GMM Model Averaging Using Higher Order Approximations

Luis F. Martins
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal

Vasco J. Gabriel
Department of Economics, University of Victoria, Canada and NIPE-UM

This version: July 2022

Abstract

Moment conditions model averaging (MA) estimators in the GMM framework are considered. Under finite sample considerations, MA estimators with optimal weights are proposed, in the sense that weights minimize the corresponding higher-order asymptotic mean squared error (AMSE). It is shown that the higher-order AMSE objective function has a closed-form expression, which makes this procedure applicable in practice. In addition, and as an alternative, different averaging schemes based on moment selection criteria are considered, in which weights for averaging across GMM estimates can be obtained by direct smoothing or by numerical minimization of a specific criterion. Asymptotic properties assuming correctly specified models are derived and the performance of the proposed averaging approaches is contrasted with existing model selection alternatives *i*) analytically, for a simple IV example, and *ii*) by means of Monte Carlo experiments in a nonlinear setting, showing that MA compares favourably in many relevant setups. The usefulness of MA methods is illustrated by revisiting Acemoglu et al.'s (2001) study on the effect of institutions on economic performance.

Keywords: Generalized Method of Moments; Model Selection; Model Averaging, Higher-Order Asymptotics; AMSE

JEL Classification: C36; C51;C52

1 Introduction

In many applications of instrumental variables and GMM estimation, there is often a large set of candidate variables that can be used as instruments. However, the properties of moment conditions and instrumental variables estimators are very sensitive to the choice and characteristics of the instrument set. Indeed, instruments might be poorly correlated with the endogenous variables, which invalidates conventional inference procedures. On the other hand, using many (potentially weak) instruments, while desirable (see Hansen, Hausman and Newey, 2008), may lead to biases and substantial deviations from the usual Gaussian asymptotic approximation (see Chao and Swanson, 2005, and Newey and Windmeijer, 2009).

Thus, much of the literature has focused on procedures for the selection of the appropriate moments/instruments. Model selection entails choosing one of the estimated competing models under consideration, possibly by deleting some of the moment conditions.[1] Andrews (1999) developed GMM analogues of model selection criteria in order to consistently select the largest set of valid moment conditions, while Hall, Inoue, Jana and Shin (2007) suggest selecting moment conditions according to the relevant moment selection criterion (RMSC), based on the entropy of the limiting distribution of the GMM estimator. On the other hand, Donald and Newey (2001) and Donald, Imbens and Newey (2009) propose a selection procedure such that the AMSE is minimized over all existing instruments deemed to be valid, while Hall and Peixe (2003) propose a canonical correlations information criteria (CCIC) for instrument selection.[2]

In this paper, we consider the alternative approach of model averaging (MA), in which parameter estimates are constructed based on a weighted average of estimates obtained using different sets of moment conditions or model specifications. Indeed, by making use of the information conveyed by otherwise discarded alternative specifications, model averaging as an estimation strategy may yield some gains in terms of bias and efficiency when compared to procedures that make use of a single set of moment conditions. The main focus of our paper is on deriving the stochastic expansion of the MA estimator and, building upon that, we study MA estimators with optimal weights, in the sense that weights minimize the MA estimator's *higher-order* AMSE. While we follow closely Donald et al. (2009), our approach contrast with theirs in that these authors employ higher-order expansions of the AMSE as a criterion for choosing instrumental variables, whereas we use the AMSE to obtain optimal weights for model averaging. For the sake of completeness, we also consider simpler alternative MA schemes for GMM in which empirical weights are obtained based on GMM moment selection criteria. This can be achieved by direct smoothing of information criteria arising from the estimation stage, or by numerical minimization of a specific criterion, as in Hansen (2007).

Our approach displays important differences when compared with the current model averaging

[1] Testing competing, non-nested formulations, in which the outcome may not be the selection of one particular model, can be carried out in a moment conditions framework, see Smith and Ramalho (2002).

[2] Shrinkage methods for GMM are an alternative to model selection and estimation, see Caner (2009), Cheng and Liao (2015) and Caner, Han and Lee (2018).

GMM literature. First, and unlike existing MA-GMM approaches based on first-order asymptotics, our main concern are finite sample considerations, i.e. gauging to what extent MA estimation can improve upon the often problematic finite sample performance of the standard GMM estimator, which leads us to focus on a higher-order AMSE criterion for MA estimation. Second, we study the more empirically relevant case in which the vector of parameters of interest is the same across different specifications (i.e., across different sets of moment conditions), thus rendering the local misspecification apparatus of Hjort and Claeskens (2003) inapplicable. Third, we suggest averaging outputs of GMM (i.e. estimates of the parameters of interest) rather than inputs, (i.e. *moment averaging*), as this allows the researcher to have an interpretable and quasi-Bayesian sense of parameter uncertainty, as well as a more informed view on the relative merits of different specifications.

We study optimally-weighted MA-GMM estimators both under exact and overidentification, and show analytically that AMSE improvements can be obtained by averaging estimators, relative to not doing so, in a standard linear setup. We then develop the asymptotic properties of the MA estimator under correct model specification, with fixed or random weights. Under fixed weights, we show that the estimator is consistent and normally distributed, whereas with random weights Gaussianity depends on the averaging scheme. Monte Carlo experiments in a nonlinear setting show that, in several setups, model averaging estimation procedures outperform the selection method of Donald et al. (2009) in terms of bias and dispersion. Finally, we apply our estimation methods to measure the effect of institutions on economic performance, based on Acemoglu, Johnson and Robinson (2001). While their estimates vary considerably across different specifications, our MA estimates are much less disperse, thus illustrating the usefulness of a model averaging approach.

Our work is a natural extension of the model averaging literature, in which averaging usually involve weights obtained from functions of model selection criteria, such as the BIC, AIC, etc (see Claeskens and Hjort, 2008 for a review). Hansen (2007) proposed a Mallows criterion for the selection of weights for averaging across least squares estimates obtained from a set of approximating models, in which regressors (or groups of regressors) are added sequentially. Kuersteiner and Okui (2010) suggest using Hansen's (2007) method as a first step to construct optimal instruments IV estimation with 2SLS, LIML and Fuller estimators. The weights are chosen to minimize the AMSE, as in Donald and Newey (2001). Lee and Zhou (2015) suggest estimating the weights from the minimization of the AMSE of the averaged estimator under the scenario of many (weak) instruments and each model having the same number of instruments. Martins and Gabriel (2014) link the choice of empirical weights to IV selection criteria by direct 'smoothing'.

In a GMM framework, Xiao (2010) and Chen, Jacho-Chávez and Linton (2016) average over estimators based on exactly identified models and the proposed weights are optimal in the sense that the MA estimator attains the (first-order) semiparametric efficiency bound. The former considers GMM estimators and the later a more general class of estimators, but imposing a cardinality of the linear combinations that increases with the sample size. In contrast with the previous references, Sueishi (2013) and DiTraglia (2016) assume model (local) misspecification: Sueishi (2013) in terms of the model's coefficients and for the class of GEL estimators, DiTraglia (2016) at the restrictions level. In both cases, the weights minimize the AMSE of the MA estimator. DiTraglia (2016) develops a new selection criteria (the Focused Moment Selection Criterion) and his MA estimator combines estimators based on valid and potentially invalid instruments/conditions to minimize MSE, thus leading to a more favourable bias-variance trade-off. Finally, Cheng, Liao and Shi (2019) combine a conservative GMM estimator based on valid moment conditions and an aggressive GMM estimator based on both valid and possibly misspecified moment conditions. That is, regardless of the dimension of the model, valid moment conditions must exist and only two point estimators are averaged. Moreover, in the case of global misspecification full weight is given to the conservative GMM estimator, thus meaning that averaging is no longer applied.

Next, section 2 introduces assumptions and definitions. In section 3, we introduce our moment conditions 'optimal' model averaging approach and show analytically their smaller higher-order AMSE when compared to standard GMM estimators. In section 4 we discuss alternative approaches to obtain empirical weights. In section 5, we derive statistical properties of the GMM model averaging estimators. Section 6 presents a summary of a Monte Carlo simulation study providing evidence in support of our MA procedures in the context of nonlinear models, against the benchmark of model selection of Donald et al. (2009). In section 7 we briefly revisit Acemoglu et al.'s (2001) study on the effect of institutions on economic performance and Section 8 concludes. All proofs are included in a supplementary appendix.

2 Definitions

Given a vector of random variables $\{y_t\}$, the estimation of a unique p-dimensional parameter vector $\theta_0 = (\theta_{0,1}, ..., \theta_{0,p}) \in \Theta \subset \Re^p$ is based on (up to) $m \geq p$ moment conditions of the form $E[g(y_t, \theta_0)] \equiv E[g_t(\theta_0)] = 0$, for all t, with corresponding empirical moments $\widehat{g}_T(\theta) = (1/T)\sum_{t=1}^{T} g(y_t, \theta)$. As in Andrews (1999), one can define a moment selection vector $c \in \Re^m$ that represents a list of "selected" moment conditions, i.e. a subset of g, denoted as $\widehat{g}_{Tc}(\theta)$. Defining

$$C = \left\{c \in \Re^m \backslash \{0\} : c_j = 0 \text{ or } 1, \forall 1 \leq j \leq m, \text{ where } c = (c_1, ..., c_m)', |c| \geq p\right\}, \tag{1}$$

c is a vector of zeros (excluded conditions) and ones (included conditions) and $|c| = \sum_j^m c_j \leq m$ for $c \in C$, with $|c|$ denoting the number of selected moments. For a particular c, the (efficient) GMM estimator is defined as

$$\widehat{\theta}_{Tc}(W) = \arg\min_{\theta \in \Theta} \widehat{g}_{Tc}(\theta)' W_{Tc} \widehat{g}_{Tc}(\theta), \tag{2}$$

where W_{Tc} is a weighting matrix such that $\text{plim}\, W_{Tc} = S_c^{-1}$, where

$$S_c = \lim_{T \to \infty} Var\left[T^{-1/2}\sum_{t=1}^{T} g_c(y_t, \theta_0)\right] \tag{3}$$

is the $|c| \times |c|$ long-run variance matrix of the process $\widehat{g}_{Tc}(\theta)$.

We adopt the standard GMM framework in which all moment conditions under consideration are valid. This is expressed in the following general result, first proved by Hansen (1982) and that can be found in any advanced econometrics textbook such as Hayashi (2000).

Assumption 1 *(Regularity conditions for a given $c \in C$). Fix the set of moment conditions to any particular $c \in C$. A.1: $\{y_t\}$ is an infinite sequence of stationary and ergodic variables; A.2: The true θ_0 belongs to the parameter space Θ which is an open subset of $\Re^p$; A.3: $g_c(\cdot, \theta)$ and $\partial g_c / \partial\theta(\cdot, \theta)$ are Borel measurable for each $\theta \in \Theta$ and $\partial g_c/\partial\theta(y, \cdot)$ is continuous on Θ for each $y \in \Re^l$; A.4: $\partial g_c(y_1, \theta)/\partial\theta$ is first moment continuous at θ_0, and the $|c| \times p$ Jacobian matrix $G_c = E\left(\left.\frac{\partial g_c(y_t,\theta)}{\partial\theta'}\right|_{\theta=\theta_0}\right)$ exists, is finite, and has full-column rank. A.5: The following CLT for stationary and ergodic variables holds: $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_c(y_t, \theta_0) \overset{d}{\to} N(0, S_c)$ where $S_c = \lim_{T\to\infty} Var\left[T^{-1/2}\sum_{t=1}^{T} g_c(y_t, \theta_0)\right]$ is a $|c| \times |c|$ positive definite matrix; and A.6: Applied to any consistent estimator $\widetilde{\theta}_{Tc}$, the following LLN for stationary and ergodic variables holds: $\frac{1}{T}\sum_{t=1}^{T} \frac{\partial g_c(y_t, \widetilde{\theta}_{Tc})}{\partial\theta'} \overset{p}{\to} G_c$.*

Lemma 1 *(Asymptotic normality of the efficient estimator $\widehat{\theta}_{Tc}$)*

Assume that Assumption 1 holds and, for any $c \in C$, W_{Tc} is *such that* $\operatorname{plim} W_{Tc} = S_c^{-1}$. *Then, for any $c \in C$,*

$$\sqrt{T}\left(\widehat{\theta}_{Tc} - \theta_0\right) \overset{d}{\to} Z_c = N(0, V_c), \tag{4}$$

where

$$V_c = \left(G_c' S_c^{-1} G_c\right)^{-1}, \tag{5}$$

where G_c and S_c are defined in Assumption 1.

