Econometrica, Vol. 78, No. 2 (March, 2010), 697–718

NOTES AND COMMENTS

CONSTRUCTING OPTIMAL INSTRUMENTS BY FIRST-STAGE PREDICTION AVERAGING

BY GUIDO KUERSTEINER AND RYO OKUI[1]

This paper considers model averaging as a way to construct optimal instruments for the two-stage least squares (2SLS), limited information maximum likelihood (LIML), and Fuller estimators in the presence of many instruments. We propose averaging across least squares predictions of the endogenous variables obtained from many different choices of instruments and then use the average predicted value of the endogenous variables in the estimation stage. The weights for averaging are chosen to minimize the asymptotic mean squared error of the model averaging version of the 2SLS, LIML, or Fuller estimator. This can be done by solving a standard quadratic programming problem.

KEYWORDS: Model averaging, instrumental variable, many instruments, two-stage least squares, LIML, Fuller's estimator, higher order theory.

1. INTRODUCTION

IN THIS PAPER, WE PROPOSE a new and flexible method of constructing optimal instruments for two stage least squares (2SLS), limited information maximum likelihood (LIML), and Fuller's estimators of linear models when there are many instruments available. Donald and Newey (2001) and Newey (2007) proposed a selection criterion to select instruments in a way that balances higher order bias and efficiency. We show that the model averaging approach of Hansen (2007) can be applied to the first stage of the 2SLS estimator as well as to a modification of LIML and Fuller's (1977) estimator. The benefits of model averaging mostly lie in a more favorable trade-off between estimator bias and efficiency relative to procedures that rely on a single set of instruments. In contrast to the existing literature on model averaging, the weights for averaging are not restricted to be positive. Our theoretical results show that for certain choices of weights, the model averaging 2SLS estimator (MA2SLS) controls higher order bias and achieves the same higher order rates of convergence as the Nagar (1959) and LIML estimators, and thus dominates conventional 2SLS procedures. Model averaging allowing for bias reduction requires a refined asymptotic approximation to the mean squared error (MSE) of the 2SLS estimator. We provide such an approximation by including terms

[1]We thank Whitney Newey, Xiaohong Chen, Bruce Hansen, Jack Porter, Kaddour Hadri, Ken Ariga, Yuichi Kitamura, Kosuke Oya, Han Hong, and seminar participants at Columbia, Georgetown, USC, Yale, CIREQ, Berkeley, Hitotsubashi, Seoul National (SETA08), Kyoto, and Osaka as well as two anonymous referees for valuable comments and suggestions. Okui acknowledges financial support from the Hong Kong University of Science and Technology under Project DAG05/06.BM16 and from the Research Grants Council of Hong Kong under Project HKUST643907. We are solely responsible for all errors.

 DOI: 10.3982/ECTA7444

of the next higher order than the leading bias term in our MSE approximation. This approach provides a criterion that directly captures the trade-off between higher order variance and bias correction.

Conventional instrument selection procedures often depend on an ad hoc ordering of the instruments. By allowing our model weights to be both positive and negative, we establish that the MA2SLS and corresponding LIML and Fuller estimators have the ability to select arbitrary subsets of instruments from an orthogonalized set of instruments. In other words, if there are certain orthogonal directions in the instrument space that are particularly useful for the first stage, our procedure is able to individually select these directions from the instrument set. The selection of relevant directions is done in a fully data-dependent and computationally efficient way without requiring any prior knowledge about the strength of the instruments.

2. DEFINITIONS AND IMPLEMENTATION

Following Donald and Newey (2001), we consider the model

$$y_i = Y_i'\beta_y + x_{1i}'\beta_x + \epsilon_i = X_i'\beta + \epsilon_i, \tag{2.1}$$
$$X_i = \begin{pmatrix} Y_i \\ x_{1i} \end{pmatrix} = f(z_i) + u_i = \begin{pmatrix} E[Y_i|z_i] \\ x_{1i} \end{pmatrix} + \begin{pmatrix} \eta_i \\ 0 \end{pmatrix}, \quad i = 1, \ldots, N,$$

where y_i is a scalar outcome variable, Y_i is a $d_1 \times 1$ vector of endogenous variables, x_{1i} is a vector of included exogenous variables, z_i is a vector of exogenous variables (including x_{1i}), ϵ_i and u_i are unobserved random variables with second moments which do not depend on z_i, and $f(\cdot)$ is an unknown function of z. Let $f_i = f(z_i)$. The set of instruments has the form $Z_{M,i} \equiv (\psi_1(z_i), \ldots, \psi_M(z_i))'$, where ψ_ks are functions of z_i such that $Z_{M,i}$ is an $M \times 1$ vector of instruments. Let $y = (y_1, \ldots, y_N)'$ and define X, ϵ, f, and u similarly.

First-stage prediction averaging is based on a weighting vector W, where $W = (w_1, \ldots, w_M)'$ and $\sum_{m=1}^{M} w_m = 1$ for some M such that $M \leq N$ for any N. We note that W is implicitly indexed by the sample size N. In Section 4, we discuss in more detail the restrictions that need to be imposed on W and M, but we point out here that w_m is allowed to take on positive and negative values. Let $Z_{m,i}$ be the vector of the first m elements of $Z_{M,i}$, let Z_m be the matrix $(Z_{m,1}, \ldots, Z_{m,N})'$ and let $P_m = Z_m(Z_m'Z_m)^{-1}Z_m'$. Define $P(W) = \sum_{i=1}^{M} w_m P_m$. The MA2SLS estimator, $\hat{\beta}$, of β is defined as

$$\hat{\beta} = (X'P(W)X)^{-1}X'P(W)y. \tag{2.2}$$

The definition of (2.2) can be extended to the LIML estimator. Let

$$\hat{\Lambda}_m = \min_{\beta} \frac{(y - X\beta)'P_m(y - X\beta)}{(y - X\beta)'(y - X\beta)}$$

and define $\hat{\Lambda}(W) = \sum_{m=1}^{M} w_m \hat{\Lambda}_m$. The model averaging LIML (MALIML) estimator, $\hat{\beta}$, of β then is defined as

$$\hat{\beta} = (X'P(W)X - \hat{\Lambda}(W)X'X)^{-1}(X'P(W)y - \hat{\Lambda}(W)X'y). \tag{2.3}$$

Similarly we consider a modification to Fuller's (1977) estimator. Let

$$\check{\Lambda}_m = \left(\frac{\hat{\Lambda}_m - \dfrac{\alpha}{N-m}(1-\hat{\Lambda}_m)}{1 - \dfrac{\alpha}{N-m}(1-\hat{\Lambda}_m)} \right),$$

where α is a constant chosen by the econometrician.[2] Define $\check{\Lambda}(W) = \sum_{m=1}^{M} w_m \check{\Lambda}_m$. The model averaging Fuller estimator (MAFuller) now is defined as

$$\hat{\beta} = (X'P(W)X - \check{\Lambda}(W)X'X)^{-1}(X'P(W)y - \check{\Lambda}(W)X'y). \tag{2.4}$$

We choose W to minimize the approximate MSE, $S_\lambda(W)$, defined in (4.1), of $\lambda'\hat{\beta}$ for some fixed $\lambda \in \mathbb{R}^d$. The optimal weight, denoted W^*, is the solution to $\min_{W\in\Omega} S_\lambda(W)$, where Ω is some set. We consider several versions of Ω which lead to different estimators. The MA2SLS, MALIML, and MAFuller estimators are unconstrained if $\Omega = \Omega_U = \{W \in l_1 | W'\mathbf{1}_M = 1\}$, where l_1 is a space of absolutely summable sequences defined in Assumption 4. From a finite sample point of view, it may be useful to further constrain the weights W to lie in a compact set. This is achieved in $\Omega = \Omega_C = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [-1, 1], \forall m \le M\}$. If we allow only positive weights, then $\Omega = \Omega_P = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [0, 1], \forall m \le M\}$.

