Instrument Selection by First Stage Prediction Averaging[*]

Guido Kuersteiner[†]and Ryo Okui[‡]

This version: April 2009

Abstract

This paper considers model averaging as a way to select instruments for the two stage least squares and limited information maximum likelihood estimators in the presence of many instruments. We propose averaging across least squares predictions of the endogenous variables obtained from many different choices of instruments and then use the average predicted value of the endogenous variables in the estimation stage. The weights for averaging are chosen to minimize the asymptotic mean squared error. 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 nests and dominates existing number-of-instrument-selection procedures.

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

JEL classification: C21, C31.

[*] We thank Whitney Newey, Xiaohong Chen, Bruce Hansen, Jack Porter, Kaddour Hadri, Ken Ariga, Yuichi Kitamura, Kosuke Oya, Han Hong and seminar participants at Columbia, Georgetown, USC, Yale, CIREQ, Berkeley, Hitotsubashi, Seoul National (SETA08), Kyoto and Osaka as well as two anonymous referees for valuable comments and suggestions. Okui acknowledges financial support from the Hong Kong University of Science and Technology under Project No. DAG05/06.BM16 and from the Research Grants Council under Project No. HKUST643907. 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) and limited information maximum likelihood (LIML) estimators of linear models when there are many instruments available. 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 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 as well as to a modification of LIML. 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 or an improved higher order mean squared error (MSE) of LIML. 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 (1959) 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. Our model averaging approach can also be applied to non-linear procedures such as the LIML estimator. We show that an averaging version of LIML dominates sequential instrument selection in terms of the higher order MSE, although in this case, model averaging LIML estimator (MALIML) does not achieve a better rate of convergence than LIML with sequentially selected instruments.

A limitation of sequential instrument selection 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 and MALIML estimators have the ability to select arbitrary subsets of instruments from an orthogonalized set of instruments. In other words, if there are certain orthogonal directions in the instrument space that are particularly useful for the first stage, our procedure is able to individually select these directions from the instrument set. Conventional sequential instrument selection on the other hand is 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.

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 (2008) and Canay (2008). 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-instruments-selection procedure in finite sample Monte Carlo experiments.

We present the general form of the MA2SLS and MALIML estimators in Section 2.1 and discuss various members of the class of MA2SLS and MALIML estimators in Section 3.2. The refined higher order MSE approximation for the MA2SLS family and the MSE approximation for MALIML are obtained in Section 3.1. Section 3.3 demonstrates that optimal members of the MA2SLS and MALIML families dominate the pure number of instrument selection method for the 2SLS, bias corrected 2SLS and LIML respectively, in terms of relative higher order MSE. In Section 4, we establish that feasible versions of the MA2SLS and MALIML estimator maintain certain optimality properties. Section 5 contains Monte Carlo evidence of the small sample properties of our estimators.

2 First Stage Model Averaging Estimators

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

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

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. The asymptotic variance of a $\sqrt{N}$-consistent regular estimator of β 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. In general, we can approximate f_i better by adding more instruments, which makes the estimator more efficient. However, the estimator might behave poorly in the presence of many instruments (Kunitomo (1980), Morimune (1983) and Bekker (1994)). This paper develops model averaging methods to handle a large number of instruments.

Let $y = (y_1, \ldots, y_N)'$. The matrices X, ϵ, u and f are defined similarly.

2.1 Model Averaging

Let W be a weighting vector such that $W = (w_{1,N}, \ldots, w_{M,N})$ and $\sum_{m=1}^{M} w_{m,N} = 1$ for some M such that $M \leq N$ for any N. We note that W is a sequence of weights $w_{m,N}$ indexed by the sample size N, but for notational convenience we use w_m where it does not create confusion. In Sections 3.2 and 4, we discuss in more details 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 $P_m = Z_m(Z_m' Z_m)^{-1} Z_m'$. Define $P(W) = \sum_{i=1}^{M} w_m P_m$. The model averaging two-stage least squares estimator (MA2SLS), $\hat{\beta}$, of β is defined as

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

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

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

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

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

Our estimators can also be extended to a modification of LIML due to Fuller (1977). Let

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

where α is a constant chosen by the econometrician.[1] The model averaging Fuller estimator (MA-Fuller) then is defined as

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

We use the term MA2SLS or MALIML 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 and LIML as well. In particular, for 2SLS, more specification bias in the reduced form leads to less estimator bias, and reduced variance in the reduced form 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 instrumental variables estimators 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.

2.2 Advantages of Model Averaging

To give a preview of our results, we focus our attention to MA2SLS, $\hat{\beta}$, in this subsection. 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 and negative. In fact, the Nagar estimator can be interpreted as a special case of MA2SLS with $M = N$, $w_j = N/\left(N - m\right)$ for $j = m$, some m,

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

$w_N = -m/(N-m)$ and $w_j = 0$ otherwise.[2] 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 (2.2) with $P(W)$ replaced by $Z_M k(Z_M' Z_M)^{-1} k Z_M'$. For expositional purposes and to relate kernel weighting to model averaging, we consider a special case in which instruments are mutually orthogonal so that $Z_M' Z_M$ is a diagonal matrix, but note that similar results hold in the general case.[3] 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

$$\sum_{m=1}^{M} w_m P_m = \sum_{j=1}^{M} k_j^2 \tilde{P}_j = Z_M k(Z_M' Z_M)^{-1} k Z_M' \tag{2.5}$$

holds. In other words, the kernel weighted 2SLS estimator corresponds to model averaging with the weights $\{w_m\}_{m=1}^{M}$ defined above.

Okui's (2008) shrinkage 2SLS estimator is also a special case of the averaged estimator (2.2). In this case, $w_L = s, w_M = 1-s, s \in [0,1], w_j = 0$ 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 (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

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

[3] 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 model averaging weights corresponding to some particular kernel become data dependent and have a more complicated formula.

of $k(\cdot)$. This later problem was recently addressed by Canay (2008) through the use of top-flat kernels (see, e.g., Politis and Romano (1995), Politis (2001) and Politis (2007)).

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.

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 for the 2SLS case. We show the asymptotic properties of the MA2SLS and LIML 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 $\alpha > 1/2$,*

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

(iii) Let $\mathbb{N}_+$ be the set of positive integers. There exists a subset $\bar{J} \subset \mathbb{N}_+$ with a finite number of elements such that for all $m \notin \bar{J}$ it follows that

$$\inf_{m \notin \bar{J}, m \leq M} m^{2\alpha+1} \left(\sup_{\lambda'\lambda=1} \lambda' f \left(P_m - P_{m+1} \right) f \lambda / N \right) > 0 \textit{ wpa1}$$

Assumption 3 *(i) Let u_{ia} be the a-th element of u_i. Then $E[\epsilon_i^r u_{ia}^s | z_i]$ are constant for all a and $r, s \geq 0$ and $r + s \leq 5$. 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) $\max_{i \leq N} P_{M,ii} \rightarrow_p 0$, where $P_{M,ii}$ signifies the (i,i)-th element of P_M. (iv) f_i is bounded.*

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

Assumption 5 *It holds either that i) $K'W^+/\sqrt{N} = \sum_{m=1}^{M} |w_m| m / \sqrt{N} \rightarrow 0$ or ii) $K'W^+/N = \sum_{m=1}^{M} |w_m| m / N \rightarrow 0$ and $M/N \rightarrow 0$.*

Assumption 6 *The eigenvalues of $E\left[Z_{k,i} Z_{k,i}' \right]$ are bounded away from zero uniformly in k. Let $\bar{H}_k = E\left[f_i Z_{k,i} \right] \left(E\left[Z_{k,i} Z_{k,i}' \right] \right)^{-1} E\left[f_i Z_{k,i}' \right]'$ and $\bar{H} = E\left[f_i f_i' \right]$. Then, $\left\| \bar{H}_k - \bar{H} \right\| = O\left(k^{-2\alpha} \right)$ for $k \rightarrow \infty$.*

Assumption 7 *$\beta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^d$.*

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

Assumptions 1-3 are similar to those imposed in Donald and Newey (2001). The set $\bar{J}$ corresponds to the set of redundant. We need to explicitly consider this set because it turns out that the optimal weight on a redundant instrument has some specific feature (see Section 7.5). Assumption 4 collects the conditions that weights must satisfy and is related to the conditions imposed by Donald and Newey (2001) on the number of instruments. The condition $K'W^+ \rightarrow \infty$ 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. For the LIML estimator, this condition can be weakened to $K'W^+/N \to 0$. The condition $\sup_{j \notin \bar{J}, j \leq L} |\sum_{m=1}^{j} w_m| = O(1/\sqrt{N})$ guarantees that small models receive asymptotically negligible weight and is needed to guarantee first order asymptotic efficiency of the MA2SLS and MALIML estimators. We also restrict W to lie in the space of absolutely summable sequences l_1. The fact that the sequences in l_1 have infinitely many elements creates no problems since one can always extend W to l_1 by setting $w_j = 0$ for all $j > M$.

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 $(\hat{r}(W) + T(W))/\text{tr}(S(W)) = o_p(1)$ as $N \to \infty$.

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 $\text{Cum}\,[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ and $E[\epsilon_i^2 u_i] = 0$,[4] we show in Corollary 7.1 that for MA2SLS

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

where $a_\sigma = \sigma_{u\epsilon}\sigma_{u\epsilon}'$, $b_\sigma = (\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')$,

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

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.[5] The first term in (3.1) represents the square of the bias, and the fourth term represents the goodness-of-fit of the first

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

[5] As was noted in Donald and Newey (2001), it is possible to use the more general criterion that allows $\text{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] \neq 0$ 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.

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 the first term $a_\sigma(K'W)^2/N$ 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.

For the MALIML estimator the approximate MSE when $\operatorname{Cum}[\epsilon_i, \epsilon_i, v_i, v_i'] = 0$ and $E[\epsilon_i^2 v_i] = 0$ with $v_i = u_i - (\sigma_{u\epsilon}/\sigma_\epsilon^2)\epsilon_i$ is found as

$$S(W) = H^{-1}\Big(\sigma_\epsilon^2 \Sigma_v \frac{W'\Gamma W}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N}\Big) H^{-1}. \qquad (3.2)$$

where $\Sigma_v = E\left[v_i v_i'\right]$. Higher order unbiasedness of MALIML is reflected in the absence of terms involving $K'W$ and parallels results for the sequential instrument selection case in Donald and Newey (2001).

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 $\min_{W\in\Omega} S_\lambda(W)$ where $S_\lambda(W) = \lambda' S(W)\lambda$ and Ω is some set.

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 $\Omega_C = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [-1,1], \forall m \le M\}$, and $\Omega_P = \{W \in l_1 | W'\mathbf{1}_M = 1; w_m \in [0,1], \forall m \le M\}$.

For the MALIML estimator we only consider cases where W is contained in Ω_U, Ω_C and Ω_P because MALIML is higher order unbiased without the constraint $K'W = 0$.

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 $\lambda' H^{-1} f'(I-$

$P(W))(I-P(W))fH^{-1}\lambda = W'UW$ such that $S(W)$ is affine in W. It then is easy to show that the optimal unconstrained weights for MA2SLS are

$$W_U^* = \arg\min_{W\in\Omega_U} S_\lambda(W) = \frac{1}{2}A^{-1}\left(K\lambda' H^{-1}B_N H^{-1}\lambda + \frac{2-\mathbf{1}_M' A^{-1}K\lambda' H^{-1}B_N H^{-1}\lambda}{\mathbf{1}_M' A^{-1}\mathbf{1}_M}\mathbf{1}_M\right), \quad (3.3)$$

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 MA2SLS 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^*.

$$W_B^* = \arg\min_{W\in\Omega_B} S_\lambda(W) = A_B^{-1}R\left(R'A_B^{-1}R\right)^{-1}b, \quad (3.4)$$

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. 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 simplified to (7.4).

The optimal weights in Ω_U for MALIML have the same form as (3.4) except that now $A_L = \left(\sigma_\epsilon^2\sigma_\lambda^2 - \sigma_{\lambda\epsilon}^2\right)\Gamma + \sigma_\epsilon^2 U$ replaces A_B, $R = \mathbf{1}_M$ and $b = 1$ such that the optimal weights are

$$W_{U,LIML}^* = A_L^{-1}\mathbf{1}_M\left(\mathbf{1}_M' A_L^{-1}\mathbf{1}_M\right)^{-1}.$$

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 $w_j = 0$ 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 $\Omega = \Omega_U, \Omega_B, ..., \Omega_P$

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

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 7.8 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, MA2SLS and MALIML dominate corresponding estimators based on sequential moment selection.

Theorem 3.1 *Assume that Assumptions 1-5 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* $C(a) = (1+a)^{-2\alpha} + o\left(|a|^{2\alpha}\right)$.
i) For $S_\lambda(W)$ *given by (3.1), it follows that*

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

Letting 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* $S_\lambda(W)$, *it follows that*

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

ii) For $S_\lambda(W)$ *given by (3.2), it follows that*

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

Remark 2 *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-instruments-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. The third part of the result shows that MALIML dominates LIML with sequential instrument selection for $W \in \Omega_U, \Omega_C$ and Ω_P.

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 (2.5)). 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 (1973) 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

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

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:

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

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 for MA2SLS and the corresponding LIML estimator for MALIML. 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,

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

Let $\hat{u}_\lambda^m = (P_M - P_m)X\hat{H}^{-1}\lambda$ and $\hat{U} = (\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)'(\hat{u}_\lambda^1, \ldots, \hat{u}_\lambda^M)$. 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 $\hat{a}_\lambda = \hat{\sigma}_{\lambda\epsilon}^2$, $\hat{b}_\lambda = \hat{\sigma}_\epsilon^2\hat{\sigma}_\lambda^2 + \hat{\sigma}_{\lambda\epsilon}^2$ and $\hat{B}_{\lambda,N} = \lambda'\hat{H}^{-1}\hat{B}_N\hat{H}^{-1}\lambda$, where $\hat{B}_N$ is some estimator of $\hat{B}_N$.[8] When the weights are only allowed to be positive, Corollary 7.2 suggests the simpler criterion

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

For MALIML we choose W based on the following criterion

$$\hat{S}_\lambda(W) = (\hat{\sigma}_\epsilon^2\hat{\sigma}_\lambda^2 - \hat{\sigma}_{\lambda\epsilon}^2)\frac{W'\Gamma W}{N} + \hat{\sigma}_\epsilon^2 \left(\frac{W'\hat{U}W - \hat{\sigma}_\lambda^2(M - 2K'W + W'\Gamma W)}{N}\right). \tag{4.3}$$

[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.

[8] When $\dim(\beta) = 1$, we have $B_N = 2(\sigma_\epsilon^2\Sigma_u + 4\sigma_{u\epsilon}^2)$ and we may use $\hat{B}_{\lambda,N} = 2(\hat{\sigma}_\epsilon^2\sigma_\lambda^2 + 4\hat{\sigma}_{\lambda\epsilon}^2)$.

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.

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

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

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

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

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

Theorem 4.1 *Let Assumptions 1-10 hold. For $\Omega = \Omega_U$, Ω_B, Ω_C, or Ω_P and $\hat{W} = \arg\min_{W\in\Omega}\hat{S}_\lambda(W)$ where $\hat{S}_\lambda(W)$ is defined in either (4.1) or (4.3) it follows that*

$$\frac{\hat{S}_\lambda\left(\hat{W}\right)}{\inf_{W\in\Omega} S_\lambda(W)} \to_p 1. \tag{4.4}$$

Theorem 4.1 complements the result in Hansen (2007). Apart from the fact that $\hat{S}_\lambda(W)$ is different from the criterion in Hansen (2007), there are more technical differences between our result and Hansen's (2007). Hansen (2007) shows (4.4) 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,[9] where we investigate the finite sample properties of the model averaging estimators. In particular, we examine the performance of the model averaging 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.

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 $i = 1, \dots, N$, 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.

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

[9] This Monte Carlo simulation was conducted with Ox 5.10 (Doornik (2007)) for Windows.

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.

5.2 MA2SLS

We first examine the performances of 2SLS-type estimators.

5.2.1 Estimators

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-All in the tables). Second, the 2SLS estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (2SLS-DN). 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 (KGMM). Let $\Omega_{KGMM} = \{W \in l_1 : w_m = L^{-1}$ if $m \leq L$ and 0 otherwise for some $L \leq M\}$. Then, the MA2SLS estimator with $W \in \Omega_{KGMM}$ corresponds to the kernel weighted 2SLS estimator with kernel $k(x) = \sqrt{\max(1 - x, 0)}$. Because the weights are always positive with Ω_{KGMM}, we use the criterion function (4.2) for KGMM. We use a grid search to find the L that minimizes the criterion. The procedure "2SLS-U" is the MA2SLS estimator with $\Omega = \Omega_U = \{W \in l_1 : W'\mathbf{1}_M = 1\}$. 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 "2SLS-C". The procedure "2SLS-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 2SLS-U, 2SLS-C and 2SLS-P is formula (4.1). The procedure "2SLS-Ps" also uses the same set Ω_P, but the criterion for computing weights is (4.2). For these MA2SLS estimators, we use the procedure SolveQP in Ox to minimize the criteria (see Doornik (2007)). 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").[10] 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 2SLS-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 2SLS-U or 2SLS-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 2SLS-U and 2SLS-C procedures.

5.2.2 Results

Tables 1-3 summarize the results of our simulation experiment. 2SLS-All performs well when the degree of endogeneity is small ($c = 0.1$). However, when $c = 0.5$ or 0.9, 2SLS-All exhibits large bias and some method to alleviate this problem is called for. The selection method of 2SLS-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, 2SLS-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), 2SLS-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), 2SLS-DN seldom outperforms 2SLS-All. In particular, in Model (c), the number of instruments selected by 2SLS-DN tends to be far less than $M/2$, which means that 2SLS-DN often employs only the instruments that are uncorrelated with the endogenous regressor. KGMM typically outperforms 2SLS-DN, which demonstrates the advantage of kernel weighting. However, the problem observed for DN also applies to KGMM. KGMM does not improve over 2SLS-All in Models (a) and (c).

All model averaging estimators perform well. 2SLS-Ps, which may be considered a natural application of Hansen's (2007) model averaging to IV estimation, outperforms 2SLS-DN and KGMM in most cases. 2SLS-P further improves over 2SLS-Ps in Models (a) and (c) substantially, which shows the benefit of taking additional higher order terms into account when choosing optimal weights. The good performance of 2SLS-P is mainly due to its low variability measured by IQR. On the other hand, in Model (b), 2SLS-P is outperformed by 2SLS-DN when $c = 0.9$. Nevertheless, the RMAD measure is never above 1.3 which is significantly lower than the RMAD measure for 2SLS-All. This result may be due partly to a trade-off between additional terms in the approximation of the MSE

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

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. 2SLS-U and 2SLS-C also perform well. Their performance is particularly remarkable when $c = 0.1$ and $n = 100$ in Models (a) and (b) and when $c = 0.9$ and $n = 1000$ in Model (c). We note that the performance of 2SLS-C and that of 2SLS-U are similar. In general, the performances of 2SLS-C and 2SLS-U are similar to 2SLS-P. However, the performance of 2SLS-U or 2SLS-C in Model (b) is not as good as that of DN when $c = 0.9$ with a RMAD measure reaching values of around 1.69. Nevertheless, 2SLS-U and 2SLS-C perform better than 2SLS-All even in these cases. Their relatively poor performance in Model (b) may be due to having too large a choice set for W. Note also that the values of "KW+" and "KW-" for 2SLS-U and 2SLS-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, 2SLS-Ps displays the most robust performance of all procedures considered. It almost never falls behind DN in terms of MAD and often outperforms it significantly. 2SLS-U, 2SLS-C and 2SLS-P show some problems in Model (b) when the degree of endogeneity is moderate to high. On the other hand, those estimators achieve even more significant improvements in terms of MAD over 2SLS-DN in Models (a) and especially (c) where DN does not perform very well.

5.3 MALIML

Next, we consider the performances of LIML-type estimators.

5.3.1 Estimators

We compare the performances of the following five estimators. First, we consider the LIML estimator with all available instruments (LIML-All in the tables). Second, the LIML estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (LIML-DN). We use the criterion function (4.3) for LIML-DN. The optimal number of instruments is obtained by a grid search. The procedures "LIML-U", "LIML-C" and "LIML-P are the MALIML estimators with $\Omega = \Omega_U = \{W \in l_1 : W'\mathbf{1}_M = 1\}$, $\Omega = \Omega_C = \{W \in l_1 : W'\mathbf{1_M} = 1; w_m \in [-1, 1], \forall m \leq M\}$ and $\Omega = \Omega_P = \{W \in l_1 : W'\mathbf{1_M} = 1; w_m \in [0, 1], \forall m \leq M\}$, respectively. For these MALIML estimators, we minimize the criterion (4.3) to obtain optimal weights. We use the procedure SolveQP

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

in Ox to minimize the criterion (see Doornik (2007)). We use the LIML 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 $S_{\lambda}(W)$.

As in the previous subsection, we compute the same six quantities to evaluate the performances of the estimators. We note that a comparison of KW+ and KW- does not provide an insight into the bias reduction because the estimators considered here do not exhibit a second-order bias. Negative weights might be present in LIML-U or LIML-C solely when the criterion indicates the existence of redundant instruments.

5.3.2 Results

Tables 4-6 summarize the results of our simulation experiment. LIML-All has a small bias in many cases as the theory indicates. Nonetheless, it is severely biased when the instruments are weak ($R_f^2 = 0.01$) and the sample size is small ($n = 100$) (see Hahn, Hausman and Kuersteiner (2004)). Compared with 2SLS-All, the MAD of LIML-All is high when $c = 0.1$ and low when $c = 0.9$. In contrast to the case of 2SLS, the instrument selection works well for LIML in all the specifications. LIML-DN outperforms LIML-All in most of cases, in particular when $c = 0.1$. When $c = 0.9$, they behave similarly.

All model averaging estimators perform well. They improve LIML-All and LIML-DN in many cases. It is interesting to note that the model averaging estimators perform particularly well when LIML-DN improves over LIML-All. For example, they perform better than LIML-DN when $c = 0.1$ and the improvement may be substantial. In Model (b), and $c = .5$ LIML-U and LIML-C perform somewhat less well than LIML-DN but better than LIML-All. Comparing LIML-C (or LIML-U) and LIML-P, we see that LIML-C improves over LIML-P when $c = 0.1$, but the performance of LIML-P is more stable so that the performance of LIML-P is the best among the model averaging estimators when they cannot perform well. We note that the good performance of the model averaging estimators appears to be due to its ability to reduce the variability (measured by IQR).

In summary, model averaging can outperform instrument selection. In particular, the improvement is substantial when instrument selection is important. Allowing negative weights can further improve the performance in those cases, while it also makes the estimator slightly less stable. From the experiment, we recommend LIML-P in particular because it can improve the estimator substantially when it works but its performance is relatively stable even when model averaging does not work well.

5.4 MAFuller

Next, we consider the performances of Fuller-type estimators.

5.4.1 Estimators

Similarly to the experiments for the LIML-type estimators, we compare the performances of the following five estimators. First, we consider the Fuller estimator with all available instruments (Fuller-All in the tables). Second, the Fuller estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (Fuller-DN). We use the criterion function (4.3) (the criterion is the same as that for LIML) for Fuller-DN. The optimal number of instruments is obtained by a grid search. The procedures "Fuller-U", "Fuller-C" and "Fuller-P are the MAFuller estimators with $\Omega = \Omega_U$, $\Omega = \Omega_C$ and $\Omega = \Omega_P$, respectively. For these MAFuller estimators, we minimize the criterion (4.3) to obtain optimal weights. We use the procedure SolveQP in Ox to minimize the criterion (see Doornik (2007)). We use the Fuller 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 $S_\lambda(W)$.

As in the previous subsection, we compute the same six quantities to evaluate the performances of the estimators.

5.4.2 Results

Tables 7-9 summarize the results of our simulation experiment. The performances of Fuller-DN and the model averaging estimators compared with Fuller-All are similar to what we observe in the experiment for LIML-type estimators. Among the Fuller-type estimators, we recommend Fuller-P by the same reason that we recommend LIML-P among the LIML-type estimators (i.e., it can improve the estimator substantially when it works but its performance is relatively stable even when model averaging does not work well.)

Fuller-All has a smaller MAD than that of LIML-All. This good performance of the Fuller estimator comes from that it has small variability (as measured by IQR). This result may be related to the well-known fact that the Fuller estimator has finite moments, but the LIML estimator does not. However, once instrument selection or model averaging is introduced, the Fuller estimator does not necessary outperform the LIML estimator. For example, when $n = 1000$ and $R_f^2 = 0.01$ in Model (a), Fuller-All has a smaller MAD than that of LIML-All, but other Fuller-type estimators are dominated by the corresponding LIML-type estimators.

5.5 MAB2SLS

Lastly, we consider the bias-corrected 2SLS estimator (B2SLS) and the model averaging version of B2SLS. B2SLS is defined as

$$\hat{\beta}_{b2sls} = \left(X' P_m X - \frac{m}{N} X' X \right)^{-1} \left(X' P_m y - \frac{m}{N} X' y \right),$$

when we use the first m instruments. The model averaging B2SLS estimator (MAB2SLS) is

$$\hat{\beta}_{mab2sls} = \left(X' P(W) X - \frac{K'W}{N} X' X \right)^{-1} \left(X' P(W) y - \frac{K'W}{N} X' y \right).$$

5.5.1 Estimators

We compare the performances of the following five estimators. First, we consider the B2SLS estimator with all available instruments (B2SLS-All in the tables). Second, the B2SLS estimator with the number of instruments chosen by Donald and Newey's (2001) procedure is examined (B2SLS-DN). The procedures “B2SLS-U”, “B2SLS-C” and “B2SLS-P are the MAB2SLS estimators with $\Omega = \Omega_U$, $\Omega = \Omega_C$ and $\Omega = \Omega_P$, respectively. The criterion used to compute optimal weights is

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

For B2SLS-DN, the optimal number of instruments is obtained by a grid search. For the MAB2SLS estimators, we use the procedure SolveQP in Ox to minimize the criterion (see Doornik (2007)). We use the B2SLS 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 $S_\lambda(W)$.

As in the previous subsection, we compute the same six quantities to evaluate the performances of the estimators.

5.5.2 Results

Tables 10-12 summarize the results of our simulation experiment. B2SLS-All eliminates bias effectively when $R_f^2 = 0.1$, however, it exhibits large bias when $R_f^2 = 0.01$. The diversity of B2SLS-All with all the instruments (measured by IQR) is smaller than that of LIML-All when $c = 0.1$ but the difference is small. On the other hand, when $c = 0.9$, LIML-All has small diversity and the difference is substantial. In general, we may recommend LIML over B2SLS when we use all the instruments. DN improves All in many cases and the improvement can be substantial. This result is similar to what we have observed in the case of LIML. However, B2SLS-DN perform substantially worse than B2SLS-All does when $c = 0.5$, 0.9 and $n = 1000$ in Models (a) and (c).

B2SLS-P typically improves B2SLS-DN. The performances of B2SLS-U and B2SLS-C are similar to each other and their performance is very unstable. When $c = 0.9$, B2SLS-U and B2SLS-C exhibit non-negligible bias and their MAD is substantially larger than that of B2SLS-DN. On the other hand, when $c = 0.1$, their improvement can be substantial. For example, when $c = 0.1$, $n = 100$ and $R_f^2 = 0.1$ in Model (a), the MAD of B2SLS-U is the smallest among the estimators considered in the experiments (including 2SLS and LIML-type estimators). We note that the good performance of the model averaging estimators appears to be due to its ability to reduce the variability (measured by IQR).

In summary, B2SLS-P demonstrates most stable performance. B2SLS-U and B2SLS-C improve B2SLS-All substantially in some case but there are also cases in which B2SLS-U and B2SLS-C do not work well. Thus, we recommend B2SLS-P when we apply model averaging to the B2SLS estimator. On the other hand, we recommend to avoid using B2SLS-type estimators. The LIML estimator has better bias property and the 2SLS estimator shows less diversity. Model averaging can improve the performances of LIML and 2SLS estimators in a more stable way than that of B2SLS.

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

This section presents the formal theorems and their proofs. In the proofs, C denotes a generic constant whose exact value depends on the context.

Theorem 7.1 *Suppose that Assumptions 1-3 are satisfied. Define $\mu_i(W) = E[\epsilon_i^2 u_i]P_{ii}(W)$ and $\mu(W) = (\mu_1(W), ..., \mu_N(W))'$. If W satisfies Assumption 4 and 5(i) then, for $\hat{\beta}$ defined in (2.2), the decomposition given by (7.7) holds with*

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

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

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

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 $\text{Cum}[\epsilon_i, \epsilon_i, u_i, u_i'] = 0$ and $E[\epsilon_i^2 u_i] = 0$ as would be the case if ϵ_i and u_i were jointly Gaussian, 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 $E[\epsilon_i^2 u_i] = 0$. Then, for $\hat{\beta}$ defined in (2.2), the decomposition given by (7.7) holds with:*

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

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 $\hat{\beta}$ defined in (2.2), the decomposition given by (7.7) holds with:*

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

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

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

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 $f'(I - P(W))\mu\,(W)\,/N$ 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 (7.7) 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* $\sum_{m=1}^{M} |w_m| = (N+k)/(N-k)$ *which is uniformly bounded if* $k = o(N)$, $K'W = 1$, $\mathbf{1}_M' W = 1$, $\sum_{m=1}^{M} |w_m|\, m = 2Nk/(N-k) \to \infty$, $\sum_{m=1}^{M} |w_m|\, m/\sqrt{N} = 2\sqrt{N}k/(N-k) \to 0$. *Further,* $\sup_{j\notin \bar{J}, j\leq L} |\sum_{m=1}^{j} w_m| = 0$ *by taking* $L \leq k$. *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* $S\,(W_N) = H^{-1}((\sigma_\epsilon^2 \Sigma_u + \sigma_{u\epsilon}\sigma_{u\epsilon}')k/N + \sigma_\epsilon^2 f'(I - P_k)f/N)H^{-1} + o\,(S\,(W_N))$. *The lead term is the same as the result in Proposition 3 of Donald and Newey (2001).*

The next theorem gives the approximate MSE of the MALIML and MAFuller estimators.