Importantly, and contrary to the usual literature, we include the higher order terms to define a criteria for selecting the weights. Newey and Smith (2004) provide the stochastic expansion for GMM/GEL (see Theorem 3.3) under i.i.d. data, correctly specified models and for fixed m. This is given by

$$\sqrt{T}\left(\widehat{\theta}_{Tc} - \theta_0\right) = \widetilde{\psi}_c + Q_{1c}/\sqrt{T} + Q_{2c}/T + O_p\left(T^{-3/2}\right), \tag{6}$$

where

$$\widetilde{\psi}_c = -\left(G_c' S_c^{-1} G_c\right)^{-1} G_c' S_c^{-1} \sqrt{T}\widehat{g}_{Tc}(\theta_0) + o_p(1) \tag{7}$$

is the leading term. They also define the higher-order MSE, but without presenting an expression for the higher-order variance: “In general, although they may be derived relatively straightforwardly from the Appendix, the expressions for Ξ for GMM and GEL are extremely complicated, and so are not given here, although some comparisons can be made.” (page 234). Results for time series data can be obtained from Anatolyev (2005) and Bao and Ullah (2007), among others, see Anatolyev and Gospodinov (2011) for a good overview.

Note that we drop the subscript c for quantities such as $\widehat{g}_T(\theta)$, V, G, S, W_T in the case of the estimator using the full set of available moment conditions (denoted as $\widehat{\theta}_T$), obtained for $c = \iota_m$, a vector of ones, such that $|c| = m$. Similarly, for any $c \in C$, the J test statistic for overidentifying restrictions is constructed as

$$J_{Tc} = T \inf_{\theta \in \Theta} \widehat{g}_{Tc}(\theta)' \, W_{Tc} \widehat{g}_{Tc}(\theta) \, . \tag{8}$$

As mentioned above, our setup assumes that only valid moment conditions are being used. If the researcher is unsure whether or not the moment conditions are correct, some selection procedures could be used to select the (sub)set of correct moment conditions.[3] Selection criteria for GMM are reviewed in section 4, when we consider alternative selection criteria-based MA estimators.

3 Optimally-Weighted Moment Conditions Model Averaging Estimators

In this section, we present a methodology whereby we average across candidate specifications to obtain an averaged estimator. Note that this differs from previous literature (namely Kuersteiner and Okui, 2010, Kapetanios and Marcellino, 2010 and Okui, 2011) in that we are not averaging across instruments to obtain an optimal set of instruments. Instead, we propose averaging different estimates of θ_0 obtained from distinct sets of moment conditions. The weights associated with each estimate are chosen according to an optimality criteria. In a particular model, we are able to show analytically that our proposed 'optimal' MA-GMM estimator has a smaller higher-order AMSE when compared to standard GMM estimators.

3.1 The Procedure

Let $\mathcal{M}$ be the collection of candidate moment conditions models. Here, $\mathcal{M}$ is a countable/finite or an uncountable set, such that model M_i belongs to the family of models $\mathcal{M} : M_i \in \mathcal{M}$. In our model averaging procedure, we specify a subset of $\mathcal{M}$ from which we define the MA estimator. For now, take any model, M_i, which is characterized by a particular set of moment conditions.

Consider m and c as defined above and the relevant objects indexed by c. Now, let $\omega = \left(\omega_1, ..., \omega_{|C|}\right)'$ be a weight vector in the unit-simplex in $\Re^{|C|}$, with $|C| = 2^m - \sum_{j=0}^{p-1} \binom{m}{j} = \sum_{j=p}^{m} \binom{m}{j}$, with the binomial coefficients $\binom{m}{j} = \frac{m!}{j!(m-j)!}$, representing the number of different elements[4] in C :

$$H_m = \{\omega \in [0,1]^{|C|} : \sum_{c \in C} \omega_c = 1\}. \tag{9}$$

[3] Potentially, our averaging approach could then take place over estimates obtained from valid specifications utilizing different combinations of the selected moment conditions, although this raises the issue of pre-testing, which we will not address here.

[4] We need to exclude $\sum_{j=0}^{p-1} \binom{m}{j}$ from the total of combinations 2^m, those for which $m < p$.

Thus, a model averaging estimator of the unknown $p \times 1$ vector θ_0 is

$$\widehat{\theta}_T(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_{Tc}. \tag{10}$$

Clearly, standard GMM estimation is a special case for which no model averaging occurs: $\omega_{c^*} = 1$ for some c^* and $\omega_{c'} = 0$ for $c' \neq c^*$ and $\widehat{\theta}_T(\omega) = \widehat{\theta}_{Tc^*}$.

Given our initial assumptions in section 2, we assume our procedure is averaging over valid specifications. However, the properties of the MA estimator will depend on whether the weights are fixed or random objects. For a given ω, the limit statistical properties of $\widehat{\theta}_T(\omega)$ depend on a linear combination of the random processes $\widehat{\theta}_{Tc}, c \in C$. Thus, under correct model specification, $\operatorname{plim} \widehat{\theta}_{Tc} = \theta_0$ for all $c \in C$ and $\widehat{\theta}_{Tc}$ is $\sqrt{T}$-gaussian with asymptotic variance

$$V_c = \left(G_c^{'} W_c G_c\right)^{-1} \left(G_c^{'} W_c S_c W_c G_c\right) \left(G_c^{'} W_c G_c\right)^{-1}. \tag{11}$$

The asymptotic variance of the efficient GMM estimator is given by

$$V_c = \left(G_c^{'} S_c^{-1} G_c\right)^{-1}. \tag{12}$$

However, we need to take into account the fact that, in our MA estimator, the moment functions $\widehat{g}_{Tc}(\theta_0)$ are different across model specifications indexed by c, which could complicate the derivation of their limiting behavior. We circumvent this problem by defining a selection matrix that contains certain rows with zeros, operating on the full moment functions, as in Domowitz and White (1982). Consider the GMM estimator obtained using the whole set of moment conditions, $c = \iota_m$, where $|c| = m$. Now, define the matrix Λ_c of dimension $|c|$ by m, such that each row $j = 1, ..., |c|$ contains zeros, except a single "1" at position i that corresponds to the moment condition as defined in model $c = \iota_m$.[5] Then,

$$\widehat{g}_{Tc}(\theta_0) = \Lambda_c \widehat{g}_T(\theta_0), \tag{13}$$

that is, we write the moment functions as a linear function of the 'full' specification, which will allow us to obtain the limiting distribution of our MA estimator, as shown in the following theorem:

Theorem 1 *(Distribution of the MA estimator for a given ω): assume that the model is correctly specified and Assumption 1 holds. As $T \to \infty$, for any $\omega \in H_m$,*

$$\widehat{\theta}_T(\omega) = \sum_{c \in C} \omega_c \widehat{\theta}_{Tc} \xrightarrow{p} \theta_0, \tag{14}$$

[5] Taking, for example, $m = 3$ (three moment conditions) and the particular specification c using conditions one and three, Λ_c is 2 by 3 with rows $(1, 0, 0)$ and $(0, 0, 1)$.

where $\widehat{\theta}_{Tc}, c \in C$, *is the GMM estimator. Moreover,*

$$\sqrt{T}\left(\widehat{\theta}_T(\omega)-\theta_0\right) \xrightarrow{d} N(0, V_\omega), \tag{15}$$

where

$$V_\omega = \left(\sum_{c\in C} \omega_c \left(G_c' W_c G_c\right)^{-1} G_c' W_c \Lambda_c\right) S \left(\sum_{c\in C} \omega_c \Lambda_c' W_c G_c \left(G_c' W_c G_c\right)^{-1}\right) \tag{16}$$

and S *denotes the long-run variance matrix employing all moment conditions (i.e.,* $c = \iota_m$*).*

In the case of efficient GMM estimation, then $V_c = (12)$, *so*

$$V_\omega = \left(\sum_{c\in C} \omega_c V_c G_c' S_c^{-1} \Lambda_c\right) S \left(\sum_{c\in C} \omega_c \Lambda_c' S_c^{-1} G_c V_c\right). \tag{17}$$

Moreover, $Z_T(\omega) = \sqrt{T}\left(\widehat{\theta}_T(\omega)-\theta_0\right)$ *convergences weakly to a zero-mean Gaussian process* $Z(\omega)$:

$$Z_T(\omega) \Rightarrow Z(\omega) \text{ on } H_m. \tag{18}$$

Remark 1. For a given ω, and noting that Λ_c is known for all $c \in C$, a consistent estimator of V_ω can be obtained using consistent estimators for G_c and W_c, for all $c \in C$, and for S as well, and inference can be carried out in the usual way.

Remark 2. The results in Theorem 1, namely the closed form expression of the asymptotic covariance of the MA estimator is very general in the context of GMM-type of estimation procedures. First, it covers the cases of linear IV and maximum likelihood estimators. Second, since model selection is indeed a special case of MA whenever $\omega_{\widetilde{c}} = 1$ and $\omega_{c'} = 0$, for all $c' \neq \widetilde{c}$, for some model $c = \widetilde{c}$, we have $\widehat{\theta}_T(\omega) = \widehat{\theta}_{T\widetilde{c}}$ and, more importantly, $V_\omega = V_{\widetilde{c}}$.

Theorem 1 is useful to understand the fact that if we only consider the first-order terms it is not difficult to derive the limiting distribution of the MA estimator, for a given ω. This implies that estimating the weight that minimizes the MSE is meaningless in our approach, since this would result in the most efficient estimator receiving full weight. Thus, we propose to include the higher order terms, which should deliver a better approximation of the properties of the MA estimator.

In general, the optimal vector ω will be unknown. As in much of the literature on model averaging, a data-dependent procedure will have to be used to determine the weights in order to implement estimation according to (10). Next, we define our weight estimator based on an optimality criterion.

3.2 Higher-Order Properties of the MA Estimator and the Optimality Criterion

The optimality criterion for estimating the model's weights follows from the higher-order AMSE of the MA estimator. On one hand, the first-order asymptotics delivers a solution that departs from typical MA schemes: it picks $\omega_{\widetilde{c}} = 1$, where $\widetilde{c} = \iota_m$, to attain the Chamberlain bound with the full model (cf. Theorem 1 above). On the other hand, the higher-order AMSE captures additional statistical properties of the estimator, especially for small sample sizes. As discussed in Hansen, Heaton and Yaron (1996), for example, and in the two special issues of the *Journal of Business and Economic Statistics* (1996 and 2002) dedicated to GMM, the standard 2-step GMM estimator may deviate substantially from its first-order asymptotic distribution.

Rilstone et al. (1996) define the same higher-order expansion as in Newey and Smith (2004),

$$\widehat{\theta}_{Tc} = \theta_0 + a_{-1/2,c} + a_{-1,c} + a_{-3/2,c} + O_p\left(T^{-2}\right), \tag{19}$$

for some $a_{-1/2,c} = O_p\left(T^{-1/2}\right)$, $a_{-1,c} = O_p\left(T^{-1}\right)$ and $a_{-3/2,c} = O_p\left(T^{-3/2}\right)$. For model c, the AMSE matrix to order $O\left(T^{-2}\right)$ is

$$\begin{aligned} AMSE\left(\widehat{\theta}_{Tc}\right) &= E\left(a_{-1/2,c}a'_{-1/2,c}\right) + E\left(a_{-1,c}a'_{-1/2,c} + a_{-1/2,c}a'_{-1,c}\right) \\ &\quad + E\left(a_{-1,c}a'_{-1,c} + a_{-3/2,c}a'_{-1/2,c} + a_{-1/2,c}a'_{-3/2,c}\right). \end{aligned} \tag{20}$$

Hence, for the MA estimator and any $\omega \in H_m$,

$$\widehat{\theta}_T(\omega) = \theta_0 + \sum_{c\in C}\omega_c a_{-1/2,c} + \sum_{c\in C}\omega_c a_{-1,c} + \sum_{c\in C}\omega_c a_{-3/2,c} + O_p\left(T^{-2}\right) \tag{21}$$

with, to the order $O\left(T^{-2}\right)$,

$$\begin{aligned} & AMSE\left(\widehat{\theta}_T(\omega)\right) \\ &= E\left(\sum_{c\in C}\omega_c a_{-1/2,c}\sum_{c\in C}\omega_c a'_{-1/2,c}\right) + E\left(\sum_{c\in C}\omega_c a_{-1,c}\sum_{c\in C}\omega_c a'_{-1/2,c} + \sum_{c\in C}\omega_c a_{-1/2,c}\sum_{c\in C}\omega_c a'_{-1,c}\right) \\ &\quad + E\left(\sum_{c\in C}\omega_c a_{-1,c}\sum_{c\in C}\omega_c a'_{-1,c} + \sum_{c\in C}\omega_c a_{-3/2,c}\sum_{c\in C}\omega_c a'_{-1/2,c} + \sum_{c\in C}\omega_c a_{-1/2,c}\sum_{c\in C}\omega_c a'_{-3/2,c}\right) \\ &= \sum_{c_1\in C}\sum_{c_2\in C}\omega_{c_1}\omega_{c_2}\left[\frac{1}{T}V_{1,c_1,c_2} + \frac{1}{T^2}\left(V_{2,c_1,c_2} + V'_{2,c_2,c_1}\right) + \frac{1}{T^2}\left(V_{3,c_1,c_2} + V_{4,c_1,c_2} + V'_{4,c_2,c_1}\right)\right], \end{aligned} \tag{22}$$

where

$$V_{1,c_1,c_2} = T.E\left(a_{-1/2,c_1}a'_{-1/2,c_2}\right); V_{2,c_1,c_2} = T^2.E\left(a_{-1,c_1}a'_{-1/2,c_2}\right) \tag{23}$$

$$V'_{2,c_2,c_1} = T^2.E\left(a_{-1/2,c_1}a'_{-1,c_2}\right); V_{3,c_1,c_2} = T^2.E\left(a_{-1,c_1}a'_{-1,c_2}\right) \tag{24}$$

$$V_{4,c_1,c_2} = T^2.E\left(a_{-3/2,c_1}a'_{-1/2,c_2}\right) \text{ and } V'_{4,c_2,c_1} = T^2.E\left(a_{-1/2,c_1}a'_{-3/2,c_2}\right). \tag{25}$$

All terms have closed form expressions (see, for example, Rilstone et al. 1996). Clearly, the AMSE of the MA estimator will only be equal to the standard (not averaged) case if we put full weight in one model c^*, say, $\omega_{c^*} = 1$ and $\omega_c = 0, c \neq c^*$.