We now discuss how to estimate $S_\lambda(W)$. Let $\tilde{\beta}$ denote some preliminary estimator of β, and define the residuals $\tilde{\epsilon} = y - X\tilde{\beta}$. As pointed out in Donald and Newey (2001), it is important that $\tilde{\beta}$ does not depend on the weighting vector W. We use the 2SLS estimator with the number of instruments selected by the first-stage Mallows criterion in simulations for MA2SLS, and use the corresponding LIML and Fuller estimator for MALIML and MAFuller. Let $\hat{H}$ be some estimator of $H = f'f/n$. Let $\tilde{u}$ be some preliminary residual vector of the first-stage regression. Let $\tilde{u}_\lambda = \tilde{u}\hat{H}^{-1}\lambda$. Define

$$\hat{\sigma}_\epsilon^2 = \tilde{\epsilon}'\tilde{\epsilon}/N, \quad \hat{\sigma}_\lambda^2 = \tilde{u}_\lambda'\tilde{u}_\lambda/N, \quad \hat{\sigma}_{\lambda\epsilon} = \tilde{u}_\lambda'\tilde{\epsilon}/N.$$

Let $\hat{u}_\lambda^m = (P_M - P_m)X\hat{H}^{-1}\lambda$ and $\hat{U} = (\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)'(\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)$.[3] Let Γ be the $M \times M$ matrix whose (i, j) element is $\min(i, j)$ and let $K = (1, 2, \ldots, M)'$. The

[2]Popular choices are $\alpha = 1$ or $\alpha = 4$. See, for example, Hahn, Hausman, and Kuersteiner (2004).

[3]Note that $\tilde{u}$ is the preliminary residual vector, but $\hat{u}_\lambda^m$s are the vectors of the differences of the residuals.

criterion $\hat{S}_\lambda(W)$ for choosing the weights for the MA2SLS defined in (2.2) is

$$\hat{S}_\lambda(W) = \hat{a}_\lambda \frac{(K'W)^2}{N} + \hat{b}_\lambda \frac{(W'\Gamma W)}{N} - \frac{K'W}{N}\hat{B}_{\lambda,N} + \hat{\sigma}_\epsilon^2\left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2(M - 2K'W + W'\Gamma W)}{N}\right) \tag{2.5}$$

with $\hat{a}_\lambda = \hat{\sigma}_{\lambda\epsilon}^2$, $\hat{b}_\lambda = \hat{\sigma}_\epsilon^2\hat{\sigma}_\lambda^2 + \hat{\sigma}_{\lambda\epsilon}^2$, and $\hat{B}_{\lambda,N} = \lambda'\hat{H}^{-1}\hat{B}_N\hat{H}^{-1}\lambda$, where $\hat{B}_N$ is some estimator[4] of B_N defined in (4.2). When the weights are only allowed to be positive, we may use the simpler criterion,

$$\hat{S}_\lambda(W) = \hat{a}_\lambda \frac{(K'W)^2}{N} + \hat{\sigma}_\epsilon^2\left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2(M - 2K'W + W'\Gamma W)}{N}\right) \tag{2.6}$$

that does not account for the terms of smaller order involving $\hat{b}_\lambda$ and $\hat{B}_{\lambda,N}$. For the MALIML and MAFuller estimators defined in (2.3) and (2.4), we choose W based on the criterion

$$\hat{S}_\lambda(W) = (\hat{\sigma}_\epsilon^2\hat{\sigma}_\lambda^2 - \hat{\sigma}_{\lambda\epsilon}^2)\frac{W'\Gamma W}{N} + \hat{\sigma}_\epsilon^2\left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2(M - 2K'W + W'\Gamma W)}{N}\right). \tag{2.7}$$

Theorem 4.4 establishes the sense in which $\hat{S}_\lambda(W)$ defined in (2.5)–(2.7) is a valid estimator of the corresponding population criterion $S_\lambda(W)$.

From a practical point of view, it is often important to report the quality of fit for the first-stage regression in 2SLS. The model averaging predictor of X is $\hat{X} = P(W)X$. When $\dim(\beta) = 1$ and X_i has mean zero (or is demeaned), a measure of the fit of the first stage is the pseudo R^2 defined as $\operatorname{corr}(X, \hat{X})^2$ and estimated by

$$\tilde{R}^2 = \widehat{\operatorname{corr}}(X, \hat{X})^2 = \frac{(X'P(W)X)^2}{X'P(W)P(W)X \cdot X'X}.$$

We have $\tilde{R}^2 \in [0, 1]$ by construction and Theorem A.5 in the Supplemental Material (Kuersteiner and Okui (2010)) shows that $\tilde{R}^2 \to_p E(f_i^2)/(E(f_i^2) + \sigma_u^2)$, which is the population R^2.

[4]When $\dim(\beta) = 1$, we have $B_N = 2(\sigma_\epsilon^2\Sigma_u + 4\sigma_{u\epsilon}^2)$, where $\Sigma_u = E[u_i u_i']$, $\sigma_{u\epsilon} = E[u_i\epsilon_i]$, and we may use $\hat{B}_{\lambda,N} = 2(\hat{\sigma}_\epsilon^2\sigma_\lambda^2 + 4\hat{\sigma}_{\lambda\epsilon}^2)$. The Supplemental Material (Kuersteiner and Okui (2010)) to this paper contains an expression for $\hat{B}_N$ for the general case.

3. MOTIVATION AND DISCUSSION OF THEORETICAL RESULTS

We refer to instrument selection procedures that take the ordering of instruments in Z_M as given and only choose the number of instruments as sequential instrument selection procedures. Sequential instrument selection is a special case of model averaging with weights chosen from the set $\Omega_{sq} \equiv \{W \in l_1 | w_m = 1$ for some m and $w_j = 0$ for $j \neq m\}$ to minimize $S_\lambda(W)$. Note that when $W \in \Omega_{sq}$, it follows that $K'W = m$ and $(I - P(W))(I - P(W)) = (I - P_m)$. Hence, $S(W)$ for 2SLS with W restricted to $W \in \Omega_{sq}$ simplifies to

$$H^{-1}\left(a_\sigma \frac{m^2}{N} + \frac{m}{N}(b_\sigma - B_N) + \sigma_\epsilon^2 \frac{f'(I - P_m)f}{N}\right)H^{-1} \tag{3.1}$$

for $m \leq M$. Because $m/N = o(m^2/N)$ as $m \to \infty$, the expression for $S(W)$ with $W \in \Omega_{sq}$ reduces to the result of Donald and Newey (2001, Proposition 1). We note that all sets $\Omega = \Omega_U, \ldots, \Omega_P$ contain the set Ω_{sq} as a subset (i.e., $\Omega_{sq} \subset \Omega$). This guarantees that MA2SLS, MALIML, and MAFuller weakly dominate the sequential instrument selection procedure in the sense that $S_\lambda(W^*) \leq \min_{W \in \Omega_{sq}} S_\lambda(W)$. Theorem 4.3 in Section 4 establishes that the inequality is strict for MA2SLS, MALIML, and MAFuller when $W \in \Omega_P$. The theorem holds under conditions where $f'(I - P_m)f$ is monotonically decaying and convex in m, and thus applies to situations where the instruments have already been optimally ordered. Since $\Omega_P \subset \Omega_C \subset \Omega_U$, Theorem 4.3 establishes strict dominance for all model averaging estimators considered in this paper.

To better understand these results, it is useful to consider the bias variance trade-off of 2SLS in more detail. It can be shown that the largest term of the higher order bias of MA2SLS, $\hat{\beta}$ in (2.2), is proportional to $K'W/\sqrt{N}$. When a specific first-stage model with exactly m instruments is selected, this result reduces to the well known result that the higher order bias is proportional to $m/\sqrt{N}$. To illustrate the bias reduction properties of MA2SLS, we consider an extreme case where the higher order bias is completely eliminated. This occurs when W satisfies the additional constraint $K'W = 0$. Thus, the higher order rate of convergence of MA2SLS can be improved relative to the rate for 2SLS by allowing w_j to be both positive and negative. In fact, the Nagar estimator can be interpreted as a special case of the MA2SLS, where $M = N$, $w_m = N/(N - m)$ for some m, $w_N = -m/(N - m)$, and $w_j = 0$ otherwise.

