilon_n^4}\left(\sup_n \sum_{j=1}^{\tau(n)} \left\|W_{nj}^0\right\|\right)^4,$$

where $\varphi_{ji} = c' W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)$. It suffices in this case that $n\epsilon_n^4 \to \infty$. Now suppose that in fact, the scalar $g_j(Z_i, \theta_0)$ are normally distributed with mean zero and variance Γ_j and mutually independent, and that the weights are equal, i.e., $W_{nj}^0 = 1/\tau(n)$ for each j. Then

$$nE[f_n(Z_i)^4] = \frac{1}{n\tau^4}\left(\sum_{j=1}^{\tau(n)} 3\Gamma_j^{-2} + 3\sum_{j\neq k}^{\tau(n)} \Gamma_j^{-1}\Gamma_k^{-1}\right) \leq \frac{3}{n\tau^2\epsilon_n^2},$$

which goes to zero provided $n\tau^2\epsilon_n^2 \to \infty$. These conditions can be weakened considerably in special cases.

Notice that we can replace Assumptions B3(a) and B5 by the condition that $\{G_{nj}(\theta_0) - G_j(\theta_0) : 1 \leq j \leq \tau(n)\}$ is a Donsker class, i.e., it satisfies the uniform CLT. This kind of assumption has been used in Portnoy (1985) for example.

The condition B7 that $\max_{1\leq j\leq \tau(n)} ||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4})$ follows from our Theorem 1(ii). It may be possible to prove our result below without a sup-norm convergence result like this, although we have not been able to find a proof based on other convergences like L_p. The usual proofs in other semiparametric estimation problems typically make use of similar results about the convergence of nuisance parameters.

Theorem 2. *Suppose that Assumptions B1-B7 hold. Then* $\sqrt{n}(\widehat{\theta} - \theta_0) \Longrightarrow \mathcal{N}(0, \Sigma)$.

The asymptotic variance matrix Σ depends on the weighting scheme and on the class of estimators considered and of course on the underlying distribution of the data. We discuss the nature of the asymptotic variance more in the next section.

To construct consistent estimates of Σ, we would compute

$$\widehat{\Sigma} = \sum_{j=1}^{\tau(n)} \sum_{l=1}^{\tau(n)} \kappa \left(\frac{|j-l|}{\tau} \right) W_{nj} \widehat{V}_{jl} W_{nl}'$$

for some weighting function κ, and

$$\widehat{V}_{jl} = \widehat{\Gamma}_j^{-1} \frac{1}{n} \sum_{i=1}^{n} g_j(Z_i, \widehat{\theta}) g_l(Z_i, \widehat{\theta})'] \widehat{\Gamma}_l^{-1\prime}.$$

The estimation of Γ_j is easy when G_{nj} are differentiable. In this case,

$$\widehat{\Gamma}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_j(Z_i, \widehat{\theta})}{\partial \theta} \to^p \Gamma_j \tag{9}$$

under some regularity conditions. The weighting function κ must satisfy some regularity conditions as in Andrews (1991). When G_{nj} are not differentiable, as for example in the LAD case, this method is not feasible. In some cases, one might be able to estimate directly the quantity Γ_j. For example, in the LAD case [with errors independent of covariates], Γ_j is proportional to the density of the errors evaluated at their median. This quantity can be estimated by a variety of nonparametric methods. A general strategy for estimating Γ_j is to use 'numerical derivatives', that is, let

$$\widehat{\Gamma}_{j;lk} = \frac{1}{n} \sum_{i=1}^{n} \frac{g_{jl}(Z_i, \widehat{\theta} + \delta e_k) - g_{jl}(Z_i, \widehat{\theta})}{\delta}, \tag{10}$$

where e_k is a vector of zeros with one in the k^{th} position, while δ is a small constant. If we let $\delta(n)$ go to zero at a certain rate as sample size increases, we can show that $\widehat{\Gamma}_{j;lk} \to^p \Gamma_{j;lk}$, and under stronger conditions $\max_{1 \le j \le \tau} ||\widehat{\Gamma}_j - \Gamma_j|| \to^p 0$. The actual derivative (9) makes δ go to zero before n, but our modified estimator (10) allows δ to go to zero with n and indeed slower than n.

Example 2 continued. Suppose that the errors are homoskedastic and the regressors are mutually orthogonal with $E(X_{ji}^2) = \sigma_j^2$. A necessary and sufficient condition for the root-n rate of convergence is that

$$\limsup_{n} \sum_{j=1}^{\tau} W_{nj}^2 \sigma_j^{-2} < \infty$$

with probability one. Since we also require $\sum_{j=1}^{\infty} \sigma_j^2 < \infty$, this rules out the equal weighting case. Nevertheless, a variety of weighting conditions satisfy the requirement. Furthermore, there is no explicit restriction on τ itself in this case.

5 Optimal Weights

5.1 Fixed τ Case

We now discuss the optimal weights. Suppose that we know only that

$$E\left[A_j(X_i)\rho(Z_i,\theta_0)\right]=0, \quad j=1,\dots,\tau, \tag{11}$$

where τ is fixed, and $A_j \in \mathbb{R}^p$. This is a standard unconditional moments estimation problem, and the optimal estimator can be arrived at by several routes.

First, GMM with optimal combination of the moment conditions: that is, we minimize the quadratic form

$$G_n^\tau(\theta)' W_n G_n^\tau(\theta) \tag{12}$$

with respect to θ, where $G_n^\tau(\theta)=n^{-1}\sum_{i=1}^n A^\tau(X_i)\rho(Z_i,\theta)$ with $A^\tau=(A_1',\dots,A_\tau')'\in\mathbb{R}^{\tau p}$ (i.e., $G_n^\tau(\theta)$ is the $\tau p\times 1$ vector containing all the sample moments). The asymptotically optimal weighting matrix is $W_{opt}=\Psi_\tau^{-1}$, where $\Psi_\tau=E[G_n^\tau(\theta_0)G_n^\tau(\theta_0)']=E[A^\tau(X)\sigma_0^2(X)A^\tau(X)']\in\mathbb{R}^{\tau p\times\tau p}$.