Theorem 7.2 *Suppose that Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied. Let* $v_i = u_i - (\sigma_{u\epsilon}/\sigma_\epsilon^2)\epsilon_i$. *Define* $\Sigma_v = \Sigma_u - \sigma_{u\varepsilon}\sigma_{u\varepsilon}'$, $\mu_v(W) = (\mu_{v,1}(W), \ldots, \mu_{v,N}(W))'$, $\mu_{v,1}(W) = E[\epsilon_i^2 v_i]P_{ii}(W)$. *If* W *satisfies Assumption 4 then, for* $\hat{\beta}$ *defined in (2.3) (MALIML) and* $\hat{\beta}$ *defined in (2.4) (MAFuller), the decomposition given by (7.7) holds with*

$$\begin{aligned} S(W) &= H^{-1}\Big(\sigma_\epsilon^2 \Sigma_v \frac{W'\Gamma W}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N} + Cum[\epsilon_i, \epsilon_i, v_i, v_i'] \frac{\sum_i (P_{ii}(W))^2}{N} \\ &\quad + \hat{\zeta} + \hat{\zeta}' - \frac{f(I-P(W))\mu_v(W)}{N} - \frac{\mu_v(W)'(I-P(W))f}{N}\Big)H^{-1}, \end{aligned}$$

where

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

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

$$S(W) = H^{-1}\Big(\sigma_\epsilon^2 \Sigma_v \frac{W'\Gamma W}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N}\Big)H^{-1}. \tag{7.6}$$

Theorem 7.3 *Assume that Assumptions 1-5 hold. Suppose that* $\dim(\beta) = 1$. *Let*

$$\tilde{R}^2 = \frac{(X'P(W)X)^2}{X'P(W)P(W)X \cdot X'X}.$$

If $\sum_{j=1}^{L} |w_j| = o(1)$ *and* $E(X_i) = 0$, *then*

$$\tilde{R}^2 \rightarrow_p \frac{E(f_i^2)}{E(f_i^2) + \sigma_u^2}.$$

7.1 Lemmas

The MA2SLS estimator 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 of MA2SLS. 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

$$\begin{aligned} N(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)' &= \hat{Q}(W) + \hat{r}(W), \\ E[\hat{Q}(W)|z] &= \sigma_\epsilon^2 H^{-1} + S(W) + T(W), \qquad (7.7) \\ (\hat{r}(W) + T(W))/tr(S(W)) &= o_p(1), \ \textit{as } K'W^+ \rightarrow \infty, N \rightarrow \infty. \end{aligned}$$

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. $\blacksquare$

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 the $j \times 1$ 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 $y'\Gamma y = \sum_{j=0}^{N-1} \left(y' b_{j,N}\right)^2 \geq 0$ and the equality holds if and only if $y = 0$. This shows that Γ is positive definite. $\blacksquare$

Lemma 7.3 *Let Γ be defined as in Lemma 7.2. If, for some sequence $L \leq M$, $L \to \infty$, $L \notin \bar{J}$ for $\bar{J}$ defined in Assumption 2(iii), $\sup_{j \notin \bar{J}, j \leq L} |\sum_{m=1}^{j} w_m| = O_p(1/\sqrt{N})$ 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. For $L \leq M$ and $L \to \infty$ it follows by the assumption that

$$1 = \left|\sum_{m=1}^{M} w_m\right| \leq \inf_{j \notin \bar{J}, j \leq L} \left(\left|\sum_{m=j+1}^{M} w_m\right| + \left|\sum_{m=1}^{j} w_m\right|\right) \leq \inf_{j \notin \bar{J}, j \leq L} \left|\sum_{m=j+1}^{M} w_m\right| + \sup_{j \notin \bar{J}, j \leq L} \left|\sum_{m=1}^{j} w_m\right|$$

such that $\inf_{j \notin \bar{J}, j \leq L} \left|\sum_{m=j+1}^{M} w_m\right| \geq 1 - O_p\left(1/\sqrt{N}\right)$. Now let $C_{\bar{J}}$ be the number of elements in $\bar{J}$ such that

$$W'\Gamma W = \sum_{j=0}^{M-1} \left(\sum_{m=j+1}^{M} w_m\right)^2 \geq \sum_{j \notin \bar{J}, j \leq L} \left(\sum_{m=j+1}^{M} w_m\right)^2 \geq (L - C_{\bar{J}})\left(1 - O_p\left(1/\sqrt{N}\right)\right)^2.$$

Since $L \to \infty$ and $C_{\bar{J}}$ is bounded and does not depend on L or N, the result follows. $\blacksquare$

Lemma 7.4 *If, for some sequence $L \leq M$, $L \to \infty$ and for $\bar{J}$ defined in Assumption 2(iii) $L \notin \bar{J}$, and $\sup_{j \notin \bar{J}, j \leq L} |\sum_{m=1}^{j} w_m| = O(1/\sqrt{N})$ as $M \to \infty$, then $\sum_{m=1, m \notin \bar{J}}^{M} \left(\sum_{s=1}^{m} w_m\right)^2 m^{-2\alpha} \to 0$.*

Proof. Note that

$$\begin{aligned}
\sum_{m=1, m \notin \bar{J}}^{M} \left(\sum_{s=1}^{m} w_m\right)^2 m^{-2\alpha} &= \sum_{m=1, m \notin \bar{J}}^{L} \left(\sum_{s=1}^{m} w_m\right)^2 m^{-2\alpha} + \sum_{m=L+1, m \notin \bar{J}}^{M} \left(\sum_{s=1}^{m} w_m\right)^2 m^{-2\alpha} \\
&\leq \left(\sup_{j \notin \bar{J}, j \leq L} \left|\sum_{s=1}^{j} w_m\right|\right)^2 \sum_{m=1}^{L} m^{-2\alpha} + \sum_{m=L+1, m \notin \bar{J}}^{M} \left(\sum_{s=1}^{m} |w_m|\right)^2 m^{-2\alpha} \\
&\leq O\left(1/N\right) \sum_{m=1}^{L} m^{-2\alpha} + C_{l1} \sum_{m=L+1}^{M} m^{-2\alpha} \to 0,
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{s=1}^{m}|w_m| \leq C_{l1} < \infty$ uniformly in N by Assumption 4. Then, $\sum_{m=L+1}^{M} m^{-2\alpha} \to 0$ because $L \to \infty$ and $\sum_{m=1}^{M} m^{-2\alpha} < \infty$ uniformly in M. ■

Lemma 7.5 *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. $\sum_{i\neq j} P_{ij}(W) = O_p\left(N - K'W\right)$*,*

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 7.5(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 $\sum_{m=1}^{M}|w_l| \leq C_{l1}$ for some $C_{l1} < \infty$ was used and the bound holds uniformly for all N by Assumption 4. 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 part 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 part 5, note that

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

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

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

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 $\Delta(W) = \mathrm{tr}(e_f(W))$.

Lemma 7.6 *Suppose that Assumptions 1-3, 4 and 5(i) are satisfied. Then*

1. $\Delta(W) = o_p(1)$,

2. $f'(I - P(W))\epsilon/\sqrt{N} = O_p(\Delta(W)^{1/2})$,

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 $g\left(W\right) : W \to \mathbb{R}$ *with* $g(W) > 0$ *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. $E[f'\left(I - P(W)\right)\epsilon\epsilon'P(W)u/N|z] = f'(I - P(W))\mu\left(W\right)/N = o_p\left((K'W^+)/N + \Delta(W)\right)$,

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. $W'\Gamma W \leq CK'W^+$.

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

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

where

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

It follows that

$$\begin{aligned} W'AW &= \left(\sum_{m=1}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 (\tilde{\gamma}_m - \tilde{\gamma}_{m+1})\right) + \tilde{\gamma}_M \qquad (7.8) \\ &= \left(\sum_{m=1, m \notin \bar{J}}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 (\tilde{\gamma}_m - \tilde{\gamma}_{m+1})\right) + \tilde{\gamma}_M + o_p\left(1\right) \end{aligned}$$

where the second equality holds by Assumption 2(ii) such that

$$\begin{aligned} W'AW &\leq \sum_{m=1,m\notin\bar{J}}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 \tilde{\gamma}_m + o_p\left(1\right) = \sum_{m=1,m\notin\bar{J}}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 \frac{\tilde{\gamma}_m}{m^{-2\alpha}} m^{-2\alpha} \\ &\leq \sup_{m\leq M}\left(m^{2\alpha}\tilde{\gamma}_m\right)\sum_{m=1,m\notin\bar{J}}^{M-1}\left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha}, \end{aligned}$$

where $\sup_{m\leq M}\left(m^{2\alpha}\tilde{\gamma}_m\right) = O_p\left(1\right)$ by Assumption 2(ii). For a sequence $L \leq M$, $L \to \infty$ and $L/N \leq M/N \to 0$ satisfying Assumption 4(ii) it follows that $\sum_{m=1,m\notin\bar{J}}^{M}\left(\sum_{s=1}^{m} w_s\right)^2 m^{-2\alpha} = o\left(1\right)$ by Lemma 7.4. 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,

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

To give part 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

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

by Lemmas 7.5(3) and 7.5(4) and noting that $\text{Cum}[\epsilon_i,\epsilon_i,u_i,u_i'] = E[\epsilon_i^2u_iu_i'] - \sigma_\epsilon^2\Sigma_u - 2\sigma_{u\epsilon}\sigma_{u\epsilon}'$.

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,

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

which gives part 5.

To prove part 6, first we consider the function of a: $g(W)/a+a$ or $a \in \mathbb{R}$ 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 $\Delta(W) \neq 0$,

$$\frac{\sqrt{\Delta(W)/N}}{g\left(W\right)/N + \Delta(W)} = \left(\frac{g\left(W\right)}{\sqrt{\Delta(W)N}} + \sqrt{\Delta(W)N}\right)^{-1} \leq \frac{1}{2\sqrt{g(W)}} \to 0 \tag{7.9}$$

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

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

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

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 $\Delta(W)$. $|\mu_b'(W)\mu_b(W)| \leq C\sum_i (P_{ii}(W))^2$ 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

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

by (7.9). In addition if we define $\mu_i\left(W\right) = E[\epsilon_i^2 u_i]P_{ii}(W)$ and $\mu\left(W\right) = \left(\mu_1\left(W\right)', ..., \mu_n\left(W\right)'\right)'$ then

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

For part 9, we have the following decomposition:

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

The boundedness of $f_i' f_i H^{-1}P_{ii}(W)$ implies that

$$\sum_i f_i f_i' H^{-1}E[\epsilon_i^2 u_i u_i']P_{ii}(W)/N^2 = O_p(1/N).$$

Let $f_{a,i}$ be the ath element of f_i. Then, we have

$$\left|\sum_{i,j} f_{a,i}f_{a,j}P_{ij}(W)/N^2\right| \le \sum_{m=1} |w_m|(f_a'P_mf_a)/N^2 \le \sum_{m=1} |w_m|(f_a'f_a)/N^2 = O_p(1/N).$$

This implies that

$$\begin{aligned}
& \sum_{i\neq j} f_if_j'H^{-1}E[\epsilon_iu_i]E[\epsilon_ju_j']P_{ij}(W)/N^2 \\
= \;& \sum_{i,j} f_if_j'H^{-1}E[\epsilon_iu_i]E[\epsilon_ju_j']P_{ij}(W)/N^2 - \sum_i f_if_i'H^{-1}E[\epsilon_iu_i]E[\epsilon_iu_i']P_{ii}(W)/N^2 \\
= \;& O_p(1/N).
\end{aligned}$$

Lastly, we have

$$\begin{aligned}
& \sum_{i\neq j} f_if_i'H^{-1}E[\epsilon_i^2]E[u_ju_j']P_{jj}(W)/N^2 \\
= \;& \left(\sum_i f_if_i'\right)H^{-1}\sigma_\epsilon^2\Sigma_u\left(\sum_j P_{jj}(W)\right)/N^2 - \sum_i f_if_i'H^{-1}\sigma_\epsilon^2\Sigma_uP_{ii}(W)/N^2 \\
= \;& \sigma_\epsilon^2\Sigma_uK'W/N + O_p(1/N).
\end{aligned}$$

Therefore, we have

$$E\left[f'\epsilon\epsilon'fH^{-1}u'P(W)u|Z\right]/N^2 = \sigma_\epsilon^2\Sigma_uK'W/N + O_p\left(1/N\right).$$

For part 10, using again Lemma 7.5(5) as before,

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

and

$$\begin{aligned}
& E[f'\epsilon\epsilon'P(W)uH^{-1}f'u|z]/N^2 \\
= \;& \sum_i f_iP_{ii}(W)E\left[\epsilon_i^2u_i'H^{-1}f_iu_i'|z\right]/N^2 + \sum_{i\neq j} f_iP_{jj}(W)E[\epsilon_ju_j']H^{-1}f_iE[u_i'\epsilon_i]/N^2 \\
& +\sigma_\epsilon^2\sum_{i\neq j} f_iP_{ij}(W)E\left[u_jH^{-1}f_ju_j'|z\right]/N^2 + \sigma_\epsilon^2\sum_{i\neq j} f_jP_{ji}(W)E\left[u_jH^{-1}f_iu_j'|z\right]/N^2 \\
= \;& O_p(1/N) + \sum_{i\neq j} f_iP_{jj}(W)E[\epsilon_ju_j']H^{-1}f_iE[u_i'\epsilon_i]/N^2 = O_p\left(1/N\right) + (K'W)/N\sum_i f_i\sigma_{u\epsilon}'H^{-1}f_i\sigma_{u\epsilon}'/N.
\end{aligned}$$

For part 11, with the same arguments, it holds that

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

and arguments similar to before give

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

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 $\sup_{k\leq M}\left|\sum_{m=k}^M w_m\right| \leq C_{l1} < \infty$ which holds uniformly in M. ■

Lemma 7.7 *Assume that Assumptions 1, 2, 3, and 4 hold. Let*

$$\Xi(W) = tr(f'(I-P(W))f/N). \tag{7.10}$$

Let $\rho_{W,N} = tr(S(W))$ where $S(W)$ is defined in (3.2). Then, we have

$$(\Xi(W))^2 = o_p(\rho_{W,N}).$$

We note that the result holds when $S(W)$ is defined in (3.1).

Remark 6 *Considering the set $\bar{J}$ in Assumption 2 is important because the optimal weighting vector has a structure such that w_j does not converge to 0 if $f'(P_m - P_{m+1})f/N = 0$. Thus, the optimal weighting vector does not satisfy $\sup_{j\leq L}\left|\sum_{s=1}^j w_s\right| = O(1/\sqrt{N})$ in general.*

Proof. Let $\tilde{\gamma}_m = \mathrm{tr}(f'(I-P_m)f/N)$ and A be the $M\times M$ matrix whose (i,j)-th element is $\min(\tilde{\gamma}_i, \tilde{\gamma}_j) = \tilde{\gamma}_{\max(i,j)}$. Let e_1 be the first unit vector. We write

$$\Xi(W) = W'Ae_1, \quad \Delta(W) = W'AW.$$

Let $W_1 = (w_1, \ldots, w_L, 0 \ldots, 0)$ and $W_2 = (0, \ldots, 0, w_{L+1}, \ldots, w_M)$. We have the following decomposition.

$$
\begin{aligned}
(\Xi(W))^2 &= W_1' A e_1 e_1' A W_1 + 2 W_1' A e_1 e_1' A W_2 + W_2' A e_1 e_1' A W_2, \\
\Delta(W) &= W_1' A W_1 + 2 W_1' A W_2 + W_2' A W_2.
\end{aligned}
$$

First, we consider

$$
\begin{aligned}
W_1' A W_1 &= \sum_{j=1}^{L-1} \left(\sum_{s=1}^{j} w_s \right)^2 (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=1}^{L} w_s \right)^2 \tilde{\gamma}_L \\
&= \sum_{j \notin \bar{J}, j \le L} \left(\sum_{s=1}^{j} w_s \right)^2 (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=1}^{L} w_s \right)^2 \tilde{\gamma}_L \\
&\le \sup_{j \notin \bar{J}, j \le L} \left(\sum_{s=1}^{j} w_s \right)^2 \left(\sum_{j \notin \bar{J}, j \le L-1} (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \tilde{\gamma}_L \right) \\
&= \sup_{j \notin \bar{J}, j \le L} \left(\sum_{s=1}^{j} w_s \right)^2 \tilde{\gamma}_1 = O_p(1/N).
\end{aligned}
$$

By Lemma 7.3, $W' \Gamma W \to \infty$ so that

$$
W_1' A W_1 = O_p(1/N) = o(W' \Gamma W / N) = o(\rho_{W,N}).
$$

Since $|W_1' A W_2| \le (W_1' A W_1)^{1/2} (W_2' A W_2)^{1/2}$ by the Cauchy–Schwartz inequality, we have $\Delta(W) = W_2' A W_2 + o_p(\rho_{W,N})$. Next, we consider

$$
\begin{aligned}
& W_1' A e_1 e_1' A W_1 \\
= & \left(\sum_{j=1}^{L-1} \left(\sum_{s=1}^{j} w_s \right) (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=1}^{L} w_s \right) \tilde{\gamma}_L \right)^2 = \left(\sum_{j=1, j \notin \bar{J}}^{L-1} \left(\sum_{s=1}^{j} w_s \right) (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=1}^{L} w_s \right) \tilde{\gamma}_L \right)^2 \\
\le & \; O_p \left(\left(\sum_{j \notin \bar{J}, j < L} \left(\sum_{s=1}^{j} w_s \right)^2 (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=1}^{L} w_s \right)^2 \tilde{\gamma}_L \right) \left(\sum_{j \notin \bar{J}, j < L} (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \tilde{\gamma}_L \right) \right) \\
= & \; O_p(W_1' A W_1) = o_p(\rho_{W,N}),
\end{aligned}
$$

where the inequality is that of Cauchy-Schwartz. We examine the order of $W_2' A e_1 e_1' A W_2$. We observe that

$$
W_2' A e_1 = \sum_{j=L+1}^{M} \left(\sum_{s=L+1}^{j} w_s \right) (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=L+1}^{M} w_s \right) \tilde{\gamma}_M,
$$

and

$$
W_2' A W_2 = \sum_{j=L+1}^{M} \left(\sum_{s=L+1}^{j} w_s \right)^2 (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=L+1}^{M} w_s \right)^2 \tilde{\gamma}_M.
$$

These formulas imply that

$$\begin{aligned}
& W_2'Ae_1 - W_2'AW_2 \\
= & \sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)\left(1-\left(\sum_{s=L+1}^{j} w_s\right)\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \left(\sum_{s=L+1}^{M} w_s\right)\left(1-\left(\sum_{s=L+1}^{M} w_s\right)\right)\tilde{\gamma}_M \\
= & \sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)\left(\sum_{s=j+1}^{M} w_s\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)\left(\sum_{s=1}^{L} w_s\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) \\
& + o_p(\rho_{W,N}),
\end{aligned}$$

where

$$\left|\left(\sum_{s=L+1}^{M} w_s\right)\left(1-\left(\sum_{s=L+1}^{M} w_s\right)\right)\right|\tilde{\gamma}_M \leq C\tilde{\gamma}_M = o_p\left(\rho_{W,N}\right).$$

We observe that, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
& \left(\sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)\left(\sum_{s=j+1}^{M} w_s\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})\right)^2 \\
\leq & \left(\sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)^2(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})\right)\left(\sum_{j=L+1}^{M}\left(\sum_{s=j+1}^{M} w_s\right)^2(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})\right) \\
\leq & W_2'AW_2 \cdot C(\tilde{\gamma}_L - \tilde{\gamma}_M) + o_p(\rho_{W,N}) = o_p(\rho_{W,N})
\end{aligned}$$

since $W_2AW_2 = O(\rho_{W,N})$ and $\tilde{\gamma}_L - \tilde{\gamma}_M = o_p(1)$. It also holds that

$$\begin{aligned}
& \left(\sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)\left(\sum_{s=1}^{L} w_s\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})\right)^2 \\
= & \left(\sum_{s=1}^{L} w_s\right)^2\left(\sum_{j=L+1}^{M}\left(\sum_{s=L+1}^{j} w_s\right)(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})\right)^2 = O_p\left(\frac{1}{N}\right) = o_p(\rho_{W,N}),
\end{aligned}$$

by Assumption 4. It therefore follows that

$$(W_2'Ae_1 - W_2'AW_2)^2 = o_p(\rho_{W,N}).$$

Therefore,

$$\begin{aligned}
W_2'Ae_1e_1A'W_2 &= (W_2'AW_2 + W_2'Ae_1 - W_2'AW_2)^2 \\
&\leq 2(W_2AW_2)^2 + 2(W_2'Ae_1 - W_2'AW_2)^2 = o_p(\rho_{W,N}).
\end{aligned}$$

Lastly, by the Cauchy-Schwartz inequality, we have

$$W_1'Ae_1e_1'AW_2 = o_p(\rho_{W,N}).$$

To sum up, we have

$$(\Xi(W))^2 = W_1'Ae_1e_1'AW_1 + 2W_1'Ae_1e_1'AW_2 + W_2'Ae_1e_1'AW_2 = o_p(\rho_{W,N}).$$

■

Lemma 7.8 *If Assumptions 1-8 hold, and for $\Omega = \Omega_U = \{W \in l_1 | W'\mathbf{1}_M = 1\}$ where M satisfies the constraints in Assumption 10 and $W = (w_1, ..., w_M)$, it follows that*

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

where $S_\lambda(W) = \lambda' S(W)\lambda$ and $S(W)$ is defined in (3.1).

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

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

where $\tilde{\gamma}_m = \operatorname{tr}(f'(I - P_m)f/N)$. This argument 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 $0 < \delta < \varepsilon/2$,

$$\operatorname{tr}\left(\frac{f'(I - P(\tilde{W}))(I - P(\tilde{W}))f}{N}\right) = \sum_{j=1}^{M}\left(\sum_{i=1}^{j} w_i\right)^2 (\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) + \tilde{\gamma}_M = O_p\left(N^{\frac{-2\alpha(1+\varepsilon)}{2\alpha+1}}\right), \qquad (7.11)$$

where we use formula (7.8). Let J_M be the set of integers j such that $1 \leq j \leq M$ for which $j^{2\alpha+1}(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}) > 0$. 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

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

which together with (7.11) implies that $\sum_{j\in J_M}\left(\sum_{i=1}^{j} w_i\right)^2 = O\left(N^{\frac{-2\alpha(\varepsilon-\delta)+1+\delta}{2\alpha+1}}\right) = o(M)$. Now, since

$$O(M) = \sum_{j\in J_M} 1^2 = \sum_{j\in J_M}\left(\left(\sum_{i=1}^{j} w_i\right)^2 + 2\left(\sum_{i=1}^{j} w_i\right)\left(\sum_{i=j+1}^{M} w_i\right) + \left(\sum_{i=j+1}^{M} w_i\right)^2\right) \qquad (7.12)$$

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 (7.12) 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 7.9 *Let*

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

If Assumptions 1-9 hold, then, for Ω as defined in Lemma 7.8, it follows that

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

where $S_\lambda(W) = \lambda'S(W)\lambda$ and $S(W)$ is defined in (3.1).

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

$$\sup_{W\in\Omega\cap\Omega_2}\frac{K'W/N}{S_\lambda(W)} \to 0 \text{ and } \sup_{W\in\Omega\cap\Omega_2}\frac{(K'W)^2/N}{S_\lambda(W)} \to 0 \tag{7.13}$$

by Lemma 7.8 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 (7.13). 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 7.10 *Let Assumptions 1-10 hold. Then, it follows that*

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

where $S_\lambda(W) = \lambda' S(W)\lambda$ and $S(W)$ is defined in (3.1).

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

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

and note that

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

by Assumption 2. Together with Lemma 7.8, this implies that

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

Combining these results with Lemma 7.9 it is then sufficient to show that

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

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 7.9 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 7.9. Now note that

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

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

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

where $\sum_{j=1}^M |w_j| \leq C_{l1}$ uniformly in M is used. It follows for some $\vartheta > 1$ from Whittle (1960, Theorem 2) that for some constant C,

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

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

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

Next, consider

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

where

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

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

$$\sup_{W\in\Omega}\frac{\sum_{i=1}^{M-1}\left(\sum_{j=1}^{i} w_j\right)^2\left|u'\left(P_{i+1} - P_i\right)f/N\right|}{S_\lambda(W)} = \sup_{W\in\Omega}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2\left|u'\left(P_{i+1} - P_i\right)f/N\right|}{S_\lambda(W)} + o_p\left(1\right) \tag{7.14}$$

because

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

where the inequality follows from Markov's inequality, Lemma 7.8, 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

$$E\left[\left|u'\left(P_{i+1} - P_i\right)f/N\right|^{2\vartheta}|z\right] \leq CE\left[\left|u_i\right|^{2\vartheta}\right]N^{-\vartheta}\left(f'\left(P_{i+1} - P_i\right)f/N\right)^\vartheta. \tag{7.15}$$

Now note that

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

which establishes (7.14). We thus turn to the lead term on the right hand side of (7.14). By the Cauchy-Schwarz inequality we have

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

It now follows that

$$\begin{aligned}
&\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 |u'\left(P_{i+1}-P_i\right)f/N| \qquad (7.16)\\
\leq\ & \left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^4 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^4 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}\\
\leq\ & \sup_{i\leq M}\left(\sum_{j=1}^{i}w_j\right)^2\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N\right)^{1/2}\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2}
\end{aligned}$$

where $\sup_{i\leq M}\left(\sum_{j=1}^{i}w_j\right)^2 \leq C_{l1}^2 < \infty$ uniformly in M such that

$$\begin{aligned}
\left(\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2 u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2} &\leq \sup_{W}\left(\sum_{j=1}^{i}|w_j|\right)^2\left(\sum_{i=1}^{K_n}u'\left(P_{i+1}-P_i\right)u/N\right)^{1/2} \qquad (7.17)\\
&\leq C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}
\end{aligned}$$

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,

$$\sup_{W\in\Omega}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2|u'\left(P_{i+1}-P_i\right)f/N|}{S_\lambda\left(W\right)} = \sup_{W\in\Omega_N}\frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i}w_j\right)^2|u'\left(P_{i+1}-P_i\right)f/N|}{S_\lambda\left(W\right)} \qquad (7.18)$$

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

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

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 7.9, 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,

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

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

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

Secondly,

$$\sup_{W\in\Omega_N\cap\Omega_{1,N}}\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 f'\left(P_{i+1}-P_i\right)f/N\leq\omega N^{(-2\alpha+\varepsilon/2)/(2\alpha+1)} \tag{7.20}$$

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

$$\begin{aligned}
&\sup_{W\in\Omega_N\cap\Omega_{1,N}} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} \qquad (7.21)\\
\leq\quad & \frac{\sup_{W\in\Omega_N\cap\Omega_{1,N}}\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{\inf_{W\in\Omega} S_\lambda\left(W\right)} \\
\leq\quad & \sqrt{\omega}N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\inf_{W\in\Omega} S_\lambda\left(W\right)}
\end{aligned}$$

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

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

Thus, setting $g_N\left(W\right)=\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|/S_\lambda(W)$ and using (7.18), (7.19) and (7.21) one obtains the bound

$$\begin{aligned}
&\sup_{W\in\Omega} \frac{\sum_{i=1}^{K_n}\left(\sum_{j=1}^{i} w_j\right)^2 \left|u'\left(P_{i+1}-P_i\right)f/N\right|}{S_\lambda\left(W\right)} \qquad (7.22)\\
\leq\quad & \frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\sqrt{\omega}N^{(-\alpha+\varepsilon/4)/(2\alpha+1)}}+\sqrt{\omega}N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\frac{C\left(u'\left(P_{K_n+1}-P_1\right)u/N\right)^{1/2}}{\inf_{W\in\Omega} S_\lambda\left(W\right)}.
\end{aligned}$$

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

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

where the inequality uses Markov's inequality, (7.22) and Lemma 7.8. Next, note that

$$\frac{C\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{(-\alpha+\varepsilon/2)/(2\alpha+1)}} = \frac{1}{\sqrt{\omega}}\frac{C\sqrt{\left(K_{n+1}-1\right)/N}}{N^{(-\alpha+\varepsilon/2)/(2\alpha+1)}} = o\left(N^{\frac{-\varepsilon/2-\varepsilon/2}{2\alpha+1}}\right) = o(1) \tag{7.23}$$

and

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

such that

$$\frac{\left(E\left[u'\left(P_{K_n+1}-P_1\right)u/N|z\right]\right)^{1/2}}{N^{-2\alpha/(2\alpha+1)}}O_p\left(N^{\frac{-\alpha+\varepsilon/4}{2\alpha+1}}\right) = o_p\left(1\right). \tag{7.24}$$