An instrument selection method related to ours is from Kuersteiner (2002), who proposed a kernel weighted form of the 2SLS estimator in the context of time series models and showed that kernel weighting reduces the bias of 2SLS. Let $k = \text{diag}(k_1, \ldots, k_M)$, where $k_j = k((j - 1)/M)$ are kernel functions $k(\cdot)$ evaluated at j/M with $k(0) = 1$. The kernel weighted 2SLS estimator then is defined as in (2.2) with $P(W)$ replaced by $Z_M k(Z_M' Z_M)^{-1} k Z_M'$. For expositional purposes and to relate kernel weighting to model averaging, we consider a special case in which instruments are mutually orthogonal so

that $Z_M'Z_M$ is a diagonal matrix. Let $\tilde{Z}_j$ be the jth column of Z_M such that $Z_M = (\tilde{Z}_1, \ldots, \tilde{Z}_M)$ and $\tilde{P}_j = \tilde{Z}_j(\tilde{Z}_j'\tilde{Z}_j)^{-1}\tilde{Z}_j'$. For a given set of kernel weights k, there exist weights W such that for $w_j = k_j^2 - k_{j+1}^2$ and $w_M = k_M^2$, the relationship $\sum_{m=1}^M w_m P_m = \sum_{j=1}^M k_j^2 \tilde{P}_j = Z_M k(Z_M'Z_M)^{-1} k Z_M'$ holds. In other words, the kernel weighted 2SLS estimator corresponds to model averaging with the weights $\{w_m\}_{m=1}^M$ defined above.

Okui's (2008) shrinkage 2SLS estimator is also a special case of the averaged estimator (2.2). In this case, $w_L = s$, $w_M = 1 - s$, $s \in [0, 1]$, and $w_j = 0$ for $j \neq L, M$ where L ($<M$) is fixed. Okui's procedure can be interpreted in terms of kernel weighted 2SLS. Letting the kernel function $k(x) = 1$ for $x \leq L/M$, $k(x) = \sqrt{s}$ for $L/M < x \leq 1$, and $k(x) = 0$ otherwise implies that the kernel weighted 2SLS estimator formulated on the orthogonalized instruments is equivalent to Okui's procedure.

The common feature of kernel weighted 2SLS estimators is that they shrink the first stage coefficient estimators toward zero. Shrinkage in the first-stage reduces bias in the second stage at the cost of reduced efficiency. While kernel weighting has been shown to reduce bias, conventional kernels with monotonically decaying "tails" cannot completely eliminate bias. Kuersteiner (2002) showed that the distortions of the optimal 2SLS weight matrix introduced by kernel weights asymptotically dominate the higher order variance of $\hat{\beta}$ for conventional choices of $k(\cdot)$. This later problem was recently addressed by Canay (2010) through the use of top-flat kernels (e.g., Politis and Romano (1995)).

Despite these advances, conventional kernel based methods have significant limitations due to the fact that once a kernel function is chosen, the weighting scheme is not flexible. The more flexible weighting schemes employed by MA2SLS guarantee that the net effect of bias reduction at the cost of decreased efficiency always results in a net reduction of the approximate MSE of the second-stage estimator. Similar improvements are possible for LIML and Fuller's estimator, where two higher order variance terms can be balanced efficiently against each other through flexible instrument weighting.

A second advantage of model averaging is its ability to eliminate irrelevant instruments with an appropriate choice of W. Let $Z_{j,M}$ be the jth column of Z_M when $j \leq M$ and define $\tilde{Z}_2 = (I - P_1)Z_{2,M}$, $\tilde{Z}_3 = (I - P_2)Z_{3,M}, \ldots, \tilde{Z}_M = (I - P_{M-1})Z_M$ such that $Z_1, \tilde{Z}_2, \ldots, \tilde{Z}_M$ are orthogonal and span Z_M. Then, $P_M = \sum_{j=1}^M \tilde{P}_j$, where $\tilde{P}_j = \tilde{Z}_j(\tilde{Z}_j'\tilde{Z}_j)^{-1}\tilde{Z}_j'$ for $j > 1$ and $\tilde{P}_1 = Z_1(Z_1'Z_1)^{-1}Z_1'$. It follows that $\sum_{m=1}^M w_m P_m = \sum_{j=1}^M \tilde{w}_j \tilde{P}_j$ for $\tilde{w}_j = \sum_{m=j}^M w_m$. Then $\tilde{w}_j$ is the weight on the mth orthogonalized instrument. We may write $W = D^{-1}\tilde{W}$, where D is an $M \times M$ matrix with elements $d_{ij} = \mathbf{1}\{j \geq i\}$ and $\tilde{W} = (\tilde{w}_1, \ldots, \tilde{w}_M)'$. The only constraint we impose on $\tilde{W}$ is $\tilde{w}_1 = 1$. Since $\tilde{W}$ is otherwise unconstrained, we can set $\tilde{w}_j = 0$ for any $1 < j \leq M$ so that the jth instrument is eliminated. The

use of negative weights thus allows MA2SLS, MALIML, and MAFuller to pick out relevant instruments from a set of instruments that contains redundant instruments.

4. REGULARITY CONDITIONS AND FORMAL RESULTS

This section covers the formal theory underlying our estimators. All proofs are available in the Supplemental Material (Kuersteiner and Okui (2010)). The choice of model weights W is based on an approximation to the higher order MSE of $\hat{\beta}$. Following Donald and Newey (2001), we approximate the MSE conditional on the exogenous variable z, $E[(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)'|z]$, by $\sigma_\epsilon^2 H^{-1} + S(W)$, where

$$N(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)' = \hat{Q}(W) + \hat{r}(W), \tag{4.1}$$
$$E[\hat{Q}(W)|z] = \sigma_\epsilon^2 H^{-1} + S(W) + T(W),$$

$H = f'f/N$, and $(\hat{r}(W)+T(W))/\operatorname{tr}(S(W)) = o_p(1)$ as $N\to\infty$. However, because of the possibility of bias elimination by setting $K'W=0$, we need to consider an expansion that contains additional higher order terms for the MA2SLS case. We show the asymptotic properties of MA2SLS, MALIML, and MAFuller under the following assumptions.

ASSUMPTION 1: *$\{y_i, X_i, z_i\}$ are independent and identically distributed (i.i.d.), $E[\epsilon_i^2|z_i] = \sigma_\epsilon^2 > 0$, and $E[\|\eta_i\|^4|z_i]$ and $E[|\epsilon_i|^4|z_i]$ are bounded.*

ASSUMPTION 2: (i) *$\bar{H} \equiv E[f_i f_i']$ exists and is nonsingular.* (ii) *For some $\alpha > 1/2$,*

$$\sup_{m\le M} m^{2\alpha}\Big(\sup_{\lambda'\lambda=1} \lambda' f'(I-P_m)f\lambda/N\Big) = O_p(1).$$

(iii) *Let N_+ be the set of positive integers. There exists a subset $\bar{J}\subset N_+$ with a finite number of elements such that $\sup_{m\in\bar{J}}\sup_{\lambda'\lambda}\lambda' f'(P_m - P_{m+1})f\lambda/N = 0$ with probability approaching 1 (wpa1) and for all $m\notin\bar{J}$, it follows that*

$$\inf_{m\notin\bar{J}, m\le M} m^{2\alpha+1}\Big(\sup_{\lambda'\lambda=1}\lambda' f'(P_m - P_{m+1})f\lambda/N\Big) > 0 \quad \text{wpa1}.$$