Second, by optimal instrumental variables with the optimal instrument in this case being a linear combination of the $A_j(X_i)$, $j=1,\dots,\tau$. That is, we solve the equations

$$\Gamma^{\tau\prime}\Psi_\tau^{-1}G_n^\tau(\widehat{\theta})=0, \tag{13}$$

where $\Gamma^\tau=\partial E[G_n^\tau(\theta_0)]/\partial\theta=E[A^\tau(X)D_0(X)']\in\mathbb{R}^{\tau p\times p}$. These two approaches provide the optimal IV (optimal GMM) estimator $\widetilde{\theta}_{oiv}^\tau$ of θ_0 for the model (11). Specifically, we have $\sqrt{n}(\widetilde{\theta}_{oiv}^\tau-\theta_0)\Longrightarrow N(0,\Sigma_{oiv}^\tau)$ as $n\to\infty$, where the asymptotic variance is given by (see e.g., Hansen (1982) for differentiable ρ, Newey and McFadden (1994) for non-differentiable ρ):

$$\begin{aligned}\Sigma_{oiv}^\tau &= \left(E[A^\tau(X)D_0(X)']'\left[E(A^\tau(X)\sigma_0^2(X)A^\tau(X)')\right]^{-1}E[A^\tau(X)D_0(X)']\right)^{-1} & (14)\\ &= \left(\Gamma^{\tau\prime}\Psi_\tau^{-1}\Gamma^\tau\right)^{-1} & (15)\end{aligned}$$

and the optimal instrument for the model (11) is:

$$\begin{aligned}A_{oiv}^\tau(x) &= \Gamma^{\tau\prime}\Psi_\tau^{-1}A^\tau(x)\\ &= E[A^\tau(X)D_0(X)']'\left[E(A^\tau(X)\sigma_0^2(X)A^\tau(X)')\right]^{-1}A^\tau(x).\end{aligned}$$

The third approach to optimal estimation here is the minimum distance method described in Rothenberg (1973). Let $\widehat{\theta}_{omd}^\tau$ minimize the criterion function

$$Q_n(\theta)=\left[\begin{pmatrix}\widehat{\theta}_1\\ \vdots\\ \widehat{\theta}_\tau\end{pmatrix}-\theta\otimes i_\tau\right]' V^{-1}\left[\begin{pmatrix}\widehat{\theta}_1\\ \vdots\\ \widehat{\theta}_\tau\end{pmatrix}-\theta\otimes i_\tau\right], \tag{16}$$

where i_τ is a $\tau \times 1$ vector of ones, and V is the $\tau p \times \tau p$ asymptotic (as $n \to \infty$ holding τ constant) variance matrix of the vector $(\sqrt{n}(\widehat{\theta}_1 - \theta_0)', \ldots, \sqrt{n}(\widehat{\theta}_\tau - \theta_0)')'$, i.e., $V = (V_{j,l})$, where $V_{jl} = \Gamma_j^{-1} E[A_j(X_i)\sigma_0^2(X_i)A_l(X_i)']\Gamma_l^{-1\prime}$ for all $j, l = 1, \ldots, \tau$. The first order condition is

$$(I_p \otimes i_\tau)' V^{-1} \begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_\tau \end{pmatrix} = (I_p \otimes i_\tau)' V^{-1} (\widehat{\theta} \otimes i_\tau),$$

which implies that the optimal estimator $\widehat{\theta}^\tau_{omd}$ is a linear combination of the $\widehat{\theta}_j$ with

$$\widehat{\theta}^\tau_{omd} = \sum_{j=1}^{\tau} W_{0j}^{opt} \widehat{\theta}_j, \tag{17}$$

where

$$\begin{aligned} W_{0j}^{opt} &= \left(\sum_{l=1}^{\tau} B_l \right)^{-1} B_j, \\ \text{and } (B_1, \ldots, B_\tau) &= (I_p \otimes i_\tau)' V^{-1}. \end{aligned}$$

Furthermore, $\sqrt{n}(\widehat{\theta}^\tau_{omd} - \theta_0) \Longrightarrow N(0, \Sigma^\tau_{omd})$, where the asymptotic [as $n \to \infty$ and τ fixed] variance is

$$\Sigma^\tau_{omd} = W_{opt} V W_{opt}' = \left((I_p \otimes i_\tau)' V^{-1} (I_p \otimes i_\tau) \right)^{-1}.$$

Proposition 1. *For each fixed τ, $\widehat{\theta}^\tau_{omd}$ is asymptotically efficient for (11) with $\Sigma^\tau_{omd} = \Sigma^\tau_{oiv}$. Moreover the optimal weighting is simply*

$$\begin{aligned} W_{0j}^{oiv} &= -\left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j' \right)^{-1} \alpha_j \Gamma_j' \text{ for } j = 1, \ldots, \tau, \\ \text{with } (\alpha_1, \ldots, \alpha_\tau) &= \Gamma^{\tau\prime} \Psi_\tau^{-1}. \end{aligned}$$

Example 1 (cont.) Recall the optimal GMM estimator in this model [i.e., under homoskedasticity etc.] is simply the two stage least squares estimator

$$\widetilde{\theta} = (Y_2' P_X Y_2)^{-1} Y_2' P_X Y_1,$$

where $P_X = X(X'X)^{-1}X'$, $Y_1 = (y_{11}, \ldots, y_{1n})'$, $Y_2 = (y_{21}, \ldots, y_{2n})'$, $X = (X_1', \ldots, X_n')$, $X_i = (X_{1i}, \ldots, X_{ki})'$. Within our class of estimators $\mathcal{E}$, the optimal estimator is

$$\widehat{\theta} = \sum_{j=1}^{k} W_{nj}^{opt} \widehat{\theta}_j = (i_k' V^{-1} i_k)^{-1} i_k' V^{-1} \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_k \end{bmatrix},$$

where $\widehat{\theta}_j = (Y_2'P_jY_2)^{-1}Y_2'P_jY_1$ for $j = 1,\ldots,k$, where $P_j = X_j(X_j'X_j)^{-1}X_j'$ and V is the $k \times k$ covariance matrix with $V_{jl} = \text{acov}(\widehat{\theta}_j, \widehat{\theta}_l)$.[10] Suppose that the instruments are mutually orthogonal, then it is easy to see that $\widehat{\theta}$ is identically equal to $\widetilde{\theta}$.[11] This gives yet another interpretation to 2SLS as being the optimal combination of exactly identified instrumental variables estimators.[12] When the reduced form is nonlinear or the errors are heteroskedastic, the equivalences are only true in an asymptotic sense.

