Estimator Averaging for Two Stage Least Squares*

Guido Kuersteiner† and Ryo Okui‡

This version: October 2007

Abstract

This paper considers model averaging as a way to select instruments for the two stage least squares estimator 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 second stage. Many existing two stage least squares estimators can be understood as model averaging methods with some restrictions on the form of the weights for averaging. Furthermore, we allow weights to be possibly negative, which yields a bias correction and instrument selection that is more robust to the ordering of instruments. The weights for averaging are chosen to minimize the asymptotic mean square error. This can be done by solving a standard quadratic programming problem and, in some cases, closed form solutions for the optimal weights are available. We demonstrate both theoretically and in Monte Carlo experiments that our method dominates existing number of instrument selection procedures for the two stage least squares estimator.

Keywords: model averaging, instrumental variable, many instruments, two stage least squares, higher order theory.

JEL classification: C21, C31.

*We thank Whitney Newey for valuable comments and suggestions. Okui acknowledges financial support from the Hong Kong University of Science and Technology under Project No. DAG05/06.BM16. We are solely responsible for all errors.

†Department of Economics, University of California, Davis. Address: One Shields Ave, Davis, CA 95616. Email: gkuerste@ucdavis.edu

‡Department of Economics, Hong Kong University of Science and Technology. Address: Clear Water Bay, Kowloon, Hong Kong. Email: okui@ust.hk

1 Introduction

In this paper we propose a new and flexible method to select the instruments for two stage least squares (2SLS) estimators of linear models when there are many instruments available. In an influential empirical study, Angrist and Krueger (1991) generated a larger instrument set from combinations of basic instruments to achieve more precise parameter estimates. However, the approach of Angrist and Krueger (1991) has raised concerns with respect to the quality of asymptotic approximations and inference procedures based on them due to weak instruments and parameter overidentification. Nelson and Startz (1990a,b) and Bound, Jaeger and Baker (1995) demonstrate that first order asymptotic theory yields poor approximations to the finite sample distribution of the 2SLS estimator when instruments are either weak or the degree of overidentification is large. Bekker (1994) provides an asymptotic analysis of the latter case by considering asymptotic sequences where the number of instruments grows at the same rate as the sample size. Staiger and Stock (1997) and Stock and Wright (2000) develop an asymptotic framework where the parameters of the first stage regression are local to zero to obtain more accurate asymptotic approximations when the instruments are weak. We also find related studies in the early literature on small sample properties of 2SLS such as the higher order expansions of Nagar (1959), Kunitomo (1980) and Morimune (1983) or the exact finite sample theory of Richardson (1968), Mariano (1972) and Phillips (1980). Weak identification and lack of identification was also covered in earlier work by Sargan (1983) and Phillips (1989). The conclusion from exact finite sample and refined asymptotic theory is that the optimality properties of 2SLS procedures may not be relevant for their actual finite sample performance and highly overidentified estimators should be used with caution.

A more recent literature has focused on improved estimators in the case of overidentified models. Donald and Newey (2001) propose a selection criterion to select the number of instruments in a way that balances higher order bias and efficiency. The higher order mean square error (MSE) approximations obtained by Donald and Newey (2001) indicate that the higher order bias of 2SLS increases linearly with the number of overidentifying conditions. Monte Carlo evidence of Donald and Newey (2001) shows that the selection criterion for the number of instruments performs particularly well when the degree of endogeneity is high but fares less well when the correlation between structural and reduced form innovations is weak to moderate. In the latter case, estimator bias tends to be overstated by the higher order MSE approximation and as a consequence too small a number of instruments is selected. This tendency for 'bias pessimism' of higher order expansions

to the distribution of the 2SLS estimator was documented by Hahn, Hausman and Kuersteiner (2004). These findings are in line with recent work on many weak instruments by Stock and Yogo (2004), Chao and Swanson (2005), Han and Phillips (2006), Hansen, Hausman and Newey (2006), Hausman, Newey and Woutersen (2006) and Newey and Windmeijer (2007). This literature shows that consistent estimation of models with many weak instruments is feasible and that under certain conditions asymptotically valid inference based on Gaussian limit distributions is possible.

The focus of this paper is to extend the results and methods proposed in Donald and Newey (2001). We show that the model averaging approach of Hansen (2007) can be applied to the first stage of the 2SLS estimator. The benefits of model averaging mostly lie in a more favorable trade off between bias and efficiency in the second stage of the 2SLS estimator. Our theoretical results show that for certain choices of weights the model averaging 2SLS estimator (MA2SLS) eliminates higher order bias and achieves the same higher order rates of convergence as the Nagar and LIML estimators and thus dominates conventional 2SLS procedures. Model averaging allowing for bias reduction requires a refined asymptotic approximation to the MSE of the 2SLS estimator. We provide such an approximation by including terms 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.

A limitation of the approach that selects the number of instruments is that the method is sensitive to the a priori ordering of instruments. By allowing our model weights to be both positive and negative, we establish that the MA2SLS estimator has 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, the MA2SLS is able to individually select these directions form the instrument set. Conventional sequential instrument selection on the other hand would be able to select these instruments only as part of a possibly much larger collection of potentially less informative instruments.

An added benefit of model averaging is that, in some cases, the optimal weights are available in closed form which lends itself to straight-forward empirical application. In Monte Carlo experiments we find that our refined selection criterion combined with a more flexible choice of instruments generally performs at least as well as only selecting the number of instruments over a wide range of models and parameter values, and performs particularly well in situations where selecting the number of instruments tends to select too few instruments.

Our model averaging approach can also be applied to non-linear procedures such as the limited

information maximum likelihood (LIML) estimator. To preserve space, we do not present results for these procedures. LIML, k-class estimators, continuous updating estimators and empirical likelihood procedures perform well in terms of their higher order properties compared to the 2SLS estimator. Higher order asymptotic properties of these estimators have recently been considered by Bekker (1994), Donald and Newey (2001), Hahn, Hausman and Kuersteiner (2004), Newey and Smith (2004) and Bekker and van der Ploeg (2005). However, simulation experiments in Donald and Newey (2001) and Hahn, Hausman and Kuersteiner (2004) indicate that the finite sample properties of LIML are not always superior to those of 2SLS and can be quite poor due to lack of finite sample moments of the estimator distribution. This moment problem is particularly pronounced when identification is weak.

In this paper we focus exclusively on the moment selection problem for 2SLS. Aside form the mixed finite sample evidence regarding the ranking of various procedures, our focus on the 2SLS procedure is also motivated by its wide use in practice. A few alternative methods to the selection approach of Donald and Newey (2001) have recently been suggested. Kuersteiner (2002) shows that kernel weighting of the instrument set can be used to reduce the 2SLS bias, an idea that was further developed by Okui (2007) and Canay (2006). The MA2SLS estimator proposed in this paper can be interpreted as a generalization of the more restrictive kernel weighted methods. While kernel weighting is shown to reduce bias, its effects on the MSE of the estimator are ambiguous. The goal of this paper therefore is to develop an instrument selection approach that is less sensitive to instrument ordering, dominates the approach of selecting the number of instruments in terms of higher order MSE and outperforms the number of instrument selection procedure in finite sample Monte Carlo experiments.

We present the general form of the MA2SLS estimator in Section 2.2 and discuss various members of the class of MA2SLS estimators in Section 3.2. The refined higher order MSE approximation for the MA2SLS family is obtained in Section 3.1. Section 3.3 demonstrates that optimal members of the MA2SLS family dominate the pure number of instrument selection method for the 2SLS estimator as well as the bias corrected version of that estimator in terms of relative higher order MSE. In Section 4 we establish that feasible versions of the MA2SLS estimator maintain certain optimality properties. Section 5 contains Monte Carlo evidence of the small sample properties of the MA2SLS estimators.

Throughout the paper, we use the following notations. For a sequence of vectors $a_1, \ldots, a_N$, the matrix a is defined as $a = (a_1, \ldots, a_N)'$. For a matrix A, A_{ij} denotes the (i, j)-th element of A.

$\sum_i$ signifies $\sum_{i=1}^N$ unless otherwise specified. "wpa1" reads "with probability approaching one".

2 First Stage Model Averaging 2SLS Estimator

2.1 Settings

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

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

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 is an unknown function of z. Let $f_i = f(z_i)$. The set of instruments has the form $Z_{M,i}' \equiv (\psi_1(z_i), \cdots, \psi_M(z_i))$, where ψ_ks are functions of z_i such that $Z_{M,i}$ is a $M \times 1$ vector of instruments.

In the current model, the asymptotic variance of a $\sqrt{N}$-consistent regular estimator cannot be smaller than $\sigma^2 \bar{H}^{-1}$, where $\sigma^2 = E[\epsilon_i^2|z_i]$ and $\bar{H} = E[f_i f_i']$ (Chamberlain (1987)). The lower bound is achieved by 2SLS if f_i can be written as a linear combination of the instruments. Likewise, if there is a linear combination of the instruments which is close to f_i, then the asymptotic variance of an instrumental variable estimator is small. This observation implies that using many instruments is desirable in terms of asymptotic variance. However, an instrumental variables estimator with many instruments may behave poorly in finite samples and can be sensitive to the number of instruments. Thus, instrument selection is critical to good finite sample performance of 2SLS estimators.

2.2 Model Averaging

Let W_N be a weighting vector such that $W_N = (w_{1,N}, \ldots, w_{M,N})$ and $\sum_{m=1}^M w_{m,N} = 1$ for some M such that $M \leq N$ for any N. We note that W_N is a sequence of weights $w_{m,N}$ indexed by the sample size N, but for notational convenience we use the symbols W and w_m. In Sections 3.2 and 4, we discuss in more detail the restrictions that need to be imposed on W and M, but 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 $Z_{M,i}$, Z_m be the matrix $(Z_{m,1}, \ldots, Z_{m,N})'$ and $P_m = Z_m(Z_m'Z_m)^{-1}Z_m'$. The model

averaging two-stage least squares estimator (MA2SLS) of β is defined as

$$\hat{\beta} = \arg\min_{\beta} = \sum_{m=1}^{M} w_m (y - X\beta)' P_m (y - X\beta). \tag{2.2}$$

Note that $(y - X\beta)' P_m (y - X\beta)$ is the objective function for the 2SLS estimator using the first m instruments. Define $P(W) = \sum_{i=1}^{M} w_m P_m$. The estimator $\hat{\beta}$ then can be written conveniently as

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

We use the term MA2SLS because $P(W)X$ is the predictor of X based on Hansen's (2007) model averaging estimator applied to the first stage regression. The model averaging estimator exploits a trade-off between specification bias and variance. In our application this trade off appears in the second stage of 2SLS as well, however with reversed implications: more specification bias in the first stage leads to less estimator bias in the second stage, and reduced variance in the first stage leads to less efficiency in the second stage. This trade off is well understood from the work of Nagar (1959), Bekker (1994) and Donald and Newey (2001) amongst others. As Hansen (2007) demonstrates, model averaging improves the bias-variance trade-off in conventional model selection contexts. These advantages translate into corresponding advantages for the second stage estimator as our theoretical analysis shows. Furthermore, we generalize the work of Hansen (2007) by allowing weights to be possibly negative while weights examined by Hansen (2007) are restricted to be positive. Allowing negative weights is important to obtain a bias correction and robustness with respect to the ordering of the instruments.

2.3 Advantages of Model Averaging

To give a preview of our results, we note that under suitable conditions on the behavior of W as a function of the sample size N it can be shown that the largest term of the higher order bias of $\hat{\beta}$ is proportional to $K'W/\sqrt{N}$ where $K = (1, 2, \ldots, M)'$. 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 $m/\sqrt{N}$. In other words, the first stage model selection approach of Donald and Newey (2001) can be nested within the class of MA2SLS estimators by choosing $w_j = 1$ for $j = m$ and $w_j = 0$ for $j \neq m$. 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 an

negative. In fact, the Nagar estimator can be interpreted as a special case of the MA2SLS where $M = N$, $w_j = N/(N-m)$ for $j = m$, some m, $w_N = -m/(N-m)$ and $w_j = 0$ otherwise.[1] As we demonstrate later, MA2SLS defines a much wider class of estimators with desirable MSE properties even when $K'W = 0$ does not hold and dominates the Nagar estimator when $K'W = 0$ is imposed.

Kuersteiner (2002) 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 $k = \text{diag}(k_1, ..., k_M)$ 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 (2.3) with $P(W)$ replaced by $Z_M k(Z_M'Z_M)^{-1}kZ_M'$. 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, but note that similar results hold in the general case.[2] Let $\tilde{Z}_j$ be the j-th 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

$$\sum_{m=1}^{M} w_m P_m = \sum_{j=1}^{M} k_j^2 \tilde{P}_j = Z_M k(Z_M'Z_M)^{-1}kZ_M' \tag{2.4}$$

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 (2007) shrinkage 2SLS estimator is also a special case of the averaged estimator (2.3). In this case, $w_L = s, w_M = 1-s, s \in [0,1], w_j = 0$ 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 estimators towards zero. Shrinkage of 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' can not completely eliminate bias. The calculations in Kuersteiner

[1] The approximate higher order MSE for the Nagar estimator is covered by Corollary 7.3 in Section 7, see Remark 4.

[2] In other words, we ortho-normalize the instruments prior to kernel weighting. Thus, that $Z_M'Z_M$ is a diagonal matrix is not really a restriction in practice. When kernel weighting is applied to the instruments that are not ortho-normalized, the MA weights corresponding to some particular kernel become data dependent and have a more complicated formula.

(2002) also show that the distortion introduced from using the weight matrix $k(Z_M'Z_M)^{-1}k$ rather than $(Z_M'Z_M)^{-1}$ asymptotically dominates the higher order variance of $\hat{\beta}$ for conventional choices of $k(\cdot)$. This later problem was recently addressed by Canay (2006) through the use of top-flat kernels (see, e.g., Politis and Romano (1995), Politis (2001) and Politis (2006)).

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 fully flexible weights 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. As we show in Section 3.3 this result holds even in cases where the bias is not fully eliminated and thus $K'W = 0$ does not hold.

A second advantage of model averaging is its ability to pick models from a wider class than sequential instrument selection can. Imagine a situation where the first m ($< M$) instruments in Z_M are redundant. In this case a sequential procedure will need to include the uninformative instruments while the model averaging procedure can in principle set weights $w_M = 1$ and $w_m = -1$ such that $P(W) = P_M - P_m$ is the projection on the orthogonalized set of the last $M - m$ instruments in Z_M. To be more specific, let z_i be the i-th column of Z_M when $i \leq M$ and define $\tilde{z}_2 = (I - P_1)z_2, \tilde{z}_3 = (I - P_2)z_3, ..., \tilde{z}_M = (I - P_{M-1})z_M$ such that $z_1, \tilde{z}_2, ...\tilde{z}_M$ are orthogonal and span Z_M. Then, $P_M = \sum_{i=1}^{M} \tilde{P}_i$ where $\tilde{P}_i = \tilde{z}_i(\tilde{z}_i'\tilde{z}_i)^{-1}\tilde{z}_i'$ for $i > 1$ and $P_1 = z_1 (z_1'z_1)^{-1} z_1'$. It follows that $\sum_{m=1}^{M} w_m P_m = \sum_{j=1}^{M} \tilde{w}_j \tilde{P}_j$ for $\tilde{w}_j = \sum_{m=j}^{M} w_m$. If D is an $M \times M$ matrix with elements $d_{ij} = \mathbf{1}\{j \geq i\}$ and $\tilde{W} = (\tilde{w}_1, ..., \tilde{w}_M)'$, then $W = D^{-1}\tilde{W}$. The only constraint we impose on $\tilde{W}$ is $\tilde{w}_1 = 1$. Since $\tilde{W}$ is otherwise unconstrained one can set $\tilde{w}_j = 0$ for any $1 < j \leq M$. In addition, an arbitrarily small but positive weight can be assigned to the first coordinate by choosing $\tilde{w}_j$ large for $j \neq 1$. The use of negative weights thus allows MA2SLS to pick out relevant instruments from a set of instruments that contains redundant instruments. In Section 5 we document the ability of MA2SLS to pick out relevant instruments through Monte Carlo experiments.

3 Higher Order Theory

3.1 Asymptotic Mean Square Error

The choice of model weights W is based on an approximation to the higher order MSE of $\hat{\beta}$. The derivations parallel those of Donald and Newey (2001). 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. We show the asymptotic properties of the MA2SLS under the following assumptions.

Assumption 1 *$\{y_i, X_i, z_i\}$ are 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 $\alpha > 1/2$,*

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

Assumption 3 *(i) $E[(\epsilon_i, u_i')'(\epsilon_i, u_i')|z_i]$, $E\left[\epsilon_i^2 u_i u_i'|z_i\right]$ and $E\left[\epsilon_i^2 u_i|z_i\right]$ are constant. Let $\sigma_{u\epsilon} = E[u_i\epsilon_i|z_i]$, $\Sigma_u = E[u_i u_i'|z_i]$. (ii) $Z_M' Z_M$ are nonsingular wpa1. (iii) $\max_{i \leq N} P_{M,ii} \to_p 0$, where $P_{M,ii}$ signifies the (i,i)-th element of P_M, (iv) f_i is bounded.*

Assumption 4 *Let $W^+ = (|w_1|, \ldots, |w_M|)'$. The following conditions hold: $\mathbf{1}_M' W = 1$; $W \in l_1$ where $l_1 = \{x = (x_1, ...) \,|\, \sum_{i=1}^\infty |x_i| < \infty\}$, $M \leq N$; and, as $N \to \infty$ and $M \to \infty$, $K'W^+ = \sum_{m=1}^M |w_m| m \to \infty$, $K'W^+/\sqrt{N} = \sum_{m=1}^M |w_m| m/\sqrt{N} \to 0$, and $\sum_{m=1}^M \left(\sum_{s=1}^m w_s\right)^2 m^{-2\alpha} \to 0$.*

Assumptions 1-3 are similar to those imposed in Donald and Newey (2001). 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. We remind the reader that W is a sequence, indexed by N, of sequences. The condition $K'W^+ \to \infty$ may be understood as the number of instruments tending to infinity. This assumption is needed to achieve the semiparametric efficiency bound and to obtain the asymptotic MSE whose leading terms depend on $K'W$. The condition $K'W^+/\sqrt{N} \to 0$ limits the rate at which the number of instruments is allowed to increase, which guarantees standard first-order asymptotic properties of the MA2SLS estimator. The condition $\sum_{m=1}^M \left(\sum_{s=1}^m w_s\right)^2 m^{-2\alpha} \to 0$ guarantees that small models receive asymptotically negligible weight and is needed to guarantee first order asymptotic efficiency of the MA2SLS estimator. 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 $j > M$.

The notion of asymptotic MSE employed here is similar to the Nagar-type asymptotic expansion (Nagar (1959)). 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

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

$H = f'f/N$ and $(\hat{r}(W) + T(W))/\mathrm{tr}(S(W)) = o_p(1)$ as $N \to \infty$. First we represent $N(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)'$ in two parts, $\hat{Q}(W)$ and $\hat{r}(W)$, and discard $\hat{r}(W)$, which goes to zero in probability more quickly

than $S(W)$. Then we take the expectation of $\hat{Q}(W)$ conditional on the exogenous variables, z, and ignore the term $T(W)$, which goes to zero more quickly than $S(W)$. The term $\sigma_\epsilon^2 H^{-1}$ corresponds to the first-order asymptotic variance. Hence, $S(W)$ is the nontrivial and dominant term in the MSE that depends on W.