Thus, we define the following optimality criterion for choosing ω :

$$\widehat{\omega} \equiv \widehat{\omega}_{T,m,p} = \arg \min_{\omega \in H_m} AMSE\left(\widehat{\theta}_T\left(\omega\right)\right). \tag{26}$$

One can consider the bias-corrected estimator instead

$$\widehat{\theta}_T^{bc}\left(\omega\right) = \widehat{\theta}_T\left(\omega\right) - B_T\left(\omega\right), \tag{27}$$

where the bias to order $O\left(T^{-1}\right)$ is given by

$$B_T\left(\omega\right) = \sum_{c \in C} \omega_c \left[E\left(a_{-1/2,c}\right) + E\left(a_{-1,c}\right)\right] = \sum_{c \in C} \omega_c B_{Tc} \tag{28}$$

(see Rilstone et al. 1996). Here, $\widehat{\theta}_T^{bc}\left(\omega\right)$ is unbiased to order $O\left(T^{-1}\right)$. In this context, to order $O\left(T^{-2}\right)$,

$$\widehat{\omega} = \arg \min_{\omega \in H_m} AMSE\left(\widehat{\theta}_T^{bc}\left(\omega\right)\right) = \arg \min_{\omega \in H_m} \left\{AMSE\left(\widehat{\theta}_T\left(\omega\right)\right) - B_T\left(\omega\right) B_T\left(\omega\right)'\right\}. \tag{29}$$

In practice, the $AMSE\left(\widehat{\theta}_T\left(\omega\right)\right)$ includes unknown quantities, namely the model's population parameters and moments, so that the feasible version of $\widehat{\omega}$ results from replacing these objects by their estimators and sample moment analogues and obtain $\widehat{AMSE}\left(\widehat{\theta}_T\left(\omega\right)\right)$, from which we calculate $\widehat{\omega}$. Notice as well that this is an optimization problem restricted to the unit simplex. The criteria we propose follows from a constrained quadratic optimization problem with a positive definite quadratic term and a linear constraint. Thus, unless some moment conditions are degenerate or perfectly correlated, this problem has a closed form solution. In the cases where this solution has too complicated an expression, we recommend obtaining it through numerical optimization or linear programming.[6] The proposed MA estimator might not be manageable in practice, particularly for a general nonlinear model with a large p (more below in Section 3.3).

Remark 3. We argue that simply minimizing the higher-order bias is not adequate. In this case, $Bias\left(\widehat{\theta}_T\left(\omega\right)\right) = \sum_{c \in C} \omega_c Bias\left(\widehat{\theta}_{Tc}\right)$, meaning that the full weight is on the model with smallest bias (or its norm for $p > 1$) and thus leading to model selection rather than averaging. In contrast, $AMSE\left(\widehat{\theta}_T\left(\omega\right)\right) = \sum_{c \in C} \omega_c AMSE\left(\widehat{\theta}_{Tc}\right)$ does not hold, which means that the model with the smallest AMSE will not necessarily get full weight (see the selection criteria of Donald and Newey, 2001, and Donald et al., 2009), so that gains can be obtained by using our proposed MA estimator.

Remark 4. The AMSE in (26) is defined for $p = 1$. In the general case (22), and following the existing MA literature, define the (scalar) target parameter of interest $\delta_0 = \delta\left(\theta_0\right) = \delta'\theta_0$, with δ known, such that $AMSE\left(\delta'\widehat{\theta}_T\left(\omega\right)\right) = \delta' AMSE\left(\widehat{\theta}_T\left(\omega\right)\right)\delta$ is a scalar; this then ensures the

[6] For example, the procedure QPROG for the software GAUSS helps solving this classic programming problem. If m is moderately large, a typical solution is to put zero weight on some of the individual models (see Martins and Gabriel, 2014).

feasibility of the optimization problem (see DiTraglia, 2016, for example).

3.3 Optimal MA-GMM Under Exact Identification

While the AMSE of the MA estimator has a closed form expression (following Rilstone et al. 1996), it includes several complicated terms. Also, when m is "large", the number of models to average becomes intractable: for example, for $m = 4$ and $p = 2$ we have $|C| = 11$ in the overidentified case. Thus, for simplicity, we first focus on optimal averaging over exactly identified models. The corresponding AMSE is now relatively easy to obtain and the number of models is manageable.

Consider the model's space $C_p \subset C$ whose elements satisfy $|c| = p$ so that the subset C_p contains $|C_p| = \binom{m}{p}$ distinct exactly identified models. If $m = 4$ and $p = 2$, we average over 6 models, i.e. combining moment conditions $1^\frown 2, 1^\frown 3, 1^\frown 4, 2^\frown 3, 2^\frown 4$, and $3^\frown 4$. In practice, it is as if one gives zero weight to all models c such that $|c| > p$. Following Lemma 3.1 in Rilstone et al. (1996), the third-order expansion of the exactly identified GMM estimator for model $c \in C_p$ satisfies

$$a_{-1/2,c} = -E\left(\left. \frac{\partial g_{tc}}{\partial \theta'} \right|_{\theta=\theta_0} \right)^{-1} (1/T) \sum_{t=1}^{T} g_{tc}(\theta_0) \equiv -G_c^{-1} \widehat{g}_{Tc} \tag{30}$$

$$a_{-1,c} = -G_c^{-1}\left(\widehat{G}_{Tc} - G_c\right) a_{-1/2,c} - \frac{1}{2} G_c^{-1} H_{2c} \left(a_{-1/2,c} \otimes a_{-1/2,c}\right) \tag{31}$$

$$\begin{aligned} a_{-3/2,c} = & -G_c^{-1}\left(\widehat{G}_{Tc} - G_c\right) a_{-1,c} - \frac{1}{2} G_c^{-1} \widehat{H}_{2Tc} \left(a_{-1/2,c} \otimes a_{-1/2,c}\right) \\ & -\frac{1}{2} G_c^{-1} H_{2c} \left[\left(a_{-1/2,c} \otimes a_{-1,c}\right) + \left(a_{-1,c} \otimes a_{-1/2,c}\right)\right] \\ & -\frac{1}{6} G_c^{-1} H_{3c} \left(a_{-1/2,c} \otimes a_{-1/2,c} \otimes a_{-1/2,c}\right), \end{aligned} \tag{32}$$

where $G_c \equiv E\left(\nabla g_{tc}(\theta_0)\right)$ is the usual $p \times p$ Jacobian, $H_{2c} \equiv E\left(\nabla^2 g_{tc}(\theta_0)\right)$ is a $p \times p^2$ matrix, $\widehat{H}_{2Tc} = (1/T) \sum_{t=1}^{T} \left(\left. \frac{\partial^2 g_{tc}}{\partial \theta \partial \theta'} \right|_{\theta=\theta_0} - H_{2c} \right)$, and $H_{3c} \equiv E\left(\nabla^3 g_{tc}(\theta_0)\right)$ is a $p \times p^3$ matrix. If the model is linear in the parameters (linear IV, say) then $H_{2c} = 0$, which simplifies further $a_{-1,c}$ and $a_{-3/2,c}$.

From these objects, we obtain the matrices $V_{\cdot,c_1,c_2}$, for any pair of exactly identified models (c_1, c_2) that are in the formula of the $AMSE\left(\widehat{\theta}_T(\omega)\right)$, which we need to minimize in order to choose the optimal empirical MA weights.[7]

Bias is given by

$$B_{Tc} = \frac{1}{T} G_c^{-1} \left\{ E\left[\left(\nabla g_{tc}(\theta_0) - G_c\right) G_c^{-1} g_{tc}(\theta_0)\right] - \frac{1}{2} H_{2c} \left[E\left(G_c^{-1} g_{tc}(\theta_0) \otimes G_c^{-1} g_{tc}(\theta_0)\right)\right] \right\}. \tag{33}$$

According to Newey and Smith (2004), the bias for exactly identified models is decomposed in two terms: the asymptotic bias of the GMM estimator with optimal weighting matrix and a term associated with the choice of the preliminary estimator.

[7] Exact expressions for these $V_{\cdot,c_1,c_2}$ matrices are given in the supplementary appendix.

The linear IV case In the standard linear model, the IV estimator is defined as $\widehat{\theta}_{Tc} = \left(\sum_{t=1}^{T} x_t z'_{tc}\right)^{-1} \left(\sum_{t=1}^{T} x_t y_t\right)$ and

$$g_{tc}(\theta) = z_{tc}\left(y_t - x'_t\theta\right) = z_{tc}u_t(\theta)\,; G_{tc}(\theta) = -x_t z'_{tc}; G_c = -E\left(x_t z'_{tc}\right); \frac{\partial^2 g_{tjc}}{\partial\theta\partial\theta'} = H_{2c} = H_{3c} = 0. \tag{34}$$

Assuming, for the sake of simplicity, homoskedastic errors, then $E\left(u_t^2(\theta_0)\,|z_{tc}\right) = \sigma^2$,

$$V_{1,c_1,c_2} = \sigma^2 E\left(x_t z'_{tc_1}\right)^{-1} E\left(z_{tc_1} z'_{tc_2}\right) E\left(z_{tc_2} x'_t\right)^{-1}, \tag{35}$$

$$V_{2,c_1,c_2} = -E\left(x_t z'_{tc_1}\right)^{-1}\left\{E\left[u_t^2(\theta_0)\left(x_t z'_{tc_1} - E\left(x_t z'_{tc_1}\right)\right)E\left(x_t z'_{tc_1}\right)^{-1} z_{tc_1} z'_{tc_2} E\left(z_{tc_2} x'_t\right)^{-1}\right]\right\} \tag{36}$$

$$\begin{aligned} &V_{3,c_1,c_2} \\ &= E\left(x_t z'_{tc_1}\right)^{-1}\left\{\begin{array}{c} E\left[\begin{array}{c} u_t^2(\theta_0)\left(x_t z'_{tc_1} - E\left(x_t z'_{tc_1}\right)\right)E\left(x_t z'_{tc_1}\right)^{-1} z_{tc_1}\cdot \\ z'_{tc_2} E\left(z_{tc_2} x'_t\right)^{-1}\left(z_{tc_2} x'_t - E\left(z_{tc_2} x'_t\right)\right) \end{array}\right] \\ +E\left[u_t^2(\theta_0)\, z_{tc_1} z'_{tc_2}\right] \end{array}\right\} E\left(z_{tc_2} x'_t\right)^{-1} \end{aligned} \tag{37}$$

and

$$\begin{aligned} &V_{4,c_1,c_2} \\ &= E\left(x_t z'_{tc_1}\right)^{-1}\left\{E\left[\begin{array}{c} u_t^2(\theta_0)\left(x_t z'_{tc_1} - E\left(x_t z'_{tc_1}\right)\right)E\left(x_t z'_{tc_1}\right)^{-1}\cdot \\ \left(x_t z'_{tc_1} - E\left(x_t z'_{tc_1}\right)\right)E\left(x_t z'_{tc_1}\right)^{-1} z_{tc_1} z'_{tc_2} \end{array}\right]\right\} E\left(z_{tc_2} x'_t\right)^{-1} \end{aligned} \tag{38}$$

and

$$B_{Tc} = -\frac{1}{T} E\left(x_t z_{tc}\right)^{-1} E\left[u_t(\theta_0)\left(x_t z_{tc} - E\left(x_t z_{tc}\right)\right)E\left(x_t z_{tc}\right)^{-1} z_{tc}\right]. \tag{39}$$

3.4 Gains in Using MA-GMM - An Illustrative Example

It is infeasible to show analytically, for the general case presented in subsection 3.2, that our proposed optimal MA-GMM estimator has an AMSE smaller than the corresponding standard (not averaged) GMM estimator. For general nonlinear models, we make comparisons using Monte Carlo simulations in section 6. Nevertheless, we are able to prove analytically the gains of using the optimal MA-GMM estimator of section 3.3 in the simplest of setups, with two candidate models.

