Relevant moment selection under mixed identification strength*

Prosper Dovonon†, Firmin Doko Tchatoka‡, and Michael Aguessy§

Concordia University and University of Adelaide

April 19, 2019

Abstract

This paper proposes a moment selection method in the presence of moment condition models with mixed identification strength. That is moment conditions including moment functions that are local to zero uniformly over the parameter set. We show that the relevant moment selection procedure of Hall et al. (2007) is inconsistent in this setting as it does not explicitly account for the rate of convergence of parameter estimation of the candidate models which may vary. We introduce a new moment selection procedure based on a criterion that sequentially evaluates the rate of convergence of the candidate model's parameter estimate and the entropy of the estimator's asymptotic distribution. The benchmark estimator that we consider is the two-step efficient generalized method of moments (GMM) estimator which is known to be efficient in this framework as well. A family of penalization functions is introduced that guarantees the consistency of the selection procedure. The finite sample performance of the proposed method is assessed through Monte Carlo simulations.

1 Introduction

The validity of the standard moment condition based inference hinges on strong/point identification. Strongly identified models are those solved by a unique parameter value. Many estimators have been proposed including the generalized method of moments (GMM) and the generalized empirical likelihood (GEL) estimators that are all consistent and asymptotically normal under further regularity conditions. Moment selection methods have also been developed under standard identification settings.

*We would like to thank Jean-Jacques Forneron, Eric Renault, and Zhongjun Qu for helpful comments. We also thank participants in the 2018 African Econometric Society meeting in Benin, the 2018 Canadian Econometric Study Group meeting in Ottawa and seminar participants at Boston University for helpful discussions. The first author gratefully acknowledges financial support from Social Sciences and Humanities Research Council (SSHRC).

†Economics Department, Concordia University, 1455 de Maisonneuve Blvd. West, H-1155, Montreal, Quebec, Canada, H3G 1M8. Email: prosper.dovonon@concordia.ca (Corresponding author).

‡School of Economics, The University of Adelaide, 10 Pulteney Street, Adelaide, Australia, SA 5005. Email: firmin.dokotchatoka@adelaide.edu.au.

§Economics Department, Concordia University, 1455 de Maisonneuve Blvd. West, H-1155, Montreal, Quebec, Canada, H3G 1M8. Email: michaelaguessy@gmail.com.

The literature on moment selection presents two main approaches. One is based on Lasso-type penalized estimation procedures in which both the parameter of interest and the *best* subset of moment restrictions are jointly estimated. This strand of the literature includes Belloni et al. (2012), Cheng and Liao (2015), Caner and Fan (2015) and Windmeijer et al. (2018).

The second strand of the literature on moment selection adopt a more classical methodology for model selection by relying on information criteria. This approach includes Andrews (1999), Donald and Newey (2001), Andrews and Lu (2001), Hall and Peixe (2003), Hall et al. (2007). The selection problem in these papers consists in selecting the *best* subset of moment restrictions among those useful to estimate a given parameter as the one minimizing an information criterion. In this framework, all the candidate models are expressed in terms of that same parameter of interest and the selection methods proposed in these papers differ by their choice of information measure. Andrews (1999) and Andrews and Lu (2001) rely on the GMM overidentification test statistic with the aim to select correct moment restrictions. Donald and Newey (2001) rely on the mean square error of some estimators including the two-stage least square estimator, its bias corrected version and the limited information maximum likelihood estimator whereas Hall et al. (2007) consider an entropy-based moment selection criterion with the focus on selecting from a set of correct moment restrictions, the relevant ones. This is a set of moment restrictions that does not contain a subset of restrictions with equal amount of information about the model parameter nor is included in a set of moment restrictions that carry more information about the parameter. In some sense, RMSC of Hall et al. (2007) and the J-statistic selection criterion of Andrews (1999) are complementary.