Formal theorems and explicit expressions for $S(W)$ are reported in Theorem 7.1 and Corollaries 7.1, 7.2 and 7.3. In this section, we briefly discuss the main findings. Under additional constraints on higher order moments such that $\text{Cum}\left[\epsilon_i, \epsilon_i, u_i, u_i'\right] = 0$ and $E[\epsilon_i^2 u_i] = 0$,[3] we show in Corollary 7.1 that

$$S\left(W\right) = H^{-1}\left(a_\sigma \frac{(K'W)^2}{N} + b_\sigma \frac{(W'\Gamma W)}{N} - \frac{K'W}{N} B_N + \sigma_\epsilon^2 \frac{f'(I - P(W))(I - P(W))f}{N}\right) H^{-1} \tag{3.1}$$

where $a_\sigma = \sigma_{u\epsilon}\sigma_{u\epsilon}'$, $b_\sigma = (\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')$, B_N is defined in (7.1) and Γ is the $M \times M$ matrix whose (i,j)-th element is $\min(i,j)$. In Section 7 we also derive results for the more general case when $\text{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] \neq 0$ and $E[\epsilon_i^2 u_i] \neq 0$. Because these formulas are substantially more complicated, we focus our discussion on the simpler case.[4] The first term in (3.1) represents the square of the bias, and the fourth term represents the goodness-of-fit of the first stage regression. These two terms appear in the existing results of the asymptotic expansions of the 2SLS estimator. The second term represents a variance inflation by including many instruments. A similar term appears in the MSE results for LIML and bias-corrected 2SLS estimators by Donald and Newey (2001). As shown in Theorem 7.1, B_N in the third term is positive definite. This shows that $a_\sigma (K'W)^2/N$ over-estimates the bias of including more instruments. We need to include the second and third terms because $K'W \to \infty$ may not hold as a result of allowing negative weights. In fact, when the weights are all positive, we have $K'W \to \infty$ (because $K'W = K'W^+$ in this case) and the second and third term then are of lower order as established in expression (7.4) of Corollary 7.2.

3.2 Estimator Classes

We choose W to minimize the approximate MSE of $\lambda'\hat{\beta}$ for some fixed $\lambda \in \mathbb{R}^d$. For this purpose define $\sigma_{\lambda\epsilon} = \lambda' H^{-1}\sigma_{u\epsilon}$ and $\sigma_\lambda^2 = \lambda' H^{-1}\Sigma_u H^{-1}\lambda$. Then, the optimal weight, denoted W^*, is the solution to $\min_{W\in\Omega} S_\lambda\left(W\right)$ where $S_\lambda\left(W\right) = \lambda' S\left(W\right)\lambda$ and Ω is some set.

[3] "Cum" signifies the fourth order cumulant so that $\text{Cum}\left[\epsilon_i, \epsilon_i, u_i, u_i'\right] = E[\epsilon_i^2 u_i u_i'] - \sigma_\epsilon^2 \Sigma_u - 2\sigma_{u\epsilon}\sigma_{u\epsilon}'$.

[4] As was noted in Donald and Newey (2001), it is possible to use the more general criterion where $\text{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] \neq 0$ and $E[\epsilon_i^2 u_i] \neq 0$ because the additional nuisance parameters for this case can be estimated. We note that in practice this seems to be rarely done.

We consider several versions of Ω which lead to different estimators. The MA2SLS estimator is unconstrained if $\Omega = \Omega_U = \{W \in l_1 | W'\mathbf{1}_M = 1\}$. More restricted versions can be constructed by considering the sets $\Omega_B = \{W \in l_1 | W'\mathbf{1}_M = 1, K'W = 0\}$ which leads to unbiased estimators. 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 the following definitions of restricted model averaging classes defined as $\Omega_C = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [-1, 1], \forall m \leq M\}$, and $\Omega_P = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [0, 1], \forall m \leq M\}$.

When Ω is equal to Ω_U or Ω_B, a closed form solution for W^* is available. Let $u_\lambda^m = (I - P_m) f H^{-1}\lambda$ and define the matrix $U = (u_\lambda^1, \ldots, u_\lambda^M)'(u_\lambda^1, \ldots, u_\lambda^M)$. It now follows that $\lambda' H^{-1} f'(I - P(W))(I - P(W)) f H^{-1}\lambda = W'UW$ such that $S(W)$ is affine in W. It then is easy to show that

$$W_U^* = \arg\min_{W \in \Omega_U} S_\lambda(W) = \frac{1}{2} A^{-1} \left(K\lambda' H^{-1} B_N H^{-1}\lambda + \frac{2 - \mathbf{1}_M' A^{-1} K\lambda' H^{-1} B_N H^{-1}\lambda}{\mathbf{1}_M' A^{-1}\mathbf{1}_M} \mathbf{1}_M \right) \quad (3.2)$$

where $A = \sigma_{\lambda\epsilon}^2 KK' + \left(\sigma_\epsilon^2 \sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2\right)\Gamma + \sigma_\epsilon^2 U$. As we show in Corollary 7.3 the approximate MSE of $\hat{\beta}$ simplifies when the constraint $K'W = 0$ is imposed. In this case, weights are chosen such as to eliminate the highest order bias term. We therefore find the following closed form solution for W_B^*.

$$W_B^* = \arg\min_{W \in \Omega_B} S_\lambda(W) = A_B^{-1} R \left(R' A_B^{-1} R\right)^{-1} b \quad (3.3)$$

where $A_B = \left(\sigma_\epsilon^2 \sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2\right)\Gamma + \sigma_\epsilon^2 U$, $b = (0, 1)'$ and $R = (K, \mathbf{1}_M)$. It is clear that $\Omega_B \subset \Omega_U$ such that $\min_{W \in \Omega_U} S_\lambda(W) \leq \min_{W \in \Omega_B} S_\lambda(W)$. Since the Nagar estimator is contained in Ω_B, it follows by construction that MA2SLS based on W_U^* weakly dominates the Nagar estimator in terms of asymptotic MSE.[5] In Section 3.3 we show that MA2SLS strictly dominates the Nagar estimator. When the optimal weights are restricted to lie in the sets Ω_C or Ω_P, no closed form solution exists. Finding the optimal weights minimizing $S_\lambda(W)$ over a constrained set is a classical quadratic programming problem for which there are readily available numerical algorithms.[6] We note that for Ω_P, it follows from Corollary 7.2 that the criterion can be simplify to (7.4).

3.3 Relative Higher Order Risk

It is easily seen that Donald and Newey's (2001) procedure can be viewed as a special case of model averaging where the weights are chosen from the set $\Omega_{DN} \equiv \{W \in l_1 | w_m = 1$ for some m and $w_j =$

[5] Our Monte Carlo results show that W_U^* has good finite sample properties. Monte Carlo evidence for W_B^*, not included in this paper to conserve space, shows that this version of the estimator has poor finite sample properties and we do not further consider it for this reason, except for a theoretical argument in Theorem 3.1.

[6] The Gauss programming language has the procedure QPROG, and the Ox programming language has the procedure SolveQP.

0 for $j \neq m\}$ to minimize $S_\lambda(W)$. Note that when $W \in \Omega_{DN}$, it follows that $K'W = m$ and $(I - P(W))(I - P(W)) = (I - P_m)$. Hence, $S(W)$ with W restricted to $W \in \Omega_{DN}$ reduces to

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

for $m \leq M$. Because $m/N = o(m^2/N)$ as $m \to \infty$, the expression for $S(W)$ with $W \in \Omega_{DN}$ reduces to the result of Donald and Newey (2001, Proposition 1). We note that all sets $\Omega = \Omega_U, \Omega_B, ..., \Omega_P$ contain the procedure of Donald and Newey (2001) as a subset (i.e., $\Omega_{DN} \subset \Omega$). This guarantees that MA2SLS weakly dominates the number of instrument selection procedure such that $S_\lambda(W^*) \leq \min_{W \in \Omega_{DN}} S_\lambda(W)$. In fact, as the argument in the proof of Lemma 7.6 shows, there are simple sequences in Ω_U and Ω_B that strongly dominate $\arg\min_{W \in \Omega_{DN}} S_\lambda(W)$ in the sense of achieving higher rates of convergence.

A stronger result is the following theorem which shows that, under some regularity conditions on the population goodness-of-fit of the first stage regression, the asymptotic MSE of MA2SLS is lower than that of 2SLS with pure number of instrument selection even when $K'W \neq 0$.

Theorem 3.1 *Assume that Assumptions 1-4 hold. Let* $\gamma_m = \lambda' H^{-1} f'(I - P_m) f H^{-1} \lambda / N$. *Assume that there exists a non-stochastic 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* $C(a) = (1 + a)^{-2\alpha} + o\left(|a|^{2\alpha}\right)$. *Then,*

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

Moreover, let W_N *be the weights with* $w_m = N/(N - m)$, $w_N = -m/(N - m)$ *and* $w_j = 0$ *for* $j \neq m$ *where* m *is chosen to minimize* $S_\lambda(W)$. *Then,*

$$\frac{\min_{W \in \Omega_B} S_\lambda(W)}{S_\lambda(W_N)} < 1 \text{ wpa1.}$$

Remark 1 *The additional conditions on* γ_m *imposed in Theorem 3.1 are satisfied if* $\gamma_m = \delta m^{-2\alpha}$, *but are also satisfied for more general specifications. For example, if* $\gamma_m = \delta(m) m^{-2\alpha} + o_p\left(m^{-2\alpha}\right)$ *as* $m \to \infty$, *where the function* $\delta(m)$ *satisfies* $\delta(m(1 + a))/\delta(m) = 1 + o\left(|a|^{2\alpha}\right)$ *wpa1, then the condition holds.*

The first part of the theorem indicates that all MA2SLS estimators considered here dominate the simple number of instrument selection procedure in terms of higher order MSE. Likewise, the second part implies that the MA2SLS estimators obtained from choosing W over the sets Ω_U and Ω_B dominate the Nagar estimator in terms of higher order MSE.

We contrast the optimality properties of MA2SLS with kernel weighted GMM. For illustration, consider the model weights $w_m = 1/M$ for $m \leq M$ and $w_m = 0$ otherwise, which correspond to the kernel weighted GMM with kernel function $k(x) = \sqrt{\max(1-x, 0)}$ (see (2.4)). Because the weights are always between 0 and 1, the MSE is given in (7.4). As a function of the kernel bandwidth M, the MSE approximation is

$$S_\lambda(M) = \sigma^2_{\lambda\epsilon}\frac{(M+1)^2}{4N} + \sigma^2_\epsilon \frac{\mathbf{1}'_M U \mathbf{1}_M}{M^2 N}.$$

The form of S_λ in this case illustrates the fact that kernel weighting generally reduces the higher order bias of 2SLS, here in this case by a factor 1/2, but that this comes at the cost of increased higher order variance. It is easily seen that $\mathbf{1}'_M U \mathbf{1}_M \geq M^2 u^{M\prime}_\lambda u^M_\lambda$. Since the difference between $\mathbf{1}'_M U \mathbf{1}_M$ and $M^2 u^{M\prime}_\lambda u^M_\lambda$ is data-dependent, it can not be established in general that kernel weighting reduces the MSE. This example illustrates that kernels do not have enough free parameters to guarantee that bias reduction sufficiently off-sets the increase in $W'UW$.

4 Implementation

Fully data dependent implementation of the estimator classes defined in Section 3.2 requires a data-dependent criterion $\hat{S}_\lambda(W)$. The non-trivial part of estimating the criterion concerns $f'(I - P(W))(I - P(W))f/N$. Donald and Newey (2001) show that the Mallows criterion can be used to estimate the term $f'(I-P_m)f/N$. This approach fits naturally in our framework of model averaging for the first stage. Hansen (2007) proposes to use the Mallows criterion $\tilde{u}'\tilde{u}/N + 2\sigma^2_\lambda K'W/N$ where $\tilde{u} = (I - P(W))XH^{-1}\lambda$ to choose the weights W for the first stage regression. The use of Mallows criterion is motivated by the fact that

$$E\left[\tilde{u}'\tilde{u}/N|z\right] = \lambda' H^{-1} f'(I - P(W))(I - P(W))fH^{-1}\lambda/N + \sigma^2_\lambda\left(W'\Gamma W - 2K'W\right)/N + \sigma^2_\lambda$$

such that $E\left[\tilde{u}'\tilde{u}/N + 2\sigma^2_\lambda K'W/N|z\right] = E[\|(f - P(W)X)H^{-1}\lambda\|^2|z]/N + \sigma^2_\lambda$. We note that in the context of instrument selection the relevant criterion is $E\left[\|(I - P(W))fH^{-1}\lambda\|^2|z\right]$ rather than $E\left[\|(f-P(W)X)H^{-1}\lambda\|^2|z\right]$ such that the criterion needs to be adjusted to $\tilde{u}'\tilde{u}/N + \sigma^2_\lambda(2K'W/N - W'\Gamma W/N)$. We also note that when $W \in \Omega_{DN}$ it holds that $W'\Gamma W = K'W = m$. Therefore, the correctly adjusted Mallows criterion in this special case is $\tilde{u}'_m\tilde{u}_m/N + \sigma^2_\lambda m/N$ which leads to the formulation used in Donald and Newey (2001, p. 1165).

We propose a slightly different criterion which is based on the difference between the residuals

$$\hat{u}_\lambda = (P_M - P(W))XH^{-1}\lambda$$

where M is a sequence increasing with N that is chosen by the statistician. In practice, M is the largest number of instruments considered for estimation. This number often is directly implied by the available data-set or determined by considerations of computational and practical feasibility. Note that $P_M \to I$, as $M \to N$, which leads to the conventional Mallows criterion. Including P_M rather than I serves two purposes. On the one hand it reduces the bias of the criterion function when W puts most weight on large models. This can be seen by considering the criterion bias where

$$E\left[\left\|(P_M - P(W))uH^{-1}\lambda\right\|^2 |z\right] = \sigma_\lambda^2\left(M - 2K'W + W'\Gamma W\right).$$

When $W \in \Omega_{DN}$, it follows that $K'W = m$ and $\sigma_\lambda^2\left(M - 2K'W + W'\Gamma W\right) = \sigma_\lambda^2\left(M - m\right)$ which tends to zero as m reaches the upper bound M. Similarly, as our theoretical analysis shows, the variability of the criterion function can be reduced by using the criterion based on P_M.

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 weight matrix W. We use the 2SLS estimator with the number of instruments selected by the first stage Mallows criterion in simulations. Let $\hat{H}$ be some estimator of H. Let $\tilde{u}$ be some preliminary residual vector of the first stage regression. Let $\tilde{u}_\lambda = \tilde{u}\hat{H}^{-1}\lambda$.[7] Define,

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

Let $\hat{u}_\lambda^m = (P_M - P_m)X\hat{H}^{-1}\lambda$ and $\hat{U} = (\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)'(\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)$. The criterion $\hat{S}_\lambda(W)$ for choosing the weights is

$$\hat{S}_\lambda(W) = \left(\hat{a}_\lambda \frac{(K'W)^2}{N} + \hat{b}_\lambda \frac{(W'\Gamma W)}{N} - \frac{K'W}{N}\hat{B}_{\lambda,N} + \hat{\sigma}_\epsilon^2\left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2\left(M - 2K'W + W'\Gamma W\right)}{N}\right)\right) \tag{4.1}$$

with $\hat{a}_\lambda = \lambda'\hat{H}^{-1}\hat{a}_\sigma\hat{H}^{-1}\lambda$, $\hat{b}_\lambda = \lambda'\hat{H}^{-1}\hat{b}_\sigma\hat{H}^{-1}\lambda$ and $\hat{B}_{\lambda,N} = \lambda'\hat{H}^{-1}\hat{B}_N\hat{H}^{-1}\lambda$. When the weights are only allowed to be positive, Corollary 7.2 suggests the simpler criterion

$$\hat{S}_\lambda(W) = \left(\hat{a}_\lambda \frac{(K'W)^2}{N} + \hat{\sigma}_\epsilon^2\left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2\left(M - 2K'W + W'\Gamma W\right)}{N}\right)\right). \tag{4.2}$$

In order 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.

[7] Note that $\tilde{u}$ is the residual vector. On the other hand, $\hat{u}_\lambda^m$s are the vectors of the differences of the residuals.

Assumption 5 *For some* α, $\sup_{m\leq M} m^{2\alpha+1}\left(\sup_{\lambda'\lambda=1}\lambda' f\left(P_m - P_{m+1}\right) f\lambda/N\right) = O_p\left(1\right)$ *and*

$$\inf_{m\in J_M} m^{2\alpha+1}\left(\sup_{\lambda'\lambda=1}\lambda' f\left(P_m - P_{m+1}\right) f\lambda/N\right) > 0 \text{ wpa1}$$

where J_M *is some set of positive integers smaller than* M *such that the cardinality* $\#J_M$ *of* J_M *is* $O\left(M\right)$.

Assumption 6 $\hat{H} - H = o_p\left(1\right)$, $\hat{a}_\sigma - a_\sigma = o_p\left(1\right)$, $\hat{b}_\sigma - b_\sigma = o_p\left(1\right)$ *and* $\hat{B}_N - B_N = o_p\left(1\right)$.

Assumption 7 *Let* α *be as defined in Assumption 5. For some* $0 < \varepsilon < \min(1/\left(2\alpha\right), 1)$, *and* δ *such that* $2\alpha\varepsilon > \delta > 0$, *it holds that* $M = O\left(N^{(1+\delta)/(2\alpha+1)}\right)$. *For some* $\vartheta > \left(1+\delta\right)/\left(1-2\alpha\varepsilon\right)$, *it holds that* $E\left(\left|u_i\right|^{2\vartheta}\right) < \infty$. *Further assume that* $\hat{\sigma}^2_\lambda - \sigma^2_\lambda = o_p\left(N^{-\delta/(2\alpha+1)}\right)$.

Remark 2 *The second part of Assumption 5 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.

The following result generalizes a result established by Li (1987) to the case of the MA2SLS estimator.

Theorem 4.1 *Let Assumptions 1-7 hold. For* $\Omega = \Omega_U$, Ω_B, Ω_C, *or* Ω_P *and* $\hat{W} = \arg\min_{W\in\Omega}\hat{S}_\lambda\left(W\right)$ *it follows that*

$$\frac{\hat{S}_\lambda\left(\hat{W}\right)}{\inf_{W\in\Omega} S_\lambda\left(W\right)} \to_p 1. \tag{4.3}$$

Theorem 4.1 complements the result in Hansen (2007). Apart from the fact that $\hat{S}_\lambda\left(W\right)$ is different from the criterion in Hansen (2007), there are more technical differences between our result and Hansen's (2007). Hansen (2007) shows (4.3) 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

This section reports the results of our Monte Carlo experiments,[8] where we investigate the finite sample properties of the MA2SLS estimators. In particular, we examine the performance of the MA2SLS estimators compared with Donald and Newey's (2001) instrument selection procedure, possible gains from considering additional higher order terms in the asymptotic MSE, and potential benefits we obtain by allowing negative weights.

[8] This Monte Carlo simulation was conducted with Ox 4.04 (Doornik (2006)) for Windows.

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:

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

for $i = 1, \ldots, N$, where Y_i is a scalar, β is the scalar parameter of interest, $Z_i \sim$iid.$N(0, I_M)$ and (ϵ_i, u_i) is iid. 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 we examine how well each procedure estimates β.