Using (7.23) and (7.24) then establishes that

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

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

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

We show that the conditions of Lemma 7.1 are satisfied and $S(W)$ has the form given in the theorem. Let $\rho_{W,N} = \mathrm{tr}(S(W))$. 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 $(K'W)^2/N$ 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 7.5(6). As

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

Lemma 7.6(2) and (3) implies that

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

and

$$T_2^h = O_p\left(\max\left(|K'W|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/\sqrt{N}\right), \tag{7.25}$$

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 7.6(1), $K'W/\sqrt{N} = o(1)$ by $|K'W|/\sqrt{N} \leq K'W^+/\sqrt{N} = o(1)$, $\sum_i (P_{ii}(W))^2 = o_p(K'W^+)$ by Lemma 7.5(2) and $W'\Gamma W = O(K'W^+)$ by Lemma 7.6(12). Therefore $T^h = o_p(1)$. Next, we observe $T_1^H = O(\Xi(W))$ by the definition. Lemmas 7.7 and 7.6(1) imply that $T_1^H = o_p(1)$. $T_2^H = O_p(1/\sqrt{N})$ by the CLT. A similar argument for T_2^h implies

$$T_3^H = O_p\left(\max\left(|K'W|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/N\right). \tag{7.26}$$

Now, we analyze

$$\|T_1^h\| \cdot \|T_1^H\| = O_p(\Delta(W)^{1/2}\Xi(W)) = o_p(\rho_{W,N}),$$

by Lemma 7.7. It holds that

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

because by Lemma 7.6(6) one can take $g(W) = N(\mathrm{tr}(S(W)) - \Delta(W))$. 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 7.6(6), it follows that

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

Next,

$$\begin{aligned} ||T_1^h|| \cdot ||T_3^H|| &= O_p\left(\Delta(W)^{1/2} \max\left(|K'W|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/N\right) \\ &= o_p\left(\max\left(|K'W|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/N\right) = o_p(\rho_{W,N}) \end{aligned}$$

by Lemma 7.6(1), (7.26) and the fact (as noted before) that $T_3^H = O\left(\mathrm{tr}(S\left(W\right)\right))$. Next, (7.25) and the definition of T_1^H imply that

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

by Lemma 7.7. By similar arguments as before it follows from Lemma 7.6(6), that

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

and $\Delta\left(W\right)^{1/2} = o_p\left(1\right)$ such that $o_p(\Delta(W)^{1/2}K'W/\sqrt{N}) = o_p\left(\rho_{W,N}\right)$ as required. Lemma 7.6(6) gives

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

Thus, we have $||T_2^h|| \cdot ||T_1^H|| = o_p(\rho_{W,N})$. From (7.25) it follows that

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

where $K'W/N = O\left(\mathrm{tr}\left(S\left(W\right)\right)\right)$ and $\sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}/N = o_p\left(\mathrm{tr}(S(W))\right)$. By (7.25) and (7.26) it follows that

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

because $\left(\left|K'W\right|/N\right)^{3/2} = o\left(\rho_{W,N}\right)$ and $\left((W'\Gamma W) + \sum_i (P_{ii}(W))^2\right)/N = O_p(\rho_{W,N})$. 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

$$\begin{aligned} \left\|T_1^H\right\|^2 &= O_p\left(\Xi(W)^2\right) = o_p\left(\rho_{W,N}\right) \text{ by Lemma 7.7,} \\ \left\|T_2^H\right\|^2 &= O_p\left(1/N\right) = o_p\left(\rho_{W,N}\right), \\ \left\|T_3^H\right\|^2 &= O_p\left(\left(\max\left(|K'W|, \sqrt{(W'\Gamma W) + \sum_i (P_{ii}(W))^2}\right)/N\right)^2\right) = o_p\left(\rho_{W,N}\right) \end{aligned}$$

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 $O_p(\Delta\left(W\right)^{1/2}/\sqrt{N}) = o_p(g\left(W\right)/N + \Delta(W)) = o_p(\rho_{W,N})$ for $g(W) = N\left(\mathrm{tr}\left(S(W)\right) - \Delta(W)\right)$ by Lemma 7.6(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

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

and

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

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

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

by Lemma 7.6(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,

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

by Lemma 7.6(8). Again, we have $E\left[T_2^h T_1^{h\prime}|z\right] = -\mu\left(W\right)'\left(I - P(W)\right)f/N$. Sixth,

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

by Lemma 7.6(4). Seventh,

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

also, we have $E\left[T_1^H H^{-1}hh'|Z\right] = -\sigma_\epsilon^2 f'(I - P(W))f/N$. Lemma 7.6(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 7.6(9). Next,

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

by Lemma 7.6(5),

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

by Lemma 7.6(10). Similarly, it follows that

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

Therefore, we have

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

where the last equality holds because $1/N = o_p(\rho_{W,N})$ 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 7.5(2) and $f'Q(W)\mu(W)/N = o_p(K'W/N + \Delta(W))$ by Lemma 7.6 (8). Therefore, we have equation (7.3).

To derive equation (7.4), we note that

$$W'\Gamma W = \sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j \min(i,j) \leq \sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j j = \sum_{i=1}^{M} w_i \sum_{j=1}^{M} w_j j = W'\mathbf{1}_M K'W = K'W,$$

which means $W'\Gamma W = O(K'W)$. Moreover, $\sum_{i=1}^{N} f_i P_{ii}(W) = O_p(K'W)$ by Lemma 7.6(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_{sq}. By Corollary 7.2, the formula of $S_\lambda(W)$ for MA2SLS when $W \in \Omega_P$ is

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

with $A = \left\|\lambda' H^{-1}\sigma_{u\epsilon}\right\|^2$ (the other two terms in (3.1) can be ignored). Let M_{sq} be the optimal number of instruments picked by the Donald and Newey (2001) algorithm. For $a \in (0,1)$, let $M_1 = (1-a)M_{sq}$ and $M_2 = (1+a)M_{sq}$ and choose W^* such that it has only two non-zero elements $w_{M_1} = w_{M_2} = 0.5$. Then, $K'W^* = M_{sq}$ and

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

Then,

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

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

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

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

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

Consider the function

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

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

When $W \in \Omega_B$, the formula of $S_\lambda(W)$ for MA2SLS is

$$S_\lambda(W) = A\frac{(W'\Gamma W)}{N} + \sigma_\epsilon^2 \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} w_j w_i \gamma_{\max(i,j)}$$

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 $AM/(N-M) + \sigma_\epsilon^2\gamma_M$. Let M_N be the choice of M that minimizes $S_\lambda(W)$ 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 $w^* = N/(N-M_N)$ and choose W^* such that W^* has only three non-zero elements $w_{M_1} = w_{M_2} = 1/2w^*$ and $w_N = -M_N/(N-M_N)$. 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 $W_N'\Gamma W_N = ((w^*)^2 + 2w^* w_N)M_N + w_N^2 N = M_N N/(N-M_N)$ and

$$\begin{aligned} W^{*\prime}\Gamma W^* &= w_1^2 M_1 + w_2^2 M_2 + 2w_1 w_2 M_1 + w_N^2 N + 2w_N(w_1 M_1 + w_2 M_2) \\ &= w_1^2 M_1 + w_2^2 M_2 + 2w_1 w_2 M_1 + w_N^2 N + 2w_N w^* M_N \\ &= ((w^*)^2 + 2w_N w^*)M_N + w_N^2 N - (1/2)(w^*)^2 a M_N \end{aligned}$$

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

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

Since the term $\left(w_N^2 + 2w_N w^*\right)\gamma_N$ is of smaller order than $S_\lambda(W_N)$ the result now follows if $(\gamma_{M_1}/4 + 3\gamma_{M_2}/4)/\gamma_{M_N} \le 1$ wpa1. But this follows from the same arguments as for the proof of the first statement of the theorem.

For MALIML, the formula of $S_\lambda(W)$ is

$$S_\lambda(W) = A\frac{(W'\Gamma W)}{N} + \sigma_\epsilon^2 \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} w_j w_i \gamma_{\max(i,j)}$$

where $A = \lambda' H^{-1}(\sigma_\epsilon^2\Sigma_u - \sigma_{u\epsilon}\sigma_{u\epsilon}')H^{-1}\lambda$. Let M_{sq} be the optimal number of instruments chosen by the DN method. The MSE of the estimator that uses M_{sq} instruments is $AM_{sq}/N + \sigma_\epsilon^2\gamma_{M_{sq}}$. For $a \in (0,1)$, let $M_1 = (1-a)M_{sq}$ and $M_2 = (1+a)M_{sq}$ and choose W^* such that it has only two non-zero elements $w_{M_1} = w_{M_2} = 0.5$. The MSE of the estimator with W^* is

$$\frac{A}{N}(0.75M_1 + 0.25M_2) + \sigma_\epsilon^2(0.25\gamma_{M_1} + 0.75\gamma_{M_2}).$$

We note that $0.75M_1 + 0.25M_2 = M_{sq} - 0.5aM_{sq} < M_{sq}$. Moreover, we have $0.25\gamma_{M_1} + 0.75\gamma_{M_2} < \gamma_{M_{sq}}$ by following from the same arguments as for the proof of the first statement of the theorem. Therefore, the desired result is shown. ■

Proof of Theorem 4.1. We follow the proof of Donald and Newey (2001, Lemma A9). We first consider the case for $S(W)$ defined in (3.1) and $\hat{S}_\lambda(W)$ defined in (4.1). Note that when $\Omega = \Omega_U$ and $\Omega = \Omega_B$, the optimal weight, W^*, is well-defined and has a closed form (see the discussion in Section 7.5). When $\Omega = \Omega_C$ or Ω_P, we note that $S_\lambda(W)$ is continuous in W and Ω is a compact set, which implies that the optimal weight, W^*, is well defined in this case too. Thus $\inf_{W\in\Omega} S_\lambda(W) = S_\lambda(W^*)$ for some $W^* \in \Omega$ holds. It then follows that

$$0 \leq 1 - \frac{\inf_{W\in\Omega} S_\lambda(W)}{S_\lambda\left(\hat{W}\right)} \leq 4 \sup_{W\in\Omega} \left| \frac{\hat{S}_\lambda(W)}{S_\lambda(W)} - 1 \right|.$$

The result now follows from Lemma 7.10.

Next, we consider the case for $S(W)$ defined in (3.2) and $\hat{S}_\lambda(W)$ defined in (4.3) (the case for MALIML). We follow the steps taken in the above argument. First, we show that $\inf_{W\in\Omega} S_\lambda(W) = O_p(N^{\frac{-2\alpha}{2\alpha+1}})$. The weighting vector, $\tilde{W}$, where $w_M = 1$ and $w_j = 0$ for $j \neq M$ for $M = O(N^{\frac{1}{2\alpha+1}})$, gives $S_\lambda(\tilde{W}) = O_p(N^{\frac{-2\alpha}{2\alpha+1}})$. The proof that this rate is sharp is exactly equivalent to the corresponding part of the proof of Lemma 7.8. We then show that $\sup_{W\in\Omega}(\tilde{S}_\lambda(W)/S_\lambda(W)) - 1 = o_p(1)$, where

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

This can be shown by following the same argument as that for the Ω_2 part of the proof of Lemma 7.9. Lastly, we show that $\sup_{W\in\Omega}(\hat{S}_\lambda(W)/S_\lambda(W)) - 1 = o_p(1)$. The proof of this statement is the same as that of Lemma 7.10. We then obtain the desired result. ■

Proof of Theorem 7.3. Since it is easy to see that $X'X/N \to_p (E(f_i^2) + \sigma_u^2)$, we need to show

$$\frac{1}{N}X'P(W)X \to_p E(f_i^2) \tag{7.27}$$

and

$$\frac{1}{N}X'P(W)P(W)X \to_p E(f_i^2) \tag{7.28}$$

to obtain the desired result.

We have the following decomposition:

$$\frac{1}{N}X'P(W)X = \frac{1}{N}f'f - \frac{1}{N}f'(I-P(W))f + \frac{1}{N}f'P(W)u + \frac{1}{N}u'P(W)f + \frac{1}{N}u'P(W)u.$$

By Lemma 7.7 and 7.6(1), it holds that

$$\frac{1}{N}f'(I-P(W))f = o_p(1).$$

Since

$$\frac{1}{N}f'P(W)u = \frac{1}{N}f'u - \frac{1}{N}f'(I-P(W))u,$$

Lemma 7.5(6) and Lemma 7.6(2) (by replacing ϵ by u) imply that

$$\frac{1}{N}f'P(W)u = o_p(1).$$

Similarly, it follows that $u'P(W)f/N = o_p(1)$. Lastly, Lemma 7.6(3) and Assumption 4 imply that

$$\frac{1}{N}u'P(W)u = o_p(1).$$

Thus, we have shown (7.27).

We now consider (7.28). We have the following decomposition:

$$\begin{aligned}\frac{1}{N}X'P(W)P(W)X &= \frac{1}{N}f'f - \frac{1}{N}f'(I-P(W)P(W))f \\ &\quad + \frac{1}{N}f'P(W)P(W)u + \frac{1}{N}u'P(W)P(W)f + \frac{1}{N}u'P(W)P(W)u.\end{aligned}$$

We have that

$$\frac{1}{N}f'(I-P(W)P(W))f = \sum_{s_1=1}^{M}\sum_{s_2=1}^{M} f'(I-P_{\min(s_1,s_2)})f = \sum_{j=1}^{M}\left(2w_j\left(\sum_{s=j+1}^{M} w_s\right) + w_j^2\right)\tilde{\gamma}_j,$$

where $\tilde{\gamma}_j = f'(I-P_j)f/N$. It follows that

$$\sum_{j=1}^{M}\left(2w_j\left(\sum_{s=j+1}^{M} w_s\right) + w_j^2\right)\tilde{\gamma}_j = \sum_{j=1}^{M} w_j\left(2 - 2\sum_{s=1}^{j} w_s + w_j\right)\tilde{\gamma}_j.$$

Take L such that $L \to \infty$. We have that

$$\begin{aligned}&\left|\sum_{j=1}^{M} w_j\left(2-2\sum_{s=1}^{j} w_s + w_j\right)\tilde{\gamma}_j\right| \\ \leq\ &\left|\sum_{j=1}^{L} w_j\left(2-2\sum_{s=1}^{j} w_s + w_j\right)\tilde{\gamma}_j\right| + \left|\sum_{j=L+1}^{M} w_j\left(2-2\sum_{s=1}^{j} w_s + w_j\right)\right|\tilde{\gamma}_L \\ =\ &\left|\sum_{j=1}^{L} w_j\left(2-2\sum_{s=1}^{j} w_s + w_j\right)\tilde{\gamma}_j\right| + o_p(1)\end{aligned}$$

since $\tilde{\gamma}_L = o_p(1)$ and $W \in l_1$ implies that $|\sum_{j=L+1}^{M} w_j(2-2\sum_{s=1}^{j} w_s + w_j)|$ is bounded. Then, since $\sum_{j=1}^{L}|w_j| = o(1)$ by the assumption, we have

$$\left|\sum_{j=1}^{L} w_j\left(2-2\sum_{s=1}^{j} w_s + w_j\right)\tilde{\gamma}_j\right| = o_p(1).$$

It follows that

$$\frac{1}{N}f'(I-P(W)P(W))f = o_p(1).$$

We have that

$$E\left(\frac{1}{N}f'P(W)P(W)u\right) = 0$$

and

$$\begin{aligned}
E\left(\left(\frac{1}{N}f'P(W)P(W)u\right)^2\right) &= E\left(\frac{1}{N^2}f'P(W)P(W)uu'P(W)P(W)f\right) \\
&= \frac{1}{N^2}\sigma_u^2 f'P(W)P(W)P(W)P(W)f \\
&= \frac{1}{N^2}\sigma_u^2 \sum_{s_1,s_2,s_3,s_4=1}^{M} w_{s_1}w_{s_2}w_{s_3}w_{s_4}f'P_{\min(s_1,s_2,s_3,s_4)}f \\
&\leq \frac{1}{N^2}\sigma_u^2 \sum_{s_1,s_2,s_3,s_4=1}^{M} |w_{s_1}w_{s_2}w_{s_3}w_{s_4}|f'f = o_p(1)
\end{aligned}$$

because $f'f/N = O_p(1)$ by Lemma 7.5(6) and $W \in l_1$ by Assumption 4. It therefore follows that

$$\frac{1}{N}f'P(W)P(W)u = o_p(1).$$

Similarly, we have that $u'P(W)P(W)f/N = o_p(1)$. Lastly, we observe that

$$E\left(\frac{1}{N}u'P(W)P(W)u\right) = \sigma_u^2\frac{W'\Gamma W}{N}$$

by Lemma 1.2 of Hansen (2007). Assumption 4 and the Markov inequality imply that

$$\frac{1}{N}u'P(W)P(W)u = o_p(1).$$

Therefore, (7.28) is shown and we have obtained the desired result. ■

7.3 Lemmas for MALIML

As the first step, we show the consistency of MALIML and derive its asymptotic distribution. Define the LIML estimator based on the first m instruments as

$$\hat{\beta}_{L,m} = \arg\min_{\beta}(y - X\beta)'P_m(y - X\beta)/((y - X\beta)'(y - X\beta)).$$

We first establish uniform consistency $\sup_{m\leq M}\left|\hat{\beta}_{L,m} - \beta_0\right| \to_p 0$ for $M/N \to 0$. This result is then used to establish the uniform convergence of $\hat{\Lambda}(W)$ over M and W satisfying Assumption 5.

Lemma 7.11 *If Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied, then*

1. $\sup_{m\leq M}\epsilon'P_m\epsilon/N = o_p(1)$,

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

3. $\sup_{m\leq M}u'P_m\epsilon/N = o_p(1)$.

Proof. For 1. we observe that

$$\sup_{k\leq M}\epsilon'P_k\epsilon/N \leq \epsilon'P_M\epsilon/N$$

and

$$E[\epsilon' P_M \epsilon | z] = \sigma_\epsilon^2 \mathrm{tr}\left(P_M\right) = \sigma_\epsilon^2 M$$

such that

$$\begin{aligned} \Pr\left(\sup_{m \le M} |\epsilon' P_m \epsilon / N| > \eta | z\right) &\le \Pr\left(|\epsilon' P_M \epsilon / N| > \eta | z\right) \\ &\le \frac{1}{\eta N} E\left[\epsilon' P_M \epsilon | z\right] \to 0. \end{aligned}$$

For part 2, note that $E\left[f'\left(I - P_m\right)\epsilon | z\right] = 0$ such that

$$\begin{aligned} \sum_{m=1}^{M} \mathrm{tr} E\left[f'\left(I-P_m\right)\epsilon\epsilon\left(I-P_m\right)f/N | z\right] &\le \sup_{m \le M}\left(m^{2\alpha}\mathrm{tr}\left(f'\left(I-P_m\right)f\right)/N\right)\sigma_\epsilon^2 \sum_{m=1}^{M} m^{-2\alpha} \\ &= O_p\left(1\right), \end{aligned}$$

which shows that $\sup_{m \le M} f'\left(I-P_m\right)\epsilon/N = O_p\left(1/\sqrt{N}\right)$.

For part 3, note that $E\left[u' P_m \epsilon/N | z\right] = E\left[v' P_m \epsilon / N | z\right] + \sigma_{u\epsilon}/\sigma_\epsilon^2 E\left[\epsilon' P_m \epsilon / N | z\right] = 0 + \sigma_{u\epsilon} m/N$ and

$$\begin{aligned} & E\left[\left\|u' P_m \epsilon/N - \sigma_{u\epsilon} m/N\right\|^2 | z\right] \qquad (7.29) \\ \le\ & M \max_{m \le M} \frac{E\left[\mathrm{tr}\left(u' P_m \epsilon\epsilon' P_m u - \sigma_{u\epsilon}\sigma_{u\epsilon}' m^2\right) | z\right]}{N^2} \\ =\ & M \max_{m \le M}\left(\frac{\mathrm{tr}\Sigma_u \mathrm{tr} P_m}{N^2} + \sum_{i_1,\ldots,i_4=1}^{N} \frac{\mathrm{tr}\left(E\left[u_{i_1}\epsilon_{i_3}\right]E\left[u_{i_4}'\epsilon_{i_2}\right]\right)P_{m,i_1 i_2}P_{m,i_3 i_4}}{N^2}\right) \\ & + M \max_{m \le M}\left(\sum_{i=1}^{N}\frac{\mathrm{tr}\left(\mathrm{Cum}\left(u_i, u_i, \epsilon_i, \epsilon_i\right)\right)P_{m,ii}^2}{N^2}\right) \\ \le\ & \left(\mathrm{tr}\Sigma_u + \mathrm{tr}\left(\mathrm{Cum}\left(u_i, u_i, \epsilon_i, \epsilon_i\right)\right)\right)\left(\frac{M}{N}\right)^2 + M \max_{m \le M} \frac{\sigma_{u\epsilon}'\sigma_{u\epsilon}}{N^2} \sum_{i_1, i_2 = 1}^{N} P_{m,i_1 i_2} P_{m, i_2 i_1} \\ =\ & o\left(1\right) + M \max_{m \le M}\left(\frac{\sigma_{u\epsilon}'\sigma_{u\epsilon} m}{N^2}\right) = o\left(1\right) \end{aligned}$$

where we used $P_{m,i_2 i_1} = P_{m,i_1 i_2}$ and $\sum_{i_1,i_2=1}^{n} P_{m,i_1 i_2}P_{m,i_2 i_1} = \sum_{i=1}^{N} P_{m,ii} = m$. Then,

$$\begin{aligned} \left\|u' P_m \epsilon / N\right\| &\le \left\|u' P_m \epsilon/N - \sigma_{u\epsilon} m/N\right\| + \left\|\sigma_{u\epsilon}\right\| m/N \le \left\|u' P_m \epsilon/N - \sigma_{u\epsilon} m/N\right\| + \left\|\sigma_{u\epsilon}\right\| M/N \\ &= \left\|u' P_m \epsilon/N - \sigma_{u\epsilon} m/N\right\| + o\left(1\right), \end{aligned}$$

where the $o\left(1\right)$ term is uniform in $m \le M$. The result now follows from (7.29). ■

Lemma 7.12 *If Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied then* $\sup_{m \le M}\left|\hat{\beta}_{L,m} - \beta_0\right| \to_p 0$.

Proof. Define $\bar{X} \equiv (y, X)$ and $D_0 \equiv (\beta_0, I)$. $\bar{X}$ can be written as $\bar{X} = X D_0 + \epsilon e_1'$, where e_1 is the first unit (column) vector. Let $\hat{A}_m = \bar{X}' P_m \bar{X}/N$ and $A_m = D_0' \bar{H}_m D_0$. Let $\hat{B} = \bar{X}'\bar{X}/N$ and $B = E\left[\bar{X}_i \bar{X}_i'\right]$ with $\bar{X}_i = (y_i, X_i)$.

Let $\tau = (1, -\beta')'$ and define the augmented parameter space $\overline{\Theta} = \{1\} \times \Theta$ such that $\tau \in \overline{\Theta}$. Then, $(1, -\hat{\beta}'_{L,m})' = \arg\min_\tau \tau' \hat{A}_m \tau / (\tau' \hat{B} \tau)$. Essentially the same argument as that in the beginning of the proof of Lemma A.5 in Donald and Newey (2001) shows that $(1, -\beta'_0)' = \arg\min_\tau \tau' A_m \tau / (\tau' B \tau)$. Then, letting $L_{n,m}(\tau) = \tau' \hat{A}_m \tau / (\tau' \hat{B} \tau)$ and $L_m(\tau) = \tau' A_m \tau / (\tau' B \tau)$ and noting that

$$\sup_{\tau \in \overline{\Theta}, m \leq M} |L_{n,m}(\tau) - L_m(\tau)| \leq \sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\tau' \left(\hat{A}_m - A_m \right) \tau}{\tau' B \tau} \right| \sup_{\tau \in \overline{\Theta}} \left| \frac{\tau' B \tau}{\tau' \hat{B} \tau} \right| + \sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\tau' A_m \tau}{\tau' B \tau} \right| \sup_{\tau \in \overline{\Theta}} \left| \frac{\tau' B \tau}{\tau' \hat{B} \tau} - 1 \right|. \tag{7.30}$$

We note that $\tau' \hat{A}_m \tau / (\tau' \hat{B} \tau) \leq 1$ uniformly in n, m and τ. It follows that

$$\sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\tau' A_m \tau}{\tau' B \tau} \right| \leq 1.$$

By a LLN, $\hat{B} - B = o_p(1)$ which implies that $\sup_{\tau \in \overline{\Theta}} \left| \frac{\tau' B \tau}{\tau' \hat{B} \tau} - 1 \right| = o_p(1)$. From Donald and Newey (2001, p.1185) it follows that B is positive definite such that $\inf_\tau \tau' B \tau > \varepsilon > 0$ for some ε and

$$\sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\tau' \left(\hat{A}_m - A_m \right) \tau}{\tau' B \tau} \right| \leq \frac{\sup_{\tau \in \overline{\Theta}, m \leq M} \left| \tau' \left(\hat{A}_m - A_m \right) \tau \right|}{\varepsilon}. \tag{7.31}$$

We now show that $\sup_{\tau \in \overline{\Theta}, m \leq M} \left| \tau' \left(\hat{A}_m - A_m \right) \tau \right| = o_p(1)$. For this purpose, we observe that

$$\sup_{m \leq M} \frac{\sigma_\epsilon^2 m}{N} \leq \frac{\sigma_\epsilon^2 M}{N} = o(1), \tag{7.32}$$

$$\begin{aligned} \sup_{\tau \in \overline{\Theta}, m \leq M} \frac{\tau' D'_0 E[u' P_m u | z] D_0 \tau}{N} &= \sup_{\tau \in \overline{\Theta}, m \leq M} \frac{\operatorname{tr}(P_m E[u D_0 \tau \tau' D'_0 u'])}{N} \\ &= \sup_{\tau \in \overline{\Theta}, m \leq M} \frac{\operatorname{tr}(P_m) \tau' D'_0 \Sigma_u D_0 \tau}{N} \\ &\leq \frac{M}{N} \sup_{\tau \in \overline{\Theta}} \tau' D'_0 \Sigma_u D_0 \tau = o(1), \end{aligned} \tag{7.33}$$

where $\sup_{\tau \in \overline{\Theta}} \tau' D'_0 \Sigma_u D_0 \tau$ is bounded by Assumption 7, and

$$\begin{aligned} \sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\tau' D'_0 E[u' P_m \epsilon | z] e'_1 \tau}{N} \right| &= \sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\operatorname{tr}(P_m E[\epsilon \tau' D'_0 u'])}{N} \right| \\ &= \sup_{\tau \in \overline{\Theta}, m \leq M} \left| \frac{\operatorname{tr}(P_m) \tau' D'_0 \sigma_{u\epsilon}}{N} \right| \\ &\leq \frac{M}{N} \sup_{\tau \in \overline{\Theta}} |\tau' D'_0 \sigma_{u\epsilon}| = o(1), \end{aligned} \tag{7.34}$$

where $\sup_{\tau \in \overline{\Theta}} \tau' D'_0 \sigma_{u\epsilon}$ is bounded by Assumption 7. The term $\hat{A}_m$ has the decomposition $\hat{A}_m - A_m = \hat{A}_{m,1} + \hat{A}_{m,2} + ... + \hat{A}_{m,9} + o(1)$, where the $o(1)$ term is uniform in $m \leq M$ and consists of (7.32), (7.33) and (7.34) and

$$\begin{aligned}
\hat{A}_{m,1} &= D_0'\left(\frac{f'P_m f}{N} - A_m\right) D_0;\ A_{m,2} = D_0'\frac{u'P_m f}{N}D_0;\ \hat{A}_{m,3} = D_0'\frac{f'P_m u}{N}D_0; \\
\hat{A}_{m,4} &= D_0'\frac{u'P_m u - E\left(u'P_m u|z\right)}{N}D_0;\ \hat{A}_{m,5} = e_1\frac{\epsilon' P_m u - m\sigma_{u\epsilon}'}{N}D_0; \\
\hat{A}_{m,6} &= D_0'\frac{uP_m\epsilon - m\sigma_{u\epsilon}}{N}e_1';\ \hat{A}_{m,7} = e_1\frac{\epsilon' P_m f}{N}D_0; \\
\hat{A}_{m,8} &= D_0'\frac{f'P_m\epsilon}{N}e_1';\ \hat{A}_{m,9} = \frac{\epsilon' P_m\epsilon - \sigma_\epsilon^2 m}{N}e_1e_1'.
\end{aligned}$$