Common to all the papers cited above is the requirement of strong identification for the consistency of the selection procedure and to ensure valid inference using the selected model. Nevertheless, strong identification is not always guaranteed for moment condition models and a still growing literature is devoted to inference in models that do not have point identification property. Identification properties are outlined by considering on the one hand strong identification and on the other hand the extreme identification pattern where the model is uninformative about the parameter of interest. In the latter, consistent estimation is not possible and identification is deemed weak. Between weak and strong identification lies a wide range of identification patterns. The strength of a moment restriction is captured by how fast it potentially vanishes over the whole parameter space as the sample size grows. The faster the moment function of the restriction does vanish, the weaker is the restriction. Weak restrictions are those vanishing at least at the rate $T^{-\frac{1}{2}}$; strong ones are those vanishing only at the true value whereas those vanishing over the parameter set at rate $T^{-\alpha}$, $\alpha \in (0, 1/2)$ are considered semi-weak (or semi-strong). More importantly, the moment restrictions defining a moment condition model can have various strengths leading the model to have mixed identification strength. In addition to the classical linear instrumental variable (IV) model with instruments of possibly mixed strength analyzed in Sections 3 and 5, further examples of such models can be found in Antoine and Renault (2012) who study inference in moment condition models with mixed strength. See also Caner (2009) and Andrews and Cheng (2012). Even though point identification fails for these models in the limit, the fact that these models, by the central limit theorem, gather information about the parameter of interest at a faster rate than they lose their potential for identification, consistent estimation becomes possible. This feature has first been pointed out by Antoine and Renault (2009) who also show - in this

setting - that consistent estimators may converge at faster rate in some directions of the parameters space.

This paper proposes a moment selection method for moment condition with mixed identification strength. We build on the work of Hall et al. (2007) and propose a relevant moment selection procedure that consistently selects the best model even if this model is of mixed strength. We argue that, in the configuration of heterogeneity of restrictions' strength, candidate models must be valued by the rate of convergence of the estimator that they deliver and, two models with the same rate of estimation should be differentiated by the amount of information they convey about the model parameter which is encapsulated in the entropy of the asymptotic distribution of the parameter estimate. The estimator that we use as benchmark is the GMM estimator which is shown to be asymptotically efficient in this framework by Dovonon et al. (2019). We propose a feasible selection criterion that has these properties. This criterion turns out to be a *modified* version of RMSC that we label mRMSC.

mRMSC conveniently scales the information part of RMSC to provide a sequential estimation of rate of convergence and entropy. In addition, new penalty terms are introduced that guarantee the consistency of the selection procedure. Conditions under which mRMSC lead to consistent selection are outlined and we show that the new selection procedure is robust to the presence of uninformative and weak models. In comparison to RMSC criterion and accounting for the scaling factor, mRMSC penalizes more strongly larger models. Indeed, the penalty term of mRMSC is proportional to $(1/\ln T)^{\alpha}$, $\alpha > 0$ while that of the BIC-RMSC - identified as the best performing the RMSC criterion - is $\ln\sqrt{T}/\sqrt{T}$. The choice of penalty for mRMSC is guided both by robustness to unknown model identification strength while guaranteeing selection consistency. In this case, stronger penalization seems to be required to dissociate possibly weak signals from noise. Simulations are performed to evaluate the finite sample properties of the proposed method. In support to our theory, the simulations reveal that, irrespective of the Monte Carlo design considered, mRMSC selects the best model with probability growing to one as the sample size increases. This exercise also highlights the limits of RMSC in settings of identification with mixed strength. Specifically, as the identification weakens, there are many instances where its probability of selecting the best model decreases to 0 with the sample size or plateaus way below 1 showing evidence of its inconsistency. This issue with RMSC is exacerbated when the number of parameters increases. Nevertheless, in standard identification settings, RMSC seems to have a slight advantage over mRMSC as it converges a bit faster. This seems to be the price for the robustness of mRMSC.

For further relation to the literature, it is worth mentioning the quasi-Bayesian model selection method recently proposed by Inoue and Shintani (2018). This method aims to select the most parsimonious model among those with the largest quasi-likelihood. Even though their approach can be adapted to moment selection, our goal differs from theirs as our quest is to find, among the models with maximum information about a parameter of interest, the one with the smallest number of moment restrictions.

The rest of the paper is organized as follows. Section 2 introduces the set up and existing asymptotic results on inference on moment condition models with mixed strength. Section 3 analyzes the performance of RMSC in this setting and reports simulation results exposing some evidence of inconsistency of this method. mRMSC is introduced in Section 4 along with its consistency properties.

Relevant choices of penalty functions are also discussed. Simulations results are reported in Section 5 while Section 6 concludes. Lengthy proofs are relegated to an Appendix.