In this framework, each experiment is indexed by the vector of specifications: $(N, M, c, \{\pi\})$, where N represents the sample size. 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 following three specifications for π.

$$\text{Model (a): } \pi_m = \sqrt{\frac{R_f^2}{\bar{K}(1 - R_f^2)}}, \quad \forall m.$$

This design is considered by Hahn and Hausman (2002) and Donald and Newey (2001). In this model, all the instruments are equally weak.

$$\text{Model (b): } \pi_m = c(M) \left(1 - \frac{m}{M+1}\right)^4, \quad \forall m.$$

This design is considered by Donald and Newey (2001). The strength of the instruments decreases gradually in this specification.

$$\text{Model (c):} \pi_m = 0 \text{ for } m \leq M/2; \ \pi_m = c(M) \left(1 - \frac{m - M/2}{M/2 + 1}\right)^4 \text{ for } m > M/2,$$

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 potential benefits of allowing for negative weights which makes the procedure more robust with respect to the ordering of 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 1000.

We compare the performances of the following seven estimators. Three of them are existing procedures and the other four procedures are the MA2SLS estimators developed in this paper. First,

we consider the 2SLS estimator with all available instruments (2SLS in the tables). Second, the 2SLS estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (DN). We use the criterion function (4.2) for DN. The optimal number of instruments is obtained by a grid search. The kernel weighted GMM of Kuersteiner (2002) is also examined (KW). Let $\Omega_{KW} = \{W \in l_1 : w_m = L^{-1} \text{ if } m \leq L \text{ and 0 otherwise for some } L \leq M\}$. Then, the MA2SLS estimator with $W \in \Omega_{KW}$ corresponds to the kernel weighted 2SLS estimator with kernel $k(x) = \sqrt{\max(1-x, 0)}$. Because the weights are always positive with Ω_{KW}, we use the criterion function (4.2) for KW. We use a grid search to find the L that minimizes the criterion. The procedure "MA-U" is the MA2SLS estimator with $\Omega = \Omega_U = \{W \in l_1 : W'\mathbf{1}_M = 1\}$. The weights for MA-U are computed using the estimated version of formula (3.2) where the matrix U in (3.2) is estimated by the modified Mallows criterion in Section 4 so that it is equivalent to minimizing (4.1). The MA2SLS estimator with $\Omega = \Omega_C = \{W \in l_1 : W'\mathbf{1_M} = 1; w_m \in [-1, 1], \forall m \leq M\}$ is denoted "MA-C". We minimize the criterion (4.1) to obtain optimal weights. The procedure "MA-P" uses the set $\Omega = \Omega_P = \{W \in l_1 : W'\mathbf{1_M} = 1; w_m \in [0, 1], \forall m \leq M\}$. The criterion for MA-P is formula (4.1). The procedure "MA-Ps" also uses the same set Ω_P, but the criterion for computing weights is (4.2). For MA-C, MA-P and MA-Ps, we use the procedure SolveQP in Ox to minimize the criterion (see Doornik (2006)). We use the 2SLS estimator 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 bias ("bias" in the tables), the inter-quantile range ("IQR"), the median absolute deviation ("MAD") and the median absolute deviation relative to that of DN ("RMAD").[9] We also compute the following two measures. The measure "KW+" is the value of $\sum_{m=1}^{M} m \max(w_m, 0)$. For 2SLS, this measure is merely the total number of instruments. For DN, it is the number of instruments chosen by the procedure. The measure "KW-" is the value of $\sum_{m=1}^{M} m|\min(w_m, 0)|$. This measure is zero for the procedures that allow only positive weights. For MA-U or MA-C, it may not be zero because of possibly negative weights. A comparison of KW+ and KW- offers some insight into the importance of bias reduction and instrument selection for the MA-U and MA-C procedures.

[9] We use these robust measures because of concerns about the existence of moments of estimators.

5.2 Results

Tables 1-6 summarize the results of our simulation experiment. The 2SLS estimator (with all available instruments) performs well when the degree of endogeneity is small ($c = 0.1$). However, when $c = 0.5$ or 0.9, 2SLS exhibits large bias and some method to alleviate this problem is called for. The selection method of DN achieves this goal only partially. In Model (b) with $c = 0.5$ and $c = 0.9$, where the rank ordering of instruments is appropriate and bias reduction is an important issue, it reduces the bias of the estimator by using a small number of instruments. However, DN tends to use too small a number of instruments and the improvement of the performance does not occur in general. Even in Model (b), DN uses too small a number of instruments when $c = 0.1$ and thus unnecessarily inflates the variability of the estimator. In Models (a) and (c), DN seldom outperforms 2SLS. In particular, in Model (c), the number of instruments selected by DN tends to be far less than $M/2$, which means that DN often employs only the instruments that are uncorrelated with the endogenous regressor. KW typically outperforms DN, which demonstrates the advantage of kernel weighting. However, the problem observed for DN also applies to KW. KW does not improve over 2SLS in Models (a) and (c).

All model averaging estimators perform well. MA-Ps, which may be considered a natural application of Hansen's (2007) model averaging to IV estimation, outperforms DN and KW in most cases. MA-P further improves over MA-Ps in Models (a) and (c) substantially, which shows the benefit of taking additional higher order terms into account when choosing optimal weights. On the other hand, in Model (b), MA-P is outperformed by DN when $c = 0.9$. Nevertheless, the RMAD measure is never above 1.28 which is significantly lower than the RMAD measure for 2SLS. This result may be due partly to a trade-off between additional terms in the approximation of the MSE and a more complicated form of the optimal weights: It provides a more precise approximation of the MSE on a theoretical level; however it also complicates estimation of the criterion $S_\lambda(W)$ which may result in larger estimation errors in the estimated criterion function. MA-U and MA-C also perform well. Their performance is particularly remarkable in Models (a) and (c) where DN tends to choose too few instruments. In particular, in Model (c) with $c = 0.9$, the performance of MA-U stands out. This result demonstrates the value of flexibility with respect to the ordering of instruments that is achieved by allowing negative weights.[10] However, the performance of MA-U or

[10] Results not presented here show that MA-C also works as well as MA-U does in Model (c) when $R_f^2 = 0.2$. These results might imply that we need sufficiently informative data for exploiting the benefit from allowing negative weights. Even for MA-U, such a benefit is more clearly seen in cases with $R_f^2 = 0.1$ than in cases with $R_f^2 = 0.01$.

MA-C in Model (b) is not as good as that of DN when $c = 0.5$ and $c = 0.9$ with a RMAD measure reaching values of around 1.3 and 1.7 respectively. Nevertheless, MA-U and MA-C perform better than 2SLS even in these cases. Their relatively poor performance in Model (b) may be due to having too large a choice set for W. The performance of MA-C is more stable over different designs than that of MA-U. Note also that the values of "KW+" and "KW-" for MA-U and MA-C indicate that they do not try to eliminate the bias completely.[11] The median biases of these estimators are similar to other MA2SLS estimators.

In summary, MA-Ps displays the most robust performance of all procedures considered. It never falls behind DN in terms of RMAD and often outperforms it significantly. The other MA2SLS estimators show some problems in Model (b) when the degree of endogeneity is moderate to high (MA-U) or high (MA-C and MA-P). On the other hand, MA-U and MA-C in particular achieve even more significant improvements in terms of MAD over DN in Models (a) and especially (c) where DN and 2SLS are unable to eliminate irrelevant instruments.

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 and in some cases have closed form expressions. 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.

7 Formal Results and Proofs

Theorem 7.1 *Suppose that Assumptions 1-3 are satisfied. Define* $\mu_i(W) = E[\epsilon_i^2 u_i]P_{ii}(W)$, $\mu(W) = (\mu_1(W), ..., \mu_N(W))'$ *and* $Q(W) = I - P(W)$. *If* W *satisfies Assumption 4 then, for* $\hat{\beta}$, *the decomposition*

[11] The higher order bias is eliminated when $K'W = 0$, which is equivalent to the case where "KW+" and "KW-" are equal.

given by (7.6) holds with

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

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

$$B_N = 2\left(\sigma_\epsilon^2\Sigma_u + d\sigma_{u\epsilon}\sigma_{u\epsilon}' + \frac{1}{N}\sum_{i=1}^N f_i\sigma_{u\epsilon}'H^{-1}\sigma_{u\epsilon}f_i' + \frac{1}{N}\sum_{i=1}^N\left(f_i\sigma_{u\epsilon}'H^{-1}f_i\sigma_{u\epsilon}' + \sigma_{u\epsilon}f_i'H^{-1}\sigma_{u\epsilon}f_i'\right)\right) \tag{7.1}$$

Remark 3 *When* $d = 1$, $B_N = 2(\sigma_\epsilon^2\Sigma_u + 4\sigma_{u\epsilon}^2)$.

Note that the term B_N is positive semi-definite. This implies that usual higher order formula that neglects the term $\frac{K'W}{N}B_N$ overestimates the effect on the bias of including more instruments. A number of special cases lead to simplifications of the above result. If $\mathrm{Cum}[\epsilon_i,\epsilon_i,u_i,u_i'] = 0$ and $E[\epsilon_i^2u_i] = 0$ as would be the case if ϵ_i and u_i were jointly Gaussian, then the following result is obtained:

Corollary 7.1 *Suppose that the same conditions as in Theorem 7.1 hold and that in addition* $Cum[\epsilon_i,\epsilon_i,u_i,u_i'] = 0$ *and* $E[\epsilon_i^2u_i] = 0$. *Then, for* $\hat\beta$, *the decomposition given by (7.6) holds with:*

$$S(W) = H^{-1}\left(\sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(K'W)^2}{N} + (\sigma_\epsilon^2\Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{(W'\Gamma W)}{N} - \frac{K'W}{N}B_N + \sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\right)H^{-1} \tag{7.2}$$

where B_N *is as defined before.*

Another interesting case arises when W is constrained such that $w_m \in [0,1]$. We have the following result.

Corollary 7.2 *Suppose that the same conditions as in Theorem 7.1 hold and that in addition* $w_m \in [0,1]$ *for all* m. *Then, for* $\hat\beta$, *the decomposition given by (7.6) holds with:*

$$\begin{aligned} S(W) &= H^{-1}\left(\sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(K'W)^2}{N} + (\sigma_\epsilon^2\Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{(W'\Gamma W)}{N} - \frac{K'W}{N}B_N\right. \\ &\quad \left. + E[\epsilon_1^2u_1']\frac{\sum_i f_i P_{ii}(W)}{N} + \frac{\sum_i f_i' P_{ii}(W)}{N}E[\epsilon_1^2u_1] + \sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\right)H^{-1} \end{aligned} \tag{7.3}$$

where B_n *is as defined before. Moreover, ignoring terms of order* $O_p(K'W)$ $(= o_p((K'W)^2))$, *to first order*

$$S(W) = H^{-1}\left(\sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(K'W)^2}{N} + \sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\right)H^{-1}. \tag{7.4}$$

A last special case arises when the constraint $K'W = 0$ is imposed on the weights. This constraint requires that w_m can be positive and negative. The expansion to higher orders than Donald and Newey is necessary

to capture the relevant trade-off between more efficiency and distortions due to additional instruments. For simplicity we also assume that $\text{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ and $E[\epsilon_i^2 u_i] = 0$. Without these additional constraints the terms involving $\sum_i (P_{ii}(W))^2/N$, $\sum_i f_i P_{ii}(W)/N$ and $f'Q(W)\mu(W)/N$ potentially matter and need to be included.

Corollary 7.3 *Suppose that the same conditions as in Theorem 7.1 hold and that in addition* $Cum[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ *and* $E[\epsilon_i^2 u_i] = 0$. *Furthermore, impose* $K'W = 0$. *Then, for* $\hat{\beta}$, *the decomposition given by (7.6) holds with*

$$S(W) = H^{-1}\left((\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{(W'\Gamma W)}{N} + \sigma_\epsilon^2 \frac{f'(I - P(W))(I - P(W))f}{N}\right) H^{-1}. \tag{7.5}$$

Remark 4 *We note that this result covers the Nagar estimator where* $M = N$, $w_m = N/(N-k)$ *for* $k = m$, $w_N = -k/(N-k)$ *and* $w_m = 0$ *otherwise for some* k *such that* $k \to \infty$ *and* $k/\sqrt{N} \to 0$. *First, we verify that all the conditions of the Corollary are satisfied where* $\sup_{k \leq M} \left|\sum_{m=k}^M w_m\right| = k/(N-k) \to 0$, $K'W = 1$, $\mathbf{1}_M'W = 1$, $\sum_{m=1}^M |w_m| m = 2Nk/(N-k) \to \infty$, $\sum_{m=1}^M |w_m| m/\sqrt{N} = 2\sqrt{N}k/(N-k) \to 0$. *Further,* $\sum_{m=1}^M \left(\sum_{s=1}^m w_m\right)^2 m^{-2\alpha} = k^{-2\alpha}/(1-k/N)^2 \to 0$. *Next, note that* $W'\Gamma W = k/(1-k/N)^2 - k^2/N(1-k/N)^2$ *and* $f'(I - P(W))(I - P(W))f = f'(I - P_k)f/(1-k/N)^2$ *noting that* $P_N = I$. *If we use* W_N *to denote the Nagar weights, then* $S(W_N) = H^{-1}\left((\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')k/N + \sigma_\epsilon^2 f'(I - P_k)f/N\right) H^{-1} + o(S(W_N))$. *The lead term is the same as the result in Proposition 3 of Donald and Newey (2001).*

7.1 Lemmas

The estimator examined has the form of $\sqrt{N}(\hat{\beta} - \beta) = \hat{H}^{-1}\hat{h}$. We define $h = f'\epsilon/\sqrt{N}$ and $H = f'f/N$. The following lemma is the key device to compute the Nagar-type MSE. This lemma is similar to Lemma A.1 in Donald and Newey (2001), but with the important difference that the expansion is valid to higher order and covers the case of higher order unbiased estimators.

Lemma 7.1 *If there is a decomposition* $\hat{h} = h + T^h + Z^h$, $\tilde{h} = h + T^h$, $\hat{H} = H + T^H + Z^H$,

$$\tilde{h}\tilde{h}' - \tilde{h}\tilde{h}'H^{-1}T^{H\prime} - T^H H^{-1}\tilde{h}\tilde{h}' = \hat{A}(W) + Z^A(W),$$

such that $T^h = o_p(1)$, $h = O_p(1)$, $H = O_p(1)$, *the determinant of* H *is bounded away from zero with probability 1,* $\rho_{W,N} = tr(S(W))$ *and* $\rho_{W,N} = o_p(1)$,

$$\begin{aligned} ||T^H||^2 &= o_p(\rho_{W,N}), \quad ||Z^h|| = o_p(\rho_{W,N}), \quad ||Z^H|| = o_p(\rho_{W,N}), \\ Z^A(W) &= o_p(\rho_{W,N}), \quad E[\hat{A}(W)|z] = \sigma^2 H + HS(W)H + o_p(\rho_{W,N}), \end{aligned}$$

then

$$\begin{aligned} N(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)' &= \hat{Q}(W) + \hat{r}(W), \\ E[\hat{Q}(W)|z] &= \sigma_\epsilon^2 H^{-1} + S(W) + T(W), \\ (\hat{r}(W) + T(W))/tr(S(W)) &= o_p(1), \ \text{as } K'W^+ \to \infty, N \to \infty. \end{aligned} \tag{7.6}$$

Remark 5 *The technical difference between our lemma and that of Donald and Newey is that we consider the interaction between T^h and T^H in the expansion and we do not require that $||T^h|| \cdot ||T^H||$ is small.*

Proof. The proof follows steps taken by Donald and Newey (2001). We observe that

$$\hat{H}^{-1}\hat{h} = H^{-1}\hat{h} - H^{-1}(\hat{H} - H)H^{-1}\hat{h} + H^{-1}(\hat{H} - H)H^{-1}(\hat{H} - H)\hat{H}\hat{h}.$$

Noting that $\hat{H} - H = T^H + Z^H$, $||T^H||^2 = o_p(\rho_{W,N})$, $||Z^H|| = o_p(\rho_{W,N})$ and $\hat{h} = \tilde{h} + Z^h = \tilde{h} + o_p(\rho_{W,N})$, we have

$$\hat{H}^{-1}\hat{h} = H^{-1}\tilde{h} - H^{-1}T^H H^{-1}\tilde{h} + o_p(\rho_{W,N}).$$

Let $\tilde{\tau} = \tilde{h} - T^H H^{-1}\tilde{h}$. Then,

$$\tilde{\tau}\tilde{\tau}' = \hat{A}(W) + Z^A(W) + T^H H^{-1}\tilde{h}\tilde{h}' H^{-1}T^H = \hat{A}(W) + o_p(\rho_{W,N}),$$

by $Z^A(W) = o_p(\rho_{W,N})$ and $||T^H|| = o_p(\rho_{W,N})$. It follows that

$$N(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = H^{-1}(\hat{A}(W) + o_p(\rho_{W,N}))H^{-1} + o_p(\rho_{W,N}) = H^{-1}\hat{A}(W)H^{-1} + o_p(\rho_{W,N}).$$

Therefore, we get the desired result. ■

Lemma 7.2 *Let Γ be the $N \times N$ matrix where $\Gamma_{ij} = \min(i, j)$. Then Γ is positive definite.*

Proof. Define the vectors $b_{j,N} = (\mathbf{0}_j', \mathbf{1}_{N-j}')'$ where $\mathbf{1}_j$ is $j \times 1$ vector of $1's$ and $\mathbf{0}_j$ is defined similarly. Then

$$\Gamma = \sum_{j=0}^{N-1} b_{j,N} b_{j,N}',$$

and for any $y \in \mathbb{R}^N$ it follows that $y'\Gamma y = \sum_{j=0}^{N-1} (y' b_{j,N})^2 \geq 0$ and the equality holds if and only if $y = 0$. This shows that Γ is positive definite. ■

Lemma 7.3 *Let Γ be defined as in Lemma 7.2. If $\sum_{m=1}^{M} \left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha} \to 0$ as $M \to \infty$ and $W'\mathbf{1}_M = 1$ for any M, then it follows that $W'\Gamma W \to \infty$ as $M \to \infty$.*

Proof. Choose any $\varepsilon > 0$, $\bar{M} > 0$ and let $\eta = \varepsilon \bar{M}^{-2\alpha}$. Let $M > \bar{M}$ be large enough such that $\sum_{m=1}^{M} (\sum_{s=1}^{m} w_s)^2 m^{-2\alpha} < \eta$. Note that

$$\sum_{m=1}^{\bar{M}} \left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha} \geq \bar{M}^{-2\alpha} \sum_{m=1}^{\bar{M}} \left(\sum_{s=1}^{m} w_s\right)^2 \geq \bar{M}^{-2\alpha} \left(\sum_{s=1}^{\bar{M}} w_s\right)^2$$

which implies that $\left(\sum_{s=1}^{\bar{M}} w_{s,M}\right)^2 < \eta \bar{M}^{2\alpha} = \varepsilon$. This implies that, for any $\bar{M}_1 < \bar{M}$, $1 = \left|\sum_{m=1}^{M} w_m\right| \leq \left|\sum_{m=\bar{M}_1+1}^{M} w_m\right| + \left|\sum_{m=1}^{\bar{M}_1} w_m\right|$ such that $\left|\sum_{m=\bar{M}_1}^{M} w_m\right| \geq 1 - \sqrt{\varepsilon}$. Now,

$$W'\Gamma W = \sum_{j=0}^{M-1} \left(\sum_{m=j+1}^{M} w_m\right)^2 \geq \sum_{j=0}^{\bar{M}} \left(\sum_{m=j+1}^{M} w_m\right)^2 \geq \bar{M}\left(1 - \sqrt{\varepsilon}\right)^2.$$

Since $\bar{M}$ was arbitrary, the result follows. ■

Lemma 7.4 *Suppose that Assumptions 1-3 are satisfied. Then we have*

1. $tr(P(W)) = \sum_{m=1}^{M} w_m m = K'W$ *(Hansen (2007) Lemma 1.1),*
2. $\sum_i (P_{ii}(W))^2 = o_p(K'W^+)$,
3. $\sum_{i \neq j} P_{ii}(W) P_{jj}(W) = (K'W)^2 + o_p(K'W^+)$,
4. $\sum_{i \neq j} P_{ij}(W) P_{ij}(W) = \sum_{m=1}^{M} \sum_{l=1}^{M} w_m w_l \min(l,m) + o_p(K'W) = W'\Gamma W + o_p(K'W)$,
5. $\sum_{i \neq j} P_{ij}(W) = O_p(N - K'W)$,
6. $h = f'\epsilon/\sqrt{N} = O_p(1)$ *and* $H = f'f/N = O_p(1)$ *(Donald and Newey (2001) Lemma A.2 (v)).*

Proof. We do not provide the proofs of parts 1 and 6, as the proofs are available in Hansen (2007) and Donald and Newey (2001). For part 2, first we note that $A_{ii} \leq B_{ii}$ if $A \leq B$, which implies that $P_{l,ii} \leq P_{M,ii}$ for $l \leq M$. Then, Assumption 3 and Lemma 7.4(1) imply

$$\begin{aligned}
\sum_i (P_{ii}(W))^2 &= \sum_{i=1}^{N} \sum_{m,l=1}^{M} w_m w_l P_{l,ii} P_{m,ii} \leq \sum_{i=1}^{N} \sum_{m,l=1}^{M} |w_m| \, |w_l| \, P_{l,ii} P_{m,ii} \\
&\leq \max_i (P_{M,ii}) \left(\sum_{m=1}^{M} |w_l| \right) \sum_{i=1}^{N} \sum_{m=1}^{M} |w_m| \, P_{m,ii} \leq C \max_i (P_{M,ii}) \mathrm{tr} P\left(W^+\right) \\
&= o_p(1)(K'W^+) = o_p(K'W^+)
\end{aligned}$$

where $\sum_{m=1}^{M} |w_l| \leq C$ for some $C < \infty$ was used. Also these results imply

$$\sum_{i \neq j} P_{ii}(W) P_{jj}(W) = \sum_i P_{ii}(W) \sum_j P_{jj}(W) - \sum_i (P_{ii}(W))^2 = (K'W)^2 + o_p(K'W^+),$$

which shows part 3.