ASSUMPTION 3: (i) *Let u_{ia} be the ath element of u_i. $E[\epsilon_i^r u_{ia}^s|z_i]$ are constant and bounded for all a, and $r, s \ge 0$ and $r+s\le 5$. Let $\sigma_{u\epsilon} = E[u_i\epsilon_i|z_i]$ and $\Sigma_u = E[u_i u_i'|z_i]$.* (ii) *$Z_M'Z_M$ are nonsingular wpa1.* (iii) *$\max_{i\le N} P_{M,ii} \to_p 0$, where $P_{M,ii}$ signifies the (i,i)th element of P_M.* (iv) *f_i is bounded.*

ASSUMPTION 4: *Let* $W^+ = (|w_{1,N}|, \ldots, |w_{M,N}|)'$. *The following conditions hold*: $1_M'W = 1$; $W \in l_1$ *for all* N, *where* $l_1 = \{x = (x_1, \ldots) | \sum_{i=1}^{\infty} |x_i| \leq C_{l1} < \infty\}$ *for some constant* C_{l1}, $M \leq N$; *and, as* $N \to \infty$ *and* $M \to \infty$, $K'W^+ = \sum_{m=1}^{M} |w_m| m \to \infty$. *For some sequence* $L \leq M$ *such that* $L \to \infty$ *as* $N \to \infty$ *and* $L \notin \bar{J}$, *where* $\bar{J}$ *is defined in Assumption* 2(iii), *it follows that* $\sup_{j \notin \bar{J}, j \leq L} |\sum_{m=1}^{j} w_m| = O(1/\sqrt{N})$ *as* $N \to \infty$.

ASSUMPTION 5: *It holds either that* (i) $K'W^+/\sqrt{N} \to 0$ *or* (ii) $K'W^+/N \to 0$ *and* $M/N \to 0$.

ASSUMPTION 6: *The eigenvalues of* $E[Z_{k,i}Z_{k,i}']$ *are bounded away from zero uniformly in* k. *Let* $\bar{H}_k = E[f_i Z_{k,i}](E[Z_{k,i}Z_{k,i}'])^{-1}E[f_i Z_{k,i}']'$ *and* $\bar{H} = E[f_i f_i']$. *Then* $\|\bar{H}_k - \bar{H}\| = O(k^{-2\alpha})$ *for* $k \to \infty$. $E[|\epsilon_i|^8|z]$ *and* $E[|u_{ia}|^8|z]$ *are uniformly bounded in* z *for all* a.

ASSUMPTION 7: $\beta \in \Theta$, *where* Θ *is a compact subset of* R^d.

REMARK 1: The second part of Assumption 2 allows for redundant instruments where $f'(P_m - P_{m+1})f/N = 0$ for some m, as long as the number of such cases is small relative to M.

Assumptions 1–3 are similar to those imposed in Donald and Newey (2001). The set $\bar{J}$ corresponds to the set of redundant instruments. Assumption 4 collects the conditions that weights must satisfy and is related to the conditions imposed by Donald and Newey (2001) on the number of instruments. The condition $K'W^+ \to \infty$ may be understood as the number of instruments tending to infinity. The condition, $\sup_{j \notin \bar{J}, j \leq L} |\sum_{s=1}^{j} w_s| = O(1/\sqrt{N})$, guarantees that small models receive asymptotically negligible weight and is needed to guarantee first-order asymptotic efficiency of the MA2SLS, MALIML, and MAFuller estimators. We also restrict W to lie in the space of absolutely summable sequences l_1. The fact that the sequences in l_1 have infinitely many elements creates no problems since one can always extend W to l_1 by setting $w_j = 0$ for all $j > M$. Assumption 5 limits the rate at which the number of instruments is allowed to increase and Assumption 5(i) guarantees standard first-order asymptotic properties of the estimators. A weaker condition, Assumption 5(ii), is sufficient for the MALIML/Fuller estimator. Assumptions 6 and 7 are used when we derive the asymptotic MSE of the MALIML/Fuller estimator.

THEOREM 4.1: *Suppose that Assumptions* 1–3 *are satisfied*. *Define* $\mu_i(W) = E[\epsilon_i^2 u_i]P_{ii}(W)$ *and* $\mu(W) = (\mu_1(W), \ldots, \mu_N(W))'$. *If* W *satisfies Assumptions* 4 *and* 5(i), *then*, *for MA2SLS* ($\hat{\beta}$ *defined in* (2.2)), *the decomposition given by* (4.1)

holds with

$$
\begin{aligned}
S(W) = H^{-1}\Bigg(& \operatorname{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] \frac{\sum_{i=1}^{N} (P_{ii}(W))^2}{N} + \sigma_{u\epsilon}\sigma_{u\epsilon}' \frac{(K'W)^2}{N} \\
& + (\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}') \frac{(W'\Gamma W)}{N} \\
& - \frac{K'W}{N} B_N + E[\epsilon_1^2 u_1] \frac{\sum_{i=1}^{N} f_i' P_{ii}(W)}{N} + \frac{\sum_{i=1}^{N} f_i P_{ii}(W)}{N} E[\epsilon_1^2 u_1'] \\
& + \frac{f'(I - P(W))\mu(W)}{N} + \frac{\mu(W)'(I - P(W))f}{N} \\
& + \sigma_\epsilon^2 \frac{f'(I - P(W))(I - P(W))f}{N} \Bigg) H^{-1},
\end{aligned}
$$

where $d = \dim(\beta)$ *and*

$$
\begin{aligned}
B_N = 2\Bigg(& \sigma_\epsilon^2 \Sigma_u + d\sigma_{u\epsilon}\sigma_{u\epsilon}' + \frac{1}{N}\sum_{i=1}^{N} f_i \sigma_{u\epsilon}' H^{-1} \sigma_{u\epsilon} f_i' \\
& + \frac{1}{N}\sum_{i=1}^{N} (f_i \sigma_{u\epsilon}' H^{-1} f_i \sigma_{u\epsilon}' + \sigma_{u\epsilon} f_i' H^{-1} \sigma_{u\epsilon} f_i') \Bigg).
\end{aligned} \tag{4.2}
$$

REMARK 2: When $d = 1$, $B_N = 2(\sigma_\epsilon^2 \Sigma_u + 4\sigma_{u\epsilon}^2)$.

Note that the term B_N is positive semidefinite. This implies that a higher order formula that neglects the term $(K'W/N)B_N$ overestimates the effect on the bias of including more instruments. A number of special cases discussed in the Supplemental Material (Kuersteiner and Okui (2010)) lead to simplifications of the above result. In particular, if $\operatorname{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ and $E[\epsilon_i^2 u_i] = 0$, as would be the case if ϵ_i and u_i were jointly Gaussian, the following result is obtained.

COROLLARY 4.1: *Suppose that the same conditions as in Theorem 4.1 hold and that, in addition,* $\operatorname{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ *and* $E[\epsilon_i^2 u_i] = 0$. *Then, for MA2SLS* ($\hat{\beta}$ *defined in* (2.2)), *the decomposition given by* (4.1) *holds with*

$$
\begin{aligned}
S(W) = H^{-1}\Bigg(& \sigma_{u\epsilon}\sigma_{u\epsilon}' \frac{(K'W)^2}{N} + (\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}') \frac{(W'\Gamma W)}{N} \\
& - \frac{K'W}{N} B_N + \sigma_\epsilon^2 \frac{f'(I - P(W))(I - P(W))f}{N} \Bigg) H^{-1},
\end{aligned} \tag{4.3}
$$

where B_N is as defined before.

For the MALIML and MAFuller estimator, we obtain the following result.