Throughout the paper, $|a|$ denotes the number of non-zero entries or the determinant of a if a is a vector or a square matrix and $\|a\|$ denotes the Frobenius norm of the matrix a, i.e., $\|a\| = \sqrt{\text{trace}(aa')}$.

2 Setup and some results

Let us consider the sample $\{Y_t : t = 1, \ldots, T\}$ described by the population moment condition

$$E\left(\phi(Y_t, \theta_0)\right) = 0, \tag{1}$$

where $\phi(\cdot, \cdot)$ is a known $\mathbb{R}^k$-valued function, θ_0 is the parameter value of interest which is unknown but lies in Θ a subset of $\mathbb{R}^p$.

The moment condition model (1) is said to globally identify θ_0 if

$$E\left(\phi(Y_t, \theta)\right) = 0, \quad \theta \in \Theta \quad \Leftrightarrow \theta = \theta_0. \tag{2}$$

This property plays an important role in the standard theory of generalized method of moments (GMM) of Hansen (1982) to claim consistency of the GMM estimator. It is also known that moment condition models are not always so strong at identifying the parameter value of interest. In particular, various level of identification strength may be expected from component of the estimating function as stressed by Antoine and Renault (2009, 2012), Caner (2009) and Andrews and Cheng (2012) among others. We refer to Antoine and Renault (2012) for further examples in addition to the classical linear IV model with instrumental variables of mixed strength studied in Sections 3 and 5.

The strong/point identification condition in (2) can be challenged in at least two ways. One may have the configuration where

$$E(\phi(Y_t, \theta)) = 0, \quad \forall \theta \in \Theta$$

reflecting the fact that the moment restrictions are uninformative about the true parameter value θ_0. Another possibility is that, instead of being nil over Θ, $E(\phi(Y_t, \theta))$ is local to 0:

$$E(\phi(Y_t, \theta)) = \frac{\rho(\theta)}{T^\delta}, \quad \rho(\theta) \in \mathbb{R}^k, \quad \delta > 0,$$

with $\rho(\theta_0) = 0$. This configuration fits into the setting of weak or nearly weak identification; see Antoine and Renault (2009).

When $0 < \delta < 1/2$, the moment condition model is referred to as nearly weak and as weak when $\delta = 1/2$. The main difference between these two settings is that consistent estimation is possible in nearly weak models and not in weak models. Of particular interest are configurations where the moment restrictions carry different level of information about the parameters of interest. A leading example of such a case is the linear instrumental variables model where the constant instrument ($z = 1$) is typically strong whereas some or all the other instruments are only weakly correlated with the included endogenous regressors.

Along this line, we consider the estimating function $\phi(\cdot)$ to be partitioned into subvectors with heterogenous strength of identification. Specifically, we assume that:

$$\phi \equiv (\phi_1', \phi_2')' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} : \quad E(\phi_i(Y_t, \theta)) = \frac{\rho_i(\theta)}{T^{\delta_i}}, \quad i = 1, 2, \quad \text{and} \quad 0 \leq \delta_1 \leq \delta_2 < 1/2. \tag{3}$$

In this representation, ϕ_1 has the potential to more strongly identify θ_0 - or some of its components - than ϕ_2. Even though this moment condition model is not informative about θ_0 in the limit if $0 < \delta_1$, Antoine and Renault (2009, 2012) show that consistent estimation is possible under the following mild conditions. The standard identification features of moment condition models pertain to the case $\delta_1 = \delta_2 = 0$.

Assumption 1 *(i)* $\rho \equiv (\rho_1', \rho_2')' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$ *is continuous on the compact parameter set* $\Theta \subset \mathbb{R}^p$ *such that* $\rho(\theta) = 0 \Leftrightarrow \theta = \theta_0$.

(ii) $\sup_{\theta \in \Theta} \sqrt{T} \left\| \bar{\phi}_T(\theta) - E(\phi(Y_t, \theta)) \right\| = O_P(1)$.

Assumption 1(i) imposes global identification of θ_0 by the suitably inflated estimating moment function while part (ii) of the assumption requires that the sample mean of the estimating function accumulates information about its population mean at a fast rate $\sqrt{T}$. Note that this is the standard rate of convergence of sample mean guaranteed by the functional central limit theorem. See Davidson (1994, Theorem 27.14). Under Assumption 1, consistent estimation is possible so long as the rate of accumulation of information outweighs the rate of dilution of information.