For $\hat{A}_{m,1}$ define $\hat{\Gamma}_{zz,m} = Z_m'Z_m/N$, $\hat{\Gamma}_{fz,m} = f'Z_m/N$, $\Gamma_{zz,k} = E[Z_{k,i}Z_{k,i}']$ and $\Gamma_{fz,k} = E[f_iZ_{k,i}']$ and choose a sequence M_1 where $M_1 \to \infty$ such that $M_1/N^3 \to 0$. It then follows for $m \leq M_1$ that

$$\begin{aligned}
E\left[\left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\|^2\right] &= N^{-2}\sum_{i,j=1}^{n}\operatorname{tr}E\left[\left(f_iZ_{m,i}' - \Gamma_{fz,m}\right)\left(f_jZ_{m,j}' - \Gamma_{fz,m}\right)'\right] \\
&= N^{-2}\sum_{i=1}^{n}\operatorname{tr}E\left[\left(f_iZ_{m,i}' - \Gamma_{fz,m}\right)\left(f_iZ_{m,i}' - \Gamma_{fz,m}\right)'\right] = O\left(\frac{m}{N}\right) = o\left(1\right)
\end{aligned}$$

and

$$E\left[\left\|\hat{\Gamma}_{zz,m} - \Gamma_{zz,m}\right\|^2\right] = N^{-2}\sum_{i=1}^{n}\operatorname{tr}E\left[\left(Z_{m,i}Z_{m,i}' - \Gamma_{zz,m}\right)\left(Z_{m,i}Z_{m,i}' - \Gamma_{zz,m}\right)'\right] = O\left(\frac{m^2}{N}\right).$$

Using the Markov inequality, one obtains

$$\Pr\left(\sup_{m\leq M_1}\left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\| \geq \varepsilon\right) \leq \frac{M_1}{\varepsilon^2}\sup_{m\leq M_1}E\left[\left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\|^2\right] = O\left(M_1^2/N\right) = o\left(1\right)$$

as well as

$$\Pr\left(\sup_{m\leq M_1}\left\|\hat{\Gamma}_{zz,m} - \Gamma_{zz,m}\right\| \geq \varepsilon\right) \leq \frac{M_1}{\varepsilon^2}\sup_{m\leq M_1}E\left[\left\|\hat{\Gamma}_{zz,m} - \Gamma_{zz,m}\right\|^2\right] = O\left(M_1^3/N\right) = o\left(1\right). \tag{7.35}$$

Let $\|C\|_2^2 = \sup \ell' C' C\ell/\ell'\ell$ for any matrix C and note that $\|C_1C_2\| \leq \|C_1\|\,\|C_2\|_2$ and $\|C_1C_2\| \leq \|C_2\|\,\|C_1\|_2$ for any conforming matrices C_1 and C_2. Now,

$$\begin{aligned}
\left\|\hat{\Gamma}_{fz,m}\hat{\Gamma}_{zz,m}^{-1}\hat{\Gamma}_{fz,m}' - A_m\right\| \leq{} & \left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\|\left\|\hat{\Gamma}_{zz,m}^{-1}\hat{\Gamma}_{fz,m}'\right\|_2 + \left\|\Gamma_{fz,m}\right\|_2\left\|\hat{\Gamma}_{zz,m}^{-1} - \Gamma_{zz,m}^{-1}\right\|\left\|\hat{\Gamma}_{fz,m}\right\|_2 \\
&+ \left\|\Gamma_{fz,m}\Gamma_{zz,m}^{-1}\right\|_2\left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\|,
\end{aligned}$$

$$\begin{aligned}
\left\|\hat{\Gamma}_{zz,m}^{-1}\hat{\Gamma}_{fz,m}'\right\|_2 &\leq \left\|\hat{\Gamma}_{zz,m}^{-1}\hat{\Gamma}_{fz,m}' - \Gamma_{zz,m}^{-1}\Gamma_{fz,m}'\right\| + \left\|\Gamma_{zz,m}^{-1}\Gamma_{fz,m}'\right\|_2 \\
&\leq \left\|\hat{\Gamma}_{fz,m} - \Gamma_{fz,m}\right\|\left\|\hat{\Gamma}_{zz,m}^{-1}\right\|_2 + \left\|\Gamma_{fz,m}\right\|_2\left\|\hat{\Gamma}_{zz,m}^{-1} - \Gamma_{zz,m}^{-1}\right\| + \left\|\Gamma_{zz,m}^{-1}\Gamma_{fz,m}'\right\|_2
\end{aligned}$$

and

$$\left\|\hat{\Gamma}_{zz,m}^{-1} - \Gamma_{zz,m}^{-1}\right\| \leq \left\|\hat{\Gamma}_{zz,m}^{-1}\right\|_2\left\|\hat{\Gamma}_{zz,m} - \Gamma_{zz,m}\right\|\left\|\Gamma_{zz,m}^{-1}\right\|_2.$$

Define F such that $\left\|\Gamma_{zz,m}^{-1}\right\|_2 \leq F$ where F is finite by Assumption 6 and let

$$\zeta_{m,N} := \left\|\hat{\Gamma}_{zz,m}^{-1} - \Gamma_{zz,m}^{-1}\right\|_2 / \left(F\left\|\hat{\Gamma}_{zz,m}^{-1} - \Gamma_{zz,m}^{-1}\right\|_2 + F^2\right) \leq \left\|\hat{\Gamma}_{zz,m} - \Gamma_{zz,m}\right\|$$

such that $\sup_{m\le M_1}\zeta_{m,N}\le\sup_{m\le M_1}\left\|\hat{\Gamma}_{zz,m}-\Gamma_{zz,m}\right\|=o_p(1)$ by (7.35). Following Lewis and Reinsel (1985, p. 397),

$$\begin{aligned}\left\|\hat{\Gamma}_{zz,m}^{-1}\right\|_2 &\le \left\|\Gamma_{zz,m}^{-1}\right\|_2+\left\|\hat{\Gamma}_{zz,m}^{-1}-\Gamma_{zz,m}^{-1}\right\|_2\\ &\le F+F^2\zeta_{m,N}/\left(1-F\zeta_{m,N}\right)\\ &\le F+F^2\left(\sup_{m\le M_1}\zeta_{m,N}\right)/\left(1-F\sup_{m\le M_1}\zeta_{m,N}\right)=O_p(1).\end{aligned}$$

It now follows that

$$\sup_{m\le M_1}\left\|\frac{f'P_mf}{N}-A_m\right\|=\sup_{m\le M_1}\left\|\hat{\Gamma}_{fz,m}\hat{\Gamma}_{zz,m}^{-1}\hat{\Gamma}'_{fz,m}-A_m\right\|=o_p(1). \tag{7.36}$$

For $M_1\le m\le M$ it follows that $A_m\to\bar{H}=E\left[f_if_i'\right]$. Then,

$$\frac{f'P_mf}{N}-A_m=-f'\left(I-P_m\right)'f/N+f'f/N-\bar{H}+\bar{H}-A_m=o_p\left(1\right), \tag{7.37}$$

where the $o_p(1)$ term is uniform in $M_1\le m\le M$ because

$$\begin{aligned}\sup_{\tau\in\overline{\Theta},M_1\le m\le M}\tau'f'\left(I-P_m\right)'f\tau/N &\le \sup_{M_1\le m\le M}\frac{m^{2\alpha}}{M_1^{\alpha}}\left(\sup_{\tau}\tau'f'\left(I-P_m\right)'f\tau/N\right)\\ &= M_1^{-\alpha}O_p\left(1\right)=o_p\left(1\right)\end{aligned}$$

by Assumption 2. By Assumption 6, and for $M_1\le m\le M$, $\bar{H}-A_m=O\left(m^{-2\alpha}\right)\le O\left(M_1^{-2\alpha}\right)=o(1)$. By a law of large numbers,

$$f'f/N-\bar{H}=O_p\left(1/\sqrt{N}\right)=o_p\left(1\right).$$

Together, (7.36) and (7.37) imply that

$$\sup_{\tau\in\overline{\Theta},M_1\le m\le M}\left\|\hat{A}_{1,m}\right\|=o_p\left(1\right).$$

Now consider, for some $\varepsilon>0$, not necessarily the same as in (7.31),

$$\begin{aligned}\Pr\left(\sup_{\tau\in\overline{\Theta},m\le M}\left|\tau'\left(\hat{A}_{m,2}+...+\hat{A}_{m,9}\right)\tau\right|>\varepsilon|z\right) &\le \sum_{j=2}^{9}\Pr\left(\sup_{\tau\in\overline{\Theta}}\|\tau\|\sum_{m=1}^{M}\left\|\hat{A}_{m,j}\right\|>\varepsilon|z\right)\\ &\le \frac{M\sup_\tau\|\tau\|}{\varepsilon}\max_{m\le M}\sum_{j=2}^{9}\sqrt{E\left[\left\|\hat{A}_{m,j}\right\|^2|z\right]}.\end{aligned} \tag{7.38}$$

To show that $M\max_{m\le M}E\left[\left(\epsilon'P_m\epsilon-\sigma_\epsilon^2m\right)^2/N^2|z\right]\to_p0$, we observe that

$$\begin{aligned}E[(\epsilon'P_m\epsilon-\sigma_\epsilon^2m)^2|z] &= \sigma_\epsilon^4\left(\mathrm{tr}P_m\right)^2+2\sigma_\epsilon^4\left(\mathrm{tr}P_m\right)-\sigma_\epsilon^4m^2+Cum[\epsilon_i,\epsilon_i,\epsilon_i,\epsilon_i]\sum_{i=1}^{N}(P_{m,ii})^2\\ &= 2\sigma_\epsilon^4m+Cum[\epsilon_i,\epsilon_i,\epsilon_i,\epsilon_i]\sum_{i=1}^{N}(P_{m,ii})^2\\ &= O(m)+o_p(m)\end{aligned}$$

because $\sum_{i=1}^{N}(P_{m,ii})^2 \leq (\max_i P_{m,ii}) \sum_{i=1}^{N} P_{m,ii} = o_p(m)$ by the same calculation as the proof of Lemma 7.6(4) and Lemma 7.5(2). Therefore

$$\begin{aligned} M \max_{m} E\left[\left(\epsilon' P_m \epsilon - \sigma_\epsilon^2 m\right)^2 / N^2 | z\right] &\leq M \max_{m \leq M} \frac{2\sigma_\epsilon^4 m + Cum[\epsilon_i, \epsilon_i, \epsilon_i, \epsilon_i] \sum_{i=1}^{N}(P_{m,ii})^2}{N^2} \qquad (7.39) \\ &\leq \frac{M \max_{m \leq M} (\max_i P_{m,ii})\, m}{N^2} + \frac{2\sigma_\epsilon^4 M^2}{N^2} \\ &= O_p\left(\frac{M^2}{N^2}\right) + \frac{2\sigma_\epsilon^4 M^2}{N^2} \rightarrow_p 0. \end{aligned}$$

Similarly, we can show that $M \max_{m \leq M} E\left[\left\|\hat{A}_{m,4}\right\|^2 |z\right] \rightarrow_p 0$, $M \max_{m \leq M} E\left[\left\|\hat{A}_{m,5}\right\|^2 |z\right] \rightarrow_p 0$ and $M \max_{m \leq M} E\left[\left\|\hat{A}_{m,6}\right\|^2 |z\right] \rightarrow_p 0$. Next,

$$\begin{aligned} M \max_{m \leq M} E\left[\left\|D_0' f' P_m u D_0 / N\right\|^2 |z\right] &\leq \|D_0\|^4 M \max_{m \leq M} E\left[\left\|f' P_m u / N\right\|^2 |z\right] \\ &= \|D_0\|^4 M \max_{m \leq M} \frac{\operatorname{tr}\left(f' P_m E\left[u u' | z\right] P_m f\right)}{N^2} \\ &= \|D_0\|^4 \operatorname{tr}(\Sigma_u) M \max_{m \leq M} \frac{\operatorname{tr}\left(f' P_m f\right)}{N^2} \\ &\leq \|D_0\|^4 \operatorname{tr}(\Sigma_u) M \frac{\operatorname{tr}(f' f)}{N^2} = O_p\left(\frac{M}{N}\right) = o_p(1), \end{aligned}$$

where $O_p\left(\sup_{m \leq M} m^{2\alpha} \left(\sup_{\lambda'\lambda=1} \lambda' f (I - P_m) f \lambda / N\right)\right) = O_p(1)$ by Assumption 2(i). Analogous calculations show that $M \max_{m \leq M} E\left[\left\|\hat{A}_{m,7}\right\|^2 |z\right] = o(1)$ and $M \max_{m \leq M} E\left[\left\|\hat{A}_{m,8}\right\|^2 |z\right] = o(1)$. Summing up, we have $\frac{M}{\varepsilon} \sup_{\tau \in \overline{\Theta}} \|\tau\| \sum_{j=1}^{9} \max_{m \leq M} E\left[\left\|\hat{A}_{m,j}\right\|^2 |z\right] \rightarrow_p 0$. Combining (7.30), (7.31), (7.32), (7.33), (7.34) (7.38) establishes that

$$\sup_{\tau \in \overline{\Theta}, m \leq M} \left|L_{n,m}(\tau) - L_m(\tau)\right| = o_p(1). \qquad (7.40)$$

From (7.31) and the fact that, if $\|\tau - \tau_0\| \geq \varepsilon$ for some $\varepsilon > 0$, there exists an $\eta > 0$ such that $\sup_{m \leq M} |L_m(\tau) - L_m(\tau_0)| \geq \eta$, it follows that

$$\Pr\left(\sup_{m \leq M} \left|\hat{\beta}_{L,m} - \beta_0\right| \geq \varepsilon | z\right) \leq \Pr\left(\sup_{m \leq M} \left|L_m(\hat{\tau}_m) - L_m(\tau_0)\right| \geq \eta | z\right)$$

with $\hat{\tau}_m = (1, -\hat{\beta}_{L,m}')'$ and by standard arguments

$$\begin{aligned} \left|L_m(\hat{\tau}_m) - L_m(\tau_0)\right| &\leq \left|L_{n,m}(\hat{\tau}_m) - L_m(\hat{\tau}_m)\right| + \left|L_{n,m}(\tau_0) - L_m(\tau_0)\right| \\ &\quad + \left|L_{n,m}(\hat{\tau}_m) - L_{n,m}(\tau_0)\right|, \end{aligned}$$

where $0 \leq L_{n,m}(\hat{\tau}_m) \leq L_{n,m}(\tau_0) + o_p(1) = o_p(1)$ uniformly in $m \leq M$ by the definition of $\hat{\tau}_m$ and Lemma 7.11 such that

$$\sup_{m \leq M} \left|L_{n,m}(\hat{\tau}_m) - L_{n,m}(\tau_0)\right| \leq 2 \sup_{m \leq M} \left|L_{n,m}(\tau_0)\right| + o_p(1) = o_p(1)$$

and

$$\begin{aligned}\Pr\left(\sup_{m\leq M}\left|L_m\left(\hat{\tau}_m\right)-L_m\left(\tau_0\right)\right|\geq\eta|z\right) &\leq \Pr\left(\sup_{m\leq M}\left|L_m\left(\hat{\tau}_m\right)-L_m\left(\tau_0\right)\right|\geq\eta|z\right) \qquad (7.41)\\ &\leq \Pr\left(2\sup_{\tau\in\overline{\Theta},m\leq M}\left|L_{n,m}\left(\tau\right)-L_m\left(\tau\right)\right|\geq\eta|z\right)\to 0\end{aligned}$$

by (7.40). ■

Lemma 7.13 *If Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied, it follows that for* $\hat{\beta}$ *defined in (2.3) (MALIML),* $\left|\hat{\beta}-\beta_0\right|\to_p 0$.

Proof. Let $A_m(\beta)\equiv(y-X\beta)'P_m(y-X\beta)/N$ and $B(\beta)\equiv(y-X\beta)'(y-X\beta)/N$. Define $\Lambda_m(\beta)\equiv A_m(\beta)/B(\beta)$.

As $\sup_{m\leq M}\left\|\hat{\beta}_{L,m}-\beta_0\right\|\to_p 0$ by Lemma 7.12, it follows that that $\sup_{m\leq M}\left|B(\hat{\beta}_{L,m})-\sigma_\epsilon^2\right|\to_p 0$, Moreover,

$$A_m(\beta_0)=\epsilon'P_m\epsilon/N\to_p 0 \qquad (7.42)$$

uniformly in $m\leq M$ by Lemma 7.11(1) which implies that $\sup_{m\leq M}|A_m(\tilde{\beta})|\to_p 0$ and therefore $\sup_{m\leq M}|\Lambda_m(\hat{\beta}_{L,m})|\to_p 0$. We also note that $\Lambda_m(\beta)=L_{n,m}\left(\tau\right)\leq 1$ uniformly in $m\leq N$ and β.

It now follows that for $\Lambda\left(W\right)=\sum_{m=1}^M w_m\Lambda_m\left(\beta_0\right)$

$$\begin{aligned}\left|\hat{\Lambda}\left(W\right)-\Lambda\left(W\right)\right| &\leq \sum_{m=1}^M\left|w_m\right|\left|L_{n,m}\left(\hat{\tau}_m\right)-L_{n,m}\left(\tau_0\right)\right|\\ &\leq 2\sup_{m,\tau}\left|L_{n,m}\left(\tau\right)-L_m\left(\tau\right)\right|\sum_{m=1}^M\left|w_m\right|+\sup_{m\leq M}\left|L_m\left(\hat{\tau}_m\right)-L_m\left(\tau_0\right)\right|\sum_{m=1}^M\left|w_m\right|,\end{aligned}$$

where $2\sup_{m,\tau}\left|L_{n,m}\left(\tau\right)-L_m\left(\tau\right)\right|=o_p\left(1\right)$ by Lemma 7.12, $\sup_{m\leq M}\left|L_m\left(\hat{\tau}_m\right)-L_m\left(\tau_0\right)\right|=o_p\left(1\right)$ by (7.41) and $\sum_{i=1}^M\left|w_m\right|=O\left(1\right)$. It now follows that

$$\hat{\beta}-\beta_0=(X'P(W)X-\hat{\Lambda}\left(W\right)X'X)^{-1}(X'P(W)\epsilon-\hat{\Lambda}\left(W\right)X'\epsilon). \qquad (7.43)$$

We have $\left(\hat{\Lambda}\left(W\right)-\Lambda\left(W\right)\right)X'X/N=o_p\left(1\right)$ and $\left|\Lambda\left(W\right)\right|\leq\sum_{m=1}^M\left|w_m\right|\left|\Lambda_m\left(\beta_0\right)\right|=o_p\left(1\right)$ such that

$$N^{-1}\left(X'P(W)X-\hat{\Lambda}\left(W\right)X'X\right)=N^{-1}X'P(W)X+o_p\left(1\right) \qquad (7.44)$$

and, similarly, $\hat{\Lambda}\left(W\right)X'\epsilon/N=o_p\left(1\right)$ such that

$$\hat{\beta}-\beta_0=\left(X'P(W)X\right)^{-1}X'P(W)\epsilon+o_p\left(1\right)$$

and the result follows from Theorem 7.1. ■

Lemma 7.14 *Suppose that Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied. Then, for* $\hat{\beta}$ *defined in (2.3) (MALIML),* $\sqrt{N}(\hat{\beta}-\beta_0)\to_d N(0,\sigma_\epsilon^2\bar{H}^{-1})$.

Proof. The result follows from (7.43), (7.44) and the fact that $X'\epsilon/\sqrt{N} = O_p(1)$ together with $\hat{\Lambda}(W) = o_p(1)$. We then have

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) = \left(X'P(W)X/\sqrt{N}\right)^{-1} \frac{X'P(W)\epsilon}{\sqrt{N}} + o_p(1)$$

such that the result again follows from Theorem 7.1. ■

Lemma 7.15 *Suppose that 1 - 4, 5(ii), 6 and 7 are satisfied. Let $\Lambda_{\beta\beta,m}(\beta)$ be the Hessian of $\Lambda_m(\beta)$. If $\sup_{m\leq M}\left\|\tilde{\beta}_m - \beta_0\right\| \rightarrow_p 0$, then*

$$\sup_{m\leq M}\left\|\Lambda_{\beta\beta,m}(\beta_0) - \Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right\| = o_p(1)$$

and

$$\sup_{m\leq M}\left\|\Lambda_{\beta\beta,m}(\beta_0) - \frac{2}{\sigma_\epsilon^2}\bar{H}_m\right\| = o_p(1).$$

Proof. Let $\Lambda_{\beta,m}(\beta)$ and $\Lambda_{\beta\beta,m}(\beta)$ be the gradient and Hessian of $\Lambda_m(\beta)$. Let $A_m(\beta) \equiv (y-X\beta)'P_m(y-X\beta)/N$ and $B(\beta) = (y-X\beta)'(y-X\beta)/N$. Let $A_{\beta,m}(\beta)$ and $B_\beta(\beta)$ be the gradients of $A_m(\beta)$ and $B(\beta)$, respectively, and $A_{\beta\beta,m}(\beta)$ and $B_{\beta\beta}(\beta)$ be the Hessians of $A_m(\beta)$ and $B(\beta)$, respectively. We have

$$\begin{aligned}
\Lambda_{\beta,m}(\beta) &= B(\beta)^{-1}(A_{\beta,m}(\beta) - \Lambda_m(\beta)B_\beta(\beta)), \\
\Lambda_{\beta\beta,m}(\beta) &= B(\beta)^{-1}(A_{\beta\beta,m}(\beta) - \Lambda_m(\beta)B_{\beta\beta}(\beta)) - B(\beta)^{-1}(B_\beta(\beta)\Lambda_{\beta,m}(\beta)' + \Lambda_{\beta,m}(\beta)B_\beta(\beta)').
\end{aligned}$$

By assumption, $\sup_{m\leq M}\left\|\tilde{\beta}_m - \beta_0\right\| \rightarrow_p 0$, which implies that $\sup_{m\leq M}|B(\tilde{\beta}_m) - \sigma_\epsilon^2| \rightarrow_p 0$, $\sup_{m\leq M}|B_\beta(\tilde{\beta}_m) - (-2\sigma_{u\epsilon})| \rightarrow_p 0$. Moreover,

$$\max_{m\leq M}|A_m(\beta_0)| = \max_{m\leq M}|\epsilon'P_m\epsilon/N| \rightarrow_p 0,$$

by Lemma 7.11(1),

$$\max_{m\leq M}\|A_{\beta,m}(\beta_0)\| = \max_{m\leq M}\|X'P_m\epsilon/N\| \leq \max_{m\leq M}\|f'P_m\epsilon/N\| + \max_{m\leq M}\|u'P_m\epsilon/N\| = o_p(1), \tag{7.45}$$

where $\max_{m\leq M}\|f'P_m\epsilon/N\| = o_p(1)$ by Lemma 7.11(2) and $\max_{m\leq M}\|u'P_m\epsilon/N\| = o_p(1)$ by Lemma 7.11(3).

From the proof of Lemma 7.12 and (7.40), it follows that $\sup_{m\leq M}\Lambda_m(\tilde{\beta}_m) \rightarrow_p 0$. Similarly, we note that

$$A_{\beta,m}(\tilde{\beta}_m) = X'P_m(y - X\tilde{\beta}_m)/N = X'P_m\epsilon/N + X'P_mX\left(\tilde{\beta}_m - \beta_0\right)/N,$$

where $\epsilon'P_mX/N = o_p(1)$ uniformly in $m \leq M$ by (7.45) and $X'P_mX/N$ is uniformly bounded by the same arguments as in the proof of Lemma 7.12. This shows that $A_{\beta,m}(\tilde{\beta}_m) \rightarrow_p 0$ and therefore $\Lambda_{\beta,m}(\tilde{\beta}_m) \rightarrow_p 0$.

Now, consider

$$\begin{aligned}\Lambda_{\beta\beta,m}(\tilde{\beta}_m) - \Lambda_{\beta\beta,m}(\beta_0) \quad = \quad & \left(\frac{1}{\epsilon'\epsilon/N} - B(\tilde{\beta}_m)^{-1}\right) 2(X'P_mX/N - \Lambda_m(\beta_0)X'X/N) \\ & - \left(\frac{1}{\epsilon'\epsilon/N} - B(\tilde{\beta}_m)^{-1}\right)\left(\frac{\epsilon'X}{N}\Lambda_{\beta,m}(\beta_0)' + \Lambda_{\beta,m}(\beta_0)\frac{X'\epsilon}{N}\right) \\ & + B(\tilde{\beta}_m)^{-1}\left(\Lambda_m(\beta_0) - \Lambda_m(\tilde{\beta}_m)\right) X'X/N \\ & - B(\tilde{\beta})^{-1}(B_\beta(\tilde{\beta}_m)\Lambda_{\beta,m}(\tilde{\beta}_m)' - B_\beta(\beta_0)\Lambda_{\beta,m}(\beta_0)') \\ & - B(\tilde{\beta}_m)^{-1}\left(\Lambda_{\beta,m}(\tilde{\beta}_m)B_\beta(\tilde{\beta}_m)' - \Lambda_{\beta,m}(\beta_0)B_\beta(\beta_0)'\right),\end{aligned}$$

where

$$\left(\frac{1}{\epsilon'\epsilon/N} - B(\tilde{\beta}_m)^{-1}\right) = \frac{B(\tilde{\beta}_m) - \epsilon'\epsilon/N}{B(\tilde{\beta}_m)\epsilon'\epsilon/N} = \frac{2\epsilon'X/N\left(\tilde{\beta}_m - \beta_0\right) + \left(\tilde{\beta} - \beta_0\right)' X'X\left(\tilde{\beta}_m - \beta_0\right)/N}{B(\tilde{\beta}_m)\epsilon'\epsilon/N} = o_p(1)$$

uniformly in $m \leq M$. Since $X'P_mX/N - \Lambda_m(\beta_0)X'X/N = O_p(1)$ uniformly in $m \leq M$ and all other terms are of smaller order, it follows that $\sup_{m\leq M}\left\|\Lambda_{\beta\beta,m}(\beta_0) - \Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right\| = o_p(1)$. Next, consider

$$\begin{aligned}\Lambda_{\beta\beta,m}(\beta_0) - \frac{2}{\sigma_\epsilon^2}\bar{H}_m \quad = \quad & \left(\frac{1}{\epsilon'\epsilon/N} - \frac{1}{\sigma_\epsilon^2}\right) 2X'P_mX/N + \frac{2}{\sigma_\epsilon^2}\left(X'P_mX/N - \bar{H}_m\right) \\ & - \frac{1}{\epsilon'\epsilon/N}\left(\Lambda_m(\beta_0)X'X/N + 2\frac{\epsilon'X}{N}\Lambda_{\beta,m}(\beta_0)' + 2\Lambda_{\beta,m}(\beta_0)\frac{X'\epsilon}{N}\right).\end{aligned}$$

Note that $2X'P_mX/N - 2\bar{H}_m \to_p 0$ where the convergence is uniform in $m \leq M$ by the same arguments as in the proof of Lemma 7.12. Also note that $B_{\beta\beta}(\beta) = 2X'X/N \to_p 2E(X_iX_i')$. It therefore follows that $\sup_{m\leq M}\left\|\Lambda_{\beta\beta,m}(\beta_0) - \frac{2}{\sigma_\epsilon^2}\bar{H}_m\right\| = o_p(1)$ uniformly in $m \leq M$. ■

Lemma 7.16 *Suppose that Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied. Then*

$$\sqrt{N}(\hat{\beta}_{L,m} - \beta_0) = \left(\bar{H}_m^{-1} + o_p(1)\right)\left(h - \frac{f'(I - P_m)\epsilon}{\sqrt{N}} + \frac{v'P_m\epsilon}{\sqrt{N}}\right) + o_p(1),$$

where both $o_p(1)$ terms are uniform in $m \leq M$.

Proof. Let $\Lambda_{\beta,m}(\beta)$ and $\Lambda_{\beta\beta,m}(\beta)$ be the gradient and Hessian of $\Lambda_m(\beta)$, respectively. A standard Taylor expansion shows that

$$\sqrt{N}(\hat{\beta}_{L,m} - \beta_0) = -\Lambda_{\beta\beta,m}(\tilde{\beta})^{-1}\sqrt{N}\Lambda_{\beta,m}(\beta_0) = \left(\frac{\tilde{\sigma}_\epsilon^2\Lambda_{\beta\beta,m}(\tilde{\beta}_m)}{2}\right)^{-1}\left(-\frac{\tilde{\sigma}_\epsilon^2\sqrt{N}\Lambda_{\beta,m}(\beta_0)}{2}\right),$$

for some mean value $\tilde{\beta}_m$, where $\tilde{\sigma}_\epsilon^2 = \epsilon'\epsilon/N$. As $\sup\left\|\hat{\beta}_{L,m} - \beta_0\right\| = o_p(1)$ by Lemma 7.12, it follows that $\sup_m\left\|\tilde{\beta}_m - \beta_0\right\| \to_p 0$, such that by Lemma 7.15 it follows that

$$\sqrt{N}(\hat{\beta}_{L,m} - \beta_0) = \left(\bar{H}_m^{-1} + o_p(1)\right)\left(-\frac{\tilde{\sigma}_\epsilon^2\sqrt{N}\Lambda_{\beta,m}(\beta_0)}{2}\right),$$

where the $o_p(1)$ term is uniform in $m \leq M$.