THEOREM 4.2: *Suppose that Assumptions 1–4, 5(ii), 6, and 7 are satisfied. Let $v_i = u_i - (\sigma_{u\epsilon}/\sigma_\epsilon^2)\epsilon_i$. Define $\Sigma_v = \Sigma_u - \sigma_{u\varepsilon}\sigma_{u\varepsilon}'$, $\mu_v(W) = (\mu_{v,1}(W), \ldots, \mu_{v,N}(W))'$, and $\mu_{v,i}(W) = E[\epsilon_i^2 v_i]P_{ii}(W)$. Then, for MALIML ($\hat\beta$ defined in (2.3)) and MAFuller ($\hat\beta$ defined in (2.4)), the decomposition given by (4.1) holds with*

$$S(W) = H^{-1}\Bigg(\sigma_\epsilon^2\Sigma_v\frac{W'\Gamma W}{N} + \sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N} + \mathrm{Cum}[\epsilon_i,\epsilon_i,v_i,v_i']\frac{\sum_{i=1}^N (P_{ii}(W))^2}{N} + \hat\zeta + \hat\zeta' - \frac{f'(I-P(W))\mu_v(W)}{N} - \frac{\mu_v(W)'(I-P(W))f}{N}\Bigg)H^{-1},$$

where

$$\hat\zeta = \sum_{i=1}^N f_i P_{ii}(W)E[\epsilon_i^2 v_i']/N - K'W/N\sum_{i=1}^N f_i E[\epsilon_i^2 v_i']/N.$$

When $\mathrm{Cum}[\epsilon_i,\epsilon_i,v_i,v_i'] = 0$ *and* $E[\epsilon_i^2 v_i] = 0$, *we have*

$$S(W) = H^{-1}\Bigg(\sigma_\epsilon^2\Sigma_v\frac{W'\Gamma W}{N} + \sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\Bigg)H^{-1}. \tag{4.4}$$

The following theorem shows that MA2SLS, MALIML, and MAFuller dominate corresponding estimators based on sequential moment selection under some regularity conditions on the population goodness-of-fit of the first-stage regression.

THEOREM 4.3: *Assume that Assumptions 1–5 hold. Let $\gamma_m = \lambda' H^{-1}f'(I - P_m)fH^{-1}\lambda/N$. Assume that there exists a nonstochastic function $C(a)$ such that $\sup_{a\in[-\varepsilon,\varepsilon]}\gamma_{m(1+a)}/\gamma_m = C(a)$ wpa1 as $N, m \to \infty$ for some $\varepsilon > 0$. Assume that $C(a) = (1+a)^{-2\alpha} + o(|a|^{2\alpha})$ as $a \to 0$.*

(i) *For $S_\lambda(W)$ given by (4.3), it follows that*

$$\frac{\min_{W\in\Omega_P} S_\lambda(W)}{\min_{W\in\Omega_{sq}} S_\lambda(W)} < 1 \quad wpa1.$$

(ii) *For $S_\lambda(W)$ given by* (4.4), *it follows that*

$$\frac{\min_{W\in\Omega_P} S_\lambda(W)}{\min_{W\in\Omega_{sq}} S_\lambda(W)} < 1 \quad wpa1.$$

REMARK 3: The additional conditions on γ_m imposed in Theorem 4.3 are satisfied if $\gamma_m = \delta m^{-2\alpha}$, but are also satisfied for more general specifications. For example, if $\gamma_m = \delta(m)m^{-2\alpha} + o_p(m^{-2\alpha})$ as $m \to \infty$, where the function $\delta(m)$ satisfies $\delta(m(1+a))/\delta(m) = 1 + o(|a|^{2\alpha})$ wpa1, then the condition holds.

To show that $\hat{W}$, which is found by minimizing $\hat{S}_\lambda(W)$, has certain optimality properties, we need to impose the following additional technical conditions.

ASSUMPTION 8: *For some α, $\sup_{m\le M} m^{2\alpha+1}(\sup_{\lambda'\lambda=1} \lambda' f'(P_m - P_{m+1}) f\lambda/N) = O_p(1)$.*

ASSUMPTION 9: *$\hat{H} - H = o_p(1)$, $\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = o_p(1)$, $\hat{\sigma}_\lambda^2 - \sigma_\lambda^2 = o_p(1)$, $\hat{\sigma}_{\lambda\epsilon} - \sigma_{\lambda\epsilon} = o_p(1)$, and $\hat{B}_N - B_N = o_p(1)$.*

ASSUMPTION 10: *Let α be as defined in Assumption 8. For some $0 < \varepsilon < \min(1/(2\alpha), 1)$ and δ such that $2\alpha\varepsilon > \delta > 0$, it holds that $M = O(N^{(1+\delta)/(2\alpha+1)})$. For some $\vartheta > (1+\delta)/(1-2\alpha\varepsilon)$, it holds that $E(|u_i|^{2\vartheta}) < \infty$. Further assume that $\hat{\sigma}_\lambda^2 - \sigma_\lambda^2 = o_p(N^{-\delta/(2\alpha+1)})$.*

Assumption 8 imposes additional smoothness on f. Assumption 9 assumes the consistency of the estimators of the parameters in the criterion function. Assumption 10 restricts the order of the number of instruments and assumes the existence of the moments of u_i. It also imposes a condition on the rate of the consistency of $\hat{\sigma}_\lambda^2$. For example, when $\alpha = 3/4$, $M = O(N^{3/5})$, $E[\|u_i\|^{16}] < \infty$, and $\hat{\sigma}_\lambda^2 - \sigma_\lambda^2 = o_p(N^{-1/5})$, Assumption 10 is satisfied by taking $\varepsilon = 1/2$, $\delta = 1/2$, and $\vartheta = 8$. We note that $\hat{\sigma}_\lambda^2 - \sigma_\lambda^2 = o_p(N^{-1/5})$ is achievable.

The following result generalizes a result established by Li (1987) to the case of the MA2SLS, MALIML, and MAFuller estimators.

THEOREM 4.4: *Let Assumptions 1–10 hold. For $\Omega = \Omega_U, \Omega_C$, or Ω_P and $\hat{W} = \arg\min_{W\in\Omega} \hat{S}_\lambda(W)$, where $\hat{S}_\lambda(W)$ is defined in either* (2.5) *or* (2.7), *it follows that*

$$\frac{\hat{S}_\lambda(\hat{W})}{\inf_{W\in\Omega} S_\lambda(W)} \to_p 1. \tag{4.5}$$

Theorem 4.4 complements the result in Hansen (2007). Apart from the fact that $\hat{S}_\lambda(W)$ is different from the criterion in Hansen (2007), there are more

technical differences between our result and Hansen's (2007). Hansen (2007) showed (4.5) only for a restricted set Ω, where Ω has a countable number of elements. We are able to remove the countability restriction and allow for more general W. However, in turn we need to impose an upper bound M on the maximal complexity of the models considered.

5. MONTE CARLO

5.1. *Design*

We use the same experimental design as Donald and Newey (2001) to ease comparability of our results with theirs. Our data-generating process is the model:

$$y_i = \beta Y_i + \epsilon_i, \qquad Y_i = \pi' Z_i + u_i$$

for $i = 1, \ldots, N$, where Y_i is a scalar, β is the scalar parameter of interest, $Z_i \sim$ i.i.d. $N(0, I_M)$, and (ϵ_i, u_i) is i.i.d. jointly normal with variances 1 and covariance c. The integer M is the total number of instruments considered in each experiment. We fix the true value of β at 0.1 and examine how well each procedure estimates β.

In this framework, each experiment is indexed by the vector of specifications: $(N, M, c, \{\pi\})$. We set $N = 100, 1000$. The number of instruments is $M = 20$ when $N = 100$ and $M = 30$ when $N = 1000$. The degree of endogeneity is controlled by the covariance c and set to $c = 0.1, 0.5, 0.9$. We consider the three specifications for π listed in Table I.

Models A and B were considered by Donald and Newey (2001). In Model C, the first $M/2$ instruments are completely irrelevant. Other instruments are relevant and the strength of them decreases gradually as in Model B. We use this model to investigate the ability of our procedure to eliminate irrelevant instruments. For each model, $c(M)$ is set so that π satisfies $\pi'\pi = R_f^2/(1 - R_f^2)$, where R_f^2 is the theoretical value of R^2 and we set $R_f^2 = \pi'\pi/(\pi'\pi + 1) = 0.1, 0.01$. The number of replications is 5000. The experiments are conducted with Ox 5.1 for Windows (see Doornik (2007)).