The linear IV regression model is specified by a structural equation of interest

$$y = x\theta + X\gamma + u, \tag{40}$$

where y is a $T \times 1$ vector, x is a $T \times n$ matrix of endogenous regressors, X is a $T \times K$ matrix of exogenous regressors, and by a reduced form equation for the endogenous x

$$x = z\Pi + X\Phi + V, \tag{41}$$

where z is a $T \times m$ matrix of instruments, with x, X and z full ranked and $m \geq n$. For the sake of simplicity, let $n = 1$, $K = 0$ and assume i.i.d. data. The error structure $w_i = (u_i, V_i)'$ satisfies the moment conditions $E(w_i|z_i) = 0$ and

$$E\left(w_i w_i'|z_i\right) = \left(\begin{array}{cc} \sigma^2 & \varphi \\ \varphi & \sigma_v^2 \end{array}\right). \tag{42}$$

Also, assume that $\sigma_3 = E\left(u_t(\theta)^3|z_{tc}\right) = 0$ and $\sigma_4 = E\left(u_t(\theta)^4|z_{tc}\right) = 1$ for all $c = 1, ..., m$.[8] Define the (scalar) parameter of interest θ and endogeneity arises if $E(x_i u_i) = \varphi \neq 0$.

For the averaging scheme, let $m = |C_p| = 2$ so that we account for two candidate models: $c = 1$ and $c = 2$ with $T \times 1$ instruments z_1 and z_2, respectively, with $z_1 \neq z_2$. Assuming further homocorrelated errors, $E[u_t(\theta_0)x_t|z_t] = E[u_t(\theta_0)x_t] = \varphi$, and higher cross-moments for error and endogenous variables, $E\left[u_t^2(\theta_0)x_t|z_t\right] = \varphi_3$ and $E\left[u_t^2(\theta_0)x_t^2|z_t\right] = \varphi_4$. Also, let z_{tc} have zero expectation and variance $E\left(z_{tc}^2\right) = \sigma_c^2$ and $E\left(z_{tc}^3\right) = \varrho_c, E\left(z_{tc}^4\right) = \kappa_c$. Moreover, let $E(z_{t1}z_{t2}) = \phi$, $E\left(z_{t1}^2 z_{t2}^2\right) = \phi_4$, $E\left(z_{t1}^3 z_{t2}\right) = \phi_4^{(1)}$, $E\left(z_{t1}z_{t2}^3\right) = \phi_4^{(2)}$, $E\left(z_{t1}^2 z_{t2}\right) = \phi_3^{(1)}$, and $E\left(z_{t1}z_{t2}^2\right) = \phi_3^{(2)}$. Finally, define the covariance of endogenous and instruments as $E(x_t z_{tc}) = \rho_c$.

The competing estimators (averaged or not) differ on the instruments each use. Thus, and to simplify calculations, we compute the AMSE's scaling them by T^2 and fixing the model coefficients that do not involve the z's: $\sigma^2 = 1$, $\varphi = 0.5$, $\varphi_3 = 0$ and $\varphi_4 = 1$.[9] Furthermore, assume z_{t1} and z_{t2} each normally distributed with correlation coefficient $\frac{\phi}{\sigma_1\sigma_2}$, so that the cokurtosis statistics equal $\frac{E\left(z_{t1}^2 z_{t2}^2\right)}{\sigma_1^2\sigma_2^2} = 1 + 2\rho^2$ and $\frac{E\left(z_{t1}^3 z_{t2}\right)}{\sigma_1^3\sigma_2} = \frac{E\left(z_{t1}z_{t2}^3\right)}{\sigma_1\sigma_2^3} = 3\rho$. The coskewness is $\phi_3^{(1)} = \phi_3^{(2)} = 0$ due to $\varphi_3 = 0$. Also, assuming $\sigma_1^2 = \sigma_2^2 = 1$, we have $\phi_4 = 1 + 2\phi^2$ and $\phi_4^{(1)} = \phi_4^{(2)} = 3\phi$. In this setup,

$$V_{1,1,2} = V_{2,1,2} = \frac{\phi}{\rho_1\rho_2}, V_{3,1,2} = \frac{1}{\rho_1^2\rho_2^2}\left(1 + 2\phi^2 + 2\rho_1\rho_2\phi\right), \tag{43}$$

$$V_{4,1,2} = \frac{\phi}{\rho_1^3\rho_2}\left(3 + \rho_1^2\right) \text{ and } B_T(\omega) = -\frac{1}{2T}\left(\omega_1\frac{1}{\rho_1^2} + (1-\omega_1)\frac{1}{\rho_2^2}\right) \tag{44}$$

and given the restriction $\omega_1 + \omega_2 = 1$,

$$AMSE\left(\widehat{\theta}_T(\omega)\right) = \omega_1^2\left[\frac{1 + \phi\left[(T+6)\rho_1^2 + 2\phi + 6\right]}{\rho_1^4}\right] + \omega_2^2\left[\frac{1 + \phi\left[(T+6)\rho_2^2 + 2\phi + 6\right]}{\rho_2^4}\right] \tag{45}$$

$$+2\omega_1\omega_2\left[\frac{\rho_1\rho_2\left[\phi\rho_1\rho_2(T+6) + 2\phi^2 + 1\right] + 3\phi\left(\rho_1^2 + \rho_2^2\right)}{\rho_1^3\rho_2^3}\right]. \tag{46}$$

[8] Hence, $E\left(u_t(\theta)^3\right) = E\left(u_t(\theta)^3 z_{tc}\right) = 0$ and $E\left(u_t(\theta)^4\right) = 1$, so that $E\left(u_t(\theta)^4 z_{tc}\right) = E(z_{tc})$.

[9] That is, the level of endogeneity is $E[u_t(\theta_0)x_t] = \varphi = 0.5$, x_t has zero mean, $E\left[u_t^2(\theta_0)x_t\right] = E(x_t) = \varphi_3 = 0$, and unit variance, $E\left[u_t^2(\theta_0)x_t^2\right] = E\left(x_t^2\right) = \varphi_4 = 1$.

As expected, the $AMSE\left(\widehat{\theta}_T(\omega)\right)$ depends on the sample size, T, the covariances of the endogenous variable and each instrument, ρ_c, and the covariance of the two instruments, ϕ. For ease of exposition, we derive next the conditions under which both models get optimal equal weights.

Proposition 1 *(Optimal equal weights and the corresponding AMSE): consider the standard linear IV regression model under the assumptions defined above and the optimal MA-GMM estimator with equal weights* $\omega_1^o = \omega_2^o = \frac{1}{2}$ *for* $\rho_1^2 = \rho_2^2$*. Let* $\psi = \rho_1^2(T+6)$ *and T fixed. For* $\rho_1 = \rho_2$*, consider* $\phi \in \left(\underline{\phi}^+, 1\right)$*, where* $\underline{\phi}^+ = -\frac{1}{16}(\psi+6) + \frac{1}{16}\sqrt{(\psi+6)^2 - 32}$*. For* $\rho_1 = -\rho_2$*, take* $\phi \in \left(\underline{\phi}^-, 1\right)$*, where* $\underline{\phi}^- = -\frac{1}{4}(\psi+10) + \frac{1}{12}\sqrt{9(\psi+10)^2 - 24}$*. Here,* $-1 < \underline{\phi}^+ < \underline{\phi}^- < 0$*. The corresponding optimal AMSE of the MA estimator is*

$$AMSE\left(\widehat{\theta}_T(\omega^o)\right) = \frac{2 + \phi(T+6)\left(\rho_1^2 + \rho_1\rho_2\right) + 4\phi^2 + 6\phi\left(1 + \frac{\rho_2}{\rho_1}\right)}{2\rho_1^4}. \qquad (47)$$

Proposition 1 provides a closed form expression for the AMSE of an optimal MA-GMM estimator. Next, we show that, under a specific set of the model's assumptions, this AMSE is the smallest compared to the standard GMM estimators (exactly identified or overidentified).

Proposition 2 *(Optimal MA-GMM estimator with the smallest AMSE): consider the standard linear IV regression model under the assumptions defined in Proposition 1 and the optimal MA-GMM estimator with equal weights* $\omega_1^o = \omega_2^o = \frac{1}{2}$ *for* $\rho_1^2 = \rho_2^2$*. Denote the GMM estimator with a single instrument as* $\widehat{\theta}_{T1}$ *and the GMM estimator with both instruments as* $\widehat{\theta}_{Toverid}$*. For any* T*,*

$$AMSE\left(\widehat{\theta}_T(\omega^o)\right) < AMSE\left(\widehat{\theta}_{Toverid}\right) < AMSE\left(\widehat{\theta}_{T1}\right), \qquad (48)$$

where $AMSE\left(\widehat{\theta}_T(\omega^o)\right) < AMSE\left(\widehat{\theta}_{Toverid}\right)$ *holds in the following cases:*

- *for* $\rho_1 = \rho_2$*, all* $\phi \in (\underline{\phi}^+, \overline{\phi}_+]$*, where* $\overline{\phi}_+ = 3\psi + 21 - \sqrt{9(\psi+7)^2 - 2\psi - 5}$*;*
- *for* $\rho_1 = -\rho_2$*, all* $\phi \in (\underline{\phi}^-, \overline{\phi}_-]$*, where* $\overline{\phi}_- = \psi - 9 - \sqrt{(\psi-9)^2 - 2\psi - 5}$*, such that* $\psi = \rho_1^2(T+6) \in \left[0, 10 - 2\sqrt{6}\right] \cup [10 + 2\sqrt{6}, +\infty)$*. Here, if* $0 \leq \psi \leq 10 - 2\sqrt{6}$ *then* $0 < \overline{\phi}_- < \overline{\phi}_+ < 1$*, whereas if* $\psi \geq 10 + 2\sqrt{6}$ *then* $0 < \overline{\phi}_+ < \overline{\phi}_- \leq 1$*.*

Proposition 2 illustrates a simple case where it can be shown analytically the AMSE gains from using the optimal MA-GMM estimator. It imposes equal weights and thus one must not rule out the possibility of further gains with non-equal optimal weights. With equal weights and opposite signs for the correlations of each instrument and the endogenous variable, $\rho_1 = -\rho_2$, the optimal MA-GMM estimator is always superior for any non-negative correlation of instruments, $\phi \in (\underline{\phi}^-, 1]$ when $\psi \geq 10 + 2\sqrt{6}$. This means that, for example, with $\rho_1^2 = 0.27$, $T = 50$ and $\phi = 0.25$, we have

$\frac{AMSE(\widehat{\theta}_{Toverid})}{AMSE(\widehat{\theta}_T(\omega^o))} = 6.165$ and $\frac{AMSE(\widehat{\theta}_{T1})}{AMSE(\widehat{\theta}_T(\omega^o))} = 22.089$. It is also straightforward to show that if $\phi = 0$,

$$\frac{AMSE\left(\widehat{\theta}_{Toverid}\right)}{AMSE\left(\widehat{\theta}_T\left(\omega^o\right)\right)} = \rho_1^2\left(\frac{T}{2}+3\right)+\frac{9}{4} > 2.25 \text{ and } \frac{AMSE\left(\widehat{\theta}_{T1}\right)}{AMSE\left(\widehat{\theta}_T\left(\omega^o\right)\right)} = \rho_1^2\left(T+5\right)+10 > 10$$

That is, our optimal MA-GMM estimator is able to, at least, halve the AMSE of the standard overidentified GMM estimator.

3.5 Optimal MA-GMM Under Overidentification

In this section, we present the optimal MA-GMM estimator with any number of moment conditions in each model, namely overidentification, for given moment conditions $g_t(\theta_0)$. We show that, for this particular class of models, the general optimal MA-GMM estimator in section 3.2 may have tractable closed-form expressions, as the MA-GMM estimator under exact identification defined in section 3.3.

Rilstone et al. (1996) show that one can write an overidentified model in terms of an exactly identified system at the expense of adding an extra (nuisance) parameter. They consider models of the form $g_t(\theta_0) = z_t u_t\left(\theta_0\right)$, with $u_t\left(\theta_0\right)$ potentially nonlinear, and homoskedastic errors $E\left(u_t^2\left(\theta_0\right)z_t z_t'\right) = \sigma^2 E\left(z_t z_t'\right)$, so that the weighting matrix is $E\left(z_t z_t'\right)^{-1}$. For overidentified GMM $(m \geq p)$, the model can be written as an exactly identified system

$$E\left(h_t\left(\theta_0,\tau_0\right)\right) = E\begin{pmatrix} z_t \otimes \varepsilon_t\left(\theta_0,\tau_0\right) \\ \tau_0 z_t u_t\left(\theta_0\right)\end{pmatrix} = 0, \tag{49}$$

where

$$\tau_0 = E\left(\nabla u_t\left(\theta_0\right)z_t'\right)E\left(z_t z_t'\right)^{-1} \text{ and } \varepsilon_t\left(\theta_0,\tau_0\right) = \nabla u_t\left(\theta_0\right) - \tau_0 z_t. \tag{50}$$

Here, $\nabla u_t\left(\theta_0\right)$ and $\varepsilon_t\left(\theta_0,\tau_0\right)$ are $p\times 1$, z_t is $m\times 1$ and the (extended) parameter vector is now (θ_0,τ_0) of size $p+pm$ (τ_0 is $p\times m$), which equals the number of equations: $z_t\otimes\varepsilon_t\left(\theta_0,\tau_0\right)$ is $mp\times 1$ and $\tau_0 z_t u_t\left(\theta_0\right)$ is $p\times 1$.