5.2 Increasing τ Case

Here we consider the more general case where τ increases with sample size. Let Σ_{oiv} be the asymptotic variance of the optimal instrumental variable estimator, and let Σ_{omd} be the asymptotic variance as $n \to \infty$ and $\tau(n) \to \infty$ of the optimal minimum distance estimator. The next theorem establishes that one can basically interchange the operations under additional assumptions.

Assumptions C

1. The matrix $D_0(X_i) = \left(\frac{\partial}{\partial\theta'}E[\rho(Z_i,\theta)|X_i]\right)|_{\theta=\theta_0}$ exists with probability one

2. $E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)']$ is finite and positive definite.

3. $D_0(X_i) = \sum_{j=1}^{\infty}\beta_{j0}\phi_j(X_i)\sigma_0^2(X_i)$, where the sequence $\{\phi_j\}$ is a complete orthonormal basis satisfying:

$$E[\sigma_0^2(X_i)\phi_j(X_i)\phi_l(X_i)'] = \begin{cases} 0_p & \text{for } j \neq l \\ I_p & \text{for } j = l. \end{cases}$$

Theorem 3. *Suppose that $E[\rho(Z_i,\theta_0)\,|X_i] = 0$ and that Assumptions C1-C3 hold. Then,*

$$\Sigma_{omd} = \Sigma_{oiv} = \left(E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)']\right)^{-1}.$$

The optimal weights in this case are any sequence like

$$W_{nj}^0 = \left(\sum_{l=1}^{\tau(n)} V_{ll}^{-1}\right)^{-1} V_{jj}^{-1},$$

[10]There is a connection with portfolio theory. Think of the estimators $\widehat{\theta}_j$ as being returns on asset j, where each asset has the same expected return but different variances. The optimal weights are the same as the weights for the global minimum variance portfolio, see Campbell, Lo, and MacKinlay (1997, pp 184-185). This is also related to the idea of combining many forecasts, see Stock and Watson (1999) and Granger (2000) for example.

[11]We are grateful to Tom Rothenberg for pointing this out to us.

[12]Interpreting 2SLS in various ways has a long history in econometrics; see Rothenberg (1974) for an early example.

where V_{jj} is the asymptotic variance matrix of $\sqrt{n}(\widehat{\theta}_j - \theta_0)$. With such a sequence of weights, $\widehat{\theta}$ has the same asymptotic variance as a comparable implementation of $\widetilde{\theta}$. Note that in the scalar homoskedastic case, the optimal weights W_{nj}^0 decrease at the same rate as β_{j0}^2 as $j \to \infty$, while the weights on the basis terms in the estimation of D_0 would decrease like β_{j0} as $j \to \infty$. This suggests that one needs to combine fewer estimators than instruments to achieve a specified variance.

In practice, one must use estimated weights. A nice property of the orthogonal case is that the weights W_{nj}^0 are proportional to the inverse of the asymptotic variance; therefore, given consistent estimators $\widehat{V}_{jj}$ of the asymptotic variances $V_{jj} = \text{var}(\sqrt{n}\widehat{\theta}_j)$, we can let

$$W_{nj} = \left(\sum_{l=1}^{\tau(n)} \widehat{V}_{ll}^{-1} \right)^{-1} \widehat{V}_{jj}^{-1}. \tag{18}$$

The issues surrounding estimating the optimal weights for similar problems have been treated in Newey (1990) and Koenker and Machado (1999). We do not pursue this further here, but refer the reader to these other papers.

6 Monte Carlo

We evaluated our methods on the sample selection model given below

$$Y_i = \beta_{10} + \beta_{20} s_i + \varepsilon_i \quad ; \quad s_i = 1\left(\alpha_{10} + \alpha_{20} X_i + \eta_i > 0\right),$$

$$\begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \right),$$

$$X_i \sim N(0,1) \quad ; \quad \alpha_{10} = \alpha_{20} = \beta_{10} = \beta_{20} = 1.$$

This model is considered in Newey (1994). The optimal instrument for s is $\pi(x) = \Pr[s = 1 | X = x]$, which makes $D(x) = (1, \pi(x))'$. We can write

$$\pi(x) = \sum_{j=1}^{\infty} \gamma_j H_j(x)$$

for some weights γ_j, where $H_j(x)$ are the Hermite polynomials computed via the recursion $H_{r+1}(x) = 2xH_r(x) - 2rH_{r-1}(x)$, where $H_1(x) = 1$, $H_2(x) = 2x$ etc. We thus consider the estimators that solve the equations $\sum_{i=1}^n (1, H_j(X_i))'(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 s_i) = 0$, which yields

$$\widehat{\beta}_j = \begin{pmatrix} \widehat{\beta}_{j1} \\ \widehat{\beta}_{j2} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n s_i \\ \sum_{i=1}^n H_j(X_i) & \sum_{i=1}^n H_j(X_i) s_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n H_j(X_i) Y_i \end{pmatrix}.$$

In this case, $\widehat{\beta}_j$ are orthogonal to $\widehat{\beta}_k$ for any $j \neq k$. We take $\tau = 10$. We evaluate the performance of the estimators $\widehat{\beta}_j$, $j = 2, \ldots, 10$, and

$$\sum_{j=1}^{\tau} W_j \widehat{\beta}_j$$

for fixed weights W_j, with $W_j = (1/\tau) I_2$ and $W_j = (j^{-\alpha} / \sum_{j=2}^{\tau} j^{-\alpha}) I_2$ for $\alpha \in \{1, 2, 3\}$. We also evaluate the performance of the 'optimal' estimator here, i.e., weights determined by (18). The numbers are computed from 5000 replications.