TABLE I

MODEL SPECIFICATIONS FOR MONTE CARLO SIMULATIONS

Model A	Model B	Model C
$\pi_m = \sqrt{\frac{R_f^2}{M(1-R_f^2)}}$	$\pi_m = c(M)(1 - \frac{m}{M+1})^4$	$\pi_m = \begin{cases} 0, & \text{for } m \le M/2 \\ c(M)(1 - \frac{m-M/2}{M/2+1})^4, & \text{for } m > M/2 \end{cases}$

5.2. *Estimators*

We compare the performances of the following 14 estimators. Seven of them are existing procedures and the other seven procedures are the MA2SLS, MALIML, and MAFuller estimators developed in this paper. First, we consider the 2SLS estimator with all available instruments (2SLS-All in the tables). Second, the 2SLS estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (2SLS-DN), where we use the criterion function (2.6). The kernel weighted generalized method of moments (GMM) of Kuersteiner (2002) is also examined (KGMM). Let $\Omega_{\mathrm{KGMM}} = \{W \in l_1 : w_m = L^{-1}$ if $m \leq L$ and 0 otherwise for some $L \leq M\}$. Then the MA2SLS estimator with $W \in \Omega_{\mathrm{KGMM}}$ corresponds to the kernel weighted 2SLS estimator with kernel $k(x) = \sqrt{\max(1-x, 0)}$. Because the weights are always positive with Ω_{KGMM}, we use the criterion function (2.6) for KGMM. The procedure 2SLS-U is the MA2SLS estimator with $\Omega = \Omega_U$. The procedure 2SLS-P uses the set $\Omega = \Omega_P$. The criterion for MA2SLS-U and MA2SLS-P is formula (2.5). The procedure 2SLS-Ps also uses the set Ω_P, but the criterion for computing weights is (2.6). For the LIML estimators, we consider the LIML estimator with all available instruments (LIML-All) and the LIML estimator with the number of instruments chosen by Donald and Newey's (2001) procedure (LIML-DN). We use the criterion function (2.7) for LIML-DN. The procedures LIML-U and LIML-P are the MALIML estimators with $\Omega = \Omega_U$ and $\Omega = \Omega_P$, respectively. For these MALIML estimators, we minimize the criterion (2.7) to obtain optimal weights. The estimators Fuller-All, Fuller-DN, Fuller-U, and Fuller-P are defined in the same way.[5]

For 2SLS-DN, LIML-DN, and Fuller-DN the optimal number of instruments is obtained by a grid search. We also use a grid search to find the L that minimizes the criterion for KGMM. For the MA2SLS, MALIML, and MA-Fuller estimators, we use the procedure SolveQP in Ox to minimize the criteria (see Doornik (2007)). We use the 2SLS, LIML, and Fuller estimators, respectively, with the number of instruments that minimizes the first-stage Mallows criterion as a first-stage estimator $\tilde{\beta}$ to estimate the parameters of the criterion function $S_\lambda(W)$.

For each estimator, we compute the median absolute deviation relative to that of 2SLS-DN (RMAD).[6] The measure KW+ is the average value of $\sum_{m=1}^{M} m \max(w_m, 0)$. The measure KW− is the average value of $\sum_{m=1}^{M} m|\min(w_m, 0)|$. This measure is zero for $W \in \Omega_P$ and $W \in \Omega_{sq}$.

[5]We do not report results for estimators based on Ω_C to preserve space. Their performance is very similar to that of estimators using Ω_U.

[6]We use this robust measure because of concerns about the existence of moments of estimators.

5.3. *Results*

Tables II–IV report simulation results for Models A–C. Table II contains results for the case when $c = 0.1$, where endogeneity is weak. 2SLS-All as well as 2SLS-U perform best. The KW+ and KW− statistics for 2SLS-U reveal that model averaging only partly attempts to offset bias, as can be expected from the theoretical analysis. 2SLS-P also performs well, but not as well as 2SLS-All, and 2SLS-U. The procedures that are based on (2.6) (2SLS-DN, KGMM, and 2SLS-Ps) do not perform well, but the performance of 2SLS-Ps is best among these three. This result reveals the importance of considering additional higher order terms in approximating the MSE. All the LIML and Fuller based estimators perform less well than their 2SLS counterparts. Nonetheless, in contrast to the results for 2SLS, both LIML-DN and the MALIML estimators improve substantially over LIML-All. The performance of LIML-U relative to LIML-All is particularly strong with similar results holding for Fuller. When identification is strong (i.e., $N = 1000$ and $R^2 = 0.1$), the difference between the different estimators disappears: all of them perform roughly at par. This is a reflection of the first-order asymptotic properties that will eventually dominate the finite sample issues in large samples.

In Table III, we consider the case of intermediate endogeneity ($c = 0.5$). When $N = 100$ and $R_f^2 = 0.01$, the 2SLS-type estimators perform well and the performance of 2SLS-P is particularly good. Fuller-P and Fuller-DN also do well in this case. On the other hand, when identification is strong ($N = 1000$ and/or $R_f^2 = 0.1$), the LIML-type and even more so the Fuller estimators perform better than the 2SLS-type estimators, and Fuller-P is the best estimator overall, sometimes outperforming 2SLS based methods by large margins. LIML-All and to a lesser extent Fuller-All perform particularly poorly across all scenarios in Table III. This underlines the importance of data-dependent instrument selection for the LIML and Fuller estimators.

Table IV contains results for the case of strong endogeneity ($c = 0.9$). In these experiments, LIML is generally the preferred method with a slight advantage over Fuller. LIML-U and LIML-P often perform equally well and at least as well as LIML-DN. In the designs with $R_f^2 = 0.1$ and $N = 100$ as well as $R_f^2 = 0.01$ and $N = 1000$, LIML-U and -P do sometimes significantly better than LIML-DN. Similar relationships hold for the Fuller-U and -P estimators relative to Fuller-DN. Overall LIML-P is the preferred procedure among the data-dependent procedures in Table IV, but Fuller-P is only slightly worse. Among the 2SLS procedures, 2SLS-Ps shows good performance over the range of cases reported on Table IV. Its RMAD value is never above 1 and in most cases significantly less. 2SLS-U outperforms 2SLS-All across all experiments. The ranking relative to 2SLS-DN depends on the data-generating process. In Model A and especially in Model C, 2SLS-U outperforms 2SLS-DN, whereas the latter has the upper hand in Model B. It is remarkable that 2SLS-U performs particularly well in Model C when $R_f^2 = 0.1$, demonstrating its ability to pick out relevant instruments.

TABLE II

MONTE CARLO RESULTS ($c = 0.1$)

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
									$N = 100$									
2SLS-All	0.38	20	0	0.66	20	0	0.42	20	0	0.74	20	0	0.37	20	0	0.68	20	0
2SLS-DN	1.00	4.47	0	1.00	9.43	0	1.00	4.35	0	1.00	7.13	0	1.00	4.52	0	1.00	9.91	0
KGMM	1.00	3.28	0	1.02	6.44	0	0.98	3.24	0	0.97	5.4	0	0.98	3.32	0	1.02	7.08	0
2SLS-U	0.40	72.4	65.6	0.68	80.5	71.6	0.44	73.2	66.4	0.75	87.1	78.4	0.39	71.2	64.4	0.70	78.4	69.8
2SLS-P	0.43	9.89	0	0.70	13.5	0	0.48	10.1	0	0.80	13.3	0	0.41	9.84	0	0.72	13.4	0
2SLS-Ps	0.75	4.9	0	0.91	8.22	0	0.80	4.96	0	0.93	7.36	0	0.75	4.92	0	0.92	8.73	0
LIML-All	1.98	20	0	1.79	20	0	1.89	20	0	2.01	20	0	1.93	20	0	1.80	20	0
LIML-DN	1.14	4.95	0	1.48	9.16	0	1.00	4.55	0	1.26	5.58	0	1.12	5.09	0	1.49	10.5	0
LIML-U	0.95	309	333	1.19	175	189	0.84	152	168	1.03	406	451	0.95	189	205	1.26	226	244
LIML-P	1.00	7.02	0	1.29	8.75	0	0.88	6.64	0	1.17	6.13	0	0.98	7.17	0	1.38	9.38	0
Fuller-All	1.42	20	0	1.60	20	0	2.23	20	0	1.78	20	0	1.36	20	0	1.61	20	0
Fuller-DN	0.62	3.72	0	1.17	8.4	0	1.00	3.54	0	1.14	5.44	0	0.61	3.76	0	1.14	9.45	0
Fuller-U	0.63	785	870	1.02	310	346	0.99	606	730	0.97	442	491	0.64	487	542	1.06	416	457
Fuller-P	0.58	5.13	0	1.07	8.05	0	0.93	5.01	0	1.05	5.9	0	0.57	5.2	0	1.11	8.52	0