Consider the gradient term. Define $\hat{\alpha} = X'\epsilon/\epsilon'\epsilon$, $\alpha = \sigma_{u\epsilon}/\sigma_\epsilon^2$ and $v = u - \epsilon\alpha'$. It holds that $\hat{\alpha} - \alpha = O_p(1/\sqrt{N})$by the CLT. We have the following decomposition:

$$\begin{aligned} -\frac{\tilde{\sigma}_\epsilon^2 \sqrt{N}\Lambda_{\beta,m}(\beta_0)}{2} &= \frac{X'P_m\epsilon}{\sqrt{N}} - \frac{\epsilon' P_m \epsilon X'\epsilon}{\sqrt{N}\epsilon'\epsilon} \\ &= h - \frac{f'(I-P_m)\epsilon}{\sqrt{N}} + \frac{v'P_m\epsilon}{\sqrt{N}} - \sqrt{N}(\hat{\alpha}-\alpha)\frac{\epsilon' P_m\epsilon}{N}. \end{aligned}$$

First, we have $h \rightarrow_d N(0, \sigma^2 \bar{H})$ by the CLT. Lemma 7.11(2) implies that $f'(I-P_m)\epsilon/\sqrt{N} = O_p(1)$ uniformly in $m \leq M$. From Lemma 7.11(1) $\sup_{m\leq M} \epsilon' P_m \epsilon/N = o_p(1)$ such that $\sqrt{N}(\hat{\alpha}-\alpha)\epsilon' P_m\epsilon/N = o_p(1)$ uniformly in $m \leq M$. In conclusion, we have

$$\sqrt{N}(\hat{\beta}_{L,m} - \beta_0) = \left(\bar{H}_m^{-1} + o_p(1)\right)\left(h - \frac{f'(I-P_m)\epsilon}{\sqrt{N}} + \frac{v'P_m\epsilon}{\sqrt{N}}\right) + o_p(1),$$

where both $o_p(1)$ terms are small uniformly in $m \leq M$. ■

Lemma 7.17 *Suppose that Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied. Then*

$$\begin{aligned} \hat{\Lambda}(W) &= \tilde{\Lambda}(W) - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) - \Lambda_q(W) + \hat{R}_\Lambda \\ &= \tilde{\Lambda}(W) + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W}{N^{3/2}}\right) + o_p\left(\frac{\rho_{W,N}}{\sqrt{N}}\right), \end{aligned}$$

where

$$\Lambda_q(W) = \frac{1}{2}\sum_{m=1}^M w_m \left(\Lambda_{\beta,m}(\beta_0)'\left(\Lambda_{\beta\beta,m}(\beta_0)\right)^{-1}\Lambda_{\beta,m}(\beta_0)\right) = O_p\left(\frac{1}{N}\right),$$

$$\tilde{\Lambda}(W) = \epsilon' P(W)\epsilon/(N\sigma_\epsilon^2) = K'W/N + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right),$$

$\tilde{\sigma}_\epsilon^2 = \epsilon'\epsilon/N$, $\sqrt{N}\hat{R}_\Lambda = O_p\left(1/\sqrt{N}\right)$, *$\hat{R}_\Lambda$ is simply the difference between $\hat{\Lambda}$ and the first three terms in the expression between two equalities, $\Lambda_{\beta,m}(\beta)$ and $\Lambda_{\beta\beta,m}(\beta)$ are the gradient and Hessian of $\Lambda_m(\beta)$ and $\rho_{W,N} = tr(S(W))$ for $S(W)$ defined in (3.2).*

Proof. We note that, in the LIML case, to show $o_p(\rho_{W,N})$, it is enough to show $o_p(W'\Gamma W/N + K'W/N + \sum_i (P_{ii}(W))^2/N + \Delta(W))$. We use the notation developed in the proof of Lemma 7.14. We expand $\hat{\Lambda}_m = \Lambda_m(\hat{\beta}_{L,m})$ around the true value β_0. By Donald and Newey (2001, p1186),

$$\Lambda_m(\beta_0) = \tilde{\Lambda}_m - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}_m + \frac{(\tilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2)^2}{\tilde{\sigma}_\epsilon^2\sigma_\epsilon^2}\tilde{\Lambda}_m,$$

where $\tilde{\Lambda}_m = \epsilon' P_m \epsilon/(N\sigma_\epsilon^2)$ such that

$$\sum_{m=1}^M w_m \Lambda_m(\beta_0) = \tilde{\Lambda}(W) - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) + \frac{(\tilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2)^2}{\tilde{\sigma}_\epsilon^2\sigma_\epsilon^2}\tilde{\Lambda}(W).$$

By a similar argument as in Lemma 7.6(4), we have

$$\begin{aligned} \tilde{\Lambda}(W) &= \frac{\epsilon' P(W)\epsilon}{N\sigma_\epsilon^2} = \frac{K'W}{N} + \frac{\epsilon' P(W)\epsilon - \sigma_\epsilon^2 K'W}{N\sigma_\epsilon^2} \qquad (7.46) \\ &= \frac{K'W}{N} + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right). \end{aligned}$$

Consider

$$\begin{aligned}\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\beta)\right]'}{\partial\beta} &= -B(\beta)^{-2}B_{\beta}(\beta)\text{vec}\left[(A_{\beta\beta,m}(\beta)-\Lambda_m(\beta)B_{\beta\beta}(\beta))-(B_{\beta}(\beta)\Lambda_{\beta,m}(\beta)'+\Lambda_{\beta,m}(\beta)B_{\beta}(\beta)')\right]' \\ & -B(\beta)^{-1}\Lambda_{\beta,m}(\beta)\text{vec}\left[B_{\beta\beta}(\beta))\right]' \\ & -B_{\beta\beta}(\beta)\left[(K_{1,n}\otimes I)\left(\Lambda_{\beta,m}(\beta)\otimes I\right)\right]'-\Lambda_{\beta\beta,m}(\beta)\left[(K_{1,n}\otimes I)\left(I\otimes B_{\beta}(\beta)\right)\right]' \\ & -B_{\beta\beta}(\beta)\left[(K_{1,n}\otimes I)\left(I\otimes \Lambda_{\beta,m}(\beta)\right)\right]'-\Lambda_{\beta\beta,m}(\beta)\left[(K_{1,n}\otimes I)\left(B_{\beta}(\beta)\otimes I\right)\right]',\end{aligned}$$

where the result follows from Magnus and Neudecker (1988, p. 185) and $K_{1,n}$ is the commutation matrix. Let $\tilde{\beta}_m$ be some mean value between $\hat{\beta}_{L,m}$ and β_0. Then,

$$\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right]'}{\partial\beta}-\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\beta_0)\right]'}{\partial\beta}=o_p\left(1\right)$$

uniformly in $m\leq M$ and $\frac{\partial \text{vec}[\Lambda_{\beta\beta,m}(\beta_0)]'}{\partial\beta}$ is bounded uniformly over $m\leq M$. A Taylor expansion then leads to

$$\begin{aligned}\sum_{m=1}^{M}w_m\hat{\Lambda}_m &= \sum_{m=1}^{M}w_m\Lambda_m(\beta_0)-\sum_{m=1}^{M}w_m\frac{1}{2}\left(\hat{\beta}_{L,m}-\beta_0\right)'\Lambda_{\beta\beta,m}(\beta_0)\left(\hat{\beta}_{L,m}-\beta_0\right) \\ & +\sum_{m=1}^{M}w_m\left(\hat{\beta}_{L,m}-\beta_0\right)'\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right]'}{\partial\beta}\left(\left(\hat{\beta}_{L,m}-\beta_0\right)\otimes\left(\hat{\beta}_{L,m}-\beta_0\right)'\right) \\ &= \sum_{m=1}^{M}w_m\Lambda_m(\beta_0)-\frac{1}{2}\sum_{m=1}^{M}w_m\Lambda_{\beta,m}\left(\beta_0\right)'\left(\Lambda_{\beta\beta,m}(\beta_0)\right)^{-1}\Lambda_{\beta,m}\left(\beta_0\right)+O_p\left(\frac{1}{N^{3/2}}\right),\end{aligned}$$

where $O_p\left(\frac{1}{N^{3/2}}\right)$ can be established by considering

$$\begin{aligned}&\left\|\sum_{m=1}^{M}w_m\left(\hat{\beta}_{L,m}-\beta_0\right)'\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right]'}{\partial\beta}\left(\left(\hat{\beta}_{L,m}-\beta_0\right)\otimes\left(\hat{\beta}_{L,m}-\beta_0\right)'\right)\right\| \\ \leq\ & \sup_{m\leq M}\left\|\frac{\partial \text{vec}\left[\Lambda_{\beta\beta,m}(\tilde{\beta}_m)\right]'}{\partial\beta}\right\|\sum_{m=1}^{M}|w_m|\left\|\hat{\beta}_{L,m}-\beta_0\right\|^3\end{aligned}$$

with

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{L,m}-\beta_0\right) &= \left(\bar{H}_m^{-1}+o_p\left(1\right)\right)\left(h-\frac{f'(I-P_m)\epsilon}{\sqrt{N}}+\frac{v'P_m\epsilon}{\sqrt{N}}\right)+o_p\left(1\right) \\ &= O_p\left(1\right)+\left(\bar{H}_m^{-1}+o_p\left(1\right)\right)\frac{v'P_m\epsilon}{\sqrt{N}},\end{aligned}$$

where the $O_p(1)$ and $o_p(1)$ terms are uniform in $m\leq M$ such that

$$\begin{aligned}\sum_{m=1}^{M}|w_m|\left\|\hat{\beta}_{L,m}-\beta_0\right\|^3 &\leq O_p\left(N^{-3/2}\right)O_p\left(1+\sum_{m=1}^{M}|w_m|\left(\left\|\bar{H}_m^{-1}\right\|^3\left\|v'P_m\epsilon/\sqrt{N}\right\|^3+\left\|\bar{H}_m^{-1}\right\|^2\left\|v'P_m\epsilon/\sqrt{N}\right\|^2\right)\right) \\ & +O_p\left(N^{-3/2}\right)O_p\left(\sum_{m=1}^{M}|w_m|\left\|\bar{H}_m^{-1}\right\|\left\|v'P_m\epsilon/\sqrt{N}\right\|\right) \\ & +o_p\left(N^{-3/2}\sum_{m=1}^{M}|w_m|\left(\left\|\bar{H}_m^{-1}\right\|^3\left\|v'P_m\epsilon/\sqrt{N}\right\|^3\right)\right).\end{aligned}$$

Consider

$$\sum_{m=1}^{M} |w_m| \left\|\bar{H}_m^{-1}\right\|^3 E\left[\left\|v'P_m\epsilon/\sqrt{N}\right\|^3 |z\right] \leq \sum_{m=1}^{M} |w_m| \left\|\bar{H}_m^{-1}\right\|^3 \left(E\left[\left\|v'P_m\epsilon/\sqrt{N}\right\|^4 |z\right]\right)^{3/4}$$

with

$$\begin{aligned}
E\left[\left\|v'P_m\epsilon/\sqrt{N}\right\|^4 |z\right] &= N^{-2}E\left[(\mathrm{tr}(v'P_m\epsilon\epsilon'P_m v))^2 |z\right] \\
&= N^{-2}\sum_{j_1,j_2}\sum_{i_1,..,i_8=1}^{N} E\left[v_{j_1,i_1}v_{j_1,i_4}v_{j_2,i_5}v_{j_2,i_8}\epsilon_{i_2}\epsilon_{i_3}\epsilon_{i_6}\epsilon_{i_7}\right] P_{m,i_1i_2}P_{m,i_3i_4}P_{m,i_5i_6}P_{m,i_7i_8} \\
&\leq CN^{-2}\sum_{j_1,j_2}\left|\sum_{i_1,..,i_4=1}^{N}\left(P_{m,i_1i_2}P_{m,i_2i_1}P_{m,i_3i_4}P_{m,i_4i_3} + 2P_{m,i_1i_2}P_{m,i_4i_1}P_{m,i_3i_4}P_{m,i_2i_3}\right)\right| \\
&\quad +CN^{-2}\sum_{j_1,j_2}\left|\sum_{i_1,..,i_4=1}^{N}\left(P_{m,i_1i_2}P_{m,i_2i_3}P_{m,i_1i_4}P_{m,i_4i_3} + 2P_{m,i_1i_2}P_{m,i_4i_3}P_{m,i_1i_4}P_{m,i_2i_3}\right)\right| \\
&\quad +CN^{-2}\sum_{j_1,j_2}\left|\sum_{i_1,..,i_4=1}^{N}\left(P_{m,i_1i_2}P_{m,i_2i_3}P_{m,i_3i_4}P_{m,i_4i_1} + 2P_{m,i_1i_2}P_{m,i_3i_2}P_{m,i_3i_4}P_{m,i_4i_1}\right)\right| \\
&\quad +\text{lower order terms,}
\end{aligned}$$

where C is a constant such that $\left|\max_{i,j}\left[\Sigma_v\right]_{i,j}\right| \sigma_\epsilon^2 \leq C$ and we use the fact that P_m is idempotent and symmetric such that $P_{m,i_1i_2} = P_{m,i_2i_1}$ and $\sum_{i_2=1}^{N} P_{m,i_1i_2}P_{m,i_2i_3} = P_{m,i_1i_3}$. This implies for example that

$$\begin{aligned}
\sum_{i_1,..,i_4=1}^{N} P_{m,i_1i_2}P_{m,i_4i_1}P_{m,i_3i_4}P_{m,i_2i_3} &= \sum_{i_1,..,i_3=1}^{N} P_{m,i_1i_2}P_{m,i_2i_3}\sum_{i_4=1}^{N} P_{m,i_1i_4}P_{m,i_4i_3} \\
&= \sum_{i_1,i_3=1}^{N} P_{m,i_1i_3}\sum_{i_2=1}^{N} P_{m,i_1i_2}P_{m,i_2i_3} = \sum_{i_1,i_3=1}^{N} P_{m,i_1i_3}^2 = \mathrm{tr}\left(P_m P_m\right) = m
\end{aligned}$$

and $\sum_{i_1,..,i_4=1}^{N} P_{m,i_1i_2}P_{m,i_2i_3}P_{m,i_1i_4}P_{m,i_4i_3} = m$, $\sum_{i_1,..,i_4=1}^{N} P_{m,i_1i_2}P_{m,i_4i_3}P_{m,i_1i_4}P_{m,i_2i_3} = m$ with the remaining terms being of lower order. This implies that

$$E\left[\left\|v'P_m\epsilon/\sqrt{N}\right\|^4 |z\right] = O\left(m/N^2\right) = o\left(1\right)$$

uniformly in $m \leq M$ and by the Markov inequality and the fact that $\left\|\bar{H}_m^{-1}\right\|$ is bounded uniformly in m that

$$\sum_{m=1}^{M} |w_m| \left(\left\|\bar{H}_m^{-1}\right\|^3 \left\|v'P_m\epsilon/\sqrt{N}\right\|^3 + \left\|\bar{H}_m^{-1}\right\|^2 \left\|v'P_m\epsilon/\sqrt{N}\right\|^2 + \left\|\bar{H}_m^{-1}\right\| \left\|v'P_m\epsilon/\sqrt{N}\right\|\right) = o_p\left(1\right).$$

Thus, we have shown that $\sum_{m=1}^{M} |w_m| \left\|\hat{\beta}_{L,m} - \beta_0\right\|^3 = O_p\left(N^{-3/2}\right)$. To summarize, it then follows that

$$\hat{\Lambda}\left(W\right) = \sum_{m=1}^{M} w_m\Lambda_m(\beta_0) - \Lambda_q\left(W\right) + O_p\left(\frac{1}{N^{3/2}}\right).$$

Since $O_p\left(N^{-3/2}\right) = N^{-1/2} o_p\left(W'\Gamma W/N\right)$, it follows that $\sqrt{N}\hat{R}_\Lambda = o_p\left(\rho_{W,N}\right)$. Now turn to $\Lambda_q\left(W\right)$, where by Lemma 7.15 and

$$\begin{aligned}
& \Lambda_{\beta,m}\left(\beta_0\right)'\left(\Lambda_{\beta\beta,m}(\beta_0)\right)^{-1}\Lambda_{\beta,m}\left(\beta_0\right) \\
= & \left(\frac{h}{\sqrt{N}} - \frac{f'(I-P_m)\epsilon}{N} + \frac{v'P_m\epsilon}{N}\right)'\left(\bar{H}_m^{-1} + o_p\left(1\right)\right)\left(\frac{h}{\sqrt{N}} - \frac{f'(I-P_m)\epsilon}{N} + \frac{v'P_m\epsilon}{N}\right) + o_p\left(1\right) \\
= & \frac{h'\bar{H}_m^{-1}h}{N} - N^{-3/2}h'\bar{H}_m^{-1}f'(I-P_m)\epsilon + N^{-3/2}h'\bar{H}_m^{-1}v'P_m\epsilon + N^{-3/2}\epsilon'(I-P_m)f\bar{H}_m^{-1}h \\
& + N^{-2}\epsilon'(I-P_m)f\bar{H}_m^{-1}f'(I-P_m)\epsilon + N^{-2}\epsilon'(I-P_m)f\bar{H}_m^{-1}v'P_m\epsilon \\
& + N^{-3/2}\epsilon'P_m v\bar{H}_m^{-1}h + N^{-2}\epsilon'P_m v\bar{H}_m^{-1}f'(I-P_m)\epsilon + N^{-2}\epsilon'P_m v\bar{H}_m^{-1}v'P_m\epsilon + \text{terms of lower order.}
\end{aligned}$$

Next, consider

$$N^{-3/2}\left\|\sum_{m=1}^{M} w_m h'\bar{H}_m^{-1}f'(I-P_m)\epsilon\right\| \leq \left\|h/N\right\|\sum_{m=1}^{M}\left|w_m\right|\left\|\bar{H}_m^{-1}\right\|\left\|f'(I-P_m)\epsilon/\sqrt{N}\right\| = O_p\left(N^{-1}\right),$$

where $\sup_{m\leq M}\left\|\epsilon'(I-P_m)f/\sqrt{N}\right\| = O_p\left(1\right)$ by Lemma 7.11 2). For $N^{-3/2}h'\bar{H}_m^{-1}v'P_m\epsilon$, note that

$$\begin{aligned}
E\left[\left\|v'P_m\epsilon/\sqrt{N}\right\|^2|z\right] &= \operatorname{tr}E\left[v'P_m\epsilon\epsilon'P_m v/N|z\right] \qquad (7.47) \\
&= \frac{m\sigma_\epsilon^2}{N}\operatorname{tr}\Sigma_v + \frac{Cum[v_i,v_i,\epsilon_i,\epsilon_i]}{N}\sum_{i=1}^{n}(P_{m,ii})^2
\end{aligned}$$

such that by the Markov inequality

$$\begin{aligned}
N^{-3/2}\left\|\sum_{m=1}^{M} w_m h'\bar{H}_m^{-1}v'P_m\epsilon\right\| &\leq \left\|h/N\right\|\sum_{m=1}^{M}\left|w_m\right|\left\|\bar{H}_m^{-1}\right\|\left\|v'P_m\epsilon/\sqrt{N}\right\| \\
&\leq O_p\left(N^{-1}\right)O_p\left(\sum_{m=1}^{M}\left|w_m\right|\sqrt{m/N}\right) = O_p\left(N^{-1}\right).
\end{aligned}$$

For $N^{-2}\epsilon'(I-P_m)f\bar{H}_m^{-1}f'(I-P_m)\epsilon$, note that

$$N^{-2}\left\|\sum_{m=1}^{M} w_m\epsilon'(I-P_m)f\bar{H}_m^{-1}f'(I-P_m)\epsilon\right\| \leq N^{-1}\sum_{m=1}^{M}\left|w_m\right|\left\|\bar{H}_m^{-1}\right\|\left\|f'(I-P_m)\epsilon/\sqrt{N}\right\|^2 = O_p\left(N^{-1}\right).$$

For $N^{-2}\epsilon'(I-P_m)f\bar{H}_m^{-1}v'P_m\epsilon$, note that

$$\begin{aligned}
N^{-2}\left\|\sum_{m=1}^{M} w_m\epsilon'(I-P_m)f\bar{H}_m^{-1}v'P_m\epsilon\right\| &\leq N^{-1}\sup_{m\leq M}\left\|\epsilon'(I-P_m)f/\sqrt{N}\right\|\sum_{m=1}^{M}\left|w_m\right|\left\|v'P_m\epsilon/\sqrt{N}\right\| \\
&= o_p\left(N^{-1}\right)
\end{aligned}$$

by Lemma 7.11 and (7.47). For $N^{-3/2}\epsilon'P_m v\bar{H}_m^{-1}h$, it follows that

$$N^{-3/2}\left\|\sum_{m=1}^{M} w_m\epsilon'P_m v\bar{H}_m^{-1}h\right\| \leq \left\|h/N\right\|\sum_{m=1}^{M}\left|w_m\right|\left\|\bar{H}_m^{-1}\right\|\left\|v'P_m\epsilon/\sqrt{N}\right\| = o_p\left(N^{-1}\right)$$

by (7.47) and the Markov inequality. For $N^{-2}\epsilon'P_m v\bar{H}_m^{-1}v'P_m\epsilon$, it holds that

$$N^{-2}\left\|\sum_{m=1}^{M} w_m\epsilon'P_m v\bar{H}_m^{-1}v'P_m\epsilon\right\| \leq N^{-1}\sum_{m=1}^{M}\left|w_m\right|\left\|\bar{H}_m^{-1}\right\|\left\|\epsilon'P_m v/\sqrt{N}\right\|^2 = o_p\left(N^{-1}\right)$$

by (7.47) and the Markov inequality. Together these results imply that

$$\sum_{m=1}^{M} w_m \left(\Lambda_{\beta,m}(\beta_0)' \left(\Lambda_{\beta\beta,m}(\beta_0)\right)^{-1} \Lambda_{\beta,m}(\beta_0)\right) = \frac{h'\bar{H}^{-1}(W)\,h}{N} + O_p\left(N^{-1}\right) = O_p\left(N^{-1}\right), \tag{7.48}$$

where $\bar{H}^{-1}(W) = \sum_{m=1}^{M} w_m \bar{H}_m^{-1}$ and $\left\|\bar{H}^{-1}(W)\right\| \leq \sum_{m=1}^{M} |w_m| \left\|\bar{H}_m^{-1}\right\| = O(1)$.

To sum up, we have

$$\begin{aligned}
\hat{\Lambda}(W) &= \sum_{m=1}^{M} w_m \Lambda_m(\beta_0) - \Lambda_q(W) + O_p\left(\frac{1}{N^{3/2}}\right) \\
&= \tilde{\Lambda}(W) - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) + \frac{(\tilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2)^2}{\tilde{\sigma}_\epsilon^2 \sigma_\epsilon^2}\tilde{\Lambda}(W) - \Lambda_q(W) + O_p\left(\frac{1}{N^{3/2}}\right) \\
&= \tilde{\Lambda}(W) - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) - \Lambda_q(W) + O_p\left(\frac{1}{N^{3/2}}\right),
\end{aligned}$$

where the last equality follows by $(\tilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2)^2 = O_p(1/N)$. This proves the first equality in the lemma.

We now consider the second equality in the lemma. We have from (7.46) that

$$\begin{aligned}
\left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) &= O_p\left(\frac{1}{\sqrt{N}}\right)\left(\frac{K'W}{N} + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right)\right) \\
&= O_p\left(\frac{K'W}{N^{3/2}}\right) + o_p\left(\frac{\rho_{W,N}}{\sqrt{N}}\right).
\end{aligned}$$

We also have that

$$\Lambda_q(W) = O_p\left(\frac{1}{N}\right)$$

form (7.48). It therefore follows that

$$\hat{\Lambda}(W) = \tilde{\Lambda}(W) + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W}{N^{3/2}}\right) + o_p\left(\frac{\rho_{W,N}}{\sqrt{N}}\right).$$

■

Lemma 7.18 *Suppose that Assumptions 1 - 4, 5(ii), 6 and 7 are satisfied. The the following statements hold:*

1. $u'P(W)u/N - \tilde{\Lambda}(W)\Sigma_u = O_p\left(\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}/N\right)$,
2. $E[h\tilde{\Lambda}(W)\epsilon'v/\sqrt{N}|z] = (K'W/N)\sum_i f_i E(\epsilon_i^2 v_i')/N + O_p(1/N) + O_p(K'W^+/N^2)$,
3. $E[hh'\bar{H}^{-1}(W)\,h/\sqrt{N}|z] = O_p(1/N)$,
4. $\sum_{m=1}^{M} w_m E\left[hh'\bar{H}_m^{-1}f'(I - P_m)\epsilon/N|z\right] = O_p(1/N)$,
5. $\sum_{m=1}^{M} w_m E\left[hh'\bar{H}_m^{-1}v'P_m\epsilon/N|z\right] = o_p(1/N)$,
6. $\sum_{m=1}^{M} w_m E\left[h\epsilon'(I - P_m)f\bar{H}_m^{-1}f'(I - P_m)\epsilon/N^{-3/2}|z\right] = O_p(1/N)$,
7. $\sum_{m=1}^{M} w_m E\left[h\epsilon'(I - P_m)f\bar{H}_m^{-1}v'P_m\epsilon/N^{-3/2}|z\right] = o_p(1/N)$,
8. $\sum_{m=1}^{M} w_m E\left[h\epsilon'P_m v\bar{H}_m^{-1}v'P_m\epsilon/N^{-3/2}|z\right] = O_p(1/N)$.

Proof. We begin with the proof of part 1. It holds that $E\left[\tilde{\Lambda}(W)|z\right] = \mathrm{tr}(P(W)E\left[\epsilon\epsilon'\right])/(N\sigma_\epsilon^2) = (K'W)/N$. We also have

$$\begin{aligned}
E\left[\left(\tilde{\Lambda}(W) - \frac{K'W}{N}\right)^2 |z\right] &= \frac{E\left[\epsilon' P(W)\epsilon\epsilon' P(W)\epsilon|z\right]}{N^2\sigma_\epsilon^4} - \left(\frac{K'W}{N}\right)^2 \\
&= \frac{\sigma_\epsilon^4 (K'W)^2 + 2\sigma^4 W'\Gamma W + O_p(\sum_i (P_{ii}(W))^2)}{N^2\sigma_\epsilon^4} - \left(\frac{K'W}{N}\right)^2 \\
&= O_p\left(\frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N^2}\right),
\end{aligned}$$

by Lemma 7.6(4) with replacing u by ϵ and Lemma 7.5(2). This gives

$$\left(\tilde{\Lambda}(W) - \frac{K'W}{N}\right)\Sigma_u = O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right).$$

Similarly, we have

$$\frac{u'P(W)u}{N} - \frac{K'W}{N}\Sigma_u = O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right).$$

Thus, part 1 is proved.

For part 2, we observe

$$\begin{aligned}
E\left[\frac{h\tilde{\Lambda}\left(W\right)\epsilon' v}{\sqrt{N}}|z\right] &= \frac{\sum_{i,j,k,l} E[f_i\epsilon_i\epsilon_j P_{jk}(W)\epsilon_k\epsilon_l v_l']}{N^2\sigma_\epsilon^2} \\
&= \frac{\sum_i f_i P_{ii}(W)E[\epsilon_i^4 v_i']}{N^2\sigma_\epsilon^2} + 2\frac{\sum_{i\neq j} f_i P_{ij}(W)E[\epsilon_j^2 v_j']}{N^2} + \frac{\sum_{i\neq j} f_i P_{jj}(W)E[\epsilon_i^2 v_i']}{N^2} \\
&= O_p\left(\frac{K'W^+}{N^2}\right) + o_p\left(\frac{K'W^+}{N^2}\right) + \frac{\sum_{i,j} f_i P_{jj}(W)E[\epsilon_i^2 v_i']}{N^2} - \frac{\sum_i f_i P_{ii}(W)E[\epsilon_i^2 v_i']}{N^2} \\
&= \frac{K'W}{N}\frac{\sum_i f_i E[\epsilon_i^2 v_i']}{N} + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W^+}{N^2}\right),
\end{aligned}$$

where Lemma 7.6(5) implies that

$$\begin{aligned}
\frac{\sum_i f_i P_{ii}(W)E[\epsilon_i^4 v_i']}{N^2\sigma_\epsilon^2} &= O_p\left(\frac{K'W^+}{N^2}\right), \\
\frac{\sum_i f_i P_{ii}(W)E[\epsilon_i^2 v_i']}{N^2\sigma_\epsilon^2} &= O_p\left(\frac{K'W^+}{N^2}\right)
\end{aligned}$$

and the fact for $f_{a,i}$ the a-th element of f_i,

$$\begin{aligned}
\left|\frac{\sum_{i\neq j} f_{a,i}P_{ij}(W)}{N^2}\right| &\leq \frac{\sum_{m=1}^M |w_m|\left|(f_a' P_m \mathbf{1}_N)\right|}{N^2} - \frac{\sum_i f_i P_{ii}(W)}{N^2} \\
&\leq \frac{\sum_{m=1}^M |w_m|\left(f' P_m f\right)^{1/2}\left(\mathbf{1}_N'\mathbf{1}_N\right)^{1/2}}{N^2} + O_p\left(\frac{K'W^+}{N^2}\right) \\
&\leq \left(\frac{f_a' f_a}{N}\right)^{1/2}\frac{\sum_{m=1}^M |w_m|}{N} + O_p\left(\frac{K'W^+}{N^2}\right) \\
&= O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W^+}{N^2}\right),
\end{aligned}$$

gives

$$\frac{\sum_{i\neq j} f_i P_{ij}(W)E[\epsilon_j^2 v_j']}{N^2} = O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W^+}{N^2}\right).$$

Part 3 follows Lemma A.8(iii) in Donald and Newey (2001). We have

$$\begin{aligned} E[hh'\bar{H}^{-1}(W)\,h/\sqrt{N}|z] &= \sum_{m=1}^{M} w_m \sum_{i_1,\ldots,i_3=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}f'_{i_2}\bar{H}_m^{-1}f_{i_3}\epsilon_{i_3}|z\right]/N^2 \\ &= \sum_{m=1}^{M} w_m \sum_{i=1}^{N} E\left[\epsilon_i^3|z\right] f_i f'_i \bar{H}_m^{-1} f_i / N^2 = O_p\left(1/N\right). \end{aligned} \tag{7.49}$$

For part 4, let $\tilde{f}'_{i,m}$ be the i-th row of $f'(I-P_m)$ such that

$$E\left[hh'\bar{H}_m^{-1}f'(I-P_m)\epsilon/N|z\right] = \sum_{i_1,\ldots,i_3=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}f'_{i_2}\bar{H}_m^{-1}\tilde{f}_{i_3,m}\epsilon_{i_3}|z\right]/N^2 = O_p\left(1/N\right)$$

by the same argument as in (7.49).