Tables Here

The performance of our optimal estimator of β_2 is roughly comparable to that shown in Newey (1990), with small biases and low variances. Note though that we have not tried to optimize over τ and so better performance might be obtained after some further experimentation. The individual estimators though can be quite erratic as can the raw average. The weighted estimator with weights proportional to $j^{-\alpha}$ seems to do quite well also, although in some cases the biases are rather large.

7 Conclusions and Extensions

Our approach has an advantage over the traditional approach to semiparametric instrumental variables in that one has a 'distribution' of estimators of the same quantity and one can view the range of values that these estimators take. If that range is not great, then it would appear that achieving efficiency is not going to be worth very much. If the range is considerable, then the efficient estimator may be very much better than any given estimator but at the same time performance might be very sensitive to how it is constructed. This information contained in the spread of the different estimators is similar to but not necessarily the same as the information contained in the standard error of an efficient estimator.[13] Also, the optimal weighting just requires the estimation of HAC matrices, at least in the orthonormal basis case, about which much has been written in econometrics.

It is quite straightforward to extend our work to produce results for the range[14]

$$\mathfrak{R}_n = \max_{1 \leq j \leq \tau(n)} \widehat{\theta}_j - \min_{1 \leq j \leq \tau(n)} \widehat{\theta}_j$$

[13] Actually, if the estimators themselves are mutually independent with the same limiting distribution, then the 95% confidence interval of a single estimator is approximately the same as the inter hemi-decile range, that is the interval $[\widehat{\theta}_{0.025 \cdot \tau}, \widehat{\theta}_{0.975 \cdot \tau}]$ of the ordered estimators. In fact, it is not possible that the estimators come from the same asymptotic distribution [since the variances must diverge along some trajectory], and so the two intervals do not coincide. Nevertheless, the connection exists.

[14] In the multiparameter case, we take the coordinate-wise ranges.

using the theory of extreme values for Gaussian processes [as in Bickel and Rosenblatt (1973) for example]. This statistic can be used as another way of measuring whether the observed range is consistent with the underlying model assumptions, i.e., as a model specification test.

A Appendix

Proof of Lemma 1. We show that

$$\Pr\left[\max_{1\leq j\leq\tau(n)}\left|\sum_{i=1}^{n}U_{ji}\right|\geq\lambda_n\right]\to 0$$

for any $\lambda_n=\delta_n\sqrt{n}$. For an array $\chi_{nj}\to\infty$ as $n\to\infty$ for each j, write

$$U_{ji}=U_{ji}1(|U_{ji}|\leq\chi_{nj})+U_{ji}1(|U_{ji}|>\chi_{nj})=\widetilde{U}_{ji}+\widetilde{\widetilde{U}}_{ji}.$$

We shall assume for simplicity that U_{ji} is symmetric about zero so that $E(\widetilde{U}_{ji})=0$. Therefore, $\widetilde{U}_{ji}$ are i.i.d. for each j with mean zero and are bounded from above by χ_{nj}. By the Bonferroni and Bernstein inequalities

$$\begin{aligned}\Pr\left[\max_{1\leq j\leq\tau(n)}\left|\sum_{i=1}^{n}\widetilde{U}_{ji}\right|\geq\lambda_n\right] &\leq \sum_{j=1}^{\tau(n)}\Pr\left[\left|\sum_{i=1}^{n}\widetilde{U}_{ji}\right|\geq\lambda_n\right]\\ &\leq \sum_{j=1}^{\tau(n)}\exp\left(\frac{-\lambda_n^2}{s_{nj}^2+2\lambda_n\chi_{nj}}\right). \qquad (19)\end{aligned}$$

We shall choose λ_n and χ_{nj} below to make this term vanish.

By the Bonferroni and Markov inequalities

$$\begin{aligned}\Pr\left[\max_{1\leq j\leq\tau(n)}\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|\geq\lambda_n\right] &\leq \sum_{j=1}^{\tau(n)}\Pr\left[\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|\geq\lambda_n\right]\\ &\leq \sum_{j=1}^{\tau(n)}\frac{E\left(\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|^{\kappa}\right)}{\lambda_n^{\kappa}}\\ &\leq \sum_{j=1}^{\tau(n)}\frac{n^{\kappa}E\left(|U_{ji}|^{\kappa}\right)\Pr\left[|U_{ji}|>\chi_{nj}\right]}{\lambda_n^{\kappa}}\\ &\leq \sum_{j=1}^{\tau(n)}\frac{n^{\kappa}[E\left(|U_{ji}|^{\kappa}\right)]^2}{\lambda_n^{\kappa}\chi_{nj}^{\kappa}}=o(1)\end{aligned}$$

provided $\sum_{j=1}^{\tau(n)} n^{\kappa}\chi_{nj}^{-\kappa}\lambda_n^{-\kappa}c_j^2 \to 0$.

Letting $\lambda_n = \delta_n\sqrt{n}$ and $\chi_{nj} = \sigma_j^2\sqrt{n}$ we need to show that:

$$\sum_{j=1}^{\tau(n)} \exp\left(\frac{-\delta_n}{\sigma_j^2}\right) \to 0 \text{ and } \frac{1}{\delta_n^{\kappa}}\sum_{j=1}^{\tau(n)} \frac{c_j^2}{\sigma_j^{2\kappa}} \to 0.$$

For the first condition it suffices that

$$\frac{\delta_n}{\max_{1\leq j\leq \tau(n)} \sigma_j^2 \log \tau(n)} \to \infty.$$

For the second condition it certainly suffices if

$$\frac{\delta_n}{\left(\sum_{j=1}^{\tau(n)} c_j^2 \sigma_j^{-2\kappa}\right)^{1/\kappa}} \to \infty.$$