Therefore, we can apply the results for exactly identified models denoting $g_t(\theta_0)$ by $g_t^*(\theta_0,\tau_0) = \left[z_t\otimes\varepsilon_t\left(\theta_0,\tau_0\right) \quad \tau_0 z_t u_t\left(\theta_0\right)\right]'$, where θ_0 is the parameter of interest and τ_0 is a nuisance parameter. If τ_0 is known, then g_t^* is itself only a function of θ_0. However, in general, τ_0 is unknown and a consistent estimator $\widehat{\tau}$ is needed for the GMM estimation of θ_0. In conclusion, averaging GMM point estimates from overidentified models is a somewhat unwieldy task and is seemingly limited to models of the form $g_t(\theta_0) = z_t u_t\left(\theta_0\right)$.

4 Non-optimal alternative MA-GMM estimators

In practice, and in particular for large nonlinear moment conditions models, the general optimal MA-GMM estimator may be difficult to obtain - there are no analytical results under exact identification, rather we evaluate its merits through Monte Carlo simulations. As an alternative, and following the standard literature in model averaging procedures, we suggest linking the problem of selecting empirical weights $\widehat{\omega}$ with moment selection criteria obtained in the estimation stage. This can be achieved either by direct 'smoothing' or by minimization of a given moment selection criterion.

4.1 Moment Selection Criteria for GMM

Given that the rejection of the J-statistic is an indicator that some moment conditions are invalid, Andrews (1999) suggests that this can be used to consistently select the correct moment conditions. Thus, a GMM moment selection criteria for a given model is defined as

$$MSC_T\left(c\right) = J_T\left(c\right) - \kappa_T\left(\left|c\right| - p\right), \tag{51}$$

where $|c| - p$ is the number of overidentifying restrictions and $\kappa_T = o\left(T\right)$ is a sequence that defines the selection criterion ($\kappa_T = 2$ for the AIC; $\kappa_T = \log T$ for the BIC; and $\kappa_T = Q \log\log T$ for some $Q > 2$ for the HQ-type criterion). Note the "bonus term" $\kappa_T\left(|c| - p\right)$ rewarding selection vectors that utilize more moment conditions.[10]

While the criteria above stress the satisfaction of *orthogonality* conditions, other procedures have been proposed in which the focus is on the *relevance* of moment conditions. Under somewhat more restrictive assumptions, Hall et al. (2007) suggest selecting a model according to the relevant moment selection criterion

$$RMSC_T\left(c\right) = \ln\left(\left|\widehat{V}_c\right|\right) + \kappa_T\left(\left|c\right| - p\right), \tag{52}$$

where the efficient GMM variance-covariance matrix $\widehat{V}_c$ is evaluated at $\widehat{\theta}_{Tc}$. On the other hand, Hall and Peixe (2003), in a generalized IV framework, consider the problem of instrument selection based on a combination of the efficiency and non-redundancy conditions

$$CCIC_T\left(c\right) = T\sum_{i=1}^{p} \ln\left[1 - r_{i,T}^2\left(c\right)\right] + \kappa_T\left(\left|c\right| - p\right), \tag{53}$$

where $r_{i,T}\left(c\right)$ is the i^{th} sample canonical correlation between $d_t(\tilde{\theta}_T)$ and $z_t\left(c\right)$, with $d_t(\theta) = \frac{\partial u_t(\theta)}{\partial \theta}$ and $\tilde{\theta}_T$ is a $\sqrt{T}-$ consistent preliminary estimator. Note that here $g(y_t, \theta) \equiv u_t(\theta)z_t(c)$, $u_t(\theta)$ is scalar and, if the model is linear, $d_t(\theta) = -x_t$.

Alternatively, given a set of moment conditions known to be valid, one can select moment

[10] Under relatively standard assumptions, Andrews (1999) shows that the moment selection criteria estimator $\widehat{c}_{msc} = \arg\min_{c\in\mathcal{C}} MSC_T\left(c\right)$ is a consistent estimator of the single "correct" selection vector c_0.

conditions that minimize a criterion based on an estimate of the AMSE, as suggested by Donald and Newey (2001) for linear IV estimation with homosdekasticity and Donald et al. (2009) for the general case. Following the latter, the criterion for GMM is of the form

$$AMSE_T(c) = \hat{\Pi}_c^2/T + \hat{\Phi}_c, \tag{54}$$

where $\hat{\Pi}_c^2/T$ is an estimate of a squared bias term, while $\hat{\Phi}_c$ is an asymptotic variance term that tends to be smaller the more instruments are used (the full notation for (54) is cumbersome, see Donald et al., 2009 for further details and Donald and Newey, 2001 for the linear IV case).

4.2 Smooth Moment Selection Criteria Weights

As suggested by Buckland, Burnham and Augustin (1997), a simple averaging scheme can be obtained by using weights proportional to the exponential form of a given GMM selection criterion *Crit* (see definitions in section 2). Thus, a smooth AIC, BIC, AMSE, etc. scheme (denoted as $\widehat{\omega}_{S\text{-}Crit}$) is based on weights for candidate model M,

$$\widehat{\omega}_M(Crit) = \frac{\exp(-\frac{1}{2}Crit_M)}{\sum_{M'\in\mathcal{M}}\exp(-\frac{1}{2}Crit_{M'\in\mathcal{M}})} \tag{55}$$

where the sum term encompasses all, not necessarily nested, $M' \in \mathcal{M}$ models of interest.[11] Other simplified weighting schemes have been explored in the literature and can potentially be employed, see Claeskens and Hjort (2008) and Martins and Gabriel (2014).

4.3 Selecting Weights by Minimizing GMM Moment Selection Criteria

In the spirit of Hansen (2007), we also propose obtaining the weight vector ω by numerical minimization of GMM moment selection criteria, which gives rise to two distinct situations. In a first case, we can evaluate a given moment selection criteria at the MA estimator $\widehat{\theta}_T(\omega)$: using Andrews's (1999) MSC, the empirical selected weight vector is defined as

$$\widehat{\omega}_{MSC} = \arg\min_{\omega\in H_m} MSC_{T\overline{c}}(\omega) = \arg\min_{\omega\in H_m}\left(J_{T\overline{c}}(\omega) - \kappa_T\left(|\overline{c}| - p\right)\right), \tag{56}$$

where $J_{T\overline{c}}(\omega) = T\widehat{g}_{T\overline{c}}\left(\widehat{\theta}_T(\omega)\right)' W_{T\overline{c}}\widehat{g}_{T\overline{c}}\left(\widehat{\theta}_T(\omega)\right)$, for a given set of moment conditions $\overline{c}$ and given $W_{T\overline{c}}$.

A second case comprises selection criteria which cannot be evaluated at $\widehat{\theta}_T(\omega)$, so therefore weights are selected as

$$\widehat{\omega}_{SC} = \arg\min_{\omega\in H_m} SC_T(\omega), \tag{57}$$

[11] For numerical stability, it is sometimes recommended that the maximum $Crit_T$ value is subtracted to each $Crit_M$.

where

$$SC_T(\omega) = \omega' diag\left(SC_1, ..., SC_{|C|}\right)\omega = \sum_{c=1}^{|C|} \omega_c^2 SC_c, \tag{58}$$

where $diag(\cdot)$ refers to a $|C| \times |C|$ diagonal matrix. In particular,

$$\widehat{\omega}_{RMSC} = \arg\min_{\omega \in H_m} \left[\omega' diag\left(\left|\widehat{V}_1\right|, ..., \left|\widehat{V}_{|C|}\right|\right)\omega\right] \tag{59}$$

$$\widehat{\omega}_{AMSE} = \arg\min_{\omega \in H_m} \left[\omega' diag\left(AMSE_{T,1}, ..., AMSE_{T,|C|}\right)\omega\right] \tag{60}$$

or, in an GIV context,

$$\widehat{\omega}_{CCIC} = \arg\min_{\omega \in H_m} \left[\omega' diag\left(\sum_{i=1}^{p} \ln\left[1 - r_{i,T,1}^2\right], ..., \sum_{i=1}^{p} \ln\left[1 - r_{i,T,|C|}^2\right]\right)\omega\right]. \tag{61}$$

Remark 5. As in Hansen (2007), the solution $\widehat{\omega}$ is found by numerical algorithms. It solves a constrained optimization problem with non-negativity and summation constraints ($\omega_c \in [0,1]$, for all c and $\sum_{c\in C} \omega_c = 1$, respectively).

Remark 6. Note that, although averaging occurs over specifications using different combinations of moment conditions, the minimization of GMM selection criteria in (56) depends on the J-statistic. This, in turn, requires the weight matrix to be chosen and therefore a set of moment conditions $\overline{c}$ to be fixed. Moreover, and unlike the least squares MA estimator of Hansen (2007) and the two-step MA instruments estimators of Kuersteiner and Okui (2010), which have distinct number of parameters to estimate for each individual model, in our case $p_c = p$ for all c. Hence,

$$\min_{\omega \in H_m} MSC_{T\overline{c}}(\omega) = \min_{\omega \in H_m} J_{T\overline{c}}(\omega) \tag{62}$$

for any penalty term κ_T. Thus, an MA estimator that minimizes a GMM selection criterion will be solely based on the $J_T(\omega)$-statistic. For the sake of efficiency, one can pick $\overline{c} = \iota_m$, a vector of ones, which implies using the whole set of moment conditions (in this case, $|\overline{c}| = m$ and, in terms of notation, "c" is dropped):

$$J_T(\omega) = T\widehat{g}_T\left(\widehat{\theta}_T(\omega)\right)' W_T \widehat{g}_T\left(\widehat{\theta}_T(\omega)\right). \tag{63}$$

For the linear IV/2SLS case, for a set of variables x_t and instruments z_t, such that $y_t = (x_t', z_t')'$, then

$$J_{T\overline{c}}(\omega) = T\left(\frac{1}{T}\sum_{t=1}^{T} z_{\overline{c},t}\left(y_t - x_t' \sum_{c\in C} \omega_c \widehat{\theta}_{Tc}\right)\right)' W_{T\overline{c}}\left(\frac{1}{T}\sum_{t=1}^{T} z_{\overline{c},t}\left(y_t - x_t' \sum_{c\in C} \omega_c \widehat{\theta}_{Tc}\right)\right). \tag{64}$$

Remark 7. The trace minimization criterion for $\widehat{\omega}_{SC}$ can be seen as a general approach to obtain weights. Liang, Zou, Wan and Zhang (2011) follow this approach, for example, although their method is based on an approximation of a MA estimator's MSE. In this vein, another potential weight selection criteria would be to find the argument $\widehat{\omega}$ that minimizes the trace of the MA variance-covariance matrix, V_ω, although this may be hard to accomplish in practice.

5 Properties of the MA-GMM Estimator

The limiting properties of the higher-order MA-GMM estimator are far from straightforward to derive. As mentioned earlier, the individual higher-order AMSEs "are extremely complicated" (Newey and Smith, 2004) for general GMM estimators and the same applies to exactly identified estimators (Rilstone et al. 1996). Thus, the averaging of these individual higher-order AMSEs makes it virtually impossible to study analytically the limiting laws of the optimal random weights and, consequently, the corresponding MA-GMM estimator. Following the existent MA-GMM literature, we conjecture that our higher-order estimator, $\sqrt{T}\left(\widehat{\theta}_T\left(\widehat{\omega}^o_{T,m,p}\right)-\theta_0\right)$, will also no longer be asymptotically normal due to a random optimal weight $\widehat{\omega}^o_{T,m,p}$ that is likely to converge in distribution to some function of the non-zero normal process that is part of the limit law of $\widehat{\theta}_{Tc}$ (see, for example, Sueishi, 2013, and DiTraglia, 2016). In particular, and similarly to what we are able to show in Theorem 2 below, we conjecture that our optimal weight $\widehat{\omega}^o_T = \arg\min_{\omega\in H_m} \widehat{AMSE\left(\widehat{\theta}_T(\omega)\right)}$ is not a consistent estimator for $\omega^o = \arg\min_{\omega\in H_m} AMSE\left(\widehat{\theta}_T(\omega)\right)$ even in the limit, due the random nature of $\widehat{\omega}$. Nevertheless, we can derive the limiting properties of our proposed non-optimal MA-GMM estimators. We can work on its first-order asymptotic distribution and regarding its higher-order distribution we can only know the AMSE which was previously defined in the paper.