Let the GMM estimator $\hat{\theta}_T$ be defined by

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \bar{\phi}_T(\theta)' W_T \bar{\phi}_T(\theta), \tag{4}$$

where W_T is a sequence of almost surely symmetric positive definite matrices converging in probability to W, a symmetric positive definite matrix. Under Assumption 1, Antoine and Renault (2009, 2012) show that

$$\rho(\hat{\theta}_T) = O_P\left(T^{\delta_2 - \frac{1}{2}}\right). \tag{5}$$

Hence, $\delta_2 < 1/2$ is sufficient condition to ensure that $\hat{\theta}_T$ converges in probability to θ_0, especially if we maintain that the parameter set Θ is compact. Note however that $\delta_2 < 1/2$ is not necessary condition for consistency in the sense that a subset of the estimating vector can even be identically 0 and consistent estimation would still be possible. Although, for this, it is important that $\delta_1 < 1/2$ and $\rho_1(\theta) = 0$ is uniquely solved by $\theta = \theta_0$.

Under further regularity conditions, the GMM estimator is asymptotically normally distributed. To introduce these conditions and the main result due to Antoine and Renault (2012), we introduce some notation. Let $s_1 = \text{Rank}\left(\frac{\partial \rho_1}{\partial \theta'}(\theta_0)\right)$ that we assume strictly smaller than p and $R = (R_1 \vdots R_2)$ be a (p,p)-non-singular matrix such that R_1 is (p, s_1)-full column rank matrix and the $s_2 = p - s_1$ columns of R_2 span the null space of $\frac{\partial \rho_1}{\partial \theta'}(\theta_0)$. Define

$$J = \begin{pmatrix} \frac{\partial \rho_1}{\partial \theta'}(\theta_0) R_1 & 0 \\ 0 & \frac{\partial \rho_2}{\partial \theta'}(\theta_0) R_2 \end{pmatrix} \quad \text{and} \quad \Lambda_T = \begin{pmatrix} T^{\frac{1}{2} - \delta_1} I_{s_1} & 0 \\ 0 & T^{\frac{1}{2} - \delta_2} I_{s_2} \end{pmatrix}. \tag{6}$$

The following assumptions are made.

Assumption 2 *(i) θ_0 is interior to Θ and $\phi(Y_t, \theta)$ is continuously differentiable on Θ.*

(ii) $\sqrt{T}\bar{\phi}_T(\theta_0) \xrightarrow{d} N(0, \Sigma)$.

(iii) There exists $C = (C_1' \vdots C_2')'$ a full column rank (k, p)-matrix such that, for $i = 1, 2$,

$$E\left(\frac{\partial \phi_i(Y_t, \theta_0)}{\partial \theta'}\right) = \frac{C_i}{T^{\delta_i}} + o(T^{-\delta_i}), \quad \text{and} \quad \sqrt{T} \sup_{\theta \in \mathcal{N}_{\theta_0}} \left\| \frac{\partial \bar{\phi}_{iT}(\theta)}{\partial \theta'} - E\left(\frac{\partial \phi_i(Y_t, \theta)}{\partial \theta'}\right) \right\| = O_P(1),$$

where $\mathcal{N}_{\theta_0}$ is a neighborhood of θ_0.

Assumption 3 *(i) $\phi_1(Y_t, \theta)$ is linear in θ or $\delta_2 < \frac{1}{4} + \frac{\delta_1}{2}$.*

(ii) $\theta \mapsto \phi(Y_t, \theta)$ is twice continuously differentiable almost everywhere in a neighborhood $\mathcal{N}_{\theta_0}$ of θ_0 and, with $i = 1, 2$, we have

$$\forall k : 1 \leq k \leq k_i, \qquad T^{\delta_i} \frac{\partial^2 \bar{\phi}_{iT,k}}{\partial \theta \partial \theta'}(\theta) \xrightarrow{P} H_{i,k}(\theta),$$

uniformly over $\mathcal{N}_{\theta_0}$, where $H_{i,k}(\theta)$ are (p, p)-matrix functions of θ.