■

Proof of Theorem 1 (i). From A3, if $\max_{1\leq j\leq \tau(n)} ||\widehat{\theta}_j - \theta_0|| > \delta$, then $\|G_j(\widehat{\theta}_j)\| \geq \epsilon_n(\delta)$ for some j. Consequently

$$\Pr\left(\max_{1\leq j\leq \tau(n)} ||\widehat{\theta}_j - \theta_0|| > \delta\right) \leq \Pr\left(\max_{1\leq j\leq \tau(n)} \|G_j(\widehat{\theta}_j)\| \geq \epsilon_n(\delta)\right), \tag{20}$$

and it is sufficient to prove that for the given $\epsilon_n(\delta) > 0$, the latter probability goes to zero. But

$$\begin{aligned}
\max_{1\leq j\leq \tau(n)} \|G_j(\widehat{\theta}_j)\| &\leq \max_{1\leq j\leq \tau(n)} \|G_j(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)\| + \max_{1\leq j\leq \tau(n)} \|G_{nj}(\widehat{\theta}_j)\|, \text{ (by the triangle inequality)} \\
&\leq \max_{1\leq j\leq \tau(n)} \sup_{\theta\in\Theta} \|G_j(\theta) - G_{nj}(\theta)\| + \max_{1\leq j\leq \tau(n)} \|G_{nj}(\widehat{\theta}_j)\|, \text{ (by set inclusion)} \\
&= o_p(\alpha_{2n}) + \max_{1\leq j\leq \tau(n)} \|G_{nj}(\widehat{\theta}_j)\|, \text{ (by A5)} \\
&= o_p(\alpha_{2n}) + \max_{1\leq j\leq \tau(n)} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\theta\in\Theta} \|G_{nj}(\theta)\|\right) + \max_{1\leq j\leq \tau(n)} \inf_{\theta\in\Theta} \|G_{nj}(\theta)\| \\
&\leq o_p(\alpha_{2n}) + \max_{1\leq j\leq \tau(n)} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\theta\in\Theta} \|G_{nj}(\theta)\|\right) + \max_{1\leq j\leq \tau(n)} \|G_{nj}(\theta_0)\| \\
&= o_p(\alpha_{2n}) + o_p(\alpha_{1n}) = o_p(\epsilon_n(\delta)), \text{ (by A4,A5 and A2).}
\end{aligned}$$

We conclude that $\max_{1\leq j\leq \tau(n)} ||\widehat{\theta}_j - \theta_0|| = o_p(1)$. Finally,

$$||\widehat{\theta} - \theta_0|| \leq \sum_{j=1}^{\tau(n)} \|W_{nj}\| \times \max_{1\leq j\leq \tau(n)} ||\widehat{\theta}_j - \theta_0|| = o_p(1)$$

by A1. ■

Proof of Theorem 1 (ii). Consistency implies that for every $\epsilon > 0$ there exists a sequence $\{\delta_n\}$, with $\delta_n \to 0$, and an N such that for all $n \geq N$,

$$\Pr\{\|\widehat{\theta} - \theta_0\| > \delta_n\} \leq \epsilon.$$

The discussion of subsequent properties can confine itself to conditions that need only hold in "shrinking neighbourhoods" of θ_0; i.e., neighbourhoods of θ_0 that can get arbitrarily small as n grows large, and still we know that our estimator will have that property with probability tending to one. Using the same proof as that of Theorem 1 (i), we have under our stronger assumption A5' that with probability tending to one

$$\begin{aligned}
\max_{1\leq j\leq\tau(n)} \|G_j(\widehat{\theta}_j)\| &\leq \max_{1\leq j\leq\tau(n)} \|G_j(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)\| + \max_{1\leq j\leq\tau(n)} \|G_{nj}(\widehat{\theta}_j)\| \\
&\leq \max_{1\leq j\leq\tau(n)} \sup_{\|\theta-\theta_0\|\leq\delta_n} \|G_j(\theta) - G_{nj}(\theta)\| + \max_{1\leq j\leq\tau(n)} \|G_{nj}(\widehat{\theta}_j)\| \\
&= o_p(n^{-1/4}) + \max_{1\leq j\leq\tau(n)} \|G_{nj}(\widehat{\theta}_j)\|, \text{ (by A5')} \\
&= o_p(n^{-1/4}), \text{ (by A4', A5' and A2)}.
\end{aligned}$$

Therefore, by A3'

$$\begin{aligned}
&\Pr\left(\max_{1\leq j\leq\tau(n)} ||\widehat{\theta}_j - \theta_0|| > \delta_n\right) \\
\leq\ &\Pr\left(\max_{1\leq j\leq\tau(n)} \|G_j(\widehat{\theta}_j)\| \geq \delta_n c_n\right) \\
\to\ &0 \quad \text{if} \quad \delta_n c_n = O(n^{-1/4})
\end{aligned}$$

Hence

$$\max_{1\leq j\leq\tau(n)} ||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4}),$$

which implies that

$$||\widehat{\theta} - \theta_0|| \leq \sum_{j=1}^{\tau(n)} \|W_{nj}\| \times \max_{1\leq j\leq\tau(n)} ||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4})$$

as required, where $\sum_{j=1}^{\tau(n)} \|W_{nj}\|$ is uniformly bounded by A1.

■

Proof of Theorem 2. Let

$$L_{nj}(\theta) = G_{nj}(\theta_0) + \Gamma_j(\theta - \theta_0)$$

for each $j = 1, 2, \ldots$ Then define θ_j^* as the minimizer of $\|L_{nj}(\theta)\|$ over $\theta \in \mathbb{R}^p$ [Note that θ_j^* minimizes over $\mathbb{R}^p$, and not over Θ. We ignore this difference below because θ_j^*will eventually be in Θ with probability going to one]. The solution satisfies

$$\sqrt{n}(\theta_j^* - \theta_0) = -\Gamma_j^{-1}\sqrt{n}G_{nj}(\theta_0) \tag{21}$$

for each j. Therefore,

$$\begin{aligned}
\sqrt{n}\sum_{j=1}^{\tau(n)} W_{nj}(\theta_j^* - \theta_0) &= \sqrt{n}\sum_{j=1}^{\tau(n)} W_{nj}^0(\theta_j^* - \theta_0) + \sqrt{n}\sum_{j=1}^{\tau(n)}(W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0) \\
&= \sum_{i=1}^{n} T_{in} + R_n,
\end{aligned}$$

where $R_n = \sqrt{n}\sum_{j=1}^{\tau(n)}(W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0)$ and $T_{in} = \frac{-1}{\sqrt{n}}\sum_{j=1}^{\tau(n)} W_{nj}^0\Gamma_j^{-1}g_j(Z_i, \theta_0)$.
The result follows after we establish (i), (ii) and (iii):

(i) $\sum_{i=1}^{n} c'T_{in} \Longrightarrow \mathcal{N}(0, c'\Sigma c)$ for any c with $||c|| = 1$;

(ii) the remainder term $R_n = o_p(1)$;

(iii) $\sqrt{n}\sum_{j=1}^{\tau(n)} W_{nj}(\theta_j^* - \widehat{\theta}_j) = o_p(1)$.