To show 4, first we observe that

$$\sum_{i \neq j} P_{ij}(W) P_{ij}(W) = \mathrm{tr}(P(W)P(W)) - \sum_i (P_{ii}(W))^2.$$

Now $\mathrm{tr}(P(W)P(W)) = \sum_{m=1}^{M} \sum_{l=1}^{M} w_m w_l \min(l,m)$ by Lemma 1.2 of Hansen (2007). Thus, combining this result with part 2 of this lemma,

$$\sum_{i \neq j} P_{ij}(W) P_{ij}(W) = \sum_{m=1}^{M} \sum_{l=1}^{M} w_m w_l \min(l,m) + o_p(K'W^+).$$

For 5, note that

$$\sum_{i \neq j} P_{ij}(W) = \mathbf{1}_N' P(W) \mathbf{1}_N - \mathrm{tr}(P(W))$$

where $\mathbf{1}_N' P_m \mathbf{1}_N \leq \mathbf{1}_N' \mathbf{1}_N = N$ by the fact that P_m is an idempotent matrix. Then note that

$$\begin{aligned}
\mathbf{1}_N' P(W) \mathbf{1}_N - \mathrm{tr}(P(W)) &= |\mathbf{1}_N' P(W) \mathbf{1}_N| - \mathrm{tr}(P(W)) \leq \sum_{m=1}^{M} |w_m| |\mathbf{1}_N' P_m \mathbf{1}_N| - \mathrm{tr}(P(W)) \\
&\leq CN - K'W
\end{aligned}$$

such that $\sum_{i\neq j} P_{ij}(W) = O_p\left(N - K'W\right) = O_p\left(N\right)$. ■

Let $e_f(W) = f'(I - P(W))(I - P(W))f/N$ and $\Delta(W) = \text{tr}(e_f(W))$.

Lemma 7.5 *Suppose that Assumptions 1-3 are satisfied. If* $\mathbf{1}_M'W = 1$, $\sup_{k\leq M}\left|\sum_{m=k}^{M} w_m\right| \leq C < \infty$ *uniformly in* M, $K'W^+ =\to \infty$, $K'W^+/\sqrt{N} \to 0$, *and* $\sum_{m=1}^{M}\left(\sum_{s=1}^{m} w_m\right)^2 m^{-2\alpha} \to 0$, *then*

1. $\Delta(W) = o_p(1)$,
2. $f'(I - P(W))\epsilon/\sqrt{N} = O(\Delta(W)^{1/2})$,
3. $E[u'P(W)\epsilon|z] = \sigma_{u\epsilon}K'W$,
4. $E[u'P(W)\epsilon\epsilon'P(W)u|z] = \sigma_{u\epsilon}\sigma_{u\epsilon}'(K'W)^2 + (\sigma_\epsilon^2\Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')(W'\Gamma W) + Cum[\epsilon_i, \epsilon_i, u_i, u_i']\sum_i (P_{ii}(W))^2$,
5. $E[f'\epsilon\epsilon'P(W)u|z] = \sum_i f_i P_{ii}(W)E[\epsilon_i^2 u_i'] = O_p(K'W)$,
6. *Let* $g(W) > 0$ *be a function of* W *such that* $g(W) \to \infty$ *as* $N \to \infty$. *Then* $\sqrt{g(W)\Delta(W)}/\sqrt{N} = O_p(g(W)/N + \Delta(W))$,
7. $E[hh'H^{-1}u'f/N|z] = \sum_i f_i f_i' H^{-1}E[\epsilon_i^2 u_i]f_i'/N^2 = O_p(1/N)$ *(Donald and Newey (2001) Lemma A.3 (vii))*,
8. $E[f'\left(I - P(W)\right)\epsilon\epsilon'P(W)u/N|z] = f'Q\left(W\right)\mu\left(W\right)/N = o_p\left((K'W^+)/N + \Delta(W)\right)$,
9. $E[f'\epsilon\epsilon'fH^{-1}u'P(W)u|z]/N^2 = O_p(1/N) + \sigma_\epsilon^2\Sigma_u K'W/N$,
10. $E[f'\epsilon\epsilon'P(W)uH^{-1}\left(u'f + f'u\right)|z]/N^2 = O_p(1/N) + (K'W/N)(\sum_i f_i\sigma_{u\epsilon}'H^{-1}\sigma_{u\epsilon}f_i/N + \sum_i f_i\sigma_{u\epsilon}'H^{-1}f_i\sigma_{u\epsilon}'/N)$,
11. $E\left[u'P(W)\epsilon\epsilon'fH^{-1}\left(u'f + f'u\right)|z\right]/N^2 = O_p\left(1/N\right) + \left(K'W/N\right)\left(d\sigma_{u\epsilon}\sigma_{u\epsilon}' + \sigma_{u\epsilon}\sum_i f_i'H^{-1}\sigma_{u\epsilon}f_i'/N\right)$,
12. $W'\Gamma W \leq CK'W^+$ *for some constant* C.

Proof. Let $\tilde{\gamma}_m = \text{tr}(f'(I - P_m)f)/N$. By construction $\tilde{\gamma}_m \geq 0$. Write

$$\text{tr}(f'(I - P(W))(I - P(W))f)/N = W'AW$$

where

$$A = \begin{pmatrix} \tilde{\gamma}_1 & \tilde{\gamma}_2 & \cdots \\ \tilde{\gamma}_2 & \tilde{\gamma}_2 & \\ \vdots & & \ddots \end{pmatrix}.$$

It follows that

$$W'AW = \left(\sum_{m=1}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2\left(\tilde{\gamma}_m - \gamma_{m+1}\right)\right) + \tilde{\gamma}_M \tag{7.7}$$

such that

$$W'AW \leq \sum_{m=1}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 \tilde{\gamma}_m = \sum_{m=1}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 \frac{\tilde{\gamma}_m}{m^{-2\alpha}} m^{-2\alpha} \leq \sup_{m\leq M}\left(m^{2\alpha}\tilde{\gamma}_m\right)\sum_{m=1}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha},$$

where $\sup_{m\leq M}\left(m^{2\alpha}\tilde{\gamma}_m\right) = O_p\left(1\right)$ by Assumption 2(ii) and $\sum_{m=1}^{M}\left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha} \to 0$ by the conditions of the Lemma. This implies that $\text{tr}(f'(I - P(W))(I - P(W))f)/N = \Delta(W) = o_p(1)$.

Next we observe that $E[f'(I-P(W))\epsilon/\sqrt{N}]=0$ and

$$E\left[\frac{f'(I-P(W))\epsilon}{\sqrt{N}}\frac{\epsilon'(I-P(W))f}{\sqrt{N}}\Big|z\right]=\sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}=\sigma_\epsilon^2 e_f(W).$$

Therefore $f'(I-P(W))\epsilon/\sqrt{N}=O_p(\Delta(W)^{1/2})$ by the Chebyshev inequality. This shows part 2.

For part 3,

$$E[u'P(W)\epsilon|z]=\sum_{i=1}^{N}P_{ii}(W)E[u_i\varepsilon_i]=\sigma_{u\epsilon}\mathrm{tr}(P(W))=\sigma_{u\epsilon}K'W.$$

To give 4, observe that $E[u_iP_{ij}(W)\epsilon_j\epsilon_kP_{kl}(W)u_l']=0$ if one of (i,j,k,l) is different from all the rest. Also $E[\epsilon_i^2u_iu_i']$ is bounded by Assumption 1. Therefore we have

$$\begin{aligned}
&E[u'P(W)\epsilon\epsilon'P(W)u|z]\\
=\;&\sum_i(P_{ii}(W))^2E[\epsilon_i^2u_iu_i']+\sum_{i\neq j}E[u_iP_{ii}(W)\epsilon_i\epsilon_jP_{jj}(W)u_j'|z]\\
&+\sum_{i\neq j}E[u_iP_{ij}(W)\epsilon_j\epsilon_iP_{ij}(W)u_j'|z]+\sum_{i\neq j}E[u_iP_{ij}(W)\epsilon_j^2P_{ji}(W)u_i'|z]\\
=\;&E[\epsilon_i^2u_iu_i']\sum_i(P_{ii}(W))^2+\sigma_{u\epsilon}\sigma_{u\epsilon}'\sum_{i\neq j}P_{ii}(W)P_{jj}(W)+(\sigma_\epsilon\Sigma_u+\sigma_{u\epsilon}\sigma_{u\epsilon}')\sum_{i\neq j}P_{ij}(W)P_{ij}(W)\\
=\;&\mathrm{Cum}[\epsilon_i,\epsilon_i,u_i,u_i']\sum_i(P_{ii}(W))^2+\sigma_{u\epsilon}\sigma_{u\epsilon}'(K'W)^2+(\sigma_\epsilon^2\Sigma_u+\sigma_{u\epsilon}\sigma_{u\epsilon}')(W'\Gamma W),
\end{aligned}$$

by Lemma 2.2, 2.3 and 2.4 and noting that $\mathrm{Cum}[\epsilon_i,\epsilon_i,u_i,u_i']=E[\epsilon_i^2u_iu_i']-\sigma_\epsilon^2\Sigma_u-2\sigma_{u\epsilon}\sigma_{u\epsilon}'$.

Assumption 1 also implies

$$E[f'\epsilon\epsilon'P(W)u|z]=\sum_{i,j,k}f_iP_{jk}(W)E[\epsilon_i\epsilon_ju_k']=\sum_i f_iP_{ii}(W)E[\epsilon_i^2u_i'].$$

and furthermore together with Assumption 3 and Lemma 2.1,

$$\left|\sum_i f_iP_{ii}(W)E[\epsilon_i^2u_i']\right|\leq\sum_i P_{ii}(W)\cdot||f_i||\cdot||E[\epsilon_i^2u_i']||=O_p(K'W),$$

which gives 5.

To prove 6, first we consider the function of a: $g(w)/a+a$, which is convex and the minimum value of which is $2\sqrt{g(W)}$ with the minimizer $a=\sqrt{g(W)}$. If $\Delta(W)=0$, then $\left(\sqrt{\Delta(W)/N}\right)/(g(W)/N+\Delta(W))=0$ and for $\Delta(W)\neq 0$,

$$\frac{\sqrt{\Delta(W)/N}}{g(W)/N+\Delta(W)}=\left(\frac{g(W)}{\sqrt{\Delta(W)N}}+\sqrt{\Delta(W)N}\right)^{-1}\leq\frac{1}{2\sqrt{g(W)}}\to 0 \tag{7.8}$$

as $g(W)\to\infty$.

For part 8, let $Q(W)=I-P(W)$ and for some a and b let $f_{i,a}=f_a(z_i)$ and $\mu_{i,b}(W)=E[\epsilon_i^2u_{ib}]P_{ii}(W)$. Now the a,b th element of $E[f'(I-P(W))\epsilon\epsilon'P(W)u|z]$ satisfies

$$\begin{aligned}
\left|E\left[\sum_{i,j,k,l}f_{i,a}Q_{ij}\epsilon_j\epsilon_kP_{kl}(W)u_{lb}\Big|z\right]\right|&=\left|\sum_{i,j}f_{i,a}Q_{ij}E[\epsilon_j^2u_{jb}]P_{jj}(W)\right|\\
&=|f_a'Q(W)\mu_b(W)|\leq|f_a'QQf_a|^{1/2}|\mu_b'(W)\mu_b(W)|^{1/2},
\end{aligned}$$

where the inequality is the Cauchy-Schwartz inequality. Now $|f_a'QQf_a|^{1/2} = O_p((N\Delta(W))^{1/2})$ by the definition of Δ_s. $|\mu_b'(W)\mu_b(W)| \leq C\sum_i (P_{ii}(W))^2$ for some constant C by Assumption 1 and applying Lemma 2(2) we have $|\mu_b'(W)\mu_b(W)| = o_p(K'W^+)$. Therefore we have

$$\begin{aligned} E\left[f'(I-P(W))\epsilon\epsilon'P(W)u/N|z\right] &= O_p((N\Delta(W))^{1/2})o_p(\sqrt{K'W^+})O_p(1/N) \\ &= o_p(\Delta(W)^{1/2}\sqrt{K'W^+}/\sqrt{N}) = o_p\left((K'W^+)/N + \Delta(W)\right) \end{aligned}$$

where the last equality follows from the fact that

$$\Delta(W)^{1/2}\sqrt{K'W^+}/\sqrt{N} \leq \left((K'W^+)/N + \Delta(W)\right\}/2$$

by (7.8). In addition if we define $\mu_i(W) = E[\epsilon_i^2 u_i]P_{ii}(W)$ and $\mu(W) = \left(\mu_1(W)', ..., \mu_n(W)'\right)'$ then

$$E\left[f'(I-P(W))\epsilon\epsilon'P(W)u/N|z\right] = f'Q(W)\mu(W)/N.$$

For part 9, note that by Lemma 7.4(5) $\sum_{i\neq j} P_{ij}(W)/N = O_p(1 - K'W/N) = O_p(1)$. Then,

$$\begin{aligned} E\left[f'\epsilon\epsilon' fH^{-1}u'P(W)u|Z\right]/N^2 &= \sum_i f_i f_i' H^{-1}E[\epsilon_i^2 u_i u_i']P_{ii}(W)/N^2 \\ &\quad +2\sum_{i\neq j} f_i f_j' H^{-1}E[\epsilon_i u_i]E[\epsilon_j u_j']P_{ij}(W)/N^2 \\ &\quad +\sum_{i\neq j} f_i f_i' H^{-1}E[\epsilon_i^2]E[u_j u_j']P_{jj}(W)/N^2 \\ &= O_p(1/N) + \sigma_\epsilon^2 \Sigma_u K'W/N. \end{aligned}$$

For part 10, using again Lemma 7.4(5) as before,

$$\begin{aligned} & E\left[f'\epsilon\epsilon'P(W)uH^{-1}u'f|z\right]/N^2 \\ = & \sum_i f_i P_{ii}(W)E[\epsilon_i^2 u_i' H^{-1}u_i|z]f_i'/N^2 + \sum_{i\neq j} f_i P_{jj}(W)E\left[\epsilon_j u_j'\right]H^{-1}E[u_i\epsilon_i]f_i'/N^2 \\ & +\sigma_\epsilon^2 \sum_{i\neq j} f_i P_{ij}(W)E\left[u_j' H^{-1}u_j|z\right]f_j'/N^2 + \sigma_\epsilon^2 \sum_{i\neq j} f_j P_{ji}(W)E\left[u_j' H^{-1}u_j\right]f_i'/N^2 \\ = & O_p(1/N) + \sum_{i\neq j} f_i P_{jj}(W)E[\epsilon_j u_j']H^{-1}E[u_i\epsilon_i]f_i'/N^2 = O_p(1/N) + (K'W/N)\sum_i f_i \sigma_{u\epsilon}' H^{-1}\sigma_{u\epsilon}f_i/N \end{aligned}$$

and

$$\begin{aligned} & E[f'\epsilon\epsilon'P(W)uH^{-1}f'u|z]/N^2 \\ = & \sum_i f_i P_{ii}(W)E\left[\epsilon_i^2 u_i' H^{-1}f_i u_i'|z\right]/N^2 + \sum_{i\neq j} f_i P_{jj}(W)E[\epsilon_j u_j']H^{-1}f_i E[u_i'\epsilon_i]/N^2 \\ & +\sigma_\epsilon^2 \sum_{i\neq j} f_i P_{ij}(W)E\left[u_j H^{-1}f_j u_j'|z\right]/N^2 + \sigma_\epsilon^2 \sum_{i\neq j} f_j P_{ji}(W)E\left[u_j H^{-1}f_i u_j'|z\right]/N^2 \end{aligned}$$

$$= O_p(1/N) + \sum_{i\neq j} f_i P_{jj}(W) E[\epsilon_j u_j'] H^{-1} f_i E[u_i'\epsilon_i]/N^2 = O_p(1/N) + (K'W)/N \sum_i f_i \sigma_{u\epsilon}' H^{-1} f_i \sigma_{u\epsilon}'/N.$$