(*Continues*)

TABLE II—*Continued*

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
								$N = 1000$										
2SLS-All	0.37	30	0	1.02	30	0	0.58	30	0	0.93	30	0	0.29	30	0	0.97	30	0
2SLS-DN	1.00	7.65	0	1.00	29	0	1.00	7.63	0	1.00	15.3	0	1.00	7.81	0	1.00	23.2	0
KGMM	0.95	5.43	0	1.14	15.5	0	0.93	5.94	0	0.98	13	0	0.96	5.68	0	1.01	15.5	0
2SLS-U	0.38	189	180	0.99	128	106	0.60	215	205	0.94	246	231	0.31	176	168	0.98	198	184
2SLS-P	0.42	14.4	0	1.01	27.9	0	0.66	16.7	0	0.97	23.9	0	0.33	13.7	0	0.99	26	0
2SLS-Ps	0.68	7.54	0	1.03	23.4	0	0.83	8.65	0	0.98	14.8	0	0.57	7.68	0	0.99	21.2	0
LIML-All	1.23	30	0	1.26	30	0	1.92	30	0	1.12	30	0	0.97	30	0	1.20	30	0
LIML-DN	0.97	8.24	0	1.24	29.2	0	1.14	5.31	0	1.06	11.2	0	0.79	10.3	0	1.17	21.8	0
LIML-U	0.75	369	392	1.26	311	311	0.90	898	959	1.05	1390	1480	0.62	524	556	1.12	1220	1310
LIML-P	0.88	10.9	0	1.24	23.2	0	1.07	7.82	0	1.05	11.4	0	0.72	12	0	1.15	19.3	0
Fuller-All	1.10	30	0	1.24	30	0	1.70	30	0	1.12	30	0	0.87	30	0	1.19	30	0
Fuller-DN	0.73	5.6	0	1.23	29.2	0	1.02	4.85	0	1.06	11.2	0	0.54	6.75	0	1.15	21.8	0
Fuller-U	0.58	682	733	1.24	314	314	0.81	1060	1130	1.04	1410	1490	0.47	1480	1580	1.12	1230	1320
Fuller-P	0.68	9.27	0	1.22	23.2	0	0.95	7.26	0	1.04	11.4	0	0.55	9.98	0	1.14	19.3	0

TABLE III

MONTE CARLO RESULTS ($c = 0.5$)

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
									$N = 100$									
2SLS-All	0.77	20	0	0.87	20	0	0.84	20	0	1.25	20	0	0.74	20	0	0.81	20	0
2SLS-DN	1.00	4.41	0	1.00	8.13	0	1.00	4.37	0	1.00	5.7	0	1.00	4.46	0	1.00	8.5	0
KGMM	0.98	3.23	0	0.99	5.87	0	0.97	3.2	0	0.93	4.51	0	0.98	3.29	0	1.05	6.3	0
2SLS-U	0.76	71.9	65.2	0.87	81.8	73.9	0.83	73.5	66.8	1.20	84.2	77.1	0.74	71.7	65	0.80	79.4	71.8
2SLS-P	0.76	9.81	0	0.84	12.4	0	0.81	9.89	0	1.02	10.7	0	0.74	9.77	0	0.79	12.1	0
2SLS-Ps	0.86	4.87	0	0.91	7.43	0	0.89	4.86	0	0.94	5.72	0	0.85	4.89	0	0.87	7.79	0
LIML-All	1.40	20	0	0.87	20	0	1.52	20	0	1.25	20	0	1.38	20	0	0.82	20	0
LIML-DN	0.97	4.94	0	0.83	9.36	0	0.95	4.56	0	0.89	5.84	0	0.95	5.05	0	0.75	10.6	0
LIML-U	0.92	2860	3140	0.93	166	176	0.94	132	144	1.05	189	207	0.88	142	157	0.76	118	124
LIML-P	0.90	7.05	0	0.81	8.96	0	0.93	6.72	0	0.92	6.89	0	0.88	7.19	0	0.72	9.68	0
Fuller-All	0.95	20	0	0.77	20	0	1.04	20	0	1.11	20	0	0.90	20	0	0.72	20	0
Fuller-DN	0.79	3.72	0	0.80	8.45	0	0.79	3.53	0	0.84	5.62	0	0.79	3.77	0	0.76	9.6	0
Fuller-U	0.81	3870	4270	0.89	254	289	0.83	383	430	0.98	217	239	0.80	664	746	0.73	287	317
Fuller-P	0.77	5.14	0	0.74	8.23	0	0.79	5.06	0	0.85	6.59	0	0.76	5.21	0	0.66	8.8	0

(*Continues*)

TABLE III—*Continued*

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
								$N = 1000$										
2SLS-All	0.82	30	0	0.70	30	0	1.32	30	0	1.50	30	0	0.62	30	0	0.95	30	0
2SLS-DN	1.00	7.1	0	1.00	12.9	0	1.00	5.76	0	1.00	7.73	0	1.00	6.74	0	1.00	16.8	0
KGMM	1.00	5.21	0	0.79	12.9	0	0.99	4.72	0	0.98	7.04	0	0.97	5.13	0	1.15	13.9	0
2SLS-U	0.82	187	179	0.93	207	197	1.28	201	193	1.33	216	210	0.61	175	168	0.85	183	177
2SLS-P	0.81	13.7	0	0.75	16.6	0	1.07	13.1	0	1.03	9.84	0	0.62	12.9	0	0.84	14.9	0
2SLS-Ps	0.89	7.13	0	0.80	13.2	0	0.97	6.55	0	0.99	7.23	0	0.74	7.09	0	0.84	12.8	0
LIML-All	0.84	30	0	0.47	30	0	1.30	30	0	0.98	30	0	0.63	30	0	0.63	30	0
LIML-DN	0.87	8.53	0	0.46	29.4	0	0.89	6.14	0	0.94	12.2	0	0.66	10.6	0	0.62	22.3	0
LIML-U	0.89	247	259	0.48	26.9	3.91	1.07	262	282	0.96	16.4	4.78	0.61	242	256	0.61	23.4	5.32
LIML-P	0.76	11.2	0	0.47	23.5	0	0.93	9.02	0	0.95	13.1	0	0.57	12.2	0	0.63	20.9	0
Fuller-All	0.73	30	0	0.47	30	0	1.15	30	0	0.97	30	0	0.55	30	0	0.63	30	0
Fuller-DN	0.89	5.57	0	0.46	29.4	0	0.86	5.3	0	0.95	12.2	0	0.72	6.77	0	0.62	22.3	0
Fuller-U	0.90	696	747	0.49	27	4.01	1.04	339	366	0.96	16.4	4.92	0.61	962	1030	0.61	23.4	5.44
Fuller-P	0.70	9.41	0	0.47	23.5	0	0.87	8.29	0	0.95	13	0	0.52	10.1	0	0.62	20.9	0

TABLE IV

MONTE CARLO RESULTS ($c = 0.9$)