For part 5, consider

$$\begin{aligned} E\left[hh'\bar{H}_m^{-1}v'P_m\epsilon/N|z\right] &= \sum_{i_1,\ldots,i_4=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}f'_{i_2}\bar{H}_m^{-1}v_{i_3}P_{m,i_3i_4}\epsilon_{i_4}|z\right]/N^2 \\ &= \sum_{i_1,\ldots,i_4=1}^{N} f_{i_1}f'_{i_2}\bar{H}_m^{-1}P_{m,i_3i_4}E\left[\epsilon_{i_1}\epsilon_{i_2}v_{i_3}\epsilon_{i_4}|z\right]/N^2 \\ &= \sum_{i=1}^{N} f_i f'_i \bar{H}_m^{-1}P_{m,ii}\mathrm{Cum}\left[\epsilon_i,\epsilon_i,v_i,\epsilon_i|z\right]/N^2 = o_p\left(N^{-1}\right). \end{aligned}$$

For part 6, let $\tilde{f}'_{i,m}$ be the i-th row of $f'(I-P_m)$ such that

$$E\left[h\epsilon'(I-P_m)f\bar{H}_m^{-1}f'(I-P_m)\epsilon/N^{-3/2}|z\right] = \sum\nolimits_{i_1,\ldots,i_3=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}\tilde{f}'_{i_2,m}\bar{H}_m^{-1}\tilde{f}_{i_3,m}\epsilon_{i_3}|z\right]/N^2 = O_p\left(1/N\right).$$

For part 7, consider

$$\begin{aligned} E\left[h\epsilon'(I-P_m)f\bar{H}_m^{-1}v'P_m\epsilon/N^{-3/2}|z\right] &= \sum_{i_1,\ldots,i_4=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}\tilde{f}'_{i_2,m}\bar{H}_m^{-1}v_{i_3}P_{m,i_3i_4}\epsilon_{i_4}|z\right]/N^2 \\ &= \sum_{i=1}^{N} f_i\tilde{f}'_i\bar{H}_m^{-1}P_{m,ii}\mathrm{Cum}\left[\epsilon_i,\epsilon_i,v_i,\epsilon_i|z\right]/N^2 = o_p\left(N^{-1}\right). \end{aligned}$$

For part 8, consider

$$\begin{aligned} & E\left[h\epsilon'P_m v\bar{H}_m^{-1}v'P_m\epsilon/N^{-3/2}|z\right] \\ = & \sum_{i_1,\ldots,i_5=1}^{N} E\left[f_{i_1}\epsilon_{i_1}\epsilon_{i_2}P_{m,i_2i_3}v'_{i_3}\bar{H}_m^{-1}v_{i_4}P_{m,i_4i_5}\epsilon_{i_5}|z\right]/N^2 \end{aligned} \tag{7.50}$$

$$
\begin{aligned}
= & \sum_{i_1,i_2=1}^{N} \sigma_\epsilon^2 f_{i_1} P_{m,i_1 i_2} P_{m,i_2 i_2} \mathrm{tr}\left(\bar{H}_m^{-1} E\left[v_{i_2} v_{i_2}' \epsilon_{i_2} | z\right]\right)/N^2 \\
& + \sum_{i_1,i_2=1}^{N} \sigma_\epsilon^2 f_{i_1} P_{m,i_1 i_1} P_{m,i_1 i_2} \mathrm{tr}\left(\bar{H}_m^{-1} E\left[v_{i_1} v_{i_1}' \epsilon_{i_1} | z\right]\right)/N^2 \\
& + \sum_{i_1,i_2=1}^{N} \sigma_\epsilon^2 f_{i_1} P_{m,i_2 i_1} P_{m,i_1 i_2} \mathrm{tr}\left(\bar{H}_m^{-1} E\left[v_{i_1} v_{i_1}' \epsilon_{i_1} | z\right]\right)/N^2 \\
& + \sum_{i_1,i_2=1}^{N} E\left[\epsilon_{i_2}^3 | z\right] f_{i_1} P_{m,i_2 i_1} P_{m,i_1 i_2} \mathrm{tr}\left(\bar{H}_m^{-1} E\left[v_{i_1} v_{i_1}' | z\right]\right)/N^2 \\
& + \sum_{i=1}^{N} f_i P_{m,ii}^2 \mathrm{tr}\left(\bar{H}_m^{-1} \mathrm{Cum}\left[\epsilon_i, \epsilon_i, v_i, v_i', \epsilon_i | z\right]\right)/N^2,
\end{aligned}
$$

where $E\left[v_i v_i' \epsilon_i | z\right]$ does not depend on z by Assumption 3 and for the first term in (7.50) we have

$$
\sum_{i_1,i_2=1}^{N} f_{i_1} P_{m,i_1 i_2} P_{m,i_2 i_2} = \sum_{i_1,i_2=1}^{N} f_{i_1} P_{m,i_1 i_2} = f' P_m \mathbf{1}_N \leq \left(f'f\right)^{1/2} \left(\mathbf{1}_N' P_m \mathbf{1}_N\right)^{1/2} \leq \sqrt{N} \left(f'f\right)^{1/2}
$$

such that

$$
\sum_{i_1,i_2=1}^{N} \sigma_\epsilon^2 f_{i_1} P_{m,i_1 i_2} P_{m,i_2 i_2} \mathrm{tr}\left(\bar{H}_m^{-1} E\left[v_{i_2} v_{i_2}' \epsilon_{i_2} | z\right]\right)/N^2 = N^{-1}\left(f'f/N\right)^{1/2} = O_p\left(N^{-1}\right)
$$

where a similar arguments shows that the second term in (7.50) is $O_p\left(N^{-1}\right)$. Next,

$$
\sum_{i_1,i_2=1}^{N} f_{i_1} P_{m,i_2 i_1} P_{m,i_1 i_2} = \sum_{i_1,i_2=1}^{N} f_{i_1} P_{m,i_1 i_2} P_{m,i_2 i_1} = \sum_{i=1}^{N} f_i P_{m,ii} \leq \sup_i \|f_i\| \sum_{i=1}^{N} P_{m,ii} = O_p(m),
$$

where $\sup_i \|f_i\| = O_p(1)$ by Assumption 3(iv) such that the third term in (7.50) is $O_p\left(m/N^2\right) = o_p\left(N^{-1}\right)$ and the same argument also shows that the fourth term in (7.50) is $o_p\left(N^{-1}\right)$. Finally,

$$
\begin{aligned}
\sum_{i=1}^{N} f_i P_{m,ii}^2 \mathrm{tr}\left(\bar{H}_m^{-1} \mathrm{Cum}\left[\epsilon_i, \epsilon_i, v_i, v_i', \epsilon_i | z\right]\right)/N^2 & \leq \left|\mathrm{tr}\left(\bar{H}_m^{-1} \mathrm{Cum}\left[\epsilon_i, \epsilon_i, v_i, v_i', \epsilon_i | z\right]\right)\right| \sup_i \|f_i\| \sum_{i=1}^{N} P_{m,ii}^2 \\
& = o_p\left(m/N^2\right) = o_p\left(N^{-1}\right).
\end{aligned}
$$

These results establish that $\sum_{m=1}^{M} w_m E\left[h \epsilon' P_m v \bar{H}_m^{-1} v' P_m \epsilon / N^{-3/2} | z\right] = O_p\left(N^{-1}\right)$ as desired. ■

7.4 Proof of Theorem 7.2

Proof. The MALIML estimator, $\hat{\beta}$ defined in (2.3), has the form:

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

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

$$\begin{aligned}
\hat{h} &= h+\sum_{j=1}^{5}T_j^h+Z^h, \\
T_1^h &= -f'(I-P(W))\epsilon/\sqrt{N}, \quad T_2^h = v'P(W)\epsilon/\sqrt{N}, \\
T_3^h &= -\tilde{\Lambda}\left(W\right)\frac{f'\epsilon}{\sqrt{N}}, \quad T_4^h = -\tilde{\Lambda}\left(W\right)\frac{v'\epsilon}{\sqrt{N}}, \quad T_5^h = \sqrt{N}\Lambda_q\left(W\right)\sigma_{u\epsilon}, \\
Z^h &= \left(\tilde{\Lambda}\left(W\right)-\hat{\Lambda}\left(W\right)+\hat{R}_\Lambda\right)\sqrt{N}\left(\frac{X'\epsilon}{N}-\sigma_{u\epsilon}\right)-\hat{R}_\Lambda\frac{X'\epsilon}{\sqrt{N}},
\end{aligned}$$

and

$$\begin{aligned}
\hat{H} &= H+\sum_{j=1}^{3}T_j^H+Z^H, \\
T_1^H &= -f'(I-P(W))f/N, \quad T_2^H=(u'f+f'u)/N, \quad T_3^H=-\tilde{\Lambda}\left(W\right)f'f/N \\
Z^H &= u'P(W)u/N-\tilde{\Lambda}\left(W\right)\Sigma_u-u'(I-P(W))f/N-f'(I-P(W))u/N \\
&\quad +\tilde{\Lambda}\left(W\right)(H+\Sigma_u)-\hat{\Lambda}\left(W\right)X'X/N.
\end{aligned}$$

Let $T^h=\sum_{j=1}^{5}T_j^h$ and $T^H=\sum_{j=1}^{3}T_j^H$. We give the order of each term. By Lemma 7.5(6), we have

$$h=O_p(1) \text{ and } H=O_p(1). \tag{7.51}$$

Lemma 7.6(2) gives

$$T_1^h=O_p(\Delta(W)^{1/2}). \tag{7.52}$$

A similar argument to Lemma 7.6(4) (note that $E[v_i\epsilon_i]=0$), we have

$$T_2^h=O_p\left(\sqrt{\frac{W'\Gamma W+\sum_i(P_{ii}(W))^2}{N}}\right). \tag{7.53}$$

Lemma 7.17 and the CLT gives

$$\begin{aligned}
T_3^h &= \left(\frac{K'W}{N}+O_p\left(\frac{\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}}{N}\right)\right)O_p(1) \\
&= O_p\left(\frac{K'W}{N}+\frac{\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}}{N}\right) \qquad (7.54)
\end{aligned}$$

and

$$T_4^h=O_p\left(\frac{K'W}{N}+\frac{\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}}{N}\right). \tag{7.55}$$

By Lemma 7.17, we have

$$T_5^h=O_p\left(\frac{1}{\sqrt{N}}\right). \tag{7.56}$$

By definition, we have

$$T_1^H=O_p(\Xi(W)), \tag{7.57}$$

where $\Xi(W)$ is defined in (7.10). By a CLT, we have

$$T_2^H = O_p\left(\frac{1}{\sqrt{N}}\right). \tag{7.58}$$

By Lemma 7.17 and Lemma 7.5(6), it holds that

$$\begin{aligned} T_3^H &= \left(\frac{K'W}{N} + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right)\right) O_p(1) \\ &= O_p\left(\frac{K'W}{N} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right). \end{aligned} \tag{7.59}$$

By Lemma 7.17 together with the CLT which implies that $\sqrt{N}\left(\frac{X'\epsilon}{N} - \sigma_{u\epsilon}\right) = O_p(1)$, as well as

$$\begin{aligned} \tilde{\Lambda}(W) - \hat{\Lambda}(W) + \hat{R}_\Lambda &= \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda}(W) + \Lambda_q(W) \\ &= O_p\left(N^{-1/2}\right) O_p\left(\frac{K'W}{N} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right) + O_p\left(N^{-1}\right), \end{aligned}$$

it follows that

$$\begin{aligned} Z^h &= O_p\left(\frac{K'W}{N} + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right)\right) O_p\left(\frac{1}{\sqrt{N}}\right) \\ &\quad + O_p\left(\frac{1}{N}\right) + o_p(\rho_{W,N}) O_p(1) \\ &= O_p\left(\frac{K'W}{N^{3/2}} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N^{3/2}}\right) + O_p\left(\frac{1}{N}\right) + o_p(\rho_{W,N}) \\ &= o_p(\rho_{W,N}), \end{aligned} \tag{7.60}$$

where $1/N = o_p(W'\Gamma W/N) = o_p(\rho_{W,N})$. Lastly, we have

$$\begin{aligned} \tilde{\Lambda}(W)(H + \Sigma_u) - \hat{\Lambda}(W) X'X/N &= \tilde{\Lambda}(W)(H + \Sigma_u - X'X/N) - \left(\hat{\Lambda}(W) - \tilde{\Lambda}(W)\right) X'X/N \\ &= o_p(\rho_{W,N}) + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W}{N^{3/2}}\right) + o_p\left(\frac{\rho_{W,N}}{\sqrt{N}}\right) = o_p(\rho_{W,N}), \end{aligned}$$

where $(H + \Sigma_u - X'X/N) = O_p\left(1/\sqrt{N}\right)$ and $\hat{\Lambda}(W) - \tilde{\Lambda}(W) = O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W}{N^{3/2}}\right) + o_p\left(\frac{\rho_{W,N}}{\sqrt{N}}\right)$ from Lemma 7.17. It then follows that

$$\begin{aligned} Z^H &= o_p(\rho_{W,N}) + O_p\left(\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right) + O_p\left(\frac{\Delta(W)^{1/2}}{\sqrt{N}}\right) \\ &= o_p(\rho_{W,N}), \end{aligned} \tag{7.61}$$

by Lemmas 7.18(1), 7.6(2), 7.17, the CLT and the LLN.

We show below that the conditions of Lemma A.1 of Donald and Newey (2001) are satisfied and $S(W)$ has the form given in the theorem.

We first have $h = O_p(1)$ and $H = O_p(1)$ by (7.51). Next, we need to show that $T^h = o_p(1)$. By (7.52), (7.53), (7.54), (7.55) and (7.56), it follows that

$$\begin{aligned} T_1^h &= O_p\left(\Delta(W)^{1/2}\right) + O_p\left(\sqrt{\frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N}}\right) \\ &\quad + O_p\left(\frac{K'W}{N} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right) + O_p\left(\frac{1}{\sqrt{N}}\right). \end{aligned}$$

Now, Lemma 7.6(2) says that $\Delta(W) = o_p(1)$. We have $|K'W/N| \leq K'W^+/N \to 0$ by Assumption 5. By Lemma 7.6(12) and Assumption 5, it holds that $W'\Gamma W/N \leq CK'W^+/N \to 0$, where C is some constant. Lemma 7.5(2) implies that $\sum_i (P_{ii}(W))^2/N = o_p(K'W^+/N) = o_p(1)$. Thus, $T_1^h = o_p(1)$ is shown.

The next step is to show that $||T^H||^2 = o_p(\rho_{W,N})$. We have, by (7.57), (7.58) and (7.59),

$$\begin{aligned} ||T^H||^2 &= O_p\Big(\Xi(W)^2 + \frac{1}{N} + \frac{\Xi(W)}{\sqrt{N}} + \frac{(K'W)^2}{N^2} + \frac{|K'W|}{N}\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N} \\ &\quad + \frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N^2} + \Xi(W)\frac{|K'W|}{N} + \Xi(W)\frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N} \\ &\quad + \frac{|K'W|}{N^{3/2}} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N^{3/2}}\Big). \end{aligned}$$

Since $\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}/N = O_p\left(\left(W'\Gamma W + \sum_i (P_{ii}(W))^2\right)/N\right) = o_p(\rho_{W,N})$, $|K'W|/N^{3/2} = o(|K'W|/N) = o_p(\rho_{W,N})$, $(K'W)^2/N^2 = o(K'W/N) = o_p(\rho_{W,N})$, $1/N = o_p(\rho_{W,N})$ and the observation that $\Xi(W)/\sqrt{N} = o_p(\rho_{W,N})$ by Lemma 7.6(6) and $\Xi(W) = O_p(\Delta(W)^{1/2})$, we have

$$||T^H||^2 = O_p\left((\Xi(W))^2\right) + o_p(\rho_{W,N}).$$

The order of $(\Xi(W))^2$ is $o_p(\rho_{W,N})$ by Lemma 7.7. Next, we consider $||T^h|| \cdot ||T^H||$. We have, by (7.52) - (7.59),

$$\begin{aligned} & ||T^h|| \cdot ||T^H|| \\ = \;& O_p\left(\Delta(W)^{1/2} + \sqrt{\frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N}} + \frac{|K'W|}{N} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N} + \frac{1}{N}\right) \\ & \cdot O_p\left(\Xi(W) + \frac{1}{\sqrt{N}} + \frac{|K'W|}{N} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N}\right) \\ = \;& O_p\left(\Delta(W)^{1/2}\Xi(W) + \frac{\Delta(W)^{1/2}}{\sqrt{N}} + \frac{\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}}{N} + \Xi(W)\sqrt{\frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N}}\right) \\ & + o_p(\rho_{W,N}) \\ = \;& O_p\left(\Delta(W)^{1/2}\Xi(W) + \Xi(W)\sqrt{\frac{W'\Gamma W + \sum_i (P_{ii}(W))^2}{N}}\right) + o_p(\rho_{W,N}) = o_p(\rho_{W,N}), \end{aligned}$$

since $o_p(1)|K'W|/N = o_p(\rho_{W,N})$, $\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}/N = o_p(\rho_{W,N})$, $1/N = o_p(\rho_{W,N})$, $\Delta(W)^{1/2}/\sqrt{N} = o_p(\rho_{W,N})$ by Lemma 7.6(6) and the order of $\Xi(W)$ is $o_p(\Delta(W)^{1/2})$ by Lemma 7.7. Lastly, it holds that $Z^h = o_p(\rho_{W,N})$ and $Z^H = o_p(\rho_{W,N})$ by (7.60) and (7.61).

We have shown that the conditions of Lemma A.1 of Donald and Newey (2001)[12] are satisfied and we apply the lemma with

$$\begin{aligned}\hat{A}(W) &= (h+T_1^h+T_2^h)(h+T_1^h+T_2^h)'+h(T_3^h+T_4^h+T_5^h)'+(T_3^h+T_4^h+T_5^h)h' \\ &\quad -hh'H^{-1}(T_1^H+T_2^H+T_3^H)-(T_1^H+T_2^H+T_3^H)H^{-1}hh'\end{aligned}$$

and

$$\begin{aligned}Z^A(W) &= (T_3^h+T_4^h+T_5^h)(T_3^h+T_4^h+T_5^h)' \\ &\quad +(T_3^h+T_4^h+T_5^h)(T_1^h+T_2^h)'+(T_1^h+T_2^h)(T_3^h+T_4^h+T_5^h).\end{aligned}$$

We show that $Z^A(W)=o_p(\rho_{W,N})$. By (7.54), (7.55) and the fact that $\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}/N = o_p(\rho_{W,N})$, it holds that

$$(T_3^h+T_4^h)(T_3^h+T_4^h)'=O_p\left(\left(\frac{K'W}{N}\right)^2\right)+o_p(\rho_{W,N})=o_p(\rho_{W,N}).$$

By (7.54), (7.55), (7.56) and the fact that $\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}/N^{3/2}=o_p(\rho_{W,N})$, it holds that

$$T_5^h(T_3^h+T_4^h)'=O_p\left(\frac{K'W}{N^{3/2}}\right)+o_p(\rho_{W,N})=o_p(\rho_{W,N}).$$

By (7.56), we have

$$T_5^h(T_5^h)'=O_p\left(\frac{1}{N}\right)=o_p(\rho_{W,N}).$$

By (7.52), (7.54), (7.55) and the fact that $\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}/N=o_p(\rho_{W,N})$, we have

$$T_1^h(T_3^h+T_4^h)=O_p\left(\Delta(W)^{1/2}\frac{K'W}{N}\right)+o_p(\rho_{W,N})=o_p(\rho_{W,N}),$$

since $\Delta(W)^{1/2}=o_p(1)$ by Lemma 7.6(2). By (7.53), (7.54), (7.55) and the fact that $\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}/N = o_p(\rho_{W,N})$, it follows that

$$T_2^h(T_3^h+T_4^h)=O_p\left(\frac{K'W}{N}\sqrt{\frac{W'\Gamma W+\sum_i(P_{ii}(W))^2}{N}}\right)+o_p(\rho_{W,N})=o_p(\rho_{W,N}).$$

Lemma 7.6(6), (7.52) and (7.56) imply that

$$T_5^h(T_1^h)'=O_p\left(\frac{\Delta(W)^{1/2}}{\sqrt{N}}\right)=o_p(\rho_{W,N}).$$

Lastly, we have

$$T_5^h(T_2^h)'=O_p\left(\frac{\sqrt{W'\Gamma W+\sum_i(P_{ii}(W))^2}}{N}\right)=o_p(\rho_{W,N}),$$

[12] We note that here we do not need to use our Lemma 7.1, which is a modified version of Lemma A.1 Donald and Newey (2001).

by (7.53), (7.56) and the fact that $\sqrt{W'\Gamma W + \sum_i (P_{ii}(W))^2}/N = o_p(\rho_{W,N})$. To sum up, we have $Z^A(W) = o_p(\rho_{W,N})$.

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, we have

$$E[hT_1^{h\prime}|z] = E\left[-\frac{f'\epsilon\epsilon'(I-P(W))f}{N}|z\right] = -\sigma_\epsilon^2\frac{f'(I-P(W))f}{N}.$$

Similarly, it holds that $E[T_1^h h'|z] = -\sigma_\epsilon^2 f'(I-P(W))f/N$. Third, Lemma 7.6(5) with replacing u by v gives

$$E[hT_2^{h\prime}|z] = \sum_{i=1}^N f_i P_{ii}(W)E[\epsilon_i^2 v_i|z]/N,$$

which is $O_p(K'W^+/N)$. Fourth,

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

Fifth, by Lemma 7.6(8) with replacing u by v, we obtain

$$E[T_1^h T_2^{h\prime}|z] = -E\left[\frac{f'(I-P(W))\epsilon\epsilon' P(W)v}{N}|z\right] = -\frac{f'(I-P(W))\mu_v(W)}{N},$$

where $\mu_v(W) = (\mu_{v,1}(W), \ldots, \mu_{v,N}(W))$ and $\mu_{v,i} = P_{ii}(W)E[\epsilon_i^2 v_i]$. Similarly, we have $E[T_2^h T_1^{h\prime}|z] = -\mu_v(W)(I-P(W))f/N$. Sixth, noting that $E[v_i\epsilon_i|z] = 0$, a similar argument as Lemma 7.6(4) gives

$$E[T_2^h T_2^{h\prime}|z] = \sigma_\epsilon^2\Sigma_v(W'\Gamma W)/N + Cum[\epsilon_i,\epsilon_i,v_i,v_i'] \sum_i (P_{ii}(W))^2/N.$$

Seventh, we have

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

Similarly, we have $E[T_1^H H^{-1}hh'|z] = -\sigma_\epsilon^2 f'(I-P(W))f/N$. Eighth, Lemma 7.6(7) implies that

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

and that $E[T_2^H H^{-1}hh'|z] = o_p(\rho_{W,N})$. Ninth, we have

$$h(T_3^h)' - hh'H^{-1}(T_3^H)' = T_3^h h' - T_3^H H^{-1}hh' = 0.$$

Tenth, we have

$$\begin{aligned} E[h(T_4^h)'|z] &= -\frac{K'W}{N}\frac{\sum_{i=1}^N f_i E[\epsilon_i^2 u_i]}{N} + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{K'W^+}{N^2}\right) \\ &= -\frac{K'W}{N}\frac{\sum_{i=1}^N f_i E[\epsilon_i^2 u_i]}{N} + o_p(\rho_{W,N}), \end{aligned}$$

by Lemma 7.18(2). Similarly, we have $E[T_4^h h'|z] = -(K'W/N)(\sum_{i=1}^N f_i E[\epsilon_i^2 u_i]/N) + o_p(\rho_{W,N})$. Lastly, Lemma 7.18(3)-(8) implies that

$$E[h(T_5^h)'|z] = O_p\left(\frac{1}{N}\right) = o_p(\rho_{W,N})$$

and that $E[T_5^h h'|z] = o_p(\rho_{W,N})$.

Let

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

Note that $\hat{\zeta} = 0$ under the third moment condition in the Theorem. Therefore, we have

$$\begin{aligned}
E(\hat{A}(K)) &= \sigma_\epsilon^2 H - \sigma_\epsilon^2 \frac{f'(I-P(W))f}{N} - \sigma_\epsilon^2 \frac{f'(I-P(W))f}{N} - \frac{f'(I-P(W))\mu_v(W)}{N} - \frac{\mu_v(W)'(I-P(W))f}{N} \\
&\quad + \sigma_\epsilon^2 \Sigma_v \frac{W'\Gamma W}{N} + Cum[\epsilon_i, \epsilon_i, v_i, v_i'] \frac{\sum_i (P_{ii}(W))^2}{N} + \hat{\zeta} + \hat{\zeta}' \\
&\quad + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))f}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))f}{N} + o_p(\rho_{W,N}) \\
&= \sigma_\epsilon^2 H + \sigma_\epsilon^2 \Sigma_v \frac{W'\Gamma W}{N} + \sigma_\epsilon^2 \frac{f'(I-P(W))(I-P(W))f}{N} \\
&\quad + Cum[\epsilon_i, \epsilon_i, v_i, v_i'] \frac{\sum_i (P_{ii}(W))^2}{N} - \frac{f'(I-P(W))\mu_v(W)}{N} - \frac{\mu_v(W)'(I-P(W))f}{N} \\
&\quad + \hat{\zeta} + \hat{\zeta}' + o_p(\rho_{W,N}).
\end{aligned}$$

By Lemma A.1 of Donald and Newey (2001), we have the desired result.

For the MAFuller estimator $\hat{\beta}$ defined in (2.4) the result can be established by noting the following. By the construction of $\hat{\Lambda}_m$, we have $0 \le 1 - \hat{\Lambda}_m \le 1$. Therefore,

$$0 < \check{\Lambda}_m - \hat{\Lambda}_m = \frac{\frac{\alpha}{N-m}(1-\hat{\Lambda}_m)^2}{1 - \frac{\alpha}{N-m}(1-\hat{\Lambda}_m)} = \frac{\alpha((1-\hat{\Lambda}_m)^2)}{N-m-\alpha(1-\hat{\Lambda}_m)} \le \frac{\alpha}{N-M-\alpha} = O\left(\frac{1}{N}\right)$$

uniformly over m. It therefore follows that

$$\check{\Lambda}(W) = \hat{\Lambda}(W) + O_p(1/N). \tag{7.62}$$

Now let $\rho_{W,N} = tr(S(W))$. We have

$$\begin{aligned}
\frac{X'P(W)X}{N} - \check{\Lambda}(W)\frac{X'X}{N} &= \frac{X'P(W)X}{N} - \hat{\Lambda}(W)\frac{X'X}{N} + O_p\left(\frac{1}{N}\right) \\
&= \frac{X'P(W)X}{N} - \hat{\Lambda}(W)\frac{X'X}{N} + o_p\left(\rho_{W,N}\right),
\end{aligned}$$

by (7.62), $X'X/N = O_p(1)$ and $1/N = o_p(\rho_{W,N})$. Similarly, we have

$$\begin{aligned}
\frac{X'P(W)\epsilon}{\sqrt{N}} - \check{\Lambda}(W)\frac{X'\epsilon}{\sqrt{N}} &= \frac{X'P(W)\epsilon}{\sqrt{N}} - \hat{\Lambda}(W)\frac{X'\epsilon}{\sqrt{N}} + O_p\left(\frac{1}{N}\right) \\
&= \frac{X'P(W)\epsilon}{\sqrt{N}} - \hat{\Lambda}(W)\frac{X'\epsilon}{\sqrt{N}} + o_p\left(\rho_{W,N}\right).
\end{aligned}$$

Therefore, the higher order mean square errors of the MALIML and the MAFuller estimator are the same.