For part 11, with the same arguments,

$$\begin{aligned}
& E\left[u'P(W)\epsilon\epsilon' f H^{-1} f' u | z\right]/N^2 \\
= & \sum_i P_{ii}(W) E\left[\epsilon_i^2 u_i f_i H^{-1} u_i f_i' | z\right]/N^2 + \sum_{i\neq j} P_{jj}(W) E[\epsilon_j u_j] f_i' H^{-1} f_i E[u_i'\epsilon_i]/N^2 \\
& + \sigma_\epsilon^2 \sum_{i\neq j} P_{ij}(W) E\left[u_j f_i' H^{-1} f_i u_i' | z\right]/N^2 + \sum_{i\neq j} P_{ij}(W) E[u_j \varepsilon_j] f_j' H^{-1} f_i E[u_i'\varepsilon_i]/N^2 \\
= & O_p\left(\frac{1}{N}\right) + \frac{K'W}{N} \sigma_{u\epsilon} \sigma_{u\epsilon}' \frac{1}{N} \sum_{i=1}^n f_i' H^{-1} f_i \\
= & O_p\left(\frac{1}{N}\right) + \frac{K'W}{N} \sigma_{u\epsilon} \sigma_{u\epsilon}' \mathrm{tr}\left(H^{-1} \frac{1}{N} \sum_i f_i f_i'\right) = O_p\left(\frac{1}{N}\right) + d\frac{K'W}{N} \sigma_{u\epsilon} \sigma_{u\epsilon}'
\end{aligned}$$

and arguments similar to before

$$\begin{aligned}
E\left[u'P(W)\epsilon\epsilon' f H^{-1} u' f' | z\right]/N^2 &= O_p(1/N) + \sum_{i\neq j} P_{jj}(W) E\left(\epsilon_j u_j\right) f_i' H^{-1} E\left(u_i \epsilon_i\right) f_i'/N^2 \\
&= O_p\left(\frac{1}{N}\right) + \frac{K'W}{N} \sigma_{u\epsilon} \frac{1}{N} \sum_i f_i' H^{-1} \sigma_{u\epsilon} f_i'.
\end{aligned}$$

For part 12, note that

$$W'\Gamma W = \sum_{m=1}^M \left(\sum_{j=m}^M w_j\right)^2 \leq \sum_{m=1}^M \sum_{j=m}^M |w_j| \left|\sum_{j=m}^M w_j\right| \leq C \sum_{m=1}^M |w_m| \, m = CK'W^+$$

where the second inequality follows from the condition $\sup_{k\leq M} \left|\sum_{m=k}^M w_m\right| \leq C < \infty$. ■

Lemma 7.6 *If Assumptions 1-5 hold, and for* $\Omega = \Omega_U = \{W \in l_1 | W'\mathbf{1}_M = 1\}$ *where* M *satisfies the constraints in Assumption 7 and* $W = (w_1, ..., w_M)$, *it follows that*

$$\inf_{W\in\Omega} S_\lambda(W) = O_p\left(N^{\frac{-2\alpha}{2\alpha+1}}\right).$$

Proof. Consider a sequence $\tilde{W}$ where $w_M = 2$, $w_{2M} = -1$ and $w_j = 0$ for $j \neq M, 2M$ and $M = \left\lfloor N^{\frac{1}{2\alpha+1}} \right\rfloor$. Clearly, $\mathbf{1}'\tilde{W} = 1$ and $\tilde{W} \in l_1$ for all N such that $\tilde{W} \in \Omega$. We note that $K'\tilde{W} = 0$. It follows that

$$S_\lambda\left(\tilde{W}\right) = \lambda' H^{-1} \left(b_\sigma \frac{(\tilde{W}'\Gamma\tilde{W})}{N} + \sigma_\epsilon^2 \frac{f'(I - P(\tilde{W}))(I - P(\tilde{W}))f}{N}\right) H^{-1}\lambda$$

where

$$\frac{(\tilde{W}'\Gamma\tilde{W})}{N} = \frac{2M}{N} = O\left(N^{\frac{-2\alpha}{2\alpha+1}}\right)$$

and

$$\lim \frac{\mathrm{tr}\left(f'(I - P(\tilde{W}))(I - P(\tilde{W}))f\right)}{N} = 4\tilde{\gamma}_M - 3\tilde{\gamma}_{2M} = O_p\left(M^{-2\alpha}\right) = O_p\left(N^{\frac{-2\alpha}{2\alpha+1}}\right)$$

which shows that $\inf_{W\in\Omega} S_\lambda(W) \leq CN^{\frac{-2\alpha}{2\alpha+1}}$.

To show that the rate is sharp, suppose that there is an $\varepsilon > 0$ such that

$$\inf_{W\in\Omega} S_\lambda(W) = O_p\left(N^{\frac{-2\alpha(1+\varepsilon)}{2\alpha+1}}\right).$$

Take any W such that, for $M = \left\lfloor N^{\frac{1+\delta}{2\alpha+1}} \right\rfloor$ where $0 < \delta < \varepsilon/2$,

$$\operatorname{tr}\left(\frac{f'(I-P(\tilde{W}))(I-P(\tilde{W}))f}{N}\right) = \sum_{j=1}^{M}\left(\sum_{i=1}^{j} w_i\right)^2\left(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}\right) + \tilde{\gamma}_M = O_p\left(N^{\frac{-2\alpha(1+\varepsilon)}{2\alpha+1}}\right), \tag{7.9}$$

where we use formula (7.7). Let J_M be the set of integers j such that $1 \leq j \leq M$ for which $j^{2\alpha+1}\left(\gamma_j - \gamma_{j+1}\right) > 0$. By the assumptions of the Lemma, w.p.a 1, $\sharp J_M = O(M)$ as $M \to \infty$ where $\sharp J_M$ is the cardinality of J_M. It follows that

$$\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)^2\left(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}\right) \geq \sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)^2 M^{-(2\alpha+1)} \geq O\left(N^{\frac{-(2\alpha+1)(1+\delta)}{2\alpha+1}}\right)\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)^2$$

which together with (7.9) implies that $\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_{i,N}\right)^2 = O\left(N^{\frac{-2\alpha(\varepsilon-\delta)+1+\delta}{2\alpha+1}}\right) = o(M)$. Now, since

$$O(M) = \sum_{j\in J_M} 1^2 = \sum_{j\in J_M}\left(\left(\sum_{i=1}^{j} w_i\right)^2 + 2\left(\sum_{i=1}^{j} w_i\right)\left(\sum_{i=j+1}^{M} w_i\right) + \left(\sum_{i=j+1}^{M} w_i\right)^2\right) \tag{7.10}$$

and by the Cauchy-Schwarz inequality

$$\begin{aligned}\left|\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)\left(\sum_{i=j+1}^{M} w_i\right)\right| &\leq \left(\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)^2\right)^{1/2}\left(\sum_{j\in J_M}\left(\sum_{i=j+1}^{M} w_i\right)^2\right)^{1/2} \\ &= o\left(\sqrt{M}\right)\left(\sum_{j\in J_M}\left(\sum_{i=j+1}^{M} w_i\right)\right)^{1/2},\end{aligned}$$

it follows that (7.10) can only hold if $\liminf_N \sum_{j\in J_M}\left(\sum_{i=j+1}^{M} w_i\right)^2/M > 0$. Then, for some $\eta > 0$ and N large enough, it follows that

$$W'\Gamma W = \sum_{j=0}^{M}\left(\sum_{m=j+1}^{M} w_m\right)^2 \geq M\eta = O\left(N^{\frac{1+\delta}{2\alpha+1}}\right)$$

such that $W'\Gamma W/N = O\left(N^{\frac{-2\alpha+\delta}{2\alpha+1}}\right)$, which implies that $S_\lambda(W) = O\left(N^{\frac{-2\alpha+\delta}{2\alpha+1}}\right)$, a contradiction to the assumption that $\inf_{W\in\Omega} S_\lambda(W) = O_p\left(N^{\frac{-2\alpha(1+\varepsilon)}{2\alpha+1}}\right)$. This argument establishes that $\inf_{W\in\Omega} S_\lambda(W) = O_p\left(N^{\frac{-2\alpha}{2\alpha+1}}\right)$ is a sharp bound. ■

Lemma 7.7 *Let*

$$\tilde{S}_\lambda(W) = \hat{H}^{-1}\left(\hat{a}_\sigma\frac{(K'W)^2}{N} + \hat{b}_\sigma\frac{(W'\Gamma W)}{N} - \frac{K'W}{N}\hat{B}_N + \hat{\sigma}_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\right)\hat{H}^{-1}$$

If Assumptions 1-6 hold, then, for Ω as defined in Lemma 7.6, it follows that

$$\sup_{W\in\Omega}\frac{\tilde{S}_\lambda(W)}{S_\lambda(W)}-1=o_p(1).$$

Proof. We define the subset $\Omega_2=\{W\in l_1\,|-\infty<\liminf_N K'W\leq\limsup_N K'W<\infty\}$. Note that

$$\sup_{W\in\Omega\cap\Omega_2}\frac{K'W/N}{S_\lambda(W)}\to 0 \text{ and } \sup_{W\in\Omega\cap\Omega_2}\frac{(K'W)^2/N}{S_\lambda(W)}\to 0 \tag{7.11}$$

by Lemma 7.6 and the fact that $\{W_N\in l_1|K'W=0\}\in\Omega_2$. It now follows immediately that

$$\lambda'\left(\hat{H}^{-1}\hat{a}_\sigma\hat{H}^{-1}-H^{-1}a_\sigma H^{-1}\right)\lambda\sup_{W\in\Omega\cap\Omega_2}\frac{(K'W)^2/N}{S_\lambda(W)}=o_p(1)$$

with the same argument holding for the term $\hat{B}_N K'W/N$. Define

$$S_{\lambda,\Omega_2}(W)=\lambda'H^{-1}\left(b_\sigma\frac{(W'\Gamma W)}{N}+\sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\right)H^{-1}\lambda$$

and note that $S_{\lambda,\Omega_2}(W)\geq\lambda'H^{-1}b_\sigma H^{-1}\lambda(W'\Gamma W)/N$ as well as $S_{\lambda,\Omega_2}(W_N)\geq\sigma_\epsilon^2\lambda'H^{-1}f'(I-P(W))(I-P(W))fH^{-1}\lambda/N$. Thus, we have

$$\begin{aligned}\sup_{W\in\Omega\cap\Omega_2}\frac{(W'\Gamma W)/N}{S_\lambda(W)} &\leq \sup_{W\in\Omega\cap\Omega_2}\frac{(W'\Gamma W)/N}{S_{\lambda,\Omega_2}(W)}\sup_{W\in\Omega\cap\Omega_2}\frac{S_{\lambda,\Omega_2}(W)}{S_\lambda(W)}\\ &\leq \frac{1}{\lambda'H^{-1}b_\sigma H^{-1}\lambda}\sup_{W\in\Omega\cap\Omega_2}\frac{S_{\lambda,\Omega_2}(W)}{S_\lambda(W)},\end{aligned}$$

where $\sup_{W\in\Omega\cap\Omega_2}S_{\lambda,\Omega_2}(W_N)/S_\lambda(W_N)\to 1$ by (7.11). This implies that

$$\lambda'\left(\hat{H}^{-1}\hat{b}_\sigma\hat{H}^{-1}-H^{-1}b_\sigma H^{-1}\right)\lambda\sup_{W\in\Omega\cap\Omega_2}\frac{(W'\Gamma W)/N}{S_\lambda(W)}=o_p(1).$$

Now consider

$$\begin{aligned}&\lambda'\left(\hat{H}^{-1}\hat{\sigma}_\epsilon^2-H^{-1}\sigma_\epsilon^2\right)\frac{f'(I-P(W))(I-P(W))f}{N}\hat{H}^{-1}\lambda\\ &+\lambda'H^{-1}\sigma_\epsilon^2\frac{f'(I-P(W))(I-P(W))f}{N}\left(\hat{H}^{-1}-H^{-1}\right)\lambda,\end{aligned}$$

where

$$\begin{aligned}&\sup_{W\in\Omega\cap\Omega_2}\frac{\left|\lambda'\left(\hat{H}^{-1}\hat{\sigma}_\epsilon^2-H^{-1}\sigma_\epsilon^2\right)f'(I-P(W))(I-P(W))f\hat{H}^{-1}\lambda/N\right|}{S_\lambda(W)}\\ \leq\ &\left\|\hat{H}^{-1}\lambda\right\|\left\|\lambda'\left(\hat{H}^{-1}\hat{\sigma}_\epsilon^2-H^{-1}\sigma_\epsilon^2\right)\right\|\sup_{W\in\Omega}\frac{\left\|(I-P(W))f/\sqrt{N}\right\|^2}{\left\|(I-P(W))fH^{-1}\lambda/\sqrt{N}\right\|^2}=o_p(1)\end{aligned}$$

where

$$\sup_{W\in\Omega}\frac{\left\|(I-P(W))f/\sqrt{N}\right\|^2}{\left\|(I-P(W))fH^{-1}\lambda/\sqrt{N}\right\|^2}=O_p(1)$$

by Assumption 2. Together, these arguments show that

$$\sup_{W\in\Omega\cap\Omega_2}\frac{\tilde{S}_\lambda(W)}{S_\lambda(W)}-1=o_p(1).$$

For $W \in \Omega \cap \Omega_2^C$ where $\Omega_2^C = \{W \in l_1 \,|\liminf_N |K'W| = \infty\}$ it follows that

$$\sup_{W \in \Omega \cap \Omega_2^C} \frac{|K'W|/N}{(K'W)^2/N} \to 0$$

such that for

$$S_{\lambda,\Omega_2^C}(W_N) = \lambda' H^{-1}\left[a_\sigma \frac{(K'W)^2}{N} + b_\sigma \frac{(W'\Gamma W)}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N}\right] H^{-1}\lambda$$

it follows that

$$\sup_{W \in \Omega \cap \Omega_2^C} \frac{S_{\lambda,\Omega_2^C}(W)}{S_\lambda(W)} \to 1 \text{ as } N \to \infty.$$

Then similar arguments as before can be used to show that

$$\sup_{W \in \Omega \cap \Omega_2} \frac{\tilde{S}_\lambda(W)}{S_\lambda(W)} - 1 = o_p(1).$$

Since $\left(\Omega_2 \cup \Omega_2^C\right) \cap \Omega = \Omega$, this establishes the claimed result. ■

Lemma 7.8 *Let Assumptions 1-7 hold. Then, it follows that*

$$\sup_{W \in \Omega} \frac{\hat{S}_\lambda(W)}{S_\lambda(W)} - 1 \to_p 0$$

Proof. Without loss of generality assume that f_i is a scalar and $\lambda' H^{-1} = 1$ so that $\sigma_\lambda^2 = \sigma_u^2$. First consider

$$\left\|(I-P(W))f/\sqrt{N}\right\|^2 - f'(I-P_M)f/N = \left\|(P_M - P(W))f/\sqrt{N}\right\|^2$$

and note that

$$f'(I-P_M)f/N = O_p\left(M^{-2\alpha}\right)$$

by Assumption 2. Together with Lemma 7.6, this implies that

$$\begin{aligned} &\sup_{W \in \Omega} \frac{\left\|(P_M - P(W))f/\sqrt{N}\right\|^2 - \left\|(I-P(W))f/\sqrt{N}\right\|^2}{S_\lambda(W)} \\ \leq\ & \frac{\sup_{W\in\Omega} f'(I-P_M)f/N}{\inf_{W\in\Omega} S_\lambda(W)} = O_p\left(M^{-2\alpha} N^{\frac{2\alpha}{2\alpha+1}}\right) = O_p\left(N^{\frac{-2\alpha\delta}{2\alpha+1}}\right) = o_p(1) \end{aligned}$$

Combining these results with Lemma 7.7 it is then sufficient to show that

$$\sup_{W \in \Omega} \frac{\left|\left\|(P_M - P(W))X/\sqrt{N}\right\|^2 - \left\|(P_M - P(W))f/\sqrt{N}\right\|^2 - \sigma_u^2(M - 2K'W + W'\Gamma W)/N\right|}{S_\lambda(W)} = o_p(1)$$

We note that in this expression we replace $\hat{\sigma}_u^2$by σ_u^2 which is justified by the same arguments as in the proof of Lemma 7.7 as long as $\hat{\sigma}_u^2 - \sigma_u^2 = o_p\left(N^{-\delta/(2\alpha+1)}\right)$ because, under the assumptions of the Lemma, it then follows that $\left(\hat{\sigma}_u^2 - \sigma_u^2\right) M/N = o_p\left(N^{-2\alpha/(2\alpha+1)}\right) = o_p\left(\inf_{W\in\Omega} S_\lambda(W)\right)$ and the remaining terms involving σ_u^2 can be handled in the same way as in the proof of Lemma 7.7. Now note that

$$\begin{aligned} &\left\|(P_M - P(W))X/\sqrt{N}\right\|^2 - \left\|(P_M - P(W))f/\sqrt{N}\right\|^2 \\ =\ & \left\|(P_M - P(W))u/\sqrt{N}\right\|^2 + 2u'(P_M - P(W))(P_M - P(W))f/N. \end{aligned}$$

It follows that

$$\begin{aligned} E\left[u'\left(P_M - P\left(W\right)\right)\left(P_M - P\left(W\right)\right)u/N|z\right] &= \sigma_u^2\left(\mathrm{tr}(P_M) - 2\mathrm{tr}(P(W)) + \mathrm{tr}\left(P(W)P(W)\right)\right)/N \\ &= \sigma_u^2\left(M - 2K'W + W'\Gamma W\right)/N, \end{aligned}$$

and

$$E\left[u'\left(P_M - P\left(W\right)\right)\left(P_M - P\left(W\right)\right)f/N|z\right] = 0.$$

Moreover, we have the bound

$$\begin{aligned} &\left|\left\|\left(P_M - P\left(W\right)\right)u\right\|^2 - \sigma_u^2\left(M - 2K'W + W'\Gamma W\right)\right| \\ \leq\ & \left|u'P_M u - \sigma_u^2 M\right| + \sup_{j\leq M}\left|u'P_j u - \sigma_u^2 j\right|\left(2\sum_{j=1}^{M}|w_j| + \sum_{j=1}^{M}\sum_{l=1}^{M}|w_j|\,|w_l|\right) \end{aligned}$$

where $\sum_{j=1}^{M}|w_j| < C$ is used. It follows for some $\vartheta > 1$ from Whittle (1960, Theorem 2) that for some constant C,

$$E\left[\left|u'P_j u - \sigma_u^2 j\right|^{2\vartheta}\right] \leq CE\left[\left|u_i\right|^{2\vartheta}\right]^2\left(\mathrm{tr}\left(P_j P_j'\right)\right)^{\vartheta} = CE\left[\left|u_i\right|^{2\vartheta}\right]^2 j^{\vartheta}$$

and thus for any $\eta > 0$ and some constant C, not necessarily the same as above,

$$\begin{aligned} &\Pr\left[\frac{\sup_{W\in\Omega}\left|\left\|\left(P_M - P\left(W\right)\right)u\right\|^2 - \sigma_u^2\left(M - 2K'W + W'\Gamma W\right)\right|/N}{\inf_{W\in\Omega} S_\lambda(W)} > \eta\right] \\ \leq\ & C\frac{E\left[\left|u'P_M u - \sigma_u^2 M\right|^{2\vartheta}\right]}{\eta^{\vartheta}N^{2\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}} + 3C\sum_{j=1}^{M}\frac{E\left[\left|u'P_j u - \sigma_u^2 j\right|^{2\vartheta}\right]}{\eta^{\vartheta}N^{2\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}} \\ \leq\ & C\frac{E\left[\left|u_i\right|^{2\vartheta}\right]^2\left(M^{\vartheta} + M^{\vartheta+1}\right)}{\eta^{\vartheta}N^{2\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}} = O\left(N^{\frac{1+\delta-\vartheta(1-\delta)}{2\alpha+1}}\right) = o_p\left(1\right) \end{aligned}$$

Next, consider

$$\left|u'\left(P_M - P\left(W\right)\right)\left(I - P\left(W\right)\right)f/N\right| = \left|\sum_{i,j=1}^{M} w_i w_j u'\left(P_M - P_{\max(i,j)}\right)f/N\right|$$

where

$$\left|\sum_{i,j=1}^{M} w_i w_j u'\left(P_M - P_{\max(i,j)}\right)f/N\right| \leq \sum_{i=1}^{M-1}\left(\sum_{j=1}^{i} w_j\right)^2\left|u'\left(P_{i+1} - P_i\right)f/N\right|.$$