Assumptions 2 and 3 are standard and impose asymptotic normality for the sample mean $\bar{\phi}_T(\theta)$ at $\theta = \theta_0$ as well as regularity conditions on its first and second-order derivatives that are useful for its Taylor series expansions. Although immaterial when ϕ_1 is linear in the parameter, the condition $\delta_2 < \frac{1}{4} + \frac{\delta_1}{2}$ in Assumption 3(i) implies that the Jacobian of the moment function is big enough to ensure that the first-order terms in the expansion of $\bar{\phi}_T(\hat{\theta}_T)$ around θ_0 dominate the higher-order terms. Note also that under some dominance conditions, the matrix C in Assumption 2 is equal to $\partial \rho(\theta_0)/\partial \theta'$. We have the following result.

Theorem 2.1 (Antoine and Renault (2009, 2012).) *If (3) holds along with Assumptions 1, 2, 3 and $0 < s_1 < p$, then*

(i) For any estimator $\tilde{\theta}_T$ of θ_0 such that $\tilde{\theta}_T - \theta_0 = O_P(T^{\delta_2 - \frac{1}{2}})$,

$$\sqrt{T} \frac{\partial \bar{\phi}_T}{\partial \theta'}(\tilde{\theta}_T) R \Lambda_T^{-1} \xrightarrow{P} J, \tag{7}$$

and

(ii)

$$\Lambda_T R^{-1}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, (J'WJ)^{-1} J'W\Sigma WJ(J'WJ)^{-1}), \tag{8}$$

where $\hat{\theta}_T$ is the GMM estimator defined by (4) and $s_1 = Rank(\partial \rho_1(\theta_0)/\partial \theta')$.

Theorem 2.1 effectively provides the asymptotic distribution of $\hat{\eta}_T = R^{-1}\hat{\theta}_T$, a linear function of $\hat{\theta}_T$ with components converging with a specific rate of convergence. In particular, the first s_1 components of $\hat{\eta}_T$ converge at $T^{\frac{1}{2} - \delta_1}$, hence are faster than the remaining $s_2 = p - s_1$ components which converge at rate $T^{\frac{1}{2} - \delta_2}$. In general, since $\hat{\theta}_T$ is typically a linear function of all components of $\hat{\eta}_T$, we expect that the slower rate of convergence would prevail for each component of $\hat{\theta}_T$. More specifically, (ii) implies that $\hat{\theta}_T - \theta_0 = O_P(T^{\delta_2 - \frac{1}{2}})$.

Remark 1 *Note that this result holds in the extreme cases where $s_1 = 0$ and $s_1 = p$. In these cases, $R = I_p$ and for $s_1 = 0$,*

$$J = \left(0 \;\middle|\; \frac{\partial \rho_2'}{\partial \theta}(\theta_0) \right)' \quad \text{and} \quad \Lambda_T = T^{\frac{1}{2}-\delta_2} I_p$$

and for $s_1 = p$,

$$J = \left(\frac{\partial \rho_1}{\partial \theta'}(\theta_0) \;\middle|\; 0 \right)' \quad \text{and} \quad \Lambda_T = T^{\frac{1}{2}-\delta_1} I_p.$$

In the case $s_1 = p$, first-order local identification is ensured by the moment restrictions determined by ϕ_1 which also determine the asymptotic distribution of the GMM estimator. In this case, ϕ_2 appears redundant in the sense that, given ϕ_1, the inclusion of the weaker moment conditions in ϕ_2 does not improve inference about θ_0. In the case however, where $s_1 = 0$, it is ϕ_2 that ensures local identification and ϕ_1 is the irrelevant set of moment restrictions.

It is not hard to see that the asymptotic variance in (8) is smallest for the choice of $W = \Sigma^{-1}$ at which value it is equal to $V_* = \left(J'\Sigma^{-1}J\right)^{-1}$. Dovonon et al. (2019) actually show that V_* stands as the semiparametric efficiency bound for the estimation of $\eta_0 = R^{-1}\theta_0$. The properly scaled two-step efficient GMM estimator using a sequence of weighting matrices W_T (converging in probability to Σ^{-1}) has V_* as asymptotic variance and they further show that this estimator is asymptotically minimax optimal with respect to a large class of loss functions.