For (i), the triangular array of random variables $c'T_{in}$ is mean zero and independent across i for each n. By B5(a) we have:

$$\begin{aligned}
\sum_{i=1}^{n} E[c'T_{in}]^2 &= E\left[\left(\sum_{j=1}^{\tau(n)} c'W_{nj}^0\Gamma_j^{-1}g_j(Z_i, \theta_0)\right)^2\right] \\
&= \sum_{j=1}^{\tau(n)}\sum_{l=1}^{\tau(n)} c'W_{nj}^0\Gamma_j^{-1}E\left[g_j(Z_i, \theta_0)g_l(Z_i, \theta_0)'\right]\Gamma_l^{-1\prime}W_{nl}^{0\prime}c \\
&\to c'\Sigma c\ .
\end{aligned}$$

And by B5(b) we have for some $\kappa > 0$,

$$\sum_{i=1}^{n} E|c'T_{in}|^{2+\kappa} \to 0.$$

Hence we obtain (i) by applying the Liapounov's triangular array central limit theorem.

For (ii), notice that Assumption B3(a) and (21) imply that $\max_{1\le j\le \tau(n)} \|\Gamma_j \sqrt{n}(\theta_j^* - \theta_0)\| = O_p(1)$. This together with Assumption B4(a) imply (ii) because:

$$\begin{aligned}\left\|\sqrt{n}\sum_{j=1}^{\tau(n)}(W_{nj}-W_{nj}^0)(\theta_j^*-\theta_0)\right\| &\le \sqrt{n}\max_{1\le j\le\tau(n)}\left\|\Gamma_j(\theta_j^*-\theta_0)\right\|\sum_{j=1}^{\tau(n)}\left\|(W_{nj}-W_{nj}^0)\Gamma_j^{-1}\right\| \\ &= O_p(1)\times o_p(1).\end{aligned}$$

For (iii), by the $n^{1/4}$-consistency result, there exists a positive sequence $\eta_n \to 0$ such that $\Pr[n^{1/4}||\widehat{\theta}-\theta_0|| > \eta_n] \to 0.$ For each j we have

$$\begin{aligned} G_{nj}(\theta) &= G_{nj}(\theta_0) + G_j(\theta) + G_{nj}(\theta) - G_j(\theta) - G_{nj}(\theta_0) \\ \text{(by B2)} &= L_{nj}(\theta) + O(||\theta-\theta_0||^2) + [G_{nj}(\theta) - G_j(\theta)] - G_{nj}(\theta_0). \end{aligned}$$

Therefore, for the above η_n we have

$$\begin{aligned} &\max_{1\le j\le\tau(n)} \sup_{||\theta-\theta_0||\le a\eta_n/n^{1/4}} \sqrt{n}\|G_{nj}(\theta)-L_{nj}(\theta)\| \\ &\le C\cdot\eta_n^2 a^2 + \max_{1\le j\le\tau(n)} \sup_{||\theta-\theta_0||\le a\eta_n/n^{1/4}} \sqrt{n}\|[G_{nj}(\theta)-G_j(\theta)]-G_{nj}(\theta_0)\| \\ \text{(by B3(b))} &= O_p(\eta_n^2) + o_p(1) = o_p(1). \end{aligned}$$

Therefore,

$$\max_{1\le j\le\tau(n)} \|\sqrt{n}[L_{nj}(\theta_j^*) - G_{nj}(\theta_j^*)]\| = o_p(1) \text{ and } \max_{1\le j\le\tau(n)} \|\sqrt{n}[L_{nj}(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)]\| = o_p(1)$$

because θ_j^* is $\sqrt{n}$-consistent and $\widehat{\theta}_j$ is $o(n^{-1/4})$-consistent. It now follows from the definition of θ_j^* and Assumption B1 and the triangular inequality that

$$\max_{1\le j\le\tau(n)} \left|\sqrt{n}\|L_{nj}(\theta_j^*)\| - \sqrt{n}\|L_{nj}(\widehat{\theta}_j)\|\right| = o_p(1). \tag{22}$$

This implies that $\max_{1\le j\le\tau(n)} \|\Gamma_j\sqrt{n}(\theta_j^* - \widehat{\theta}_j)\| = o_p(1)$, because of the properties of least squares residuals. Then we have

$$\begin{aligned} &\sqrt{n}\sum_{j=1}^{\tau(n)} W_{nj}(\theta_j^* - \widehat{\theta}_j) \\ &\le \sum_{j=1}^{\tau(n)} \left\|W_{nj}\Gamma_j^{-1}\right\| \cdot \max_{1\le j\le\tau(n)} \|\Gamma_j\sqrt{n}(\theta_j^* - \widehat{\theta}_j)\| \\ &\le O_p(1)\times o_p(1) = o_p(1), \end{aligned}$$

where the last inequality is due to Assumption B4(a)(b) since

$$\begin{aligned}\sum_{j=1}^{\tau(n)} \left\|W_{nj}\Gamma_j^{-1}\right\| &\leq \sum_{j=1}^{\tau(n)} \left\|W_{nj}^0\Gamma_j^{-1}\right\| + \sum_{j=1}^{\tau(n)} \left\|(W_{nj} - W_{nj}^0)\Gamma_j^{-1}\right\| \\ &= O(1) + o_p(1) = O_p(1),\end{aligned}$$

the result (iii) follows.