Let $K_n = N^{\lfloor(1-\varepsilon)/(2\alpha+1)\rfloor}$. Then,

$$\sup_{W\in\Omega}\frac{\sum_{i=1}^{M-1}\left(\sum_{j=1}^{i} w_j\right)^2\left|u'\left(P_{i+1} - P_i\right)f/N\right|}{S_\lambda(W)} = \sup_{W\in\Omega}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2\left|u'\left(P_{i+1} - P_i\right)f/N\right|}{S_\lambda(W)} + o_p\left(1\right) \tag{7.12}$$

because

$$\begin{aligned}
& \Pr\left(\sup_{W\in\Omega}\frac{\sum_{i=K_n+1}^{M-1}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda(W)}>\eta|z\right)\\
\leq\ & \Pr\left(\frac{\sup_{W\in\Omega}\sum_{i=K_n+1}^{M-1}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{\inf_{W\in\Omega}S_\lambda(W)}>\eta|z\right)\\
\leq\ & \frac{CE\left[\left|u_i\right|^{2\vartheta}\right]\sum_{j=K_n+1}^{M}\left(f'\left(P_{j+1}-P_j\right)f/N\right)^{\vartheta}}{\eta^{\vartheta}N^{\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}}
\end{aligned}$$

where the inequality follows from Markov's inequality, Lemma 7.6, the fact that $\left|\sum_{j=1}^{i}w_j\right|$ is uniformly bounded on Ω, and Theorem 1 of Whittle (1960) which implies that

$$E\left|u'\left(P_{i+1}-P_i\right)f/N\right|^{2\vartheta}\leq CE\left[\left|u_i\right|^{2\vartheta}\right]N^{-\vartheta}\left(f'\left(P_{i+1}-P_i\right)f/N\right)^{\vartheta}. \tag{7.13}$$

Now note that

$$\begin{aligned}
\frac{CE\left[\left|u_i\right|^{2\vartheta}\right]\sum_{j=K_n+1}^{M}\left(f'\left(P_{j+1}-P_j\right)f/N\right)^{\vartheta}}{\eta^{\vartheta}N^{\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}} &\leq \frac{CE\left[\left|u_i\right|^{2\vartheta}\right]\left(f'\left(I-P_{K_N}\right)f/N\right)^{\vartheta}M}{\eta^{\vartheta}N^{\vartheta}N^{-4\alpha\vartheta/(2\alpha+1)}}\\
&= O_p\left(K_n^{-2\alpha\vartheta}M/N^{\vartheta}N^{4\alpha\vartheta/(2\alpha+1)}\right)\\
&= O_p\left(N^{-\frac{2(1-\epsilon)\alpha\vartheta}{2\alpha+1}-\vartheta+\frac{1+\delta}{2\alpha+1}+\frac{4\alpha\vartheta}{2\alpha+1}}\right)=o_p(1)
\end{aligned}$$

which establishes (7.12). We thus turn to the lead term on the right hand side of (7.12). By the Cauchy-Schwarz inequality we have

$$\left|u'\left(P_{i+1}-P_i\right)f/N\right|\leq\left(f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}\left(u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}.$$

It now follows that

$$\begin{aligned}
& \sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right| \qquad (7.14)\\
\leq\ & \left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^4 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^4 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}\\
\leq\ & \sup_{i\leq M}\left(\sum_{j=1}^{i}w_j\right)^2\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}\left(\sum\nolimits_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}
\end{aligned}$$

where $\sup_{i\leq M}\left(\sum_{j=1}^{i}w_j\right)^2<C<\infty$ for some C such that

$$\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}\leq\sup_{W}\left(\sum_{j=1}^{i}\left|w_j\right|\right)^2\left(\sum_{i=1}^{K_n}u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2} \tag{7.15}$$

$$\leq \quad C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}$$

where $W \in l_1$ was used to bound $\sup_W \left(\sum_{j=1}^{i}|w_j|\right)^2$. Let $\Omega_N \subset \Omega$ be the sequence of subsets of sequences in Ω for which $w_i = 0$ for all $i > N$. Clearly,

$$\sup_{W\in\Omega}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} = \sup_{W\in\Omega_N}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} \tag{7.16}$$

Now, fix an arbitrary $\omega > 0$ and define the sequence of sets

$$\Omega_{1,N} = \left\{ W \in \Omega_N \left| \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N}{N^{(-2\alpha+\varepsilon/2)/(2\alpha+1)}} \leq \omega \right. \right\}$$

and let $\Omega_{1,N}^C$ be the complement of $\Omega_{1,N}$ in Ω_N, such that $\Omega_N = \left(\Omega_N \cap \Omega_{1,N}\right) \cup \left(\Omega_N \cap \Omega_{1,N}^c\right)$. We note that $\Omega_{1,N}$ depends on the realizations for the instruments z.

As was demonstrated in the proof of Lemma 7.7, as N tends to infinity, $S_\lambda\left(W\right) \geq \sigma_\epsilon^2 \lambda' H^{-1} f'(I - P(W))(I - P(W))fH^{-1}\lambda/N$. Also note that

$$f'(I-P(W))(I-P(W))f/N \geq \sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N$$

Therefore, for N sufficiently large,

$$\begin{aligned}
& \sup_{W\in\Omega_N\cap\Omega_{1,N}^C}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} \\
\leq & \sup_{W\in\Omega_N\cap\Omega_{1,N}^C}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N} \\
\leq & \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\inf_{W\in\Omega_N\cap\Omega_{1,N}^C}\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}}
\end{aligned}$$

where

$$\inf_{W\in\Omega_N\cap\Omega_{1,N}^C}\frac{\left(\sum_{i\in J_{K_n}}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}}{N^{(-\alpha+\varepsilon/4)/(2\alpha+1)}} \geq \sqrt{\omega}$$

by the construction of $\Omega_{1,N}$. It then follows that

$$\sup_{W\in\Omega_N\cap\Omega_{1,N}^C}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2\left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} \leq \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\sqrt{\omega}N^{(-\alpha+\varepsilon/4)/(2\alpha+1)}}. \tag{7.17}$$

Secondly,

$$\sup_{W\in\Omega_N\cap\Omega_{1,N}}\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N \leq \omega N^{(-2\alpha+\varepsilon/2)/(2\alpha+1)} \tag{7.18}$$

by the definition of $\Omega_{1,N}$ such that

$$\begin{aligned} & \sup_{W\in\Omega_N\cap\Omega_{1,N}} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda(W)} \\ \leq & \frac{\sup_{W\in\Omega_N\cap\Omega_{1,N}}\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{\inf_{W\in\Omega} S_\lambda(W)} \\ \leq & \sqrt{\omega}N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}} \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\inf_{W\in\Omega} S_\lambda(W)} \end{aligned} \tag{7.19}$$

It now follows for any random function $g_N(W)$ that

$$\begin{aligned} \sup_{W\in\Omega_N} g_N(W) &= \max\left(\sup_{W\in\Omega_N\cap\Omega_{1,N}} g_N(W), \sup_{W\in\Omega_N\cap\Omega_{1,N}^C} g_N(W)\right) \\ &\leq \sup_{W\in\Omega_N\cap\Omega_{1,N}} g_N(W) + \sup_{W\in\Omega_N\cap\Omega_{1,N}^C} g_N(W). \end{aligned}$$

Thus, setting $g_N(W) = \sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right| / S_\lambda(W)$ and using (7.16), (7.17) and (7.19) one obtains the bound

$$\begin{aligned} & \sup_{W\in\Omega} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda(W)} \\ \leq & \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\sqrt{\omega}N^{(-\alpha+\varepsilon/4)/(2\alpha+1)}} + \sqrt{\omega}N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}} \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\inf_{W\in\Omega} S_\lambda(W)}. \end{aligned} \tag{7.20}$$

It then follows that for any $\eta_1 > 0$,

$$\begin{aligned} & \Pr\left[\left|\sup_{W\in\Omega} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda(W)}\right| > \eta_1 \middle| z\right] \\ \leq & \frac{1}{\sqrt{\omega}} \frac{C\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{(-\alpha+\varepsilon/2)/(2\alpha+1)}} + \frac{\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{-2\alpha/(2\alpha+1)}} O_p\left(N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\right), \end{aligned}$$

where the inequality uses Markov's inequality, (7.20) and Lemma 7.6. Next, note that

$$\frac{C\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{(-\alpha+\varepsilon/2)/(2\alpha+1)}} = \frac{1}{\sqrt{\omega}} \frac{C\sqrt{\left(K_{n+1}-1\right)/N}}{N^{(-\alpha+\varepsilon/2)/(2\alpha+1)}} = o\left(N^{\frac{-\varepsilon/2-\varepsilon/2}{2\alpha+1}}\right) = o(1) \tag{7.21}$$

and

$$\begin{aligned} E\left[u'P_{K_{n+1}}u/N|Z\right]^{1/2} O_p\left(N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\right) &= O_p\left(K_n^{1/2}N^{(-\alpha+\varepsilon/4)/(2\alpha+1)-1/2}\right) \\ &= O_p\left(N^{\frac{-2\alpha-\varepsilon/4}{2\alpha+1}}\right) = o_p\left(N^{\frac{-2\alpha}{2\alpha+1}}\right) \end{aligned}$$

such that

$$\frac{\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{-2\alpha/(2\alpha+1)}} O_p\left(N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\right) = o_p(1). \tag{7.22}$$

Using (7.21) and (7.22) then establishes that

$$\Pr\left[\left|\sup_{W\in\Omega} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda(W)}\right| > \eta_1 \middle| z\right] = o(1) + o_p(1).$$

This completes the proof of the Lemma. ■

7.2 Proofs of Theorems and Corollaries

Proof of Theorem 7.1. The MA2SLS estimator has the form:

$$\sqrt{N}(\hat{\beta}-\beta_0)=\hat{H}^{-1}\hat{h},\quad \hat{H}=X'P(W)X/N,\quad \hat{h}=X'P(W)\epsilon/\sqrt{N}.$$

Also $\hat{H}$ and $\hat{h}$ are decomposed as

$$\begin{aligned}
\hat{h} &= h+T_1^h+T_2^h,\\
T_1^h &= -f'(I-P(W))\epsilon/\sqrt{N},\quad T_2^h=u'P(W)\epsilon/\sqrt{N}\\
\hat{H} &= H+T_1^H+T_2^H+T_3^H+Z^H\\
T_1^H &= -f'(I-P(W))f/N,\quad T_2^H=(u'f+f'u)/N,\quad T_3^H=u'P(W)u/N\\
Z^H &= (u'(I-P(W))f+f'(I-P(W))u)/N.
\end{aligned}$$

We show that the conditions of Lemma 1 are satisfied and $S(W)$ has the form given in the theorem. Differently from Donald and Newey (2001) we extend the MA2SLS to order $K'W/N$. It is important to point out that since W can contain negative weights, it is possible that $K'W/N$ is not the dominating term in $S(W)$. For example, $K'W=0$ is allowed. However, $K'W/N=O(S(W))$ by construction.

Now $h=O_p(1)$ and $H=O_p(1)$ by Lemma 2(5). As

$$T^h=T_1^h+T_2^h=-f'(I-P(W))\epsilon/\sqrt{N}+u'P(W)\epsilon/\sqrt{N},$$

Lemma 7.5(2) and (3) implies that

$$T_1^h=O_p(\Delta(W)^{1/2})$$

and

$$T_2^h=O_p\left(\max\left(|K'W|,\sqrt{(W'\Gamma W)+\sum_i(P_{ii}(W))^2}\right)/\sqrt{N}\right) \tag{7.23}$$

so

$$T^h=O_p(\Delta(W)^{1/2})+O_p\left(\max\left(|K'W|,\sqrt{(W'\Gamma W)+\sum_i(P_{ii}(W))^2}\right)/\sqrt{N}\right)$$

where $\Delta(W)=o_p(1)$ by Lemma 7.5(1), $K'W/\sqrt{N}=o(1)$ by $K'W/\sqrt{N}\leq K'W^+/\sqrt{N}=o(1)$, $\sum_i(P_{ii}(W))^2=o_p\left(K'W^+\right)$ by Lemma 7.4(2) and $W'\Gamma W=O\left(K'W^+\right)$ by Lemma 7.5(12). Therefore $T^h=o_p(1)$. Next, we observe

$$f'(I-P(W))f=\sum_{m=1}^{M}w_m f'(I-P_m)f.$$

Since $f'(I-P_m)f/N=O_p(m^{-2\alpha})$ by Assumption 2(ii), we get $T_1^H=O_p(\Delta(W))$. $T_2^H=O_p(1/\sqrt{N})$ by the CLT. Also

$$T_3^H=u'P(W)u/N=O_p(K'W/N) \tag{7.24}$$

because $E\left[u'P(W)u|z\right] = \text{tr}\left(P\left(W\right)\right)\text{tr}\left(\Sigma_u\right) = \text{tr}\left(\Sigma_u\right)K'W$ where last equality follows from Lemma 7.4(1). Now we analyze

$$\|T_1^h\| \cdot \|T_1^H\| = O_p(\Delta(W)^{3/2}) = o_p\left(\rho_{W,N}\right),$$
$$||T_1^h|| \cdot ||T_2^H|| = O_p\left(\Delta(W)^{1/2}/\sqrt{N}\right) = o_p(\rho_{W,N})$$

because by Lemma 7.5(6) one can take $g(W) = N\left(\text{tr}\left(S(W)\right) - \Delta(W)\right)$. From Lemma 7.3 it follows that $W'\Gamma W \to \infty$ as $N \to \infty$. This implies that $g(W) \to \infty$. Then, by Lemma 7.5(6), it follows that

$$\Delta(W)^{1/2}/\sqrt{N} = o_p\left(\frac{g\left(W\right)}{N} + \Delta\left(w\right)\right) = o_p\left(\text{tr}\left(S(W)\right)\right) = o_p\left(\rho_{W,N}\right)$$

Next,

$$||T_1^h|| \cdot ||T_3^H|| = O_p\left(\Delta(W)^{1/2}K'W/N\right) = o_p\left(K'W/N\right) = o_p\left(\rho_{W,N}\right)$$

by Lemma 7.5(1), (7.24) and the fact (as noted before) that $K'W/N = O\left(\text{tr}(S\left(W\right))\right)$. Next, note that by Lemma 7.5(3) and (4) if $K'W \neq 0$,

$$||T_2^h|| \cdot ||T_1^H|| = O_p\left(\Delta(W)K'W/\sqrt{N}\right).$$

By similar arguments as before it follows from Lemma 7.5(6), that

$$\Delta(W)^{1/2}K'W/\sqrt{N} \leq \Delta(W)^{1/2}\left|K'W\right|/\sqrt{N} \leq \left(K'W\right)^2/N + \Delta(W) = O\left(\rho_{W,N}\right)$$

and $\Delta\left(W\right)^{1/2} = o_p\left(1\right)$ such that $\Delta(W)K'W/\sqrt{N} = o_p\left(\rho_{W,N}\right)$ as required. If $K'W = 0$ then the proof of Lemma 7.5(3) and (4) implies that $||T_2^h|| = O\left(\sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)$ such that

$$\begin{aligned} ||T_2^h|| \cdot ||T_1^H|| &= O_p\left(\Delta(W)/\sqrt{N}\sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right) \\ &= O_p\left(\Delta(W)^{1/2}\right) O_p\left(\frac{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}{N} + \Delta\left(W\right)\right) = o_p\left(\rho_{W,N}\right). \end{aligned}$$

From (7.23) it follows that

$$||T_2^h|| \cdot ||T_2^H|| = O_p\left(\max\left(\left|K'W\right|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/N\right),$$

where $K'W/N = O\left(\text{tr}\left(S\left(W\right)\right)\right)$ and $\sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}/N = o_p\left(\text{tr}(S(W))\right)$. By (7.23) and (7.24) it follows that

$$||T_2^h|| \cdot ||T_3^H|| = O_p\left(\max\left(\left|K'W\right|^{3/2}, \left(\sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)\left|K'W\right|\right)/N^{3/2}\right) = o_p\left(\rho_{W,N}\right).$$

because $\left(\left|K'W\right|/N\right)^{3/2} = o\left(K'W/N\right) = o\left(\rho_{W,N}\right)$ and $\sqrt{\left((W'\Gamma W) + \sum_i (P_{ii}(W))^2\right)/N} = o\left(1\right)$. Similarly, $||T_2^h||^2||T_1^H|| = o_p\left(\rho_{W,N}\right)$, $||T_2^h||^2||T_2^H|| = o_p\left(\rho_{W,N}\right)$ and $||T_2^h||^2||T_3^H|| = o_p\left(\rho_{W,N}\right)$. For $\left\|T^H\right\|^2$, we have

$$\begin{aligned} \left\|T_1^H\right\|^2 &= O_p\left(\Delta(W)^2\right) = o_p\left(\rho_{W,N}\right), \\ \left\|T_2^H\right\|^2 &= O_p\left(1/N\right) = o_p\left(\rho_{W,N}\right), \\ \left\|T_3^H\right\|^2 &= O_p((K'W/N)^2) = o_p\left(\rho_{W,N}\right) \end{aligned}$$

so that by the Cauchy Schwartz inequality $\left\|T^H\right\|^2 = o_p\left(\rho_{W,N}\right)$.

As $||Z^h|| = 0$ in our case, $||Z^h|| = o_p(\rho_{W,N})$. The last part, which we need to show $o_p(\rho_{W,N})$, is $||Z^H||$. Now $Z^H = u'(I - P(W))f/N + f'(I - P(W))u/N$ and both terms are $O_p(\Delta(W)^{1/2}/\sqrt{N}) = o_p(g(W)/N + \Delta(W)) = o_p(\rho_{W,N})$ for $g(W) = N(\mathrm{tr}(S(W)) - \Delta(W))$ by Lemma 7.5(6). Therefore we have $||Z^H|| = o_p(\rho_{W,N})$.