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
									$N = 100$									
2SLS-All	0.97	20	0	0.93	20	0	1.06	20	0	1.81	20	0	0.94	20	0	0.71	20	0
2SLS-DN	1.00	4.18	0	1.00	3.81	0	1.00	3.78	0	1.00	2.65	0	1.00	4.31	0	1.00	3.34	0
KGMM	1.00	3.12	0	0.97	3.38	0	0.98	2.95	0	0.91	2.48	0	1.00	3.18	0	0.97	3.01	0
2SLS-U	0.97	71.7	65.2	0.94	77.4	71.6	1.06	68.6	62.5	1.69	62.4	57.5	0.94	72.1	65.6	0.69	73.4	67.7
2SLS-P	0.97	9.52	0	0.92	8.65	0	1.03	8.75	0	1.30	6.1	0	0.94	9.69	0	0.69	8.65	0
2SLS-Ps	0.97	4.65	0	0.93	4.47	0	0.96	4.18	0	0.97	2.94	0	0.96	4.82	0	0.75	4.69	0
LIML-All	0.94	20	0	0.38	20	0	1.03	20	0	0.73	20	0	0.93	20	0	0.29	20	0
LIML-DN	0.94	5.61	0	0.56	10.8	0	0.89	5.74	0	0.73	8.21	0	0.95	5.58	0	0.38	11.4	0
LIML-U	0.93	140	152	0.46	36.5	29.8	0.89	90.4	95.7	0.67	9.93	0.319	0.93	189	217	0.33	61.3	57.6
LIML-P	0.91	7.4	0	0.44	10.3	0	0.87	7.88	0	0.67	9.71	0	0.91	7.36	0	0.32	10.7	0
Fuller-All	0.78	20	0	0.33	20	0	0.86	20	0	0.65	20	0	0.76	20	0	0.26	20	0
Fuller-DN	0.98	3.87	0	0.66	9.4	0	0.94	3.9	0	0.73	7.71	0	0.97	3.83	0	0.46	10.2	0
Fuller-U	0.97	565	623	0.52	66.6	66.7	0.95	423	475	0.61	10.1	0.856	0.96	555	642	0.41	201	215
Fuller-P	0.95	5.46	0	0.49	9.51	0	0.93	6.09	0	0.61	9.49	0	0.95	5.33	0	0.39	9.84	0

(*Continues*)

TABLE IV—*Continued*

	Model A						Model B						Model C					
	$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$			$R_f^2 = 0.01$			$R_f^2 = 0.1$		
	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−	RMAD	KW+	KW−
									$N = 1000$									
2SLS-All	0.94	30	0	0.71	30	0	1.72	30	0	2.23	30	0	0.75	30	0	0.21	30	0
2SLS-DN	1.00	3.69	0	1.00	2.98	0	1.00	2.15	0	1.00	5.02	0	1.00	4.11	0	1.00	1.15	0
KGMM	0.99	3.06	0	0.83	3.56	0	0.91	2.07	0	0.90	4.62	0	0.99	3.17	0	0.98	1.23	0
2SLS-U	0.95	172	166	1.13	177	171	1.65	135	131	1.77	139	134	0.73	167	161	0.15	125	120
2SLS-P	0.93	9.7	0	0.81	8.73	0	1.25	6	0	1.03	6.63	0	0.74	10.8	0	0.17	9.05	0
2SLS-Ps	0.93	4.46	0	0.86	4.97	0	0.91	2.69	0	0.98	4.71	0	0.82	5.33	0	0.18	6.12	0
LIML-All	0.36	30	0	0.25	30	0	0.66	30	0	0.79	30	0	0.29	30	0	0.07	30	0
LIML-DN	0.66	12.3	0	0.25	29.6	0	0.68	10.9	0	0.79	16.9	0	0.59	12.2	0	0.07	24	0
LIML-U	0.47	126	123	0.26	23.9	0	0.59	14.1	0.251	0.78	16.3	0	0.39	325	337	0.07	22	0
LIML-P	0.44	13.4	0	0.26	23.9	0	0.59	13.9	0	0.78	16.3	0	0.36	13.5	0	0.07	21.9	0
Fuller-All	0.31	30	0	0.25	30	0	0.57	30	0	0.78	30	0	0.25	30	0	0.07	30	0
Fuller-DN	0.87	7.8	0	0.25	29.6	0	0.70	9.6	0	0.79	16.8	0	0.91	7.47	0	0.07	24	0
Fuller-U	0.61	346	366	0.26	23.9	0	0.52	25.9	14.8	0.79	16.3	0	0.53	1350	1430	0.07	22	0
Fuller-P	0.53	11.3	0	0.26	23.9	0	0.52	13.3	0	0.79	16.3	0	0.46	11	0	0.07	21.9	0

Overall, our results show that when the correlation between structural and reduced form errors is low, 2SLS-U is the preferred procedure among all data-dependent estimators considered. With increasing correlation, LIML-P and Fuller-P perform best in most cases. In addition, the LIML and Fuller estimators are quite sensitive to the number of instruments, and data-dependent methods clearly dominate LIML-All and Fuller-All in the majority of cases. The simulation results also confirm our theoretical findings. They indicate the importance of considering additional higher order terms in approximating the MSE of the estimator: 2SLS-P generally performs better than 2SLS-Ps except in Model B when $c = 0.9$. Our results also document the ability of the model averaging procedure with possibly negative weights to pick out irrelevant instruments. In Model C, our estimators based on Ω_U perform better than the sequential selection methods, and 2SLS-U significantly outperforms 2SLS-All when $R^2 = 0.1$, $c \geq 0.5$, and $N = 1000$. Finally, we note that 2SLS-Ps systematically outperforms 2SLS-DN across all Monte Carlo designs.

6. CONCLUSIONS

For models with many overidentifying moment conditions, we show that model averaging of the first-stage regression can be done in a way that reduces the higher order MSE of the 2SLS estimator relative to procedures that are based on a single first-stage model. The procedures we propose are easy to implement numerically. Monte Carlo experiments document that the MA2SLS estimators perform at least as well as conventional moment selection approaches, and perform particularly well when the degree of endogeneity is low to moderate and when the instrument set contains uninformative instruments. When endogeneity is moderate to strong, the MALIML and MAFuller estimators with positive weights are the preferred estimators.

REFERENCES

CANAY, I. A. (2010): "Simultaneous Selection and Weighting of Moments in GMM Using a Trapezoidal Kernel," *Journal of Econometrics* (forthcoming). [702]

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191. [697-699,701,703,704,708,709]

DOORNIK, J. A. (2007): *Ox 5—An Object-Oriented Matrix Programming Language*. London: Timberlake Consultants Ltd. [708,709]

FULLER, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–954. [697,699]

HAHN, J., J. HAUSMAN, AND G. KUERSTEINER (2004): "Estimation With Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations," *Econometrics Journal*, 7, 272–306. [699]

HANSEN, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [697, 707,708]

KUERSTEINER, G. M. (2002): "Mean Squared Error Reduction for GMM Estimators of Linear Time Series Models," Unpublished Manuscript, University of California, Davis. [701,702,709]

KUERSTEINER, G., AND R. OKUI (2010): "Supplement to 'Constructing Optimal Instruments by First Stage Prediction Averaging': Auxiliary Appendix," *Econometrica Supplemental Material*, 78, http://www.econometricsociety.org/ecta/Supmat/7444_proofs.pdf; http://www.econometricsociety.org/ecta/Supmat/7444_data and programs.zip. [700,703,705]

LI, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [707]

NAGAR, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595. [697]

NEWEY, W. K. (2007): "Choosing Among Instruments Sets," Unpublished Manuscript, Massachusetts Institute of Technology. [697]

OKUI, R. (2008): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics* (forthcoming). [702]

POLITIS, D. N., AND J. P. ROMANO (1995): "Bias-Corrected Nonparametric Spectral Estimation," *Journal of Time Series Analysis*, 16, 67–103. [702]

Dept. of Economics, University of California, Davis, 1 Shields Avenue, Davis, CA 95616, U.S.A.; gkuerste@ucdavis.edu

and

Institute of Economic Research, Kyoto University, Yoshida-Hommachi, Sakyo, Kyoto, Kyoto, 606-8501, Japan; okui@kier.kyoto-u.ac.jp.

Manuscript received September, 2007; final revision received October, 2009.