■

Proof of Proposition 1. On the one-hand, by the results in Hansen (1982), the optimal GMM (or IV) estimator is asymptotically efficient among all regular root-n asymptotic normal estimators for the moment restrictions (11), hence $\Sigma_{oiv}^{\tau} \leq \Sigma_{omd}^{\tau}$ in the positive semi-definite matrix sense. On the other hand, we notice that the optimal IV (optimal GMM) estimator has the expansion

$$\sqrt{n}(\widetilde{\theta}_{oiv}^{\tau} - \theta_0) = -(\Gamma^{\tau\prime}\Psi_{\tau}^{-1}\Gamma^{\tau})^{-1}\Gamma^{\tau\prime}\Psi_{\tau}^{-1}\sqrt{n}G_n^{\tau}(\theta_0) + o_p(1),$$

which can be rewritten as

$$\sqrt{n}(\widetilde{\theta}_{oiv}^{\tau} - \theta_0) = -\left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j'\right)^{-1} \sum_{j=1}^{\tau} \alpha_j \sqrt{n} G_{nj}(\theta_0) + o_p(1), \tag{23}$$

where $\Gamma^{\tau\prime}\Psi_{\tau}^{-1} = (\alpha_1, \ldots, \alpha_\tau)$ with $\alpha_j \in \mathbb{R}^{p\times p}$, and $\Gamma^{\tau} = (\Gamma_1', \ldots, \Gamma_\tau')'$ with $\Gamma_j = E[A_j(X)D_0(X)']$, and $G_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A_j(X_i)\rho(Z_i, \theta)$ for $j = 1, \ldots, \tau$. That is, the optimal GMM (or IV) estimator $\widetilde{\theta}_{oiv}^{\tau}$ belongs to the class of linear combinations of the $\widehat{\theta}_j, j = 1, \ldots, \tau$ with

$$\widetilde{\theta}_{oiv}^{\tau} = \sum_{j=1}^{\tau} W_{0j}^{oiv}\widehat{\theta}_j + o_p(n^{-1/2})$$

and

$$W_{0j}^{oiv} = -\left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j'\right)^{-1} \alpha_j \Gamma_j' \text{ for } j = 1, \ldots, \tau.$$

However, by the results in Rothenberg (1973), $\widehat{\theta}_{omd}^{\tau} = \sum_{j=1}^{\tau} W_{0j}^{opt}\widehat{\theta}_j$ is asymptotically efficient among the regular class of estimators of the form $\sum_{j=1}^{\tau} W_{0j}\widehat{\theta}_j$ with $\sum_{j=1}^{\tau} W_{0j} = I_p$, hence $\Sigma_{omd}^{\tau} \leq \Sigma_{oiv}^{\tau}$ in the positive semi-definite matrix sense. Therefore $\Sigma_{omd}^{\tau} = \Sigma_{oiv}^{\tau}$ in (14).

■

Proof of Theorem 3. Assumption C3 implies that $\beta_{j0} = E[D_0(X_i)\phi_j(X_i)']$. We have:

$$\begin{aligned}\Sigma_{oiv} &= \left(E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)']\right)^{-1} = \left(E\left[\sum_{j=1}^{\infty} \beta_{j0}\phi_j(X_i)\sigma_0^2(X_i)\sigma_0^{-2}(X_i)D_0(X_i)'\right]\right)^{-1} \\ &= \left(\sum_{j=1}^{\infty} \beta_{j0} E\left[\phi_j(X_i)D_0(X_i)'\right]\right)^{-1} = \left(\sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}'\right)^{-1}\end{aligned}$$

$$= \left(\sum_{j=1}^{\infty} E[D_0(X_i)'\phi_j(X_i)']E\left[\phi_j(X_i)D_0(X_i)\right] \right)^{-1}.$$

Assumptions C2 and C3 imply that $0 < \sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}' < \infty$.

By Assumptions C1-C3, we have

$$\begin{aligned} V_{jl} &= 0 \text{ for all } j \neq l \\ V_{jj} &= \{\Gamma_j'\Gamma_j\}^{-1} = \{E[\phi_j(X_i)D_0(X_i)]'E[\phi_j(X_i)D_0(X_i)]\}^{-1} \end{aligned}$$

$$\begin{aligned} \Sigma_{omd} &= \lim_{\tau\to\infty} \left((I_p \otimes i_\tau)'V^{-1}(I_p \otimes i_\tau)\right)^{-1} \\ &= \lim_{\tau\to\infty} \left(\sum_{j=1}^{\tau} V_{jj}^{-1} \right)^{-1} \\ &= \lim_{\tau\to\infty} \left(\sum_{j=1}^{\tau} \{E[\phi_j(X_i)D_0(X_i)]'E[\phi_j(X_i)D_0(X_i)]\} \right)^{-1} \\ &= \left(\sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}' \right)^{-1}. \end{aligned}$$

■

References

Ai, C. and X. Chen (1999): "Efficient estimation of models with conditional moment restrictions containing unknown functions", working paper, London School of Economics.

Amemiya, T., (1974): "The non-linear two-stage least squares estimator," *Journal of Econometrics* 2, 105-110.

Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica 59, 307-346.

Bickel, P.J., and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. Annals of Statistics 1, 1071-1095.

Bierens, H.J. (1987). Kernel estimators of regression functions. Cambridge University Press: Advances in Econometrics.

Breiman, L. (1996). Bagging Predictors. Machine Learning 26, 123-140.

Breiman, L. (1996). Using Adaptive Bagging to Debias Regressions. University of California Berkeley, Department of Statistics, Technical Report 547.

Campbell, J., A. Lo and A.C. Mackinlay (1997): *The Econometrics of Financial Markets*, Princeton University Press: Princeton.

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

Granger, C.W.J. (2000): "Thick modeling", UCSD Dept. of Economics Working Paper.

Gray, H.L. and W.R. Schucany (1972). The generalized jacknife statistic. New York: Dekker.

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

Hansen, L.P. (1985): "A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators," *Journal of Econometrics* 30, 203-238.

Huber, P. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Fifth Berkeley Sysposium on Mathematical Statistics and Probability,* 221-233. Berkeley, CA: University of California.

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and J. Machado (1999): "GMM inference when the number of moment conditions is large," *Journal of Econometrics* 93, 327-344.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica* 57, 995-1026.

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

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

Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators", *Econometrica* 62, 1349-1382.

Newey, W.K., (1999): "Efficient Semiparametric Estimation Via moment restrictions", MIT.

Newey, W.K. and D.F. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics, vol. IV*, eds. D.F. McFadden and R.F. Engle III. North Holland.