Note that we have shown $\hat{H} = H + o_p(1)$ and $\hat{h} = h + o_p(1)$. Lemma 7.1 can now be applied where the discussion above indicates

$$\begin{aligned} Z^A(W) &= -hT_1^{h\prime}H^{-1}\left(\sum_{j=1}^3 T_j^H\right)' - \left(\sum_{j=1}^3 T_j^H\right)H^{-1}T_1^h h' - T_1^h h' H^{-1}\left(\sum_{j=1}^3 T_j^H\right)' - \left(\sum_{j=1}^3 T_j^H\right)H^{-1}hT_1^{h\prime} \\ &\quad -hT_2^{h\prime}H^{-1}T_3^{H\prime} - T_3^H H^{-1}T_2^h h' - T_2^h h' H^{-1}T_3^{H\prime} - T_3^H H^{-1}hT_2^{h\prime} \\ &\quad -(T_1^h + T_2^h)(T_1^h + T_2^h)'H^{-1}\left(\sum_{j=1}^3 T_j^H\right)' - \left(\sum_{j=1}^3 T_j^H\right)H^{-1}(T_1^h + T_2^h)(T_1^h + T_2^h)' \\ &= o_p(\rho_{W,N}) \end{aligned}$$

and

$$\begin{aligned} \hat{A}(W) &= (h + T_1^h + T_2^h)(h + T_1^h + T_2^h)' - hh'H^{-1}\left(\sum_{j=1}^3 T_j^H\right)' - \left(\sum_{j=1}^3 T_j^H\right)H^{-1}hh' \\ &\quad -hT_2^{h\prime}H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}T_2^h h' - T_2^h h' H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}hT_2^{h\prime}. \end{aligned}$$

Now we calculate the expectation of each term in $\hat{A}(W)$. First of all, $E[hh'|z] = E[f\epsilon\epsilon' f'/N|z] = \sigma_\epsilon^2 H$. Second, $E\left[hT_1^{h\prime}|z\right] = E[-f\epsilon\epsilon(I - P(W))f'/N|z] = -\sigma_\epsilon^2 f(I - P(W))f'/N$. Similarly $E\left[T_1^h h'|z\right] = -\sigma_\epsilon^2 f(I - P(W))f'/N$. Third,

$$E\left[hT_2^{h\prime}|z\right] = E[f\epsilon\epsilon' P(W)u/N|z] = E\left[\epsilon_1^2 u_1'\right]\sum_i f_i P_{ii}(W)/N = O_p(K'W/N),$$

by Lemma 7.5(5). This implies that $E\left[T_2^h h'|Z\right] = O_p(K'W/N)$ too. Fourth,

$$E\left[T_1^h T_1^{h\prime}|z\right] = E\left[\frac{f'(I - P(W))\epsilon\epsilon'(I - P(W))f}{N}|z\right] = \sigma_\epsilon^2\frac{f'(I - P(W))(I - P(W))f}{N}.$$

Fifth,

$$E\left[T_1^h T_2^{h\prime}|z\right] = -E[f'(I - P(W))\epsilon\epsilon' P(W)u/N|z] = f'Q(W)\mu(W)/N$$

by Lemma 7.5(8). Again, we have $E\left[T_2^h T_1^{h\prime}|z\right] = \mu(W)'Q(W)f/N$. Sixth,

$$\begin{aligned} E\left[T_2^h T_2^{h\prime}|z\right] &= E\left[\frac{u'P(W)\epsilon\epsilon' P(W)u}{N}|z\right] \\ &= \sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(K'W)^2}{N} + (\sigma_\epsilon^2\Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{(W'\Gamma W)}{N} + \mathrm{Cum}[\epsilon_i, \epsilon_i, u_i, u_i']\sum_i (P_{ii}(W))^2, \end{aligned}$$

by Lemma 7.5(4). Seventh,

$$E\left[hh'H^{-1}T_1^H|z\right] = -E\left[\frac{f'\epsilon\epsilon' f H^{-1}f'(I - P(W))f}{N^2}|z\right] = -\sigma_\epsilon^2\frac{f'(I - P(W))f}{N},$$

also, we have $E\left[T_1^H H^{-1} hh'|Z\right] = -\sigma_\epsilon^2 f'(I - P(W))f/N$. Lemma 7.5(7) implies

$$E\left[hh'H^{-1}T_2^H|z\right] = E\left[\frac{hh'H^{-1}(u'f + f'u)}{N}|z\right] = O_p\left(\frac{1}{N}\right)$$

and $E\left[T_2^H H^{-1} hh'|z\right] = O_P(1/N)$. Also,

$$E\left[hh'H^{-1}T_3^H|z\right] = E\left[\frac{f'\epsilon\epsilon' f H^{-1}u'P(W)u}{N^2}|z\right] = \sigma_\epsilon^2\Sigma_u\frac{K'W}{N} + O_p\left(\frac{1}{N}\right)$$

by Lemma 7.5(9). Next,

$$\begin{aligned} E\left[hT_2^{h'}H^{-1}T_1^H|z\right] &= -E\left[\frac{f'\epsilon\epsilon' P(W)uH^{-1}f'(I - P(W))f}{N^2}|z\right] \\ &= \frac{1}{N}\sum_i f_i P_{ii}(W)E\left[\epsilon_i^2 u_i'\right]H^{-1}\frac{f'(I - P(W))f}{N} \\ &= O_p\left(K'W/N\Delta\left(W\right)\right) = o_p\left(\rho_{W,N}\right). \end{aligned}$$

by Lemma 7.5(5),

$$\begin{aligned} E\left[hT_2^{h'}H^{-1}T_2^H|z\right] &= E\left[\frac{f'\epsilon\epsilon' P(W)uH^{-1}(u'f + f'u)}{N^2}|z\right] \\ &= O_p\left(\frac{1}{N}\right) + \frac{K'W}{N}\left(\frac{1}{N}\sum_i f_i\sigma_{u\epsilon}'H^{-1}\sigma_{u\epsilon}f_i' + \frac{1}{N}\sum_i f_i\sigma_{u\epsilon}'H^{-1}f_i\sigma_{u\epsilon}'\right) \end{aligned}$$

by Lemma 7.5(10). Similarly, it follows that

$$\begin{aligned} E\left[T_2^h h'H^{-1}T_2^H|z\right] &= E\left[\frac{u'P(W)\epsilon\epsilon' f H^{-1}(u'f + f'u)}{N^2}|z\right] \\ &= O_p\left(\frac{1}{N}\right) + \frac{K'W}{N}\left(d\sigma_{u\epsilon}\sigma_{u\epsilon}' + \sigma_{u\epsilon}\frac{1}{N}\sum_i f_i'H^{-1}\sigma_{u\epsilon}f_i'\right) \end{aligned}$$

Therefore, we have

$$\begin{aligned} & E\left[\hat{A}(K)|z\right] \\ = \; & \sigma_\epsilon^2 H - 2\sigma_\epsilon^2\frac{f'(I - P(W))f}{N} + \sigma_\epsilon^2\frac{f'(I - P(W))(I - P(W))f}{N} \\ & + E[\epsilon_1^2 u_1']\sum_i f_i P_{ii}(W)/N + E[\epsilon_1^2 u_1]\sum_i f_i' P_{ii}(W)/N + f'Q\left(W\right)\mu\left(W\right)/N + \mu\left(W\right)'Q(W)f/N \\ & + \sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(K'W)^2}{N} + (\sigma_\epsilon^2\Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{(W'\Gamma W)}{N} + o_p\left(\frac{K'W}{N}\right) \\ & + 2\sigma_\epsilon^2\frac{f'(I - P(W))f}{N} + O_p\left(\frac{1}{N}\right) - 2\sigma_\epsilon^2\Sigma_u\frac{K'W}{N} \\ & - \frac{K'W}{N}2\left(d\sigma_{u\epsilon}\sigma_{u\epsilon}' + \frac{1}{N}\sum_{i=1}^n f_i\sigma_{u\epsilon}'H^{-1}\sigma_{u\epsilon}f_i' + \frac{1}{N}\sum_{i=1}^n\left(f_i\sigma_{u\epsilon}'H^{-1}f_i\sigma_{u\epsilon}' + \sigma_{u\epsilon}f_i'H^{-1}\sigma_{u\epsilon}f_i'\right)\right) \\ & + o_p\left(\rho_{W,N}\right) \\ = \; & \sigma_\epsilon^2 H + \sigma_\epsilon^2\frac{f'(I - P(W))(I - P(W))f}{N} \\ & + E[\epsilon_1^2 u_1']\sum_i f_i P_{ii}(W)/N + E[\epsilon_1^2 u_1]\sum_i f_i' P_{ii}(W)/N + f'Q(W)\mu(W)/N + \mu(W)'Q(W)f/N \end{aligned}$$

$$
\begin{aligned}
&+\sigma_{u\epsilon}\sigma'_{u\epsilon}\frac{(K'W)^2}{N}+(\sigma_\epsilon^2\Sigma_u+\sigma_{u\epsilon}\sigma'_{u\epsilon})\frac{(W'\Gamma W)}{N}\\
&-2\frac{K'W}{N}\left(\sigma_\epsilon^2\Sigma_u+d\sigma_{u\epsilon}\sigma'_{u\epsilon}+\frac{1}{N}\sum_{i=1}^{n}f_i\sigma'_{u\epsilon}H^{-1}\sigma_{u\epsilon}f_i'+\frac{1}{N}\sum_{i=1}^{n}\left(f_i\sigma'_{u\epsilon}H^{-1}f_i\sigma'_{u\epsilon}+\sigma_{u\epsilon}f_i'H^{-1}\sigma_{u\epsilon}f_i'\right)\right)\\
&+o_p\left(\rho_{W,N}\right)
\end{aligned}
$$

where the last equality holds because $1/N=o_p(\rho_{W,N})$, $K'W/N=o_p(\rho_{W,N})$ and $o_p((\Delta(W)K'W/N)^{1/2})=o_p(\rho_{W,N})$ by the fact that $(\Delta(W)K'W/N)^{1/2}\leq K'W/N+\Delta(W)$. ■

We omit the proofs of Corollary 7.1 and 7.3 because they are trivial given Theorem 7.1.

Proof of Corollary 7.2. We note that in this case $K'W=K'W^+$. Thus, $\sum_i\{P_{ii}(W)\}^2=o_p(K'W)$ by Lemma 7.4(2) and $f'Q(W)\mu(W)/N=o_p(K'W/N+\Delta(W))$ by Lemma 7.5 (8). Therefore, we have equation (7.3).

To derive equation (7.4), we note that

$$
W'\Gamma W=\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\min(i,j)\leq\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_jj=\sum_{i=1}^{M}w_i\sum_{j=1}^{M}w_jj=W'\mathbf{1}_MK'W=K'W,
$$

which means $W'\Gamma W=O(K'W)$. Moreover, $\sum_{i=1}^{N}f_iP_{ii}(W)=O_p(K'W)$ by Lemma 7.5(5). Therefore, we have equation (7.4). ■

Proof of Theorem 3.1. The result is established by constructing a sequence in Ω_P that dominates the optimal choice in Ω_{DN}. By Corollary 7.2 the approximate MSE of MA2SLS when $W\in\Omega_P$ is

$$
A\frac{(K'W)^2}{N}+\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}w_jw_i\gamma_{\max(i,j)}.
$$

for $A=\left\|\lambda'H^{-1}\sigma_{u\epsilon}\right\|^2$. Let M_{DN} be the optimal number of instruments picked by the DN algorithm. For $a\in(0,1)$, let $M_1=(1-a)M_{DN}$ and $M_2=(1+a)M_{DN}$ and choose W^* such that has only two non-zero elements $w_{M_1}=w_{M_2}=0.5$. Then, $K'W^*=M_{DN}$ and

$$
\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}w_jw_i\gamma_{\max(i,j)}=0.25\gamma_{M_1}+0.75\gamma_{M_2}
$$

Then,

$$
\begin{aligned}
\frac{\min_{W\in\Omega_P}S_\lambda(W)}{\min_{W\in\Omega_{DN}}S_\lambda(W)} &\leq \frac{S_\lambda(W^*)}{S_{DN}(M_{DN})}=\frac{A(K'W^*)^2/N+0.25\gamma_{M_1}+0.75\gamma_{M_2}}{A(M_{DN})^2/N+\gamma_{M_{DN}}}\\
&= \frac{A(M_{DN})^2/(N\gamma_{M_{DN}})+0.25\left(\gamma_{M_1}/\gamma_{M_{DN}}\right)+0.75\left(\gamma_{M_2}/\gamma_{M_{DN}}\right)}{A(M_{DN})^2/(N\gamma_{M_{DN}})+1}
\end{aligned}
$$

where $\gamma=\limsup_{N\to\infty}A(M_{DN})^2/(N\gamma_{M_{DN}})<\infty$ because M^{DN} sets the rates of the bias and the variance equal. The above expression is bounded by 1 if

$$
0.25\left(\gamma_{M_1}/\gamma_{M_{DN}}\right)+0.75\left(\gamma_{M_2}/\gamma_{M_{DN}}\right)<1.
$$

By assumption, for N large enough, it follows that, with probability close to one,

$$0.25\left(\gamma_{M_1}/\gamma_{M_{DN}}\right)+0.75\left(\gamma_{M_2}/\gamma_{M_{DN}}\right)=0.25\left(1-a\right)^{-2\alpha}+0.75\left(1+a\right)^{-2\alpha}+o\left(\left|a\right|^{2\alpha}\right).$$

Consider the function

$$h\left(a\right)=0.25\left(1-a\right)^{-2\alpha}+0.75\left(1+a\right)^{-2\alpha}$$

where $h\left(0\right)=1$, $\partial h\left(a\right)/\partial a=0.5\alpha\left(1-a\right)^{-2\alpha-1}-1.5\alpha\left(1+a\right)^{-2\alpha-1}$ such that $\partial h\left(0\right)/\partial a=-1\alpha$. This implies that for some a, possibly close to zero, $h\left(a\right)<1$ and thus $0.25\left(\gamma_{M_1}/\gamma_{M_{DN}}\right)+0.75\left(\gamma_{M_2}/\gamma_{M_{DN}}\right)<1$.

When $W\in\Omega_B$ the MSE of the MA2SLS is

$$S_\lambda(W)=A\frac{(W'\Gamma W)}{N}+\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}w_jw_i\gamma_{\max(i,j)}$$

where $A=\lambda'H^{-1}(\sigma_\epsilon^2\Sigma_u+\sigma_{u\epsilon}\sigma_{u\epsilon}')H^{-1}\lambda$ while the MSE for the Nagar estimator with M instruments is $AM/\left(N-M\right)+\gamma_M$. Let M_N be the choice of M that minimizes $S_\lambda\left(W\right)$ when $W=W_N$ as defined in Remark 4. For $a\in(0,1)$ let $M_1=(1-a)M_N$ and $M_2=(1+a)M_N$. Define $w^*=N/\left(N-M_N\right)$ and choose W^* such that W^* has only three non-zero elements $w_{M_1}=w_{M_2}=1/2w^*$ and $w_N=-M_N/\left(N-M_N\right)$. For brevity write w_1 and w_2 instead of w_{M_1} and w_{M_2}. Then $w_1+w_2+w_N=1$ and $K'W^*=0$ such that $W^*\in\Omega_B$. Note that $W_N'\Gamma W_N=((w^*)^2+2w^*w_N)M_N+w_N^2N=M_NN/(N-M_N)$ and

$$\begin{aligned}W^{*\prime}\Gamma W^*&=w_1^2M_1+w_2^2M_2+2w_1w_2M_1+w_N^2N+2w_N\left(w_1M_1+w_2M_2\right)\\&=w_1^2M_1+w_2^2M_2+2w_1w_2M_1+w_N^2N+2w_Nw^*M_N\\&=((w^*)^2+2w_Nw^*)M_N+w_N^2N-(1/2)(w^*)^2aM_N\end{aligned}$$

such that $W^{*\prime}\Gamma W^*<W_N'\Gamma W_N$. In the same way it follows that, for W^*,

$$\begin{aligned}\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}w_jw_i\gamma_{\max(i,j)}&=w_1^2\gamma_{M_1}+\left(w_2^2+2w_1w_2\right)\gamma_{M_2}+\left(w_N^2+2w_N\left(w_1+w_2\right)\right)\gamma_N\\&=\left(w^*\right)^2\left(\gamma_{M_1}/4+3\gamma_{M_2}/4\right)+\left(w_N^2+2w_Nw^*\right)\gamma_N.\end{aligned}$$

Since the term $\left(w_N^2+2w_Nw^*\right)\gamma_N$ is of smaller order than $S_\lambda\left(W_N\right)$ the result now follows if $(\gamma_{M_1}/4+3\gamma_{M_2}/4)/\gamma_{M_N}\leq1$ wpa1. But this follows from the same arguments as for the proof of the first statement of the theorem. ■

Proof of Theorem 4.1. We follow the proof of Donald and Newey (2001, Lemma A9). Note that $\inf_{W\in\Omega}S_\lambda\left(W\right)=S_\lambda\left(W^*\right)$ for W^* given by (3.2) when $\Omega=\Omega_U$, or W^* given by (3.3) when $\Omega=\Omega_B$. When $\Omega=\Omega_C$ or Ω_P we note that $S_\lambda\left(W\right)$ is continuous in W and Ω is a compact set. Thus $\inf_{W\in\Omega}S_\lambda\left(W\right)=S_\lambda\left(W^*\right)$ for some $W^*\in\Omega$ holds. It then follows that

$$0\leq1-\frac{\inf_{W\in\Omega}S_\lambda\left(W\right)}{S_\lambda\left(\hat{W}\right)}\leq4\sup_{W\in\Omega}\left|\frac{\hat{S}_\lambda\left(W\right)}{S_\lambda\left(W\right)}-1\right|.$$

The result now follows from Lemma 7.8. ■

References

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?, *Quarterly Journal of Economics* **106**(4): 979–1014.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators, *Econometrica* **62**(3): 657–681.

Bekker, P. A. & van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data, *Statistica Neerlandica* **59**(3): 239–267.

Bound, J., Jaeger, D. A. & Baker, R. M. (1996). Problems with instrumental variables estimation when correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* **90**(430): 443–450.

Canay, I. A. (2006). Simultaneous selection and weighting of moments in GMM using a trapezoidal kernel. unpublished manuscript.

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

Chao, J. C. & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments, *Econometrica* **73**(5): 1673–1692.

Donald, S. G. & Newey, W. K. (2001). Choosing the number of instruments, *Econometrica* **69**(5): 1161–1191.

Doornik, J. A. (2006). *An Object-oriented Matrix Programming Language - Ox 4*, Timberlake Consultants Ltd.

Hahn, J. & Hausman, J. (2002). A new specification test of the validity of instrumental variables, *Econometrica* **70**(1): 163–189.

Hahn, J., Hausman, J. & Kuersteiner, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations, *Econometrics Journal* **7**: 272–306.

Han, C. & Phillips, P. C. B. (2006). GMM with many moment conditions, *Econoemtrica* **74**(1): 147–192.

Hansen, B. E. (2007). Least squares model averaging, *Econometrica* **75**(4): 1175–1189.

Hansen, C., Hausman, J. & Newey, W. K. (2006). Estimation with many instrumental varialbes. forthcoming in the Journal of Business and Economic Statistics.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica* **50**(4): 1029–1053.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators, *Journal of Econometrics* **30**(1): 203–238.

Hausman, J., Newey, W. K. & Woutersen, T. (2006). Instrumental variable estimation with heteroskedasticity and many instruments. unpublished manuscript.

Kuersteiner, G. M. (2002). Mean squared error reduction for gmm estimators of linear time series models, *unpublished manuscript* .

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations, *Journal of the American Statistical Association* **75**: 693–700.

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**(3): 958–975.

Mallows, C. L. (1973). Some comments on c_p, *Technometics* **15**: 661–675.

Mariano, R. S. (1972). The existence of moments of the ordinary least squares and two-stage least squares estimators, *Econometrica* **40**: 643 – 652.

Morimune, K. (1983). Approximate distribution of the k-class estimators when the degree of overidentifiability is large compared with the sample size, *Econometrica* **51**(3): 821–841.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations, *Econometrica* **27**(4): 575–595.

Nelson, C. R. & Startz, R. (1990a). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one, *Journal of Business* **63**(1): S125–S140.

Nelson, C. R. & Startz, R. (1990b). Some further results on the exact small sample properties of the instrumental variable estimator, *Econometrica* **58**(4): 967–976.

Newey, W. K. & Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica* **72**(1): 219–255.

Newey, W. K. & Windmeijer, F. (2007). GMM with many weak moment conditions. unpublished manuscript.

Okui, R. (2007). Instrumental variable estimation in the presence of many moment conditions. unpublished manuscript.

Phillips, P. C. B. (1980). The exact finite sample density of instrumental variables estimators in an equation with n+1 endogenous variables, *Econometrica* **48**: 861–878.

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model, *in* Z. Griliches & M. D. Intriligator (eds), *Handbook of Econometris*, Vol. 1, North-Holland Publishing Company, chapter 8.