Correspondingly, the MA estimator with smooth weights based on criterion $Crit$ as in (55) is denoted as $\widehat{\theta}_T(\widehat{\omega}^{S\text{-}Crit}_T)$, $\widehat{\theta}_T(\widehat{\omega}^{MSC}_T)$ with weights based on (56) and $\widehat{\theta}_T(\widehat{\omega}^{SC}_T)$, where SC denotes RMSC, CCIC or AMSE criteria. For simplicity, and following much of the model averaging literature, we will focus on results for selection criteria with the AIC penalty (see Claeskens and Hjort, 2008). Given that the randomness properties of the weights follow from the limiting behavior of the selection criteria, in order to study the properties of the MA-GMM estimator for each criterion, we need the following additional assumption:

Assumption 2 *(Regularity conditions for GMM selection criteria)*
Depending on the chosen MA approach, assume either the conditions for ***(A2-MSC)*** *the MSC as in Andrews (1999); or* ***(A2-RMSC)*** *the RMSC as in Hall et al. (2007); or* ***(A2-CCIC)*** *the CCIC as in Hall and Peixe (2003); or* ***(A2-AMSE)*** *the AMSE as in Donald et al. (2009).*[12]

The asymptotic distribution of the MA estimator depends on the limiting law of the weights. On one hand, the limit result for a smoothed scheme follows directly from convergence of the selection criterion. On the other hand, an 'arg min'-based approach provides a limit quantity that follows from weak convergence of the objective function for selecting the weights. Also, it should be noticed

[12] See Supplementary Appendix for details.

that the random MSC weights converge in distribution to a specific variable due to the random nature of the J-statistic, whereas for the RMSC, CCIC and AMSE cases we have convergence in probability. Thus, we are able to derive the following theorem:

Theorem 2 *(Distribution of the MA estimator for random $\widehat{\omega}$): assume that the model is correctly specified, the Assumptions 1 and 2 hold and consider the AIC penalty for the smoothed RMSC and CCIC MA procedures.*

a) The distributions for MSC-based estimators. As $T \to \infty$,

$$\sqrt{T}\left(\widehat{\theta}_T\left(\widehat{\omega}_T^{MSC}\right) - \theta_0\right) \overset{d}{\to} \sum_{c \in C} \widetilde{\omega}_c Z_c \tag{65}$$

and

$$\sqrt{T}\left(\widehat{\theta}_T\left(\widehat{\omega}_T^{S\text{-}MSC}\right) - \theta_0\right) \overset{d}{\to} \sum_{c \in C} \omega^{S\text{-}MSC}\left(c, p\right) Z_c \tag{66}$$

such that

$$\widehat{\omega}_T^{MSC} \overset{d}{\to} \widetilde{\omega} = \left(\widetilde{\omega}_1, ..., \widetilde{\omega}_{|C|}\right)' = \arg\min_{\omega \in H_m} \{b' Z(\omega) + \frac{1}{2} Z(\omega)' A Z(\omega)\} \text{ and} \tag{67}$$

$$\widehat{\omega}_{Tc}^{S\text{-}MSC} = \frac{\exp(-\frac{1}{2} J_T\left(c\right) + (|c| - p))}{\sum_{c' \in C} \exp(-\frac{1}{2} J_T\left(c'\right) + (|c'| - p))} \overset{d}{\to} \omega^{S\text{-}MSC}\left(c, p\right), c = 1, ..., |C|, \tag{68}$$

where $Z(\omega)$ was defined in Theorem 1, b is a zero-mean normal random vector, $A = \operatorname{plim}_{T\to\infty} T^{-1} \nabla^2 J_T^*(\theta_0)$ with $\nabla^2 J_T^*(\theta) = \partial^2 J_T^*(\theta)/(\partial\theta\partial\theta')$ denoting the matrix of second partial derivatives of $J_T^*(\theta) = T \widehat{g}_{T\overline{c}}\left(\theta\right)' W_{T\overline{c}} \widehat{g}_{T\overline{c}}\left(\theta\right)$, $J_T\left(c\right)$ was defined in Section 2,

$$\omega^{S\text{-}MSC}\left(c, p\right) = \frac{\exp(-\frac{1}{2}\chi_{(|c|-p)} + |c|)}{\sum_{c' \in C} \exp(-\frac{1}{2}\chi_{(|c'|-p)} + |c'|)}, \tag{69}$$

and the normal random variable Z_c was defined in Lemma 1.

b) The distributions for RMSC, CCIC, and AMSE-based estimators. As $T \to \infty$,

$$\sqrt{T}\left(\widehat{\theta}_T\left(\widehat{\omega}_T\right) - \theta_0\right) \overset{d}{\to} N\left(0, V_{\omega^{o*}}\right), \tag{70}$$

where $V_{\omega^{o*}}$ is the matrix V_ω of Theorem 1 evaluated at $\omega = \omega^{o*}$, corresponding to either one of the *quantities*

$$\omega^{o,RMSC} = \arg\min_{\omega \in H_m} \left\{ \omega' diag\left(|V_1|, ..., \left| V_{|C|} \right| \right) \omega \right\}, \tag{71}$$

$$\omega^{o,CCIC} = \arg\min_{\omega \in H_m} \left\{ \omega' diag\left(\sum_{i=1}^{p} \ln\left[1 - r_{i,1}^2\right], ..., \sum_{i=1}^{p} \ln\left[1 - r_{i,|C|}^2\right] \right) \omega \right\}, \tag{72}$$

$$\omega^{o,AMSE} = \arg\min_{\omega \in H_m} \left\{ \omega' diag\left(AMSE_1, ..., AMSE_{|C|} \right) \omega \right\} \tag{73}$$

$$\omega_c^{o,S\text{-}RMSC} = \frac{|V_c|^{-\frac{1}{2}} \exp\left(p - |c|\right)}{\sum_{c' \in C} |V_{c'}|^{-\frac{1}{2}} \exp\left(p - |c'|\right)}, \tag{74}$$

$$\omega_c^{o,S\text{-}CCIC} = \frac{\exp(-\frac{1}{2}\sum_{i=1}^{p} \ln\left[1 - r_i^2\left(c\right)\right] - (|c| - p))}{\sum_{c' \in C} \exp(-\frac{1}{2}\sum_{i=1}^{p} \ln\left[1 - r_i^2\left(c'\right)\right] - (|c'| - p))}, \tag{75}$$

$$\omega_c^{o,S\text{-}AMSE} = \frac{\exp(-\frac{1}{2} AMSE\left(c\right))}{\sum_{c' \in C} \exp(-\frac{1}{2} AMSE\left(c'\right))}, \tag{76}$$

where V_c was defined in Section 2, $r_{i,c}$ is the i^{th} corresponding population canonical correlation and $AMSE\left(c\right)$ is the AMSE derived by Donald et al. (2009) (see also Donald and Newey, 2001 for linear IV estimators).

Remark 8. In the case of MSC-based MA estimators, the asymptotic distribution of the MA estimator will often be a scale mixture of normal densities (as suggested by simulations not reported here, but available upon request), but is not necessarily always normal. Bootstrap methods can be employed to obtain an approximate distribution of the MA estimator in this case. The nonstandard asymptotic distributions of the random weights and MA estimators are not new in the literature - see, for example, DiTraglia (2016) and Zhang and Liu (2019) in the context of two other different types of MA estimators.

6 Monte Carlo Study

In this section, we report results from a Monte Carlo study assessing the finite sample properties of the proposed MA estimators, using the selection method of Donald et al. (2009) as our benchmark and contrasting their performance along distinct dimensions, namely sample size (T) and number of moment restrictions (M). To do so, we use the fairly general nonlinear design used of Schennach (2007) as the DGP, given by

$$g(y_t, \theta) = \left[r_t(\theta) \quad r_t(\theta) y_{t2} \quad r_t(\theta) \left(y_{t3} - 1\right) \quad ... \quad r_t(\theta) \left(y_{tM} - 1\right) \right]' \tag{77}$$

where

$$r_t(\theta) = \exp\left(-0.72 - \left(y_{t1} + y_{t2}\right)\theta + 3y_{t2}\right) - 1. \tag{78}$$

Here, we have $M \geq 2$ moment restrictions and a single parameter θ_0 that takes the value of 3, i.e., $E\left[g(y_t, \theta_0)\right] = 0$ if and only if $\theta_0 = 3$ with

$$(y_{t1}, y_{t2})' \sim N\left(0, (0.16)\, I_2\right) \tag{79}$$

$$y_{tj} \sim \chi_1^2, \text{ for } j = 3, ..., K. \tag{80}$$

The third moments of all elements of $g_t(\theta_0)$ are non-zero.

The overall purpose is to examine how well each MA procedure estimates θ_0 along distinct dimensions, namely sample size (T) and number of moment restrictions (M), hence we cover the cases of small and large models and samples, i.e. $M = 2, 4, 10, 20$, and $T = 50, 100, 200$.[13] The number of replications is $10,000$. Following Donald et al. (2009), we compute their estimator using the reference model selection criterion (DIN) and, for the sake of completeness, we compute the estimators based on the MSC and $RMSC$ selection criteria with BIC penalty. Furthermore, we also estimate θ_0 by GMM and Empirical Likelihood (EL) using the full set of restrictions (GMM-*all* and EL-*all*).

We consider optimal-weights MA-GMM estimators averaging over exactly identified models, smooth-weights MA estimators using MSC-BIC, denoted as S-MSC, and RMSC-BIC, denoted as S-RMSC, and MA estimators that make use of $\widehat{\omega}_{MSC}$, denoted MA-MSC. The MA-MSC estimator is computed using $W_{T\bar{c}} = \left(\frac{Z'Z}{T}\right)^{-1}$ and with $\bar{c} = \iota_j$ (all restrictions). Non-optimal methods using other criteria produced similar results. In terms of the optimal MA estimators we consider the full expression of its higher-order AMSE (MA-GMM-*ho*), assuming that $V_4 = 0$ (MA-GMM-*ho4*) and imposing $V_3 = V_4 = 0$ (MA-GMM-*ho3,4*). For the non-optimal MA estimators, we compare three different averaging schemes: *i*) taking all combinations of models (*-*all*); *ii*) models adding one moment restriction at a time (*-*add*), i.e., models g_1, $g_1{}^\frown g_2$, $g_1{}^\frown g_2{}^\frown g_3, ...,$ $g_1{}^\frown ...{}^\frown g_M$; *iii*) models that are only exactly identified (*-*ex*). For the optimal MA estimators, as well as the *-*add* and *-*ex* schemes, we average over M models; on the other hand, for the *-*all* scheme we have $|C| = 3$ $(M = 2)$, $|C| = 15$ $(M = 4)$, $|C| = 1023$ $(M = 10)$ and $|C| = 1048575$ $(M = 20)$. The MA procedure S-MSC-*ex* can be interpreted as an equal weighted scheme because $\widehat{\omega} = 1/M$.

For each estimator, we compute the median bias (MB), the median absolute deviation (MAD), and interdecile ranges (DR) (q90-q10) to measure dispersion. We also examine statistical inference by computing the coverage rate for 90% confidence intervals using a consistent estimator for $V_{\omega^{o*}}$ and under normality. Notice that according to Theorem 2, normality rarely applies to MA estimators, namely MA-GMM and MA-MSC. This way, we will also be able to draw some conclusions about the inference properties of the MA estimators by (wrongly) assuming normality of the distribution.[14]

[13] For conciseness, here we focus on $M = 2$ and $M = 4$; results for $M = 10$ and $M = 20$ are qualitatively similar and are reported in the Supplementary Appendix.

[14] The literature on post-model selection inference (e.g. Pötscher, 1991) argues that the conditional and unconditional distribution of post-model selection estimators cannot be uniformly consistently estimated and that the coverage probability of the confidence interval is lower than the nominal level.

< Table 1 >

The results for bias, absolute deviation and interdecile ranges are in Table 1. A few general conclusions should be highlighted. First, there is always at least one MA approach, regardless of the specific averaging scheme, that performs better than the selection procedure of Donald et al. (2009) - this is particularly evident for small M. Second, non-optimal MA procedures tend to dominate over optimal ones. Third, averaging models by adding one moment restriction at a time seems to be the best MA approach. Fourth, the EL-*all* outperforms the GMM-*all* only for small M if all moment conditions are used.

< Table 2 >

The results for coverage rates are presented in Table 2. In general, MA estimators are reasonably accurate, even (wrongly) assuming normality, especially for large T and moderate M, followed by the RMSC (for small M) selection procedure, which displays relatively good coverage rates. For small M the MA methods based on exactly identified models are the most accurate ones, especially the S-MSC-*ex* and MA-MSC-*ex*, while the S-RMSC-*add* is clearly the best for large M. On the other hand, higher-order 'optimal' MA estimators behave well for large T.

< Table 3 >

As a final exercise, we analyze the distributions of the estimated weights $\widehat{\omega}$, displayed in Table 3 for $M = 2$. The most notable result is that there is a non-negligible probability of optimal and MA-MSC estimators giving full weight to a single competing model, namely for large T. In contrast, and as expected, smoothing schemes tend not to drop any model from estimation. For $M = 2$, the models are equally weighted for the higher-order case and S-RMSC-*ex* (besides S-MSC-*ex*, obviously). For the remaining *-*ex* scheme, MA-MSC-*ex*, and the two MSC-*add* procedures (MA-MSC-*add* and S-MSC-*add*), significantly more weight is given to the second restriction, while the opposite is true for S-RMSC-*add*. Taking all possible combinations produces a variety of results: S-MSC-*all* gives almost all weight to the model using both restrictions, S-RMSC-*all* equally weights the two exactly identified models and neglects the full model, while MA-MSC-*all* gives most of the weight to the model with the second restriction only. Moreover, averaging estimators tend to favour models with a minimum number of conditions. The exception is the S-MSC, typically giving most of the weight to the model using the full set of restrictions.