Regarding inference about θ_0 within the GMM framework, one may expect, in the light of Theorem 2.1 that knowing s_1, δ_i's, R and also the moment function's partition in (3) is essential. Interestingly however, Antoine and Renault (2009, 2012) have shown that such a knowledge is not required. In particular, inference about θ_0 using the two-step efficient GMM estimator can validly be carried out using the standard formula. Specifically, the standard GMM inference is robust to the sorts of deviations encapsulated in the conditions of Theorem 2.1. (See Antoine and Renault (2009, p.S151).)

This makes relevant the question of moment selection in the context of nearly weak moment restrictions which is the focus of this paper. Below, we first consider the relevant moment selection methodology introduced by Hall et al. (2007) and investigate its performance in the presence of nearly weak moment equalities. We then propose a modified relevant moment selection criterion that robustly select the best model even when this model does not enjoy a strong identification property.

3 Performance of the standard relevant moment selection procedure

This section investigates the performance of RMSC model selection procedure when the best model might not be strongly identifying. This is done through Monte Carlo simulations of instrumental variable (IV) models and we provide some intuition about potential shortcomings that paves the way for a *modified RMSC* selection criterion that we introduce in the next section. Before introducing the simulation setup, let us first introduce the RMSC criterion.

RMSC is a penalized entropy measure that is minimized over candidate models to obtain the most relevant model. Let ϕ denote the estimating function of the moment condition model

$$E(\phi(Y_t, \theta_0)) = 0 \tag{9}$$

which is supposed to have standard identification properties. RMSC uses the entropy of the asymptotic distribution of the efficient estimator $\hat{\theta}_T(\phi)$ of θ_0 in (9) which, up to a constant, is

$$ent_\theta(\phi) \equiv \frac{1}{2}\ln|V(\phi)| = -\frac{1}{2}\ln\left|G(\phi)'\Sigma(\phi)^{-1}G(\phi)\right|,$$

with $V(\phi) = \left(G(\phi)'\Sigma(\phi)^{-1}G(\phi)\right)^{-1}$, where $G(\phi) = E(\partial\phi(Y_t,\theta_0)/\partial\theta')$, $\Sigma(\phi) = \lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_T(\theta_0))$ and $|A|$ stands for the determinant of A if A is a square matrix or the size of A if A is a vector. The sample estimate of $ent_\theta(\phi)$ yields RMSC:

$$RMSC(\phi) = -\ln\left|\hat{G}_T(\phi)'\hat{\Sigma}(\phi)^{-1}\hat{G}_T(\phi)\right| + \kappa(|\phi|,T) = \frac{1}{2}\ln\left|\hat{V}_T(\phi)\right| + \kappa(|\phi|,T), \tag{10}$$

where $\hat{G}_T(\phi)$, $\hat{\Sigma}(\phi)$ and $\hat{V}_T(\phi)$ are consistent estimators of $G(\phi)$, $\Sigma(\phi)$ and $V(\phi)$, respectively and κ the penalty function. Throughout this section, we will consider the BIC-type penalty function:

$$\kappa(k,T) = (k-p)\frac{\ln\sqrt{\tau_T}}{\sqrt{\tau_T}} \tag{11}$$

which has been identified by Hall et al. (2007) the best performing one compared to other alternatives including the Hannan-Quinn penalty. In (11), τ_T represents the rate of convergence of the estimator $\hat{V}_T(\phi)$. In particular,

$$\hat{V}_T(\phi) - V(\phi) = O_P(\tau_T^{-1}).$$

Under some regularity conditions, if the process $\{\phi(Y_t,\theta_0) : t = 1,\ldots,n\}$ is at most finite lag-dependent, $\tau_T = \sqrt{T}$ but if the estimator $\hat{V}_T(\phi)$ involves a kernel estimation of the long run variance, then $\tau_T = \sqrt{T/\ell_T}$ where ℓ_T is the kernel bandwidth. See Andrews (1991).

We now consider the classical linear IV model with possibly nearly weak instrumental variables:

$$Y = X\theta + U \tag{12}$$

$$X = Z\Pi + V, \tag{13}$$

with Y the T-vector of realizations of the dependent variable, X the (T,p)-matrix of p explanatory variables, some of which may be endogenous, Z the (T,k)-matrix of instrumental variables (IVs); U and V are T-vector and (T,p)-matrix of errors, respectively; θ and Π, p-vector and (k,p)-matrix of parameters, respectively.