Phillips, P. C. B. (1989). Partially identified models, *Econometric Theory* **5**: 181–240.

Politis, D. N. (2001). On nonparametric function estimation with infinite-order flat-top kernels, *in* C. C. et al. (ed.), *Probability and Statistical Models with applications*, Chapman and Hall, pp. 469–483.

Politis, D. N. (2006). Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. mimeo.

Politis, D. N. & Romano, J. P. (1995). Bias-corrected nonparametric spectral estimation, *Journal of Time Series Analysis* **16**.

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator, *Journal of the American Statistical Association* **63**: 1214–1226.

Sargan, J. D. (1983). Identification and lack of identification, *Econometrica* **51**(6): 1605–1633.

Staiger, D. & Stock, J. H. (1997). Instrumental variables regression with weak instruments, *Econometrica* **65**(3): 557–586.

Stock, J. H. & Wright, J. H. (2000). GMM with weak identification, *Econometrica* **68**: 1097–1126.

Stock, J. H. & Yogo, M. (2005). Asymptotic distribution of instrumental variables statistics with many weak instruments, *in* D. W. K. Andrews & J. H. Stock (eds), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, pp. 109–120.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables, *Theory of Probability and its Applications* **5**(3): 302–305.

Table 1: Monte Carlo results: Model (a), $R_f^2 = 0.1$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0658	0.0548	0.0549	0.0659	0.0695	0.0701	0.0596
$K = 20$	IQR	0.24	0.372	0.382	0.246	0.256	0.261	0.339
	MAD	0.132	0.204	0.197	0.133	0.135	0.142	0.183
	RMAD	0.651	1	0.966	0.654	0.665	0.697	0.899
	KW+	20	9.32	6.35	80.5	39	9.85	5.74
	KW-	0	0	0	71.8	30.9	0	0
$n = 1000$								
$K = 30$	bias	0.0145	0.0136	0.0148	0.0158	0.0166	0.0144	0.0155
	IQR	0.113	0.114	0.125	0.11	0.115	0.115	0.114
	MAD	0.0569	0.0568	0.0658	0.0563	0.0583	0.0572	0.0578
	RMAD	1	1	1.16	0.991	1.03	1.01	1.02
	KW+	30	29	15.5	129	67.3	23.5	21.7
	KW-	0	0	0	108	47.5	0	0
$c = 0.5$								
$n = 100$	bias	0.324	0.313	0.308	0.323	0.319	0.317	0.316
$K = 20$	IQR	0.212	0.35	0.354	0.221	0.23	0.232	0.316
	MAD	0.325	0.376	0.364	0.323	0.32	0.317	0.346
	RMAD	0.865	1	0.967	0.859	0.851	0.843	0.919
	KW+	20	8.16	5.85	82.6	27.4	9.53	5.4
	KW-	0	0	0	74.6	18.9	0	0
$n = 1000$								
$K = 30$	bias	0.0991	0.136	0.11	0.139	0.118	0.11	0.118
	IQR	0.106	0.16	0.129	0.107	0.0909	0.108	0.115
	MAD	0.102	0.147	0.118	0.139	0.118	0.111	0.12
	RMAD	0.695	1	0.807	0.947	0.806	0.756	0.816
	KW+	30	12.6	12.8	206	51.3	15.5	12.2
	KW-	0	0	0	196	37.5	0	0
$c = 0.9$								
$n = 100$	bias	0.573	0.566	0.561	0.587	0.565	0.572	0.567
$K = 20$	IQR	0.153	0.404	0.384	0.154	0.179	0.168	0.257
	MAD	0.573	0.617	0.6	0.587	0.565	0.572	0.57
	RMAD	0.929	1	0.973	0.951	0.916	0.927	0.924
	KW+	20	3.85	3.35	77.8	13.9	7.5	3.6
	KW-	0	0	0	72.1	4.63	0	0
$n = 1000$								
$K = 30$	bias	0.185	0.234	0.203	0.299	0.189	0.215	0.23
	IQR	0.0866	0.268	0.209	0.089	0.0928	0.093	0.115
	MAD	0.185	0.271	0.222	0.299	0.189	0.215	0.23
	RMAD	0.68	1	0.818	1.1	0.695	0.792	0.846
	KW+	30	2.88	3.51	175	17	8.6	4.86
	KW-	0	0	0	168	3.77	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.

Table 2: Monte Carlo results: Model (a), $R_f^2 = 0.01$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0981	0.035	0.05	0.0974	0.102	0.0951	0.0886
$K = 20$	IQR	0.279	0.773	0.775	0.3	0.322	0.322	0.66
	MAD	0.166	0.393	0.399	0.166	0.178	0.183	0.3
	RMAD	0.422	1	1.01	0.422	0.454	0.464	0.763
	KW+	20	4.6	3.29	72.7	37.4	7.46	3.17
	KW-	0	0	0	65.9	31	0	0
$n = 1000$								
$K = 30$	bias	0.0681	0.0642	0.0576	0.0643	0.0818	0.0716	0.0956
	IQR	0.218	0.662	0.638	0.231	0.25	0.266	0.471
	MAD	0.114	0.351	0.34	0.123	0.134	0.135	0.23
	RMAD	0.326	1	0.971	0.352	0.383	0.385	0.657
	KW+	30	7.48	5.24	187	84	10.5	4.61
	KW-	0	0	0	179	75.3	0	0
$c = 0.5$								
$n = 100$	bias	0.481	0.447	0.452	0.476	0.485	0.484	0.493
$K = 20$	IQR	0.25	0.803	0.798	0.267	0.292	0.285	0.614
	MAD	0.481	0.593	0.591	0.476	0.491	0.486	0.521
	RMAD	0.811	1	0.998	0.804	0.828	0.82	0.878
	KW+	20	4.43	3.22	71.4	26.7	7.24	2.98
	KW-	0	0	0	64.8	19.6	0	0
$n = 1000$								
$K = 30$	bias	0.371	0.361	0.364	0.376	0.382	0.371	0.393
	IQR	0.197	0.621	0.569	0.212	0.244	0.231	0.432
	MAD	0.371	0.478	0.476	0.376	0.385	0.371	0.413
	RMAD	0.777	1	0.995	0.786	0.805	0.777	0.864
	KW+	30	6.72	5.02	187	46.8	10.1	4.32
	KW-	0	0	0	179	35.6	0	0
$c = 0.9$								
$n = 100$	bias	0.855	0.847	0.854	0.854	0.85	0.856	0.868
$K = 20$	IQR	0.14	0.438	0.438	0.148	0.162	0.163	0.367
	MAD	0.855	0.878	0.876	0.854	0.852	0.856	0.869
	RMAD	0.973	1	0.997	0.972	0.971	0.975	0.989
	KW+	20	4.02	3.04	71.1	15.1	7.02	2.74
	KW-	0	0	0	64.7	6.26	0	0
$n = 1000$								
$K = 30$	bias	0.677	0.682	0.675	0.682	0.681	0.675	0.686
	IQR	0.122	0.456	0.426	0.129	0.14	0.142	0.283
	MAD	0.677	0.725	0.709	0.682	0.681	0.675	0.688
	RMAD	0.934	1	0.978	0.942	0.939	0.932	0.949
	KW+	30	3.66	3.04	170	28.7	8.1	3.04
	KW-	0	0	0	164	15.3	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.

Table 3: Monte Carlo results: Model (b), $R_f^2 = 0.1$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0647	0.0303	0.0369	0.0649	0.0583	0.0481	0.0401
$K = 20$	IQR	0.239	0.335	0.326	0.242	0.254	0.264	0.324
	MAD	0.13	0.171	0.169	0.133	0.134	0.142	0.163
	RMAD	0.758	1	0.991	0.78	0.784	0.831	0.953
	KW+	20	7.06	5.39	87	35.7	9.06	4.33
	KW-	0	0	0	78.2	28	0	0
$n = 1000$								
$K = 30$	bias	0.0184	0.00917	0.00934	0.0181	0.0129	0.013	0.00838
	IQR	0.108	0.127	0.122	0.114	0.117	0.118	0.125
	MAD	0.058	0.0638	0.0624	0.0574	0.0586	0.0591	0.0635
	RMAD	0.909	1	0.979	0.901	0.918	0.926	0.996
	KW+	30	15	12.8	245	66.2	16.6	10.3
	KW-	0	0	0	231	51.9	0	0
$c = 0.5$								
$n = 100$	bias	0.309	0.191	0.167	0.3	0.275	0.253	0.185
$K = 20$	IQR	0.223	0.345	0.344	0.233	0.231	0.265	0.33
	MAD	0.309	0.25	0.237	0.3	0.276	0.258	0.241
	RMAD	1.23	1	0.947	1.2	1.1	1.03	0.962
	KW+	20	5.53	4.48	85.2	28.2	8.24	3.93
	KW-	0	0	0	78.2	20.3	0	0
$n = 1000$								
$K = 30$	bias	0.106	0.0392	0.0341	0.0903	0.0652	0.0493	0.0379
	IQR	0.107	0.132	0.133	0.111	0.103	0.126	0.133
	MAD	0.108	0.071	0.0749	0.0942	0.0722	0.0738	0.0735
	RMAD	1.52	1	1.06	1.33	1.02	1.04	1.04
	KW+	30	7.71	7.02	215	58.3	9.32	6.82
	KW-	0	0	0	209	48	0	0
$c = 0.9$								
$n = 100$	bias	0.572	0.249	0.227	0.536	0.439	0.406	0.278
$K = 20$	IQR	0.146	0.372	0.359	0.172	0.183	0.205	0.313
	MAD	0.572	0.316	0.278	0.536	0.439	0.406	0.304
	RMAD	1.81	1	0.881	1.69	1.39	1.28	0.962
	KW+	20	2.68	2.5	61.9	15.2	5.79	2.66
	KW-	0	0	0	57	7.37	0	0
$n = 1000$								
$K = 30$	bias	0.189	0.064	0.052	0.15	0.0957	0.077	0.0669
	IQR	0.0891	0.133	0.134	0.104	0.104	0.119	0.126
	MAD	0.189	0.0893	0.0788	0.15	0.0975	0.0866	0.0864
	RMAD	2.12	1	0.883	1.68	1.09	0.97	0.967
	KW+	30	5.02	4.62	138	33.5	6.57	4.65
	KW-	0	0	0	133	23.7	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.

Table 4: Monte Carlo results: Model (b), $R_f^2 = 0.01$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0953	0.0361	0.0452	0.0976	0.095	0.0845	0.0656
$K = 20$	IQR	0.291	0.712	0.692	0.3	0.325	0.341	0.562
	MAD	0.167	0.359	0.351	0.173	0.181	0.184	0.293
	RMAD	0.466	1	0.979	0.481	0.504	0.513	0.817
	KW+	20	4.59	3.33	72.3	36.8	7.47	3.09
	KW-	0	0	0	65.2	30.4	0	0
$n = 1000$								
$K = 30$	bias	0.0679	0.0267	0.0387	0.0683	0.0693	0.0575	0.0496
	IQR	0.209	0.416	0.371	0.226	0.236	0.252	0.335
	MAD	0.116	0.211	0.197	0.119	0.127	0.129	0.173
	RMAD	0.55	1	0.932	0.564	0.602	0.613	0.817
	KW+	30	7.8	6.04	213	75.2	11.3	4.55
	KW-	0	0	0	204	65.7	0	0
$c = 0.5$								
$n = 100$	bias	0.475	0.392	0.387	0.47	0.474	0.462	0.419
$K = 20$	IQR	0.26	0.71	0.699	0.261	0.296	0.301	0.578
	MAD	0.475	0.55	0.534	0.47	0.478	0.463	0.454
	RMAD	0.864	1	0.97	0.855	0.87	0.842	0.827
	KW+	20	4.44	3.24	74	26.8	7.36	3.03
	KW-	0	0	0	67.3	19.4	0	0
$n = 1000$								
$K = 30$	bias	0.377	0.196	0.193	0.365	0.331	0.303	0.227
	IQR	0.195	0.383	0.392	0.203	0.218	0.242	0.354
	MAD	0.377	0.284	0.278	0.365	0.332	0.306	0.271
	RMAD	1.33	1	0.98	1.29	1.17	1.08	0.957
	KW+	30	5.59	4.64	200	42.3	9.8	3.94
	KW-	0	0	0	192	30.3	0	0
$c = 0.9$								
$n = 100$	bias	0.852	0.741	0.724	0.848	0.828	0.824	0.758
$K = 20$	IQR	0.137	0.578	0.545	0.145	0.174	0.172	0.415
	MAD	0.852	0.798	0.784	0.848	0.828	0.824	0.761
	RMAD	1.07	1	0.982	1.06	1.04	1.03	0.954
	KW+	20	3.6	2.81	68	16.6	6.73	2.7
	KW-	0	0	0	62	7.94	0	0
$n = 1000$								
$K = 30$	bias	0.67	0.296	0.288	0.644	0.561	0.496	0.341
	IQR	0.121	0.426	0.402	0.141	0.151	0.185	0.318
	MAD	0.67	0.395	0.362	0.644	0.562	0.496	0.354
	RMAD	1.69	1	0.914	1.63	1.42	1.25	0.895
	KW+	30	2.12	2	135	28	5.82	2.39
	KW-	0	0	0	130	14.9	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.

Table 5: Monte Carlo results: Model (c), $R_f^2 = 0.1$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0652	0.0636	0.0728	0.0631	0.0684	0.0621	0.0758
$K = 20$	IQR	0.245	0.351	0.36	0.252	0.26	0.254	0.326
	MAD	0.133	0.188	0.19	0.138	0.139	0.137	0.171
	RMAD	0.705	1	1.01	0.733	0.737	0.729	0.91
	KW+	20	9.92	7.13	77.1	36.7	9.96	6.29
	KW-	0	0	0	68.4	28.8	0	0
$n = 1000$								
$K = 30$	bias	0.0205	0.0175	0.0241	0.0202	0.019	0.0193	0.0182
	IQR	0.112	0.117	0.113	0.114	0.116	0.114	0.116
	MAD	0.0578	0.0576	0.0578	0.0579	0.0582	0.0571	0.0573
	RMAD	1	1	1	1	1.01	0.992	0.995
	KW+	30	23.2	15.5	193	61.1	21	18.2
	KW-	0	0	0	179	46	0	0
$c = 0.5$								
$n = 100$	bias	0.324	0.347	0.368	0.322	0.317	0.318	0.333
$K = 20$	IQR	0.22	0.35	0.34	0.227	0.249	0.241	0.29
	MAD	0.324	0.398	0.413	0.323	0.323	0.318	0.355
	RMAD	0.814	1	1.04	0.812	0.811	0.799	0.893
	KW+	20	8.49	6.38	79.5	26	9.45	5.76
	KW-	0	0	0	71.9	17.5	0	0
$n = 1000$								
$K = 30$	bias	0.101	0.103	0.124	0.0874	0.086	0.087	0.0857
	IQR	0.107	0.117	0.111	0.115	0.138	0.112	0.11
	MAD	0.104	0.112	0.126	0.0923	0.0949	0.0912	0.0919
	RMAD	0.932	1	1.13	0.825	0.848	0.815	0.822
	KW+	30	16.6	13.9	183	42.7	14.1	12
	KW-	0	0	0	177	30.3	0	0
$c = 0.9$								
$n = 100$	bias	0.576	0.742	0.746	0.556	0.591	0.567	0.613
$K = 20$	IQR	0.158	0.536	0.456	0.176	0.178	0.174	0.231
	MAD	0.576	0.809	0.797	0.556	0.591	0.567	0.613
	RMAD	0.712	1	0.985	0.687	0.731	0.701	0.758
	KW+	20	3.33	2.92	72.8	12.5	7.42	3.64
	KW-	0	0	0	67.1	2.96	0	0
$n = 1000$								
$K = 30$	bias	0.187	0.783	0.798	0.136	0.209	0.159	0.167
	IQR	0.092	0.844	0.772	0.101	0.115	0.101	0.102
	MAD	0.187	0.901	0.886	0.136	0.209	0.159	0.167
	RMAD	0.207	1	0.984	0.15	0.232	0.176	0.185
	KW+	30	1.12	1.19	125	18.8	8.96	6
	KW-	0	0	0	120	6.41	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.

Table 6: Monte Carlo results: Model (c), $R_f^2 = 0.01$

		2SLS	DN	KW	MA-U	MA-C	MA-P	MA-Ps
$c = 0.1$								
$n = 100$	bias	0.0871	0.0695	0.0873	0.0995	0.117	0.113	0.121
$K = 20$	IQR	0.284	0.861	0.886	0.313	0.329	0.328	0.687
	MAD	0.167	0.447	0.451	0.178	0.19	0.184	0.348
	RMAD	0.373	1	1.01	0.397	0.425	0.411	0.778
	KW+	20	4.4	3.17	70.4	36.5	7.33	3.08
	KW-	0	0	0	63.7	30.1	0	0
$n = 1000$								
$K = 30$	bias	0.0699	0.0875	0.0847	0.0637	0.0666	0.0765	0.108
	IQR	0.21	0.803	0.789	0.212	0.227	0.25	0.465
	MAD	0.12	0.433	0.409	0.119	0.123	0.135	0.243
	RMAD	0.277	1	0.945	0.275	0.284	0.312	0.561
	KW+	30	7.91	5.64	171	79.4	10.3	5.02
	KW-	0	0	0	163	70.6	0	0
$c = 0.5$								
$n = 100$	bias	0.485	0.505	0.5	0.483	0.495	0.494	0.535
$K = 20$	IQR	0.265	0.775	0.756	0.278	0.292	0.306	0.604
	MAD	0.485	0.653	0.634	0.483	0.502	0.494	0.572
	RMAD	0.744	1	0.971	0.74	0.77	0.757	0.876
	KW+	20	4.2	3.1	71.5	26.5	7.21	2.93
	KW-	0	0	0	64.9	19.4	0	0
$n = 1000$								
$K = 30$	bias	0.362	0.405	0.425	0.356	0.373	0.374	0.417
	IQR	0.183	0.737	0.729	0.199	0.213	0.211	0.46
	MAD	0.362	0.572	0.575	0.356	0.379	0.374	0.439
	RMAD	0.633	1	1	0.622	0.662	0.654	0.766
	KW+	30	6.79	5.02	174	46.4	10	4.71
	KW-	0	0	0	167	34.7	0	0
$c = 0.9$								
$n = 100$	bias	0.855	0.878	0.878	0.854	0.859	0.861	0.879
$K = 20$	IQR	0.145	0.385	0.374	0.156	0.171	0.163	0.332
	MAD	0.855	0.893	0.899	0.854	0.859	0.861	0.879
	RMAD	0.958	1	1.01	0.957	0.962	0.965	0.984
	KW+	20	4.12	3.04	72	13.7	7.14	2.74
	KW-	0	0	0	65.3	4.61	0	0
$n = 1000$								
$K = 30$	bias	0.67	0.805	0.817	0.652	0.68	0.666	0.734
	IQR	0.123	0.597	0.564	0.139	0.14	0.147	0.302
	MAD	0.67	0.898	0.884	0.652	0.68	0.666	0.735
	RMAD	0.746	1	0.984	0.726	0.758	0.742	0.818
	KW+	30	3.92	3.13	165	28.3	8.47	3.32
	KW-	0	0	0	159	14.6	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN2SLS; "KW+" = $\sum_{m=1}^{M} \max(w_m, 0)m$; "KW-" = $\sum_{m=1}^{M} |\min(w_m, 0)|m$.