7 Empirical Application

To further illustrate the usefulness of our MA methods in small sample cases, we revisit Acemoglu et al.'s (2001) study on the effect of institutions on post-colonial development. These authors uncover a strong negative reduced-form relationship between GDP per capita today and settler mortality rates, purportedly reflecting the effect of settler mortality working through the institutions brought by Europeans. For each model, their IV estimates are relatively precisely estimated and large, results changing little when additional controls are included,[15] although there is variation across specifications.

We use the same dataset as in Acemoglu et al. (2001), and, for the sake of simplicity, we focus on a common sample of 59 countries for which data on mortality, protection against expropriation and GDP is available, estimating their baseline model

$$\log y_i = \alpha + \beta R_i + u_i, \tag{81}$$

where $\log y_i$ is the logarithm of 1995 per capita GDP (on a PPP basis) for country i and R is the "Risk of Expropriation" index from Political Risk Services, averaged over the period 1985-1995, measured on a scale from 0 to 10, with a higher value indicating lower risk.

The instruments for R_i include $mort_i$, the logarithm of an estimate of the mortality rate experienced by European settlers during the period in which the country was colonized, but also measures of European migration to the colonies and early institutions (see their Table A1 for details). We focus on six different cases of exactly identified models considered by Acemoglu et al. (2001) which we then average: *i)* $mort_i$, *ii)* European settlements in 1900 ("es1900"), *iii)* constraint on the executive in 1900 ("c1900"), *iv)* democracy in 1900 ("d1900"), *v)* constraint on the executive in the first year of independence ("cindep"), and *vi)* democracy in the first year of independence ("dindep"), either one of these as the only instrument for institutions.

< Table 4 >

Our results for the base sample are in the first panel of Table 4. The six IV point estimates of Acemoglu et al. (2001) for β ($\widehat{\beta}^*$, first row) range from 0.55 to 0.94, which, although qualitatively similar, indicates a considerable quantitative difference; our IV estimates for the common sample (second row) are equally wide. In turn, the MA estimates (first column, each model's estimated weight in italics) of the effect of protection against expropriation on GDP per capita are obviously smaller than the baseline estimate $\widehat{\beta}^* = 0.86$ ($\widehat{\beta}^* = 0.94$ if n=64). These range from 0.68 to 0.85, depending on the weight given to the baseline estimate $\widehat{\beta}^*$ - the largest is for MA-Smooth$_{CCIC}$, with a weight of 0.91 to the model "es1900", which has the same point estimate as the baseline. The higher-order MA estimates are relatively close to $\widehat{\beta}^*$ (0.75 and 0.72) giving zero weight to the estimates from models "d1900" and "cindep". Noticeably, the weights given to the $\widehat{\beta}^*$ depend on the averaging scheme, but it never receives the highest weight out of the six estimates (it is quite close for the optimal criterion). In all MA cases, "es1900" has the highest estimated weight.

Acemoglu et al. (2001) further consider two country groups, one without the 'Neo-Europes' (United States, Canada, Australia and New Zealand), for which they found larger estimated effects

[15] Such as the identity of the main colonizer, legal origin, climate, religion, geography, natural resources, soil quality, and measures of ethno-linguistic fragmentation, among others, which may be correlated with mortality and growth.

compared to their baseline estimate, and one where all the African countries are dropped from the base sample, with smaller estimated effects. Computing the MA estimates in this case (middle panel of Table 4), these are larger than the one found by Acemoglu et al. (2001), ranging from 1.26 to 1.74, compared to 1.26. The weights given to the baseline estimate are the largest for all MA schemes, except for the higher-order ones (in this MA procedure, we obtain almost equal weights for all models). With respect to the MA estimates for the base sample without Africa (bottom panel of Table 4), they are essentially the same as the baseline estimated effect of 0.52, partly because this model's estimate receives the largest weight in all MA schemes, with the exception of the MA-Numerical$_{MSC}$.

Overall, the above suggests that in empirical situations where the range of estimates across specifications can be quite wide, a model averaging approach can offer us not only a more balanced perspective, but also give an indication, through the estimated model weights, of how competing models are favoured by the data.

8 Conclusion

This paper develops new GMM-based model averaging estimators. We propose optimally-weighted estimators in the sense that weights minimize the higher-order AMSE of the MA estimator. We use a variety of moment selection criteria to select weights for averaging across GMM estimates. This can be achieved by direct smoothing of information criteria arising from the estimation stage, or by numerical minimization of a specific criterion. We study the asymptotic properties of the resulting estimators for correctly specified models and we illustrate our methods by revisiting Acemoglu et al.'s (2001) study on the effect of institutions on economic performance. As shown, it is quite useful to understand which specifications are favoured by the MA criteria.

Monte Carlo experiments using a standard nonlinear model show that our MA estimation procedures outperform the optimal instrument selection method of Donald et al. (2009) in many relevant setups, including models with weaker instruments. An interesting outcome of these simulations is that averaging (by smoothing or by numerical minimization) based on the moment selection criteria used as benchmark often leads to better results than selection based on that MSC itself.

There are several aspects that merit further attention. First, post-averaging inference for this type of estimators remains an unresolved issue - simulation-based as in Zhang and Liu (2019) can be perhaps be extended to our framework. Moreover, we note that in this paper we focused on the case of estimating a parameter vector of fixed dimension, averaging over different estimates obtained with different instrument or moment condition sets, which is perhaps the more empirically relevant case. However, it would be interesting to consider a local misspecification setup similar to Hjort and Claeskens (2003) or to study the behavior of the estimator under misspecification of the moment conditions (see Cheng and Liao, 2015, Caner et al., 2018). Also, we take m as fixed, but conjecture that the results hold true for m growing with, albeit at a smaller rate than the sample size, similarly to Donald and Newey (2001, pp. 1161-1162).

Furthermore, model averaging within the GEL class of estimators would be an important extension, given their good properties in finite samples. Deriving the statistical properties of the MA version of GEL estimators is however far from straightforward, particularly due to the dependence between the point estimator and the Lagrange multiplier, and therefore beyond the scope of this paper. In fact, the GEL class of estimators share the main problems of overidentified GMM, although we can assume general functions for the moment conditions and the nuisance coefficient plays the particular role of the Lagrange multiplier. These important topics on model averaging for moment conditions models are left for future research.

References

[1] Acemoglu, D., Johnson, S. and Robinson J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, *91*, 1369–1401.

[2] Anatolyev, S. (2005). GMM, GEL, serial correlation, and asymptotic bias. *Econometrica*, *73*, 983–1002.

[3] Anatolyev, S. and Gospodinov, N. (2011). *Methods for Estimation and Inference in Modern Econometrics*. Boca Raton: Chapman and Hall.

[4] Andrews, D. W. K. (1999). Consistent Moment Selection Procedures for Generalized Method of Moments Estimation. *Econometrica, 67*, 543–564.

[5] Bao, Y. and Ullah, A. (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics, 140*, 650–669.

[6] Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model Selection: an Integral Part of Inference. *Biometrics, 53*, 603–618.

[7] Caner, M. (2009). LASSO Type GMM Estimator. *Econometric Theory*, *25*, 1–23.

[8] Caner, M., Han, X. and Lee, Y. (2018). Adaptive Elastic Net GMM Estimation With Many Invalid Moment Conditions: Simultaneous Model and Moment Selection. *Journal of Business and Economic Statistics*, *36*, 24–46.

[9] Chao, J. C. and Swanson, N. R. (2005). Consistent Estimation with a Large Number of Weak Instruments. *Econometrica*, *73*, 1673–1692.

[10] Chen, X., Jacho-Chávez, D. T. and Linton, O. (2016). Averaging of an Increasing Number of Moment Condition Estimators. *Econometric Theory*, *32*, 30–70.

[11] Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *Journal of Econometrics*, *186*, 443–464.

[12] Cheng, X., Liao, Z. and Shi, R. (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics*, *10*, 931–979.

[13] Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

[14] DiTraglia, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics*, *195*, 187–208.

[15] Domowitz, I. and White, H. (1982). Misspecified Models with Dependent Observations. *Journal of Econometrics*, *20*, 35–58.

[16] Donald, S. G., Imbens, G. W. and Newey, W. K. (2009). Choosing Instrumental Variables in Conditional Moment Restriction Models. *Journal of Econometrics, 152*, 28–36.

[17] Donald, S. G. and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica, 69*, 1161–1192.

[18] Hall, A. R., Inoue, A., Jana, K. and Shin, C. (2007). Information in Generalized Method of Moments Estimation and Entropy Based Moment Selection. *Journal of Econometrics, 138*, 488–512.

[19] Hall, A. R. and Peixe, F. P .M. (2003). A Consistent Method for the Selection of Relevant Instruments. *Econometric Reviews, 22*, 269-287.

[20] Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica*, *75*, 1175–1189.

[21] Hansen, C., Hausman, J. and Newey, W. K. (2008). Estimation with Many Instrumental Variables. *Journal of Business and Economics Statistics*, *26*, 398–422.

[22] Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, *50*, 1029–1054.

[23] Hansen, L. P., Heaton, J. and Yaron, A. (1996), Finite-Sample Properties of Some Alternative GMM estimators, *Journal of Business and Economic Statistics*, 1996, 14, 262-80.

[24] Hayashi, F. (2000). Econometrics. *Princeton University Press.*

[25] Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association*, *98*, 879–899.

[26] Kapetanios, G. and Marcellino, M. (2010). Cross-sectional Averaging and Instrumental Variable Estimation with Many Weak Instruments. *Economics Letters*, *108*, 36–39.

[27] Kuersteiner, G. and Okui, R. (2010). Constructing Optimal Instruments by First-Stage Prediction Averaging. *Econometrica*, *78*, 697–718.

[28] Lee, Y. and Zhou, Y. (2015). Averaged Instrumental Variables Estimator. Center for Policy Research, 212, Syracuse University.

[29] Liang, H., Zou, G., Wan, A. T. K. and Zhang, X. (2011). Optimal Weight Choice for Frequentist Model Average Estimators. *Journal of the American Statistical Association*, *106*, 1053–1066.

[30] Martins, L. F. and Gabriel, V. J. (2014). Linear Instrumental Variables Model Averaging Estimation. *Computational Statistics and Data Analysis*, *71*, 709–724

[31] Newey, W. K. and Smith, R. J. (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica, 72*, 219–255.

[32] Newey, W. K. and Windmeijer, F. (2009). Generalized Method of Moments with Many Weak Moment Conditions. *Econometrica*, *77*, 687–719.

[33] Okui, R. (2011). Instrumental Variable Estimation in the Presence of Many Moment Conditions. *Journal of Econometrics* , *165*, 70–86.

[34] RilstoneSrivastavaUllah Rilstone, P., Srivastava, V. K. and Ullah, A. (1996). The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics* , *96*, 369–395.

[35] Schennach, S. M. (2007). Point Estimation with Exponentially Tilted Empirical Likelihood. *The Annals of Statistics*, *35*, 634–672.

[36] Smith, R. J. and Ramalho, J. S. (2002). Generalized Empirical Likelihood Non-Nested Tests. *Journal of Econometrics*, *107*, 99–125.

[37] Sueishi, N. (2013). Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging. *Econometrics* (Open Access), *1*, 141–156.

[38] Xiao, Z. (2010). The Weighted Method of Moments Approach for Moment Condition Models. *Economics Letters, 107*, 183–186.

[39] Zhang, X. and Liu, C-A. (2019). Inference After Model Averaging in Linear Regression Models. *Econometric Theory*, *35*, 816–841.