To allow for variability in the strength of the instruments we set

$$\Pi = \mathbb{L}_T^{-1}C \equiv \begin{pmatrix} T^{-\delta_1}C_1 \\ T^{-\delta_2}C_2 \end{pmatrix}, \quad \text{with} \quad \mathbb{L}_T = \begin{pmatrix} T^{\delta_1}I_{k_1} & 0 \\ 0 & T^{\delta_2}I_{k_2} \end{pmatrix}$$

for some $0 \le \delta_1 \le \delta_2 < 1/2$, and C_i, (k_i,p)-matrix for $i = 1,2$; and $k_1 + k_2 = k$. Partition $Z = [Z_1 \vdots Z_2]$ according to the partition of Π, i.e., Z_i, (T,k_i)-matrix for $i = 1,2$. Thus we can write the system (12)-(13) as:

$$Y = X\theta + U \tag{14}$$

$$X = Z_1\frac{C_1}{T^{\delta_1}} + Z_2\frac{C_2}{T^{\delta_2}} + V. \tag{15}$$

When $\delta_1 = \delta_2$ the instruments in Z_1 and Z_2 have equal strength while those in Z_1 are stronger than those in Z_2 if $\delta_1 < \delta_2$. We maintain the following assumption.

Assumption 4 *(i)* $\{w_t \equiv (Y_t, X_t, Z_t) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^k : t = 1, \ldots, T\}$ *is a sample of independent and identically distributed random vectors with finite second moments.*

(ii) C *is full column rank and*

$$\Delta \equiv \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} = \begin{pmatrix} E(Z_{1t}Z_{1t}') & E(Z_{1t}Z_{2t}') \\ E(Z_{2t}Z_{1t}') & E(Z_{2t}Z_{2t}') \end{pmatrix}$$

is nonsingular.

(iii) $E(Z_tU_t) = 0, \quad E(Z_tV_t) = 0,$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_tU_t \xrightarrow{d} N(0, \sigma_u^2\Delta), \quad \textit{and} \quad \frac{Z'V}{\sqrt{T}} = O_P(1),$$

where $\sigma_u^2 = E(U_t^2)$.

Assumption 4(i) restrict the sample to be independent and identically distributed. While this assumption may look restrictive it is only made for simplification purposes. The main points in this section continue to hold for stationary and ergodic time-dependent data. Assumption 4(ii) is standard. Nonsingularity of Δ imposes no linear duplication of instruments while the rank condition on C amounts to the standard rank condition on $E(Z_tX_t')$. Assumption 4(iii) requires homoskedasticity for U_t and exogeneity for Z_t as well as some limit properties useful to derive the asymptotic distribution of the estimators that we will consider. We do not restrict the correlation between U_t and V_t which is typically different from 0 in presence of endogenous regressors.

The linear IV model (14)-(15) under Assumption 4 implies that the true parameter θ_0 solves

$$E\left(Z_t(Y_t - X_t'\theta)\right) = 0. \tag{16}$$

As shown by Antoine and Renault (2009), this moment restriction fits into the framework introduced in Section 2 if the instruments Z_{1t} and Z_{2t} are orthogonal, that is $\Delta_{12} = 0$. Actually, write

$$E[Z_t(Y_t - X_t'\theta)] = \begin{pmatrix} E[Z_{1t}(Y_t - X_t'\theta)] \\ E[Z_{2t}(Y_t - X_t'\theta)] \end{pmatrix} = \begin{pmatrix} T^{-\delta_1}\rho_1(\theta) + T^{-\delta_2}\nu_1(\theta) \\ T^{-\delta_2}\rho_2(\theta) + T^{-\delta_1}\nu_2(\theta) \end{pmatrix}, \tag{17}$$

$$\tag{18}$$

with

$$\rho_1(\theta) = \Delta_{11}C_1(\theta_0 - \theta), \quad \nu_1(\theta) = \Delta_{12}C_2(\theta_0 - \theta),$$
$$\rho_2(\theta) = \Delta_{22}C_2(\theta_0 - \theta), \quad \nu_2(\theta) = \Delta_{21}C_1(\theta_0 - \theta).$$

If $\Delta_{12} = \Delta_{21}' = 0$, (18) becomes:

$$E[\phi_i(w_t, \theta)] = T^{-\delta_i}\rho_i(\theta),\ t = 1, \ldots, T;\ i = 1, 2, \tag{19}$$

which has the form in (3) with $\phi_i(w_t, \theta) = Z_{it}(Y_t - X_t'\theta)$ and $\rho_i(\theta)$ given in (18); $i = 1, 2$.