■

7.5 Verification of Regularity Conditions for Unconstrained Optimal Weights

In order to demonstrate that the regularity conditions imposed are not too stringent, it is useful to consider various optimal weights and verify that the conditions hold. We note that when Ω is equal to Ω_U or Ω_B, a closed form solution for W^* is available. Let $\tilde{\gamma}_m = \lambda' H^{-1} f'(I - P_m) f H^{-1}\lambda / N$ and U be the matrix whose (i,j)-element is $\tilde{\gamma}_{\max(i,j)}$ so that $\lambda' H^{-1} f'(I - P(W))(I - P(W)) f H^{-1}\lambda/N = W'UW$. This implies that $S_\lambda(W)$ is quadratic function in W and the optimal weight is given by solving the first order condition. For the MALIML estimator with $\Omega = \Omega_U$, we have

$$W^* = (\mathbf{1}_M'(U + \sigma_v^2\Gamma)^{-1}\mathbf{1}_M)^{-1}(U + \sigma_v^2\Gamma)^{-1}\mathbf{1}_M = \begin{pmatrix} \frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_1 - \tilde{\gamma}_2)} \\ -\frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_1 - \tilde{\gamma}_2)} + \frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_2 - \tilde{\gamma}_3)} \\ \vdots \\ -\frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_{M-2} - \tilde{\gamma}_{M-1})} + \frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_{M-1} - \tilde{\gamma}_M)} \\ -\frac{\sigma_v^2}{\sigma_v^2 + N(\tilde{\gamma}_{M-1} - \tilde{\gamma}_M)} + 1 \end{pmatrix}$$

such that

$$\sum_{s=1}^{j} w_s = \frac{\sigma_v^2}{\sigma_v^2 + N\left(\tilde{\gamma}_j - \tilde{\gamma}_{j+1}\right)}.$$

It follows that for some $\varepsilon > 0$

$$\left|\sum_{s=1}^{j} w_s\right| \leq \frac{j^{2\alpha+1}}{N}\frac{\sigma_v^2}{j^{2\alpha+1}\sigma_v^2/N + \varepsilon} \text{ } wpa1 \text{ for } j \notin \bar{J}$$

and

$$\left|\sum_{s=1}^{j} w_s\right| \leq \frac{L^{2\alpha+1}}{N}\frac{\sigma_v^2}{\varepsilon} \text{ for } j \notin \bar{J}, j \leq L$$

such that, for $L = O\left(N^{\frac{1}{2(2\alpha+1)}}\right)$, it follows that

$$\sup_{j \notin \bar{J}, j \leq L}\left|\sum_{s=1}^{j} w_s\right| = O_p\left(1/\sqrt{N}\right).$$

The case of MA2SLS with $\Omega = \Omega_U$ is handled next. The optimal weight is given by

$$\begin{aligned} & W_U^* \\ = \; & \arg\min_{W \in \Omega_U} S_\lambda(W) = \frac{1}{2}A^{-1}\left(K\lambda' H^{-1} B_N H^{-1}\lambda + \frac{2 - \mathbf{1}_M' A^{-1} K \lambda' H^{-1} B_N H^{-1}\lambda}{\mathbf{1}_M' A^{-1}\mathbf{1}_M}\mathbf{1}_M\right) \\ = \; & e_M + \frac{1}{2}\frac{2(\sigma_\epsilon^2\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2 + M\sigma_{\lambda\epsilon}^2) - B_\lambda}{\sigma_\lambda^2\sigma_\epsilon^2 + \sigma_{\lambda\epsilon}^2 + \sigma_{\lambda\epsilon}^2\sum_{j=1}^{M-1}\frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})}}\begin{pmatrix} \frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_1 - \tilde{\gamma}_2)} \\ -\frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_1 - \tilde{\gamma}_2)} + \frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_2 - \tilde{\gamma}_3)} \\ \vdots \\ -\frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_{M-1} - \tilde{\gamma}_M)} \end{pmatrix}. \end{aligned}$$

First, consider

$$\sum_{j=1}^{M-1}\frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2 + N(\tilde{\gamma}_j - \tilde{\gamma}_{j+1})} \leq O_p\left(\frac{1}{N}\sum_{j=1}^{M-1} j^{2\alpha+1}\frac{\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{j^{2\alpha+1}\left(\sigma_\lambda^2 + \sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2\right)/N + \varepsilon}\right) = O_p\left(\frac{M^{2\alpha+2}}{N}\right)$$

such that

$$\frac{2(\sigma_\epsilon^2\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2+M\sigma_{\lambda\epsilon}^2)-B_\lambda}{\sigma_\lambda^2\sigma_\epsilon^2+\sigma_{\lambda\epsilon}^2+\sigma_{\lambda\epsilon}^2\sum_{j=1}^{M-1}\frac{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2+N(\tilde{\gamma}_j-\tilde{\gamma}_{j+1})}}=\begin{cases} O_p\left(M\right) & \text{if } \frac{M^{2\alpha+2}}{N}=O\left(1\right), \\ O_p\left(M^{-2\alpha+1}N\right) & \text{otherwise.}\end{cases}$$

By the same argument as before we have

$$\left|\sum_{s=1}^{j}\frac{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2+N(\tilde{\gamma}_1-\tilde{\gamma}_2)}\right|\leq\frac{L^{2\alpha+1}}{N}\frac{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2}{\sigma_\lambda^2+\sigma_{\lambda\epsilon}^2/\sigma_\epsilon^2+\varepsilon}\text{ for } j\notin\bar{J}, j\leq L$$

such that

$$\sup_{j\notin\bar{J},j\leq L}\left|\sum_{s=1}^{j}w_s\right|=\begin{cases} O_p\left(\frac{ML^{2\alpha+1}}{N}\right) & \text{if } \frac{M^{2\alpha+2}}{N}=O\left(1\right), \\ O_p\left(M^{-2\alpha+1}L^{2\alpha+1}\right) & \text{otherwise,}\end{cases}$$

where in the first case the desired rate obtains if

$$L=o\left(\left(\frac{N}{M}\right)^{1/(2\alpha+1)}\right)$$

and in the second case if

$$L=o\left(M^{\frac{2\alpha-1}{2\alpha+1}}\right).$$

Note that when $M=N$ it follows that $\frac{M^{2\alpha+2}}{N}\to\infty$ such that the second case applies and $L=o\left(N^{\frac{2\alpha-1}{2\alpha+1}}\right)$ delivers the desired result. The condition $\frac{M^{2\alpha+2}}{N}=O\left(1\right)$ may be too restrictive in practice because the optimal rate in the case of Donald and Newey (2001) is $M=O\left(N^{1/(2a+2)}\right)$, indicating that the upper bound M should grow faster for MA type estimators. This indicates that the second case is more relevant in practice. The constraint on L in the second case is mild since L can go to infinity at arbitrarily slow rates for Lemma 7.7 to hold.

References

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

Canay, I. A. (2008). 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.

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

Doornik, J. A. (2007). *Ox 5 - An Object-oriented Matrix Programming Language*, Timberlake Consultants Ltd.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator, *Econometrica* **45**: 939–954.

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

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

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

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.

Lewis, R. and Reinsel, G. C. (1985). Prediction of mulitivariate time series by autoregressive model fitting, *Journal of Multivariate Analysis* **16**(3): 393–411.

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.

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.

Okui, R. (2008). Instrumental variable estimation in the presence of many moment conditions, forthcoming in the Journal of Econometrics.

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. (2007). Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices, unpublished manuscript.

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

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), 2SLS

		2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps	2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps
$c = 0.1$					$R_f^2 = 0.01$							$R_f^2 = 0.1$			
$n = 100$	bias	0.0954	0.0904	0.0919	0.0943	0.0954	0.0957	0.0918	0.0597	0.0591	0.061	0.0593	0.0593	0.0594	0.0589
$K = 20$	IQR	0.294	0.836	0.843	0.312	0.316	0.335	0.635	0.24	0.367	0.378	0.248	0.248	0.263	0.339
	MAD	0.163	0.43	0.429	0.17	0.171	0.183	0.324	0.129	0.194	0.198	0.132	0.132	0.136	0.176
	RMAD	0.379	1	0.997	0.395	0.398	0.425	0.753	0.666	1	1.02	0.681	0.682	0.7	0.905
	KW+	20	4.47	3.28	72.4	54	9.89	4.9	20	9.43	6.44	80.5	54	13.5	8.22
	KW-	0	0	0	65.6	46.2	0	0	0	0	0	71.6	43.4	0	0
$n = 1000$															
$K = 30$	bias	0.0731	0.064	0.0678	0.0746	0.0739	0.0731	0.0731	0.0224	0.0222	0.0212	0.0235	0.023	0.0226	0.0208
	IQR	0.218	0.641	0.619	0.228	0.231	0.259	0.43	0.115	0.115	0.132	0.112	0.113	0.115	0.12
	MAD	0.122	0.333	0.318	0.126	0.129	0.139	0.227	0.0598	0.0588	0.0672	0.0584	0.0584	0.0594	0.0604
	RMAD	0.365	1	0.953	0.378	0.386	0.416	0.682	1.02	1	1.14	0.992	0.992	1.01	1.03
	KW+	30	7.65	5.43	189	124	14.4	7.54	30	29	15.5	128	68.8	27.9	23.4
	KW-	0	0	0	180	114	0	0	0	0	0	106	43.6	0	0
$c = 0.5$															
$n = 100$	bias	0.475	0.473	0.475	0.473	0.474	0.472	0.467	0.317	0.317	0.317	0.319	0.319	0.308	0.305
$K = 20$	IQR	0.262	0.765	0.755	0.277	0.277	0.304	0.567	0.222	0.353	0.356	0.227	0.226	0.235	0.308
	MAD	0.475	0.62	0.607	0.473	0.474	0.472	0.535	0.317	0.366	0.361	0.32	0.32	0.309	0.334
	RMAD	0.766	1	0.979	0.763	0.765	0.761	0.864	0.865	1	0.987	0.873	0.874	0.843	0.912
	KW+	20	4.41	3.23	71.9	54.1	9.81	4.87	20	8.13	5.87	81.8	56	12.4	7.43
	KW-	0	0	0	65.2	46.4	0	0	0	0	0	73.9	46.4	0	0
$n = 1000$															
$K = 30$	bias	0.374	0.358	0.36	0.377	0.378	0.368	0.359	0.105	0.14	0.113	0.142	0.15	0.113	0.12
	IQR	0.194	0.588	0.559	0.206	0.207	0.22	0.391	0.109	0.158	0.127	0.108	0.106	0.109	0.114
	MAD	0.374	0.457	0.459	0.377	0.378	0.368	0.407	0.106	0.152	0.12	0.142	0.15	0.114	0.121
	RMAD	0.819	1	1.01	0.826	0.828	0.806	0.892	0.702	1	0.79	0.936	0.987	0.754	0.8
	KW+	30	7.1	5.21	187	125	13.7	7.13	30	12.9	12.9	207	120	16.6	13.2
	KW-	0	0	0	179	115	0	0	0	0	0	197	107	0	0
$c = 0.9$															
$n = 100$	bias	0.858	0.851	0.852	0.858	0.858	0.854	0.849	0.571	0.566	0.563	0.581	0.584	0.565	0.563
$K = 20$	IQR	0.142	0.396	0.398	0.153	0.153	0.162	0.3	0.152	0.402	0.374	0.15	0.149	0.164	0.239
	MAD	0.858	0.881	0.877	0.858	0.858	0.854	0.853	0.571	0.615	0.599	0.581	0.584	0.565	0.571
	RMAD	0.974	1	0.996	0.974	0.975	0.97	0.969	0.929	1	0.973	0.944	0.949	0.918	0.928
	KW+	20	4.18	3.12	71.7	54.4	9.52	4.65	20	3.81	3.38	77.4	59.4	8.65	4.47
	KW-	0	0	0	65.2	46.9	0	0	0	0	0	71.6	52.8	0	0
$n = 1000$															
$K = 30$	bias	0.673	0.674	0.669	0.681	0.683	0.666	0.663	0.188	0.233	0.202	0.3	0.314	0.215	0.229
	IQR	0.123	0.432	0.422	0.125	0.125	0.137	0.26	0.0898	0.255	0.205	0.0877	0.086	0.0976	0.121
	MAD	0.673	0.718	0.709	0.681	0.683	0.666	0.671	0.188	0.266	0.222	0.3	0.314	0.215	0.229
	RMAD	0.937	1	0.988	0.949	0.952	0.928	0.934	0.707	1	0.833	1.13	1.18	0.807	0.859
	KW+	30	3.69	3.06	172	129	9.7	4.46	30	2.98	3.56	177	127	8.73	4.97
	KW-	0	0	0	166	122	0	0	0	0	0	171	120	0	0

Note: “bias” = median bias; “IQR” = inter-quantile range; “MAD” = median absolute deviation; “RMAD” = MAD relative to that of DN; “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 (b), 2SLS

		2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps	2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps
$c = 0.1$					$R_f^2 = 0.01$							$R_f^2 = 0.1$			
$n = 100$	bias	0.0927	0.0726	0.0745	0.0925	0.0927	0.0909	0.0766	0.061	0.0328	0.0369	0.0573	0.0566	0.0494	0.0382
$K = 20$	IQR	0.298	0.761	0.758	0.312	0.315	0.344	0.6	0.239	0.342	0.331	0.247	0.252	0.274	0.314
	MAD	0.164	0.387	0.38	0.171	0.171	0.184	0.308	0.128	0.174	0.169	0.13	0.131	0.14	0.161
	RMAD	0.423	1	0.982	0.443	0.442	0.476	0.795	0.736	1	0.968	0.747	0.753	0.804	0.925
	KW+	20	4.35	3.24	73.2	54.7	10.1	4.96	20	7.13	5.4	87.1	56.9	13.3	7.36
	KW-	0	0	0	66.4	46.8	0	0	0	0	0	78.4	46.5	0	0
$n = 1000$															
$K = 30$	bias	0.0764	0.0383	0.0453	0.0734	0.0738	0.0637	0.0487	0.0227	0.0139	0.0128	0.0211	0.0204	0.019	0.0132
	IQR	0.215	0.406	0.375	0.225	0.227	0.258	0.338	0.114	0.125	0.122	0.115	0.117	0.119	0.123
	MAD	0.121	0.208	0.193	0.125	0.127	0.137	0.172	0.0597	0.0641	0.0628	0.0602	0.0604	0.062	0.0627
	RMAD	0.579	1	0.925	0.602	0.608	0.659	0.828	0.933	1	0.98	0.939	0.943	0.968	0.979
	KW+	30	7.63	5.94	215	129	16.7	8.65	30	15.3	13	246	115	23.9	14.8
	KW-	0	0	0	205	117	0	0	0	0	0	231	96	0	0
$c = 0.5$															
$n = 100$	bias	0.476	0.411	0.416	0.47	0.47	0.458	0.427	0.316	0.194	0.173	0.304	0.298	0.255	0.188
$K = 20$	IQR	0.264	0.705	0.691	0.277	0.277	0.309	0.547	0.222	0.349	0.345	0.235	0.237	0.27	0.335
	MAD	0.476	0.568	0.553	0.47	0.471	0.458	0.503	0.317	0.254	0.236	0.305	0.299	0.26	0.238
	RMAD	0.838	1	0.972	0.827	0.828	0.806	0.885	1.25	1	0.93	1.2	1.18	1.03	0.939
	KW+	20	4.37	3.2	73.5	54.8	9.89	4.86	20	5.7	4.51	84.2	59.2	10.7	5.72
	KW-	0	0	0	66.8	47.1	0	0	0	0	0	77.1	50.9	0	0
$n = 1000$															
$K = 30$	bias	0.376	0.211	0.2	0.364	0.361	0.304	0.22	0.107	0.0424	0.0362	0.0915	0.0839	0.0531	0.0426
	IQR	0.189	0.373	0.392	0.198	0.2	0.25	0.355	0.108	0.126	0.128	0.111	0.112	0.123	0.125
	MAD	0.376	0.284	0.282	0.364	0.361	0.305	0.276	0.108	0.0722	0.0708	0.096	0.0898	0.0747	0.0714
	RMAD	1.32	1	0.993	1.28	1.27	1.07	0.97	1.5	1	0.982	1.33	1.24	1.03	0.989
	KW+	30	5.76	4.72	201	133	13.1	6.55	30	7.73	7.04	216	145	9.84	7.23
	KW-	0	0	0	193	124	0	0	0	0	0	210	138	0	0
$c = 0.9$															
$n = 100$	bias	0.855	0.753	0.739	0.853	0.852	0.829	0.754	0.571	0.25	0.231	0.532	0.527	0.408	0.276
$K = 20$	IQR	0.142	0.534	0.517	0.153	0.153	0.183	0.412	0.15	0.367	0.359	0.171	0.171	0.208	0.321
	MAD	0.855	0.804	0.791	0.853	0.852	0.829	0.769	0.571	0.315	0.286	0.532	0.527	0.408	0.305
	RMAD	1.06	1	0.984	1.06	1.06	1.03	0.957	1.81	1	0.908	1.69	1.67	1.3	0.969
	KW+	20	3.78	2.95	68.6	53.7	8.75	4.18	20	2.65	2.48	62.4	54.9	6.1	2.94
	KW-	0	0	0	62.5	46.8	0	0	0	0	0	57.5	49.8	0	0
$n = 1000$															
$K = 30$	bias	0.67	0.305	0.288	0.642	0.638	0.486	0.326	0.189	0.0625	0.0498	0.149	0.143	0.0775	0.0664
	IQR	0.12	0.435	0.396	0.138	0.139	0.184	0.318	0.0884	0.127	0.13	0.0999	0.0993	0.113	0.122
	MAD	0.67	0.389	0.354	0.642	0.638	0.486	0.353	0.189	0.0846	0.0764	0.15	0.143	0.0872	0.0825
	RMAD	1.72	1	0.911	1.65	1.64	1.25	0.906	2.23	1	0.902	1.77	1.69	1.03	0.975
	KW+	30	2.15	2.07	135	121	6	2.69	30	5.02	4.62	139	123	6.63	4.71
	KW-	0	0	0	131	116	0	0	0	0	0	134	118	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN; "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 (c), 2SLS

		2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps	2SLS-All	2SLS-DN	KGMM	2SLS-U	2SLS-C	2SLS-P	2SLS-Ps
$c = 0.1$					$R_f^2 = 0.01$							$R_f^2 = 0.1$			
$n = 100$	bias	0.0931	0.101	0.0981	0.0976	0.0962	0.0924	0.102	0.0644	0.0633	0.0665	0.0641	0.0629	0.0626	0.0633
$K = 20$	IQR	0.291	0.874	0.863	0.312	0.314	0.342	0.644	0.242	0.361	0.361	0.251	0.251	0.259	0.324
	MAD	0.163	0.443	0.435	0.173	0.175	0.181	0.334	0.129	0.19	0.194	0.133	0.134	0.137	0.174
	RMAD	0.367	1	0.98	0.391	0.396	0.408	0.754	0.679	1	1.02	0.701	0.706	0.722	0.917
	KW+	20	4.52	3.32	71.2	53.6	9.84	4.92	20	9.91	7.08	78.4	52.5	13.4	8.73
	KW-	0	0	0	64.4	45.8	0	0	0	0	0	69.8	42.1	0	0
$n = 1000$															
$K = 30$	bias	0.0749	0.0767	0.0824	0.0754	0.0751	0.0762	0.0722	0.02	0.0173	0.0221	0.0197	0.0187	0.0183	0.0166
	IQR	0.216	0.822	0.781	0.221	0.224	0.251	0.44	0.112	0.117	0.118	0.113	0.114	0.114	0.116
	MAD	0.119	0.413	0.397	0.126	0.124	0.136	0.236	0.0581	0.0602	0.061	0.0588	0.0589	0.0593	0.0593
	RMAD	0.288	1	0.961	0.305	0.301	0.328	0.57	0.965	1	1.01	0.976	0.978	0.984	0.984
	KW+	30	7.81	5.68	176	120	13.7	7.68	30	23.2	15.5	198	99.4	26	21.2
	KW-	0	0	0	168	110	0	0	0	0	0	184	80.5	0	0
$c = 0.5$															
$n = 100$	bias	0.474	0.498	0.492	0.475	0.475	0.476	0.487	0.319	0.344	0.362	0.313	0.31	0.309	0.319
$K = 20$	IQR	0.258	0.76	0.745	0.269	0.271	0.299	0.572	0.218	0.361	0.344	0.228	0.23	0.233	0.297
	MAD	0.474	0.64	0.627	0.475	0.475	0.476	0.547	0.319	0.392	0.411	0.313	0.311	0.31	0.341
	RMAD	0.74	1	0.98	0.742	0.742	0.744	0.855	0.813	1	1.05	0.798	0.792	0.791	0.87
	KW+	20	4.46	3.29	71.7	54	9.77	4.89	20	8.5	6.3	79.4	55.1	12.1	7.79
	KW-	0	0	0	65	46.3	0	0	0	0	0	71.8	46	0	0
$n = 1000$															
$K = 30$	bias	0.373	0.404	0.429	0.367	0.362	0.368	0.394	0.104	0.101	0.122	0.0878	0.0799	0.0856	0.0859
	IQR	0.198	0.803	0.767	0.21	0.213	0.226	0.41	0.109	0.117	0.111	0.112	0.112	0.113	0.11
	MAD	0.373	0.598	0.578	0.367	0.362	0.368	0.44	0.105	0.11	0.126	0.0932	0.0869	0.0921	0.0921
	RMAD	0.624	1	0.967	0.613	0.606	0.615	0.736	0.957	1	1.14	0.847	0.79	0.838	0.837
	KW+	30	6.74	5.13	175	122	12.9	7.09	30	16.8	13.9	183	128	14.9	12.8
	KW-	0	0	0	168	112	0	0	0	0	0	177	121	0	0
$c = 0.9$															
$n = 100$	bias	0.856	0.881	0.881	0.857	0.857	0.856	0.869	0.574	0.751	0.745	0.554	0.548	0.558	0.604
$K = 20$	IQR	0.141	0.391	0.385	0.146	0.146	0.157	0.29	0.153	0.532	0.454	0.172	0.174	0.168	0.21
	MAD	0.856	0.909	0.905	0.857	0.857	0.856	0.873	0.574	0.807	0.782	0.554	0.548	0.558	0.608
	RMAD	0.942	1	0.996	0.944	0.944	0.942	0.96	0.712	1	0.969	0.687	0.68	0.691	0.753
	KW+	20	4.31	3.18	72.1	54.4	9.69	4.82	20	3.34	3.01	73.4	57.8	8.65	4.69
	KW-	0	0	0	65.6	46.9	0	0	0	0	0	67.7	51.4	0	0
$n = 1000$															
$K = 30$	bias	0.671	0.818	0.833	0.652	0.646	0.663	0.735	0.188	0.821	0.812	0.136	0.13	0.157	0.168
	IQR	0.126	0.601	0.554	0.147	0.15	0.149	0.286	0.0921	0.812	0.731	0.101	0.101	0.0972	0.0984
	MAD	0.671	0.894	0.889	0.652	0.646	0.663	0.736	0.188	0.917	0.897	0.136	0.13	0.157	0.168
	RMAD	0.75	1	0.995	0.73	0.723	0.742	0.824	0.205	1	0.978	0.148	0.142	0.171	0.183
	KW+	30	4.11	3.17	167	124	10.8	5.33	30	1.15	1.23	125	112	9.05	6.12
	KW-	0	0	0	161	116	0	0	0	0	0	120	107	0	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN; "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 (a), LIML

		LIML					LIML				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.293	0.128	0.0958	0.0916	0.0873	0.037	0.05	0.051	0.0497	0.0337
$K = 20$	IQR	2	0.975	0.809	0.838	0.853	0.693	0.572	0.453	0.464	0.495
	MAD	0.851	0.49	0.408	0.424	0.43	0.347	0.288	0.23	0.237	0.25
	RMAD	1.74	1	0.833	0.865	0.879	1.2	1	0.801	0.825	0.868
	KW+	20	4.95	309	22.8	7.02	20	9.16	175	31.5	8.75
	KW-	0	0	333	20.7	0	0	0	189	28	0
$n = 1000$											
$K = 30$	bias	0.061	0.0833	0.0716	0.0665	0.0526	0.001	0.0017	0.015	0.0149	0.00371
	IQR	0.826	0.633	0.492	0.505	0.578	0.147	0.146	0.147	0.144	0.146
	MAD	0.409	0.323	0.251	0.258	0.292	0.074	0.0731	0.074	0.0731	0.0729
	RMAD	1.27	1	0.777	0.8	0.903	1.01	1	1.01	1	0.998
	KW+	30	8.24	369	49.8	10.9	30	29.2	311	83.5	23.2
	KW-	0	0	392	46.2	0	0	0	311	65.7	0
$c = 0.5$											
$n = 100$	bias	0.289	0.46	0.458	0.449	0.432	0.021	0.198	0.227	0.222	0.171
$K = 20$	IQR	1.61	0.868	0.713	0.734	0.759	0.651	0.527	0.514	0.517	0.496
	MAD	0.868	0.601	0.568	0.571	0.559	0.319	0.302	0.342	0.336	0.296
	RMAD	1.45	1	0.945	0.951	0.931	1.06	1	1.13	1.11	0.979
	KW+	20	4.94	2860	22.4	7.05	20	9.36	166	25.7	8.96
	KW-	0	0	3140	20.1	0	0	0	176	20.8	0
$n = 1000$											
$K = 30$	bias	0.043	0.322	0.315	0.308	0.227	0.002	0.00661	0.016	0.0162	0.00984
	IQR	0.769	0.572	0.548	0.552	0.553	0.142	0.141	0.146	0.146	0.142
	MAD	0.384	0.396	0.409	0.402	0.348	0.071	0.0705	0.074	0.0735	0.0719
	RMAD	0.971	1	1.03	1.01	0.878	1.01	1	1.04	1.04	1.02
	KW+	30	8.53	247	47.2	11.2	30	29.4	26.9	26.9	23.5
	KW-	0	0	259	42.7	0	0	0	3.91	3.91	0
$c = 0.9$											
$n = 100$	bias	0.139	0.708	0.793	0.79	0.777	-0.01	0.281	0.187	0.186	0.176
$K = 20$	IQR	9.46	0.561	0.514	0.518	0.505	0.507	0.399	0.435	0.435	0.413
	MAD	0.831	0.83	0.822	0.819	0.802	0.234	0.343	0.282	0.282	0.27
	RMAD	1	1	0.99	0.986	0.966	0.683	1	0.823	0.823	0.786
	KW+	20	5.61	140	20.3	7.4	20	10.8	36.5	13.4	10.3
	KW-	0	0	152	17	0	0	0	29.8	4.07	0
$n = 1000$											
$K = 30$	bias	0.006	0.413	0.244	0.244	0.226	0	0.00686	0.011	0.0101	0.0101
	IQR	0.58	0.48	0.503	0.502	0.455	0.135	0.131	0.136	0.136	0.136
	MAD	0.261	0.477	0.341	0.341	0.319	0.067	0.066	0.069	0.069	0.0689
	RMAD	0.547	1	0.714	0.714	0.669	1.01	1	1.05	1.05	1.04
	KW+	30	12.3	126	23.8	13.4	30	29.6	23.9	23.9	23.9
	KW-	0	0	123	12.6	0	0	0	0.008	0.008	0

Note: “bias” = median bias; “IQR” = inter-quantile range; “MAD” = median absolute deviation; “RMAD” = MAD relative to that of DN; “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 (b), LIML

		LIML					LIML				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$		$R_f^2 = 0.01$					$R_f^2 = 0.1$				
$n = 100$	bias	0.302	0.105	0.0949	0.0905	0.0805	0.031	0.0223	0.044	0.0413	0.0248
$K = 20$	IQR	2.01	0.92	0.745	0.773	0.793	0.693	0.438	0.352	0.36	0.403
	MAD	0.859	0.454	0.381	0.393	0.4	0.35	0.22	0.179	0.183	0.203
	RMAD	1.89	1	0.839	0.866	0.881	1.59	1	0.816	0.835	0.924
	KW+	20	4.55	152	23.6	6.64	20	5.58	406	35.8	6.13
	KW-	0	0	168	22.4	0	0	0	451	38	0
$n = 1000$											
$K = 30$	bias	0.073	0.0317	0.0622	0.0554	0.0342	0.002	0.00446	0.018	0.0153	0.00489
	IQR	0.797	0.477	0.352	0.372	0.44	0.144	0.136	0.131	0.128	0.134
	MAD	0.399	0.237	0.187	0.196	0.223	0.072	0.0681	0.067	0.065	0.0675
	RMAD	1.68	1	0.789	0.828	0.942	1.05	1	0.989	0.954	0.99
	KW+	30	5.31	898	64.5	7.82	30	11.2	1390	115	11.4
	KW-	0	0	959	69.1	0	0	0	1480	119	0
$c = 0.5$											
$n = 100$	bias	0.292	0.375	0.419	0.417	0.391	0.015	0.11	0.134	0.132	0.0921
$K = 20$	IQR	1.6	0.821	0.718	0.742	0.737	0.65	0.398	0.471	0.466	0.442
	MAD	0.864	0.542	0.535	0.536	0.528	0.317	0.225	0.266	0.264	0.233
	RMAD	1.59	1	0.986	0.989	0.973	1.41	1	1.19	1.18	1.04
	KW+	20	4.56	132	22.6	6.72	20	5.84	189	22.4	6.89
	KW-	0	0	144	21	0	0	0	207	20.3	0
$n = 1000$											
$K = 30$	bias	0.053	0.152	0.17	0.17	0.12	0.001	0.0146	0.012	0.0117	0.00961
	IQR	0.739	0.428	0.511	0.504	0.47	0.141	0.133	0.136	0.136	0.135
	MAD	0.369	0.253	0.304	0.303	0.264	0.071	0.0682	0.07	0.0696	0.0683
	RMAD	1.46	1	1.2	1.2	1.04	1.03	1	1.02	1.02	1
	KW+	30	6.14	262	40.7	9.02	30	12.2	16.4	16.4	13.1
	KW-	0	0	282	39.3	0	0	0	4.78	4.78	0
$c = 0.9$											
$n = 100$	bias	0.142	0.563	0.648	0.645	0.633	-0.01	0.113	0.033	0.0325	0.0323
$K = 20$	IQR	7.52	0.685	0.659	0.654	0.63	0.504	0.387	0.426	0.426	0.426
	MAD	0.826	0.717	0.717	0.712	0.696	0.229	0.231	0.211	0.211	0.211
	RMAD	1.15	1	1	0.993	0.971	0.993	1	0.917	0.917	0.915
	KW+	20	5.74	90.4	16.7	7.88	20	8.21	9.93	9.88	9.71
	KW-	0	0	95.7	11.7	0	0	0	0.319	0.261	0
$n = 1000$											
$K = 30$	bias	0.01	0.163	0.0425	0.0425	0.042	0.002	0.013	0.004	0.0042	0.0042
	IQR	0.554	0.392	0.455	0.455	0.455	0.135	0.13	0.133	0.133	0.133
	MAD	0.255	0.264	0.229	0.229	0.228	0.067	0.0665	0.066	0.0662	0.0662
	RMAD	0.967	1	0.866	0.866	0.865	1	1	0.996	0.996	0.996
	KW+	30	10.9	14.1	14.1	13.9	30	16.9	16.3	16.3	16.3
	KW-	0	0	0.251	0.251	0	0	0	0	0.0001	0