Tables

Table 1: Median Bias (*MD*), Median Absolute Deviation (*MAD*) and Decile Range (*DR*), $M = 2, 4$

$M = 2$	$T = 50$			$T = 100$			$T = 200$		
	MB	MAD	DR	MB	MAD	DR	MB	MAD	DR
GMM-*all*	-0.0334	0.2843	1.2055	-0.0046	0.1916	0.7500	0.0068	0.1339	0.5121
EL-*all*	0.0462	0.2548	1.0077	0.0293	0.1816	0.7065	0.0197	0.1317	0.5015
DIN	-0.0370	0.3003	1.1700	0.0086	0.2113	0.8217	0.0219	0.1508	0.5897
MSC	-0.0334	0.2843	1.2055	-0.0046	0.1916	0.7500	0.0068	0.1339	0.5121
RMSC	-0.1291	0.3345	1.2245	-0.1171	0.2483	0.8669	-0.0805	0.1789	0.6251
MA-GMM-*ho*	-0.0768	0.2949	1.1124	-0.0440	0.2177	0.8192	-0.0177	0.1554	0.5912
MA-GMM-*ho4*	-0.0527	0.2791	1.0681	-0.0400	0.2020	0.7738	-0.0201	0.1458	0.5527
MA-GMM-*ho3,4*	0.0230	0.2540	1.0018	0.0045	0.1822	0.7091	0.0011	0.1319	0.5030
S-MSC-*all*	-0.0073	0.2712	1.0763	0.0106	0.1864	0.7255	0.0099	0.1317	0.5049
S-MSC-*add*	-0.0006	0.2837	1.1512	0.0185	0.1961	0.7740	0.0179	0.1364	0.5282
S-MSC-*ex*	0.0547	0.2492	0.9969	0.0375	0.1826	0.7169	0.0254	0.1330	0.5075
S-RMSC-*all*	-0.0058	0.2538	0.9897	-0.0106	0.1813	0.7037	-0.0080	0.1318	0.4982
S-RMSC-*add*	0.0976	0.3256	1.2899	0.0533	0.2407	0.9243	0.0318	0.1718	0.6631
S-RMSC-*ex*	-0.0034	0.2552	0.9921	-0.0104	0.1810	0.7031	-0.0082	0.1318	0.4982
MA-MSC-*all*	-0.0146	0.2791	1.0882	-0.0207	0.2173	0.8055	-0.0137	0.1602	0.5937
MA-MSC-*add*	0.0069	0.2610	1.0280	0.0043	0.1860	0.7225	0.0076	0.1335	0.5097
MA-MSC-*ex*	-0.0137	0.2788	1.0801	-0.0207	0.2171	0.8031	-0.0137	0.1601	0.5932
$M = 4$									
GMM-*all*	-0.1154	0.2956	1.1566	-0.0360	0.1913	0.7455	0.0010	0.1332	0.5070
EL-*all*	0.0944	0.2721	1.0912	0.0498	0.1884	0.7347	0.0357	0.1350	0.5038
DIN	-0.1101	0.3141	1.2199	-0.0204	0.2125	0.8374	0.0171	0.1538	0.5953
MSC	-0.1670	0.3361	1.6045	-0.0642	0.2228	0.9352	-0.0159	0.1468	0.5831
RMSC	-0.1843	0.3470	1.2133	-0.1324	0.2524	0.8593	-0.0886	0.1789	0.6131
MA-GMM-*ho*	-0.2458	0.3515	1.1581	-0.1413	0.2497	0.8924	-0.0690	0.1680	0.6368
MA-GMM-*ho4*	-0.1985	0.3164	1.0736	-0.1121	0.2206	0.8125	-0.0517	0.1502	0.5649
MA-GMM-*ho3,4*	-0.2357	0.3613	1.3343	-0.1919	0.2718	1.0639	-0.1364	0.1900	0.8159
S-MSC-*all*	-0.2203	0.3406	1.4326	-0.0993	0.2186	0.9303	-0.0351	0.1477	0.5849
S-MSC-*add*	-0.0741	0.2815	1.1428	-0.0201	0.1916	0.7497	0.0047	0.1327	0.5065
S-MSC-*ex*	-0.4867	0.5147	1.2792	-0.5128	0.5204	1.1134	-0.5290	0.5313	0.9997
S-RMSC-*all*	-0.2124	0.3045	1.0049	-0.1452	0.2182	0.7171	-0.0868	0.1492	0.5183
S-RMSC-*add*	0.0925	0.3132	1.2155	0.0494	0.2391	0.8962	0.0257	0.1692	0.6394
S-RMSC-*ex*	-0.2095	0.3035	1.0080	-0.1456	0.2183	0.7170	-0.0871	0.1492	0.5186
MA-MSC-*all*	-0.0695	0.3204	1.2537	-0.0406	0.2391	0.9431	-0.0112	0.1745	0.6871
MA-MSC-*add*	-0.0530	0.2669	1.0295	-0.0236	0.1870	0.7302	-0.0006	0.1339	0.5068
MA-MSC-*ex*	-0.0709	0.3171	1.2344	-0.0441	0.2378	0.9319	-0.0168	0.1739	0.6854

Notes: "*all*" employs the full set of restrictions; "*ho*" denotes full 'optimal' higher-order weights, "*ho4*" assumes $V_4 = 0$, "*ho3,4*" imposes $V_3 = V_4 = 0$; "*add*" means adding one moment restriction at a time; "*ex*" denotes averaging only exactly identified models.

Table 2: Coverage Rate, $M = 2, 4$

	$T = 50$		$T = 100$		$T = 200$	
	$M = 2$	$M = 4$	$M = 2$	$M = 4$	$M = 2$	$M = 4$
GMM-*all*	0.7355	0.7008	0.7980	0.7748	0.8415	0.8310
EL-*all*	0.7629	0.6417	0.8097	0.7443	0.8437	0.8111
DIN	0.7650	0.7220	0.8064	0.7830	0.8330	0.8206
MSC	0.7355	0.6551	0.7980	0.7180	0.8415	0.7861
RMSC	0.8030	0.7613	0.8234	0.8132	0.8398	0.8384
MA-GMM-*ho*	0.8186	0.7675	0.7552	0.8227	0.7761	0.8665
MA-GMM-*ho4*	0.8117	0.7670	0.7838	0.8203	0.8094	0.8657
MA-GMM-*ho3,4*	0.7914	0.7763	0.8282	0.8242	0.8533	0.8770
S-MSC-*all*	0.1699	0.0507	0.1397	0.0420	0.1056	0.0278
S-MSC-*add*	0.7348	0.7025	0.7866	0.7718	0.8282	0.8337
S-MSC-*ex*	0.7903	0.8382	0.8243	0.8721	0.8461	0.9096
S-RMSC-*all*	0.8387	0.6932	0.8646	0.7435	0.8920	0.7891
S-RMSC-*add*	0.7668	0.7547	0.8171	0.8085	0.8472	0.8565
S-RMSC-*ex*	0.8038	0.7697	0.8394	0.8300	0.8699	0.8822
MA-MSC-*all*	0.7688	0.4597	0.7984	0.4829	0.8253	0.5190
MA-MSC-*add*	0.7607	0.7262	0.8071	0.7898	0.8440	0.8406
MA-MSC-*ex*	0.8044	0.8571	0.8360	0.8993	0.8591	0.9347

See notes to Table 1.

Table 3: Weights (ω) Distribution, M= 2

$M = 2$	$T = 50$			$T = 100$			$T = 200$		
	ω mean	ω sd	% $\omega = 1$	ω mean	ω sd	% $\omega = 1$	ω mean	ω sd	% $\omega = 1$
MA-GMM-*ho*	0.5184	0.3778	0.2035	0.5980	0.3728	0.2768	0.6670	0.3480	0.3181
	0.4816	0.3778	0.1919	0.4020	0.3728	0.1494	0.3330	0.3480	0.1020
MA-GMM-*ho4*	0.5104	0.3078	0.0777	0.5788	0.3048	0.1171	0.6317	0.2831	0.1413
	0.4896	0.3078	0.0953	0.4212	0.3048	0.0716	0.3683	0.2831	0.0439
MA-GMM-*ho3,4*	0.4659	0.1351	0.0002	0.4829	0.1212	0.0001	0.5054	0.1070	0.0000
	0.5341	0.1351	0.0020	0.5171	0.1212	0.0008	0.4946	0.1070	0.0003
S-MSC-*all*	0.0896	0.1160	0.0000	0.0688	0.1161	0.0000	0.0459	0.1008	0.0000
	0.0896	0.1160	0.0000	0.0688	0.1161	0.0000	0.0459	0.1008	0.0000
	0.8208	0.2320	0.0000	0.8625	0.2321	0.0000	0.9081	0.2016	0.0000
S-MSC-*add*	0.1241	0.2049	0.0027	0.0994	0.2079	0.0020	0.0670	0.1803	0.0006
	0.8759	0.2049	0.0000	0.9006	0.2079	0.0000	0.9330	0.1803	0.0000
S-MSC-*ex*	-	-	-	-	-	-	-	-	-
S-RMSC-*all*	0.4852	0.1463	0.0000	0.5183	0.1345	0.0000	0.5409	0.1216	0.0000
	0.4914	0.1531	0.0000	0.4702	0.1387	0.0000	0.4535	0.1236	0.0000
	0.0234	0.0157	0.0000	0.0115	0.0066	0.0000	0.0057	0.0029	0.0000
S-RMSC-*add*	0.9529	0.0354	0.0000	0.9786	0.0086	0.0000	0.9898	0.0033	0.0000
	0.0471	0.0354	0.0000	0.0214	0.0086	0.0000	0.0102	0.0033	0.0000
S-RMSC-*ex*	0.4979	0.1547	0.0000	0.5249	0.1390	0.0000	0.5442	0.1238	0.0000
	0.5021	0.1547	0.0000	0.4751	0.1390	0.0000	0.4558	0.1238	0.0000
MA-MSC-*all*	0.1269	0.1917	0.0035	0.0970	0.1535	0.0022	0.0860	0.1339	0.0034
	0.6738	0.2973	0.3332	0.7143	0.2700	0.3453	0.7266	0.2514	0.3261
	0.1993	0.1912	0.0185	0.1887	0.1731	0.0104	0.1874	0.1588	0.0052
MA-MSC-*add*	0.0665	0.2051	0.0036	0.0310	0.1424	0.0023	0.0220	0.1147	0.0034
	0.9335	0.2051	0.8871	0.9690	0.1424	0.9464	0.9780	0.1147	0.9603
MA-MSC-*ex*	0.1985	0.2137	0.0037	0.1868	0.2028	0.0024	0.1879	0.1988	0.0039
	0.8015	0.2137	0.3650	0.8132	0.2028	0.3650	0.8121	0.1988	0.3380

See notes to Table 1; '% $\omega = 1$' denotes the percentage of replications for which a particular model is given full weight; recall that S-MSC-*ex* corresponds to equal weights $\widehat{\omega} = 1/M$.

Table 4: MA-IV Regressions of log GDP per capita

Instruments (z)		*mort*	*es1900*	*c1900*	*d1900*	*cindep*	*dindep*
Base Sample							
IV $\widehat{\beta}^*$ Acemoglu et al.'s (2001)		0.94 (n=64)	0.87 (n=63)	0.71 (n=60)	0.72 (n=59)	0.60 (n=60)	0.55 (n=60)
IV ($n = 59$)		0.86	0.86	0.70	0.72	0.34	0.42
MA-HigherOrderOptimal	0.75	*0.30*	*0.33*	*0.18*	*0.00*	*0.00*	*0.18*
MA-HigherOrderOptimal-*bc*	0.72	*0.30*	*0.33*	*0.17*	*0.00*	*0.00*	*0.19*
MA-Smooth$_{DN}$	0.80	*0.19*	*0.43*	*0.14*	*0.22*	*0.00*	*0.02*
MA-Smooth$_{RMSC}$	0.68	*0.17*	*0.19*	*0.18*	*0.19*	*0.11*	*0.16*
MA-Smooth$_{CCIC}$	0.85	*0.06*	*0.91*	*0.01*	*0.02*	*0.00*	*0.00*
MA-Numerical$_{MSC}$	0.71	*0.22*	*0.22*	*0.18*	*0.18*	*0.09*	*0.11*
Base Sample without Neo-Europes							
IV $\widehat{\beta}^*$ Acemoglu et al.'s (2001)		1.28 (n=60)	-	-	-	-	-
IV ($n = 55$)		1.19	1.98	0.99$^{\#}$	1.51$^{\#}$	-1.03$^{\#}$	0.08$^{\#}$
IV ($n = 59$)		1.26	2.18	-	-	-	-
MA-HigherOrderOptimal	1.74	*0.47*	*0.53*	-	-	-	-
MA-HigherOrderOptimal-*bc*	1.53	*0.46*	*0.54*	-	-	-	-
MA-Smooth$_{DN}$	1.26	*1.00*	*0.00*	-	-	-	-
MA-Smooth$_{RMSC}$	1.47	*0.77*	*0.23*	-	-	-	-
MA-Smooth$_{CCIC}$	1.31	*0.94*	*0.06*	-	-	-	-
MA-Numerical$_{MSC}$	1.26	*1.00*	*0.00*	-	-	-	-
Base Sample without Africa							
IV $\widehat{\beta}^*$ Acemoglu et al.'s (2001)		0.58 (n=37)	-	-	-	-	-
IV ($n = 33$)		0.52	0.73	0.68	0.66	0.32	0.31
MA-HigherOrderOptimal	0.54	*0.42*	*0.20*	*0.14*	*0.00*	*0.00*	*0.23*
MA-HigherOrderOptimal-*bc*	0.53	*0.42*	*0.20*	*0.14*	*0.00*	*0.00*	*0.23*
MA-Smooth$_{DN}$	0.54	*0.27*	*0.19*	*0.13*	*0.12*	*0.11*	*0.17*
MA-Smooth$_{RMSC}$	0.53	*0.23*	*0.16*	*0.15*	*0.15*	*0.15*	*0.17*
MA-Smooth$_{CCIC}$	0.52	*0.98*	*0.01*	*0.00*	*0.00*	*0.00*	*0.00*
MA-Numerical$_{MSC}$	0.61	*0.16*	*0.25*	*0.23*	*0.22*	*0.08*	*0.07*

Notes: "*mort*" is the log European settler mortality, "*es1900*" is European settlements in 1900, "*c1900*" is constraint on executive in 1900, "*d1900*" is democracy in 1900, "*cindep*" is constraint on executive in first year of independence, and "*dindep*" is democracy in first year of independence; $\widehat{\beta}^*$ denotes the baseline estimate; Weight estimates in italics; "*bc*" is the bias-corrected optimal MA estimator; "#" means that β is not statistical significant for that model (using that z).