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

Newey, W.K. and R.J. Smith (2001). Higher order properties of GMM and Generalized Empirical Likelihood Estimators. Unpublished Manuscript available at http://emlab.berkeley.edu/symposia/nsf01/ag

Olley S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica* 64, 1263-1297.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica* 57, 1027-1057.

Portnoy, S. (1985): "Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: II normal approximation". *Ann. Statist.* 17, 382-400.

Robinson, P. M. (1991): "Best nonlinear three-stage least squares estimation of certain econometric models," *Econometrica*, 59, 755-786.

Rothenberg, T.J., (1973). *Efficient estimation with a priori information.* Cowles monograph.

Rothenberg, T.J. (1974). A note on Two-Stage Least Squares. Unpublished manuscript, UC Berkeley.

Sawa, T. (1973): "Almost unbiased estimator in simultaneous equations systems," *International Economic Review* 14, 97-106.

Stock, J.W. and M.W. Watson (1999): "Forecasting inflation", *Journal of Monetary Economics,* 44, 293-335.

Watson, M.W. (2000). Macroeconomic Forecasting Using Many Predictors. Princeton University.

Table 1

Method	β_1					β_2				
j	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$
2	0.99	0.21	0.99	0.75	1.24	1.01	0.23	1.01	0.69	1.31
3	0.66	0.40	1.01	0.29	1.71	1.38	0.69	0.98	0.07	1.91
4	-0.08	0.82	1.00	-0.33	2.38	2.43	1.45	0.99	-0.82	2.74
5	1.39	0.39	1.01	-0.23	2.29	0.49	0.67	0.97	-0.68	2.64
6	2.28	1.17	1.03	-0.27	2.37	-0.64	1.96	0.95	-0.81	2.68
7	0.43	0.46	0.99	-0.54	2.45	1.65	0.74	1.01	-0.92	3.04
8	1.71	2.14	1.02	-0.26	2.34	-0.05	3.63	0.97	-0.76	2.66
9	-0.40	2.61	1.00	-0.42	2.46	2.93	5.24	0.99	-0.92	2.88
10	-1.97	5.66	1.01	-0.36	2.44	4.90	9.25	0.96	-0.89	2.83
avg	0.56	0.18	0.77	0.68	1.28	1.62	0.33	1.31	0.73	1.44
$\alpha = 1$	1.12	0.19	1.05	0.98	1.14	0.80	0.21	0.88	0.79	0.96
$\alpha = 2$	1.05	0.22	1.02	0.97	1.12	0.88	0.23	0.91	0.83	0.95
$\alpha = 3$	0.98	0.18	0.94	0.88	1.03	0.97	0.26	0.99	0.79	1.10
α_{opt}	1.02	0.16	1.00	0.73	1.25	0.99	0.18	0.99	0.67	1.32

Notes: n=100

Table 2

Method	β_1					β_2				
j	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$
2	1.00	0.17	1.00	0.83	1.17	1.00	0.21	0.99	0.79	1.22
3	1.14	0.12	1.01	0.50	1.51	0.82	0.19	0.98	0.35	1.66
4	0.68	1.55	1.00	-0.30	2.29	1.48	2.65	1.00	-0.71	2.73
5	2.06	2.39	1.01	-0.29	2.33	-0.49	4.13	0.99	-0.73	2.69
6	1.67	0.39	0.98	-0.15	2.17	0.12	0.66	1.04	-0.53	2.49
7	-0.43	0.60	0.98	-0.28	2.26	2.87	1.01	1.02	-0.67	2.68
8	1.52	2.92	0.98	-0.39	2.37	-0.57	3.02	1.02	-0.85	2.84
9	-1.70	2.34	1.01	-0.35	2.41	0.80	3.57	0.98	-0.85	2.78
10	0.77	0.70	0.95	-0.43	2.34	1.26	1.19	1.07	-0.79	2.86
avg	0.75	0.25	1.01	0.68	1.52	0.81	0.39	0.90	0.68	0.97
$\alpha = 1$	1.10	0.18	1.16	1.10	1.23	0.84	0.28	0.78	0.72	0.85
$\alpha = 2$	1.16	0.12	1.20	1.07	1.26	0.79	0.18	0.74	0.68	0.86
$\alpha = 3$	1.23	0.16	1.28	1.13	1.30	0.72	0.14	0.68	0.66	0.79
α_{opt}	0.99	0.08	1.00	0.82	1.19	1.01	0.13	1.00	0.77	1.24

Notes: n=200

Table 3

Method	β_1					β_2				
j	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$	Mean	Var	$q_{0.5}$	$q_{0.25}$	$q_{0.75}$
2	0.99	0.13	0.99	0.89	1.09	1.01	0.14	1.00	0.87	1.22
3	0.99	0.12	1.02	0.71	1.29	1.02	0.12	0.97	0.61	1.66
4	3.54	0.75	0.94	-0.07	1.97	-2.28	1.26	1.09	-0.28	2.73
5	3.60	3.30	1.01	-0.32	2.33	-2.42	2.81	0.99	-0.74	2.69
6	1.39	0.21	1.00	0.23	1.80	0.47	0.18	0.99	-0.04	2.49
7	0.84	2.47	1.04	-0.25	2.32	1.22	1.26	0.95	-0.74	2.68
8	1.49	0.42	1.02	-0.11	2.23	0.36	0.70	0.96	-0.61	2.84
9	-2.68	2.32	0.99	-0.19	2.19	2.74	3.35	0.99	-0.57	2.78
10	-0.89	1.18	0.95	-0.35	2.29	1.17	2.11	1.06	-0.70	2.86
avg	1.42	0.30	1.63	0.97	1.76	0.44	0.51	0.17	0.01	1.02
$\alpha = 1$	1.02	0.17	1.03	0.95	1.07	0.97	0.13	0.99	0.91	1.02
$\alpha = 2$	1.03	0.08	1.05	0.98	1.05	0.96	0.07	0.93	0.89	1.02
$\alpha = 3$	1.03	0.12	1.04	0.98	1.07	0.95	0.13	0.93	0.88	1.01
α_{opt}	0.99	0.05	1.03	0.96	1.07	1.00	0.06	0.99	0.97	1.10

Notes: n=500