Note: "bias" = median bias; "IQR" = inter-quantile range; "MAD" = median absolute deviation; "RMAD" = MAD relative to that of DN; "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), LIML

		LIML					LIML				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.305	0.136	0.0848	0.0883	0.0893	0.039	0.0354	0.038	0.0361	0.0299
$K = 20$	IQR	1.99	0.986	0.834	0.866	0.861	0.685	0.562	0.481	0.487	0.52
	MAD	0.855	0.498	0.419	0.432	0.435	0.342	0.283	0.24	0.243	0.262
	RMAD	1.72	1	0.842	0.867	0.874	1.21	1	0.848	0.86	0.928
	KW+	20	5.09	189	22.6	7.17	20	10.5	226	31	9.38
	KW-	0	0	205	20.1	0	0	0	244	27.3	0
$n = 1000$											
$K = 30$	bias	0.055	0.0953	0.0599	0.0525	0.0457	-0	470.0009	110	5870	5430.002
	IQR	0.82	0.634	0.499	0.509	0.579	0.144	0.141	0.135	0.13	0.139
	MAD	0.4	0.326	0.255	0.261	0.298	0.072	0.0702	0.068	0.0661	0.0693
	RMAD	1.23	1	0.784	0.801	0.916	1.03	1	0.962	0.941	0.987
	KW+	30	10.3	524	48	12	30	21.8	1220	102	19.3
	KW-	0	0	556	43.8	0	0	0	1310	105	0
$c = 0.5$											
$n = 100$	bias	0.298	0.471	0.471	0.469	0.453	0.02	0.183	0.175	0.166	0.159
$K = 20$	IQR	1.63	0.883	0.736	0.759	0.766	0.653	0.515	0.499	0.492	0.482
	MAD	0.882	0.605	0.565	0.575	0.561	0.322	0.295	0.297	0.292	0.282
	RMAD	1.46	1	0.933	0.95	0.927	1.09	1	1.01	0.99	0.956
	KW+	20	5.05	142	22.3	7.19	20	10.6	118	24.1	9.68
	KW-	0	0	157	19.8	0	0	0	124	18.4	0
$n = 1000$											
$K = 30$	bias	0.037	0.311	0.255	0.241	0.215	-0	0.00909	0	0	0.00389
	IQR	0.758	0.602	0.531	0.528	0.55	0.14	0.137	0.135	0.135	0.138
	MAD	0.375	0.395	0.362	0.351	0.338	0.07	0.0687	0.068	0.0676	0.0693
	RMAD	0.951	1	0.916	0.889	0.856	1.01	1	0.984	0.984	1.01
	KW+	30	10.6	242	44.4	12.2	30	22.3	23.4	23.4	20.9
	KW-	0	0	256	39.1	0	0	0	5.32	5.32	0
$c = 0.9$											
$n = 100$	bias	0.154	0.746	0.829	0.825	0.812	-0.01	3 0.246	0.21	2 0.2	1 0.203
$K = 20$	IQR	8.08	0.541	0.475	0.486	0.478	0.511	0.429	0.355	0.352	0.34
	MAD	0.845	0.864	0.846	0.843	0.831	0.232	0.309	0.263	0.261	0.255
	RMAD	0.978	1	0.979	0.976	0.961	0.749	1	0.852	0.844	0.825
	KW+	20	5.58	189	21.5	7.36	20	11.4	61.3	15.7	10.7
	KW-	0	0	217	18.6	0	0	0	57.6	6.41	0
$n = 1000$											
$K = 30$	bias	-0	0.435	0.39	0.306	0.287	-0	0.0082	0.004	0.0036	0.00357
	IQR	0.572	0.694	0.434	0.423	0.385	0.131	0.129	0.129	0.129	0.129
	MAD	0.257	0.525	0.351	0.348	0.325	0.065	0.0652	0.064	0.0642	0.0643
	RMAD	0.49	1	0.669	0.663	0.62	0.995	1	0.985	0.985	0.987
	KW+	30	12.2	325	33.2	13.5	30	24	22	22	21.9
	KW-	0	0	337	23.8	0	0	0	0.05	0.0501	0

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

Table 7: Monte Carlo results: Model (a), Fuller

		Fuller					Fuller				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.081	0.0938	0.0894	0.0824	0.0917	0.011	0.0521	0.0577	0.0579	0.0438
$K = 20$	IQR	1.21	0.508	0.519	0.568	0.489	0.625	0.446	0.386	0.398	0.407
	MAD	0.609	0.265	0.272	0.294	0.251	0.31	0.227	0.198	0.206	0.208
	RMAD	2.29	1	1.03	1.11	0.945	1.37	1	0.873	0.904	0.915
	KW+	20	3.72	785	34.6	5.13	20	8.4	310	36.5	8.05
	KW-	0	0	870	38.6	0	0	0	346	35	0
$n = 1000$											
$K = 30$	bias	0.02	0.0826	0.0788	0.0751	0.0594	0.002	0.00264	0.016	0.0159	0.0048
	IQR	0.736	0.468	0.364	0.396	0.444	0.146	0.145	0.145	0.143	0.144
	MAD	0.367	0.243	0.194	0.208	0.227	0.073	0.0721	0.0731	0.0721	0.0719
	RMAD	1.51	1	0.8	0.857	0.934	1.01	1	1.01	1	0.996
	KW+	30	5.6	682	67.6	9.27	30	29.2	314	83.8	23.2
	KW-	0	0	733	69.1	0	0	0	314	66	0
$c = 0.5$											
$n = 100$	bias	0.397	0.489	0.479	0.473	0.47	0.067	0.261	0.284	0.275	0.226
$K = 20$	IQR	1.09	0.468	0.482	0.515	0.44	0.549	0.436	0.417	0.421	0.398
	MAD	0.588	0.492	0.501	0.501	0.478	0.282	0.293	0.327	0.32	0.272
	RMAD	1.19	1	1.02	1.02	0.97	0.962	1	1.12	1.09	0.93
	KW+	20	3.72	3870	33.8	5.14	20	8.45	254	30.3	8.23
	KW-	0	0	4270	37.6	0	0	0	289	27.5	0
$n = 1000$											
$K = 30$	bias	0.093	0.398	0.394	0.376	0.284	0.006	0.0107	0.0218	0.0218	0.0156
	IQR	0.649	0.419	0.388	0.4	0.409	0.14	0.139	0.143	0.143	0.14
	MAD	0.335	0.408	0.41	0.397	0.319	0.071	0.0705	0.0743	0.0743	0.0717
	RMAD	0.822	1	1	0.975	0.782	1	1	1.05	1.05	1.02
	KW+	30	5.57	696	63.7	9.41	30	29.4	27	27	23.5
	KW-	0	0	747	64.4	0	0	0	4.01	4	0
$c = 0.9$											
$n = 100$	bias	0.69	0.863	0.852	0.846	0.841	0.082	0.401	0.315	0.315	0.297
$K = 20$	IQR	0.679	0.266	0.312	0.319	0.267	0.372	0.345	0.333	0.331	0.291
	MAD	0.69	0.863	0.858	0.852	0.841	0.201	0.404	0.32	0.319	0.301
	RMAD	0.799	1	0.994	0.987	0.974	0.498	1	0.791	0.789	0.746
	KW+	20	3.87	565	31	5.46	20	9.4	66.6	16	9.51
	KW-	0	0	623	33.7	0	0	0	66.7	8.33	0
$n = 1000$											
$K = 30$	bias	0.098	0.624	0.435	0.435	0.383	0.009	0.0149	0.0198	0.0198	0.0197
	IQR	0.405	0.391	0.42	0.404	0.289	0.131	0.128	0.131	0.131	0.131
	MAD	0.226	0.625	0.436	0.436	0.383	0.066	0.0654	0.0688	0.0688	0.0688
	RMAD	0.362	1	0.698	0.698	0.614	1.01	1	1.05	1.05	1.05
	KW+	30	7.8	346	36.3	11.3	30	29.6	23.9	23.9	23.9
	KW-	0	0	366	30.1	0	0	0	0.0084	0.0084	0

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

Table 8: Monte Carlo results: Model (b), Fuller

		Fuller					Fuller				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.072	0.0786	0.0912	0.089	0.0823	0.006	0.0284	0.0464	0.0439	0.0325
$K = 20$	IQR	1.18	0.53	0.508	0.539	0.489	0.617	0.393	0.323	0.33	0.361
	MAD	0.601	0.269	0.267	0.285	0.25	0.31	0.198	0.169	0.171	0.182
	RMAD	2.23	1	0.991	1.06	0.927	1.57	1	0.855	0.867	0.923
	KW+	20	3.54	606	33.7	5.01	20	5.44	442	38.1	5.9
	KW-	0	0	730	37.8	0	0	0	491	41.2	0
$n = 1000$											
$K = 30$	bias	0.033	0.0373	0.0672	0.0608	0.0417	0.003	0.00522	0.0188	0.016	0.0057
	IQR	0.702	0.419	0.31	0.332	0.391	0.142	0.134	0.13	0.127	0.132
	MAD	0.354	0.212	0.169	0.178	0.197	0.072	0.0678	0.0669	0.0642	0.0666
	RMAD	1.67	1	0.799	0.841	0.927	1.06	1	0.987	0.948	0.983
	KW+	30	4.85	1060	71.4	7.26	30	11.2	1410	116	11.4
	KW-	0	0	1130	78.2	0	0	0	1490	120	0
$c = 0.5$											
$n = 100$	bias	0.398	0.44	0.442	0.437	0.434	0.064	0.144	0.17	0.169	0.131
$K = 20$	IQR	1.09	0.487	0.497	0.52	0.456	0.546	0.361	0.409	0.403	0.383
	MAD	0.59	0.45	0.473	0.469	0.447	0.283	0.214	0.25	0.249	0.217
	RMAD	1.31	1	1.05	1.04	0.992	1.32	1	1.17	1.16	1.02
	KW+	20	3.53	383	32.3	5.06	20	5.62	217	23.9	6.59
	KW-	0	0	430	36.1	0	0	0	239	22.7	0
$n = 1000$											
$K = 30$	bias	0.106	0.192	0.22	0.218	0.168	0.006	0.0185	0.0166	0.0166	0.0144
	IQR	0.627	0.37	0.443	0.433	0.401	0.139	0.131	0.134	0.134	0.133
	MAD	0.328	0.245	0.295	0.291	0.248	0.07	0.0684	0.0695	0.0695	0.0687
	RMAD	1.34	1	1.2	1.18	1.01	1.03	1	1.01	1.01	1
	KW+	30	5.3	339	45.5	8.29	30	12.2	16.4	16.4	13
	KW-	0	0	366	46.2	0	0	0	4.92	4.92	0
$c = 0.9$											
$n = 100$	bias	0.692	0.754	0.754	0.752	0.746	0.079	0.188	0.119	0.119	0.119
$K = 20$	IQR	0.673	0.332	0.391	0.39	0.35	0.367	0.304	0.318	0.318	0.315
	MAD	0.692	0.754	0.763	0.756	0.746	0.205	0.229	0.193	0.193	0.193
	RMAD	0.918	1	1.01	1	0.99	0.897	1	0.845	0.845	0.845
	KW+	20	3.9	423	22.8	6.09	20	7.71	10.1	9.79	9.49
	KW-	0	0	475	22.4	0	0	0	0.856	0.453	0
$n = 1000$											
$K = 30$	bias	0.106	0.249	0.14	0.14	0.138	0.01	0.0208	0.0126	0.0126	0.0126
	IQR	0.396	0.297	0.318	0.318	0.318	0.132	0.127	0.129	0.129	0.129
	MAD	0.221	0.271	0.204	0.204	0.203	0.066	0.0665	0.0666	0.0666	0.0666
	RMAD	0.816	1	0.753	0.753	0.748	0.998	1	1	1	1
	KW+	30	9.6	25.9	14.1	13.3	30	16.8	16.3	16.3	16.3
	KW-	0	0	14.8	1.08	0	0	0	0.0002	0.0002	0

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

Table 9: Monte Carlo results: Model (c), Fuller

		Fuller					Fuller				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.075	0.0961	0.0978	0.0938	0.0922	0.009	0.0483	0.0523	0.048	0.0416
$K = 20$	IQR	1.2	0.508	0.531	0.58	0.48	0.61	0.425	0.397	0.408	0.417
	MAD	0.602	0.269	0.285	0.308	0.253	0.306	0.217	0.201	0.208	0.211
	RMAD	2.24	1	1.06	1.15	0.94	1.41	1	0.927	0.958	0.97
	KW+	20	3.76	487	34.5	5.2	20	9.45	416	36.8	8.52
	KW-	0	0	542	38.1	0	0	0	457	35.7	0
$n = 1000$											
$K = 30$	bias	0.022	0.084	0.0673	0.0602	0.0555	0	0.00185	0.0171	0.019	0.0032
	IQR	0.722	0.44	0.363	0.388	0.457	0.143	0.139	0.133	0.129	0.137
	MAD	0.36	0.224	0.194	0.207	0.229	0.072	0.0694	0.0672	0.0653	0.0685
	RMAD	1.61	1	0.865	0.923	1.02	1.03	1	0.968	0.941	0.987
	KW+	30	6.75	1480	69.8	9.98	30	21.8	1230	103	19.3
	KW-	0	0	1580	72.1	0	0	0	1320	106	0
$c = 0.5$											
$n = 100$	bias	0.392	0.498	0.491	0.489	0.48	0.066	0.258	0.244	0.228	0.218
$K = 20$	IQR	1.09	0.449	0.48	0.524	0.436	0.546	0.439	0.412	0.408	0.389
	MAD	0.579	0.503	0.509	0.511	0.486	0.282	0.296	0.286	0.275	0.259
	RMAD	1.15	1	1.01	1.01	0.965	0.955	1	0.965	0.931	0.875
	KW+	20	3.77	664	34	5.21	20	9.6	287	29.9	8.8
	KW-	0	0	746	37.6	0	0	0	317	26.8	0
$n = 1000$											
$K = 30$	bias	0.088	0.427	0.333	0.304	0.277	0.003	0.0135	0.0054	0.0054	0.009
	IQR	0.634	0.423	0.408	0.412	0.423	0.138	0.135	0.133	0.133	0.136
	MAD	0.329	0.433	0.365	0.337	0.308	0.069	0.0678	0.0668	0.0668	0.0681
	RMAD	0.761	1	0.843	0.779	0.711	1.02	1	0.985	0.985	1
	KW+	30	6.77	962	64.5	10.1	30	22.3	23.4	23.4	20.9
	KW-	0	0	1030	65.8	0	0	0	5.44	5.44	0
$c = 0.9$											
$n = 100$	bias	0.693	0.883	0.87	0.865	0.86	0.083	0.359	0.322	0.315	0.314
$K = 20$	IQR	0.673	0.24	0.286	0.305	0.25	0.367	0.548	0.31	0.291	0.265
	MAD	0.693	0.883	0.875	0.87	0.86	0.207	0.37	0.327	0.319	0.314
	RMAD	0.784	1	0.99	0.985	0.974	0.559	1	0.883	0.862	0.848
	KW+	20	3.83	555	33.4	5.33	20	10.2	201	20.7	9.84
	KW-	0	0	642	36.7	0	0	0	215	13.9	0
$n = 1000$											
$K = 30$	bias	0.095	0.817	0.473	0.45	0.407	0.007	0.0164	0.0123	0.0123	0.0124
	IQR	0.4	0.463	0.411	0.339	0.274	0.128	0.126	0.125	0.125	0.125
	MAD	0.226	0.817	0.475	0.453	0.407	0.065	0.0647	0.0638	0.0638	0.0637
	RMAD	0.277	1	0.582	0.555	0.498	1	1	0.986	0.986	0.986
	KW+	30	7.47	1350	57.1	11	30	24	22	22	21.9
	KW-	0	0	1430	55.7	0	0	0	0.0525	0.0525	0

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

Table 10: Monte Carlo results: Model (a), B2SLS

		B2SLS					B2SLS				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.107	0.0809	0.0965	0.0932	0.0803	0.015	0.0408	0.0809	0.067	0.0289
$K = 20$	IQR	1.41	0.823	0.201	0.283	1.04	0.782	0.58	0.178	0.272	0.619
	MAD	0.715	0.417	0.123	0.16	0.519	0.391	0.293	0.11	0.143	0.311
	RMAD	1.72	1	0.296	0.384	1.25	1.34	1	0.377	0.489	1.06
	KW+	20	3.19	1680	68.7	3.64	20	7.72	1810	70.9	6.98
	KW-	0	0	1850	77	0	0	0	2000	74.7	0
$n = 1000$											
$K = 30$	bias	0.0289	0.0853	0.0922	0.0863	0.0576	0.000893	0.00226	0.0428	0.0169	0.00237
	IQR	0.893	0.741	0.115	0.213	0.709	0.148	0.148	0.117	0.137	0.149
	MAD	0.449	0.369	0.101	0.123	0.352	0.0739	0.0734	0.0698	0.0702	0.0746
	RMAD	1.22	1	0.275	0.334	0.954	1.01	1	0.951	0.958	1.02
	KW+	30	4.4	4350	158	6.95	30	29.2	3040	113	23.1
	KW-	0	0	4650	169	0	0	0	3230	97.5	0
$c = 0.5$											
$n = 100$	bias	0.5	0.483	0.489	0.486	0.464	0.0735	0.258	0.41	0.365	0.174
$K = 20$	IQR	1.2	0.723	0.177	0.257	0.916	0.823	0.55	0.171	0.255	0.591
	MAD	0.778	0.592	0.489	0.486	0.648	0.416	0.368	0.41	0.365	0.348
	RMAD	1.32	1	0.827	0.821	1.1	1.13	1	1.11	0.991	0.946
	KW+	20	3.21	7740	68.8	3.62	20	7.46	1650	70.4	6.87
	KW-	0	0	8520	77.1	0	0	0	1810	74	0
$n = 1000$											
$K = 30$	bias	0.151	0.389	0.472	0.442	0.249	-0.00223	0.0046	0.123	0.0815	0.00873
	IQR	0.89	0.637	0.114	0.198	0.647	0.151	0.148	0.136	0.133	0.148
	MAD	0.476	0.508	0.472	0.442	0.409	0.075	0.0746	0.126	0.0918	0.0747
	RMAD	0.937	1	0.929	0.869	0.804	1.01	1	1.69	1.23	1
	KW+	30	4.28	3760	157	6.92	30	29	248	113	22.4
	KW-	0	0	4000	168	0	0	0	242	97.1	0
$c = 0.9$											
$n = 100$	bias	0.872	0.869	0.884	0.877	0.836	0.127	0.457	0.705	0.649	0.323
$K = 20$	IQR	0.692	0.42	0.0987	0.139	0.512	0.843	0.474	0.169	0.199	0.509
	MAD	0.942	0.893	0.884	0.877	0.879	0.43	0.506	0.705	0.649	0.412
	RMAD	1.05	1	0.989	0.982	0.984	0.85	1	1.39	1.28	0.813
	KW+	20	3.06	1700	68.4	3.57	20	6.3	643	68.5	6.47
	KW-	0	0	1860	76.7	0	0	0	715	71.2	0
$n = 1000$											
$K = 30$	bias	0.213	0.676	0.816	0.776	0.426	-0.00471	0.0148	0.178	0.157	0.0154
	IQR	0.845	0.542	0.119	0.142	0.514	0.161	0.148	0.109	0.109	0.152
	MAD	0.483	0.731	0.816	0.776	0.503	0.0787	0.0756	0.178	0.157	0.078
	RMAD	0.661	1	1.12	1.06	0.688	1.04	1	2.36	2.08	1.03
	KW+	30	3.53	1920	152	6.56	30	28.3	139	109	20.3
	KW-	0	0	2050	161	0	0	0	127	95.2	0

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

Table 11: Monte Carlo results: Model (b), B2SLS

		B2SLS					B2SLS				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$				$R_f^2 = 0.01$					$R_f^2 = 0.1$		
$n = 100$	bias	0.112	0.0756	0.0976	0.0962	0.0721	0.015	0.0211	0.077	0.0619	0.0185
$K = 20$	IQR	1.41	0.849	0.201	0.284	1.01	0.76	0.471	0.172	0.253	0.465
	MAD	0.71	0.423	0.124	0.156	0.511	0.383	0.238	0.104	0.134	0.232
	RMAD	1.68	1	0.293	0.369	1.21	1.61	1	0.437	0.562	0.976
	KW+	20	3.09	1870	69.6	3.54	20	5.04	2670	73.8	5.1
	KW-	0	0	2060	78.6	0	0	0	2940	82.4	0
$n = 1000$											
$K = 30$	bias	0.0459	0.0308	0.0922	0.0813	0.0273	0.00189	0.00279	0.0578	0.0181	0.00384
	IQR	0.883	0.541	0.0937	0.202	0.519	0.145	0.137	0.0829	0.122	0.136
	MAD	0.442	0.271	0.0975	0.119	0.259	0.0727	0.0683	0.0648	0.0611	0.0677
	RMAD	1.63	1	0.36	0.438	0.955	1.07	1	0.949	0.895	0.992
	KW+	30	4.31	8350	167	5.8	30	11.2	11300	166	11.2
	KW-	0	0	8870	184	0	0	0	11900	173	0
$c = 0.5$											
$n = 100$	bias	0.479	0.454	0.482	0.475	0.422	0.0819	0.107	0.381	0.31	0.0857
$K = 20$	IQR	1.22	0.768	0.178	0.253	0.925	0.806	0.475	0.184	0.234	0.464
	MAD	0.777	0.58	0.482	0.475	0.614	0.411	0.259	0.381	0.311	0.245
	RMAD	1.34	1	0.831	0.819	1.06	1.59	1	1.47	1.2	0.948
	KW+	20	2.99	1380	69.4	3.51	20	4.82	1890	72.5	4.97
	KW-	0	0	1540	78.4	0	0	0	2070	79.9	0
$n = 1000$											
$K = 30$	bias	0.158	0.144	0.455	0.398	0.128	-0.00127	0.0115	0.135	0.0807	0.0131
	IQR	0.849	0.53	0.116	0.18	0.495	0.149	0.136	0.117	0.114	0.134
	MAD	0.457	0.303	0.455	0.398	0.278	0.0731	0.0692	0.136	0.0876	0.0681
	RMAD	1.51	1	1.5	1.31	0.919	1.06	1	1.96	1.27	0.985
	KW+	30	4.21	3110	164	5.68	30	10.6	453	159	10.7
	KW-	0	0	3330	179	0	0	0	475	163	0
$c = 0.9$											
$n = 100$	bias	0.873	0.793	0.874	0.857	0.736	0.118	0.173	0.59	0.534	0.148
$K = 20$	IQR	0.712	0.582	0.108	0.145	0.634	0.868	0.475	0.177	0.172	0.455
	MAD	0.942	0.845	0.874	0.857	0.804	0.428	0.297	0.59	0.534	0.274
	RMAD	1.11	1	1.03	1.01	0.952	1.44	1	1.99	1.8	0.922
	KW+	20	2.96	1040	67.4	3.42	20	4.1	142	65.1	4.5
	KW-	0	0	1170	75.3	0	0	0	154	68.9	0
$n = 1000$											
$K = 30$	bias	0.215	0.232	0.723	0.684	0.216	-0.00327	0.0178	0.144	0.125	0.0225
	IQR	0.837	0.475	0.127	0.133	0.445	0.16	0.133	0.101	0.102	0.132
	MAD	0.471	0.343	0.723	0.684	0.321	0.0786	0.0692	0.145	0.126	0.0693
	RMAD	1.37	1	2.11	1.99	0.934	1.13	1	2.09	1.81	1
	KW+	30	3.38	1100	144	4.98	30	9.38	167	132	9.67
	KW-	0	0	1390	152	0	0	0	168	131	0

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

Table 12: Monte Carlo results: Model (c), B2SLS

		B2SLS					B2SLS				
		-All	-DN	-U	-C	-P	-All	-DN	-U	-C	-P
$c = 0.1$		$R_f^2 = 0.01$					$R_f^2 = 0.1$				
$n = 100$	bias	0.0977	0.0961	0.101	0.0957	0.102	0.0209	0.0439	0.0756	0.0641	0.0335
$K = 20$	IQR	1.42	0.822	0.206	0.288	1.04	0.787	0.543	0.174	0.263	0.622
	MAD	0.726	0.419	0.127	0.159	0.531	0.395	0.273	0.109	0.139	0.31
	RMAD	1.73	1	0.304	0.379	1.27	1.45	1	0.399	0.508	1.14
	KW+	20	3.28	1690	68.6	3.69	20	8.74	2420	70.3	7.42
	KW-	0	0	1850	76.7	0	0	0	2660	75	0
$n = 1000$											
$K = 30$	bias	0.0355	0.0854	0.0905	0.0785	0.0408	-0.00138	0.0004	0.0596	0.0156	0.00089
	IQR	0.899	0.683	0.113	0.2	0.705	0.145	0.141	0.0777	0.115	0.14
	MAD	0.446	0.353	0.1	0.116	0.354	0.0722	0.0705	0.0654	0.0591	0.0698
	RMAD	1.26	1	0.284	0.33	1	1.02	1	0.928	0.839	0.99
	KW+	30	5.36	4110	154	7.36	30	21.7	11600	148	19.1
	KW-	0	0	4440	166	0	0	0	12500	156	0
$c = 0.5$											
$n = 100$	bias	0.512	0.489	0.492	0.49	0.467	0.0956	0.251	0.397	0.326	0.174
$K = 20$	IQR	1.22	0.717	0.176	0.25	0.934	0.819	0.53	0.171	0.241	0.613
	MAD	0.785	0.595	0.492	0.49	0.66	0.417	0.354	0.397	0.326	0.348
	RMAD	1.32	1	0.827	0.824	1.11	1.18	1	1.12	0.921	0.983
	KW+	20	3.21	1280	68.6	3.69	20	8.45	1510	69.9	7.31
	KW-	0	0	1440	76.7	0	0	0	1670	74.5	0
$n = 1000$											
$K = 30$	bias	0.146	0.439	0.454	0.397	0.234	-0.00437	0.00673	0.141	0.075	0.00678
	IQR	0.887	0.659	0.125	0.192	0.684	0.153	0.144	0.117	0.112	0.144
	MAD	0.472	0.523	0.454	0.397	0.422	0.0756	0.0721	0.141	0.0837	0.0718
	RMAD	0.902	1	0.868	0.759	0.807	1.05	1	1.96	1.16	0.995
	KW+	30	5.1	3150	154	7.32	30	21.5	439	143	18.5
	KW-	0	0	3370	165	0	0	0	469	148	0
$c = 0.9$											
$n = 100$	bias	0.869	0.885	0.886	0.879	0.855	0.126	0.522	0.653	0.57	0.302
$K = 20$	IQR	0.679	0.381	0.0954	0.139	0.493	0.851	0.571	0.202	0.207	0.534
	MAD	0.935	0.9	0.886	0.879	0.888	0.432	0.56	0.653	0.57	0.404
	RMAD	1.04	1	0.984	0.976	0.986	0.772	1	1.17	1.02	0.722
	KW+	20	3.11	1320	68.8	3.63	20	7.12	1200	68.1	6.87
	KW-	0	0	1490	77.1	0	0	0	1290	71.9	0
$n = 1000$											
$K = 30$	bias	0.2	0.855	0.789	0.708	0.399	-0.00755	0.0105	0.141	0.116	0.0131
	IQR	0.872	0.484	0.159	0.162	0.595	0.16	0.149	0.098	0.101	0.147
	MAD	0.485	0.877	0.789	0.708	0.498	0.0787	0.0747	0.141	0.117	0.0747
	RMAD	0.553	1	0.9	0.807	0.567	1.05	1	1.88	1.57	1
	KW+	30	3.65	3030	152	7.11	30	20.9	161	124	17.1
	KW-	0	0	3260	163	0	0	0	165	124	0

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