The efficient GMM estimator of θ_0 from the moment condition (16) is the two-stage least square estimator:

$$\hat{\theta}_T = \left(X'P_Z X\right)^{-1} X'P_Z Y = \theta_0 + \left(X'P_Z X\right)^{-1} X'P_Z U. \tag{20}$$

where $P_Z = Z(Z'Z)^{-1}Z'$. Its asymptotic distribution can be obtained readily from Theorem 2.1 if the instruments are orthogonal. The following proposition gives this distribution without such a restriction.

To introduce this result, let $s_1 \equiv Rank(C_1)$, and if $0 < s_1 < p$, let $R = (R_1 \vdots R_2)$ be a (p, p)-nonsingular rotation matrix such that $R'R = I_p$ and R_2 a $(p, p - s_1)$-matrix satisfying $C_1 R_2 = 0$.

Proposition 3.1 *Under Assumption 4, the following statements hold.*

(i) If $0 < s_1 < p$,

$$\Lambda_T R'(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V), \quad \textit{with} \quad V = \sigma_u^2 \left[\begin{pmatrix} R_1'C_1' & 0 \\ 0 & R_2'C_2' \end{pmatrix} \Delta \begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix} \right]^{-1}.$$

(ii) If $s_1 = p$,

$$T^{\frac{1}{2}-\delta_1}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V), \quad \textit{with} \quad V = \sigma_u^2 (C_1'\Delta_{11}C_1)^{-1}.$$

(iii) In cases (i) and (ii), the asymptotic variance is consistently estimated by

$$\tilde{V}_T = \hat{\sigma}_u^2 \left(\Lambda_T^{-1} R' X' P_Z X R \Lambda_T^{-1}\right)^{-1}, \ \textit{and} \ \tilde{V}_T = \hat{\sigma}_u^2 \left(T^{2\delta_1 - 1} X' P_Z X\right)^{-1},$$

respectively, with $\hat{\sigma}_u^2 = (Y - X\hat{\theta}_T)'(Y - X\hat{\theta}_T)/T$.

This proposition highlights the expected mixture of rate of convergence of the GMM estimator when instruments have mixed strength. It also shows that if the stronger instruments locally identify the parameter of interest, consistency is achieved at a faster rate and the weaker IVs become irrelevant as they do not affect the asymptotic variance. However, if the stronger set does not identify the true parameter in all directions (this is the case for instance if we have two endogenous variables and only one stronger IV), the weaker set of IVs appears relevant to estimate the remaining directions, albeit at a slower rate of convergence.

The linear IV model offers a suitable framework to investigate the performance of the RMSC procedure in the presence of moment restrictions with nonstandard or mixed strength. We consider the following data generating process.

$$Y = X\theta + U, \quad X = z_1\pi_{1T} + z_2\pi_{2T} + V, \quad \pi_{iT} = \frac{c_i}{T^{\delta_i}}, \quad i = 1, 2.$$

The instruments $z_1, z_2 \in \mathbb{R}^T$ are independent with common distribution $N(0, I_T)$ and are independent of U and V which lie in $\mathbb{R}^T$ with common distribution $N(0, I_T)$ and $Cov(U_t, V_t) = \rho$ for all

$t = 1, \ldots, T$. We consider cases of equal strength for the instruments with $\delta_1 = \delta_2 = 0, 0.2, 0.3, 0.4$ and cases of mixed strength with $(\delta_1, \delta_2) = (0, 0.4), (0.1, 0.4), (0.2, 0.4), (0.3, 0.4)$.

We then consider the case of single endogenous variable and set $\theta_0 = 0.1$ and $c_1 = 1.48$ and $c_2 = 1.48$ and the case of two endogenous variables with $\theta_0 = (0.1, 0.1)'$, $c_1 = (1.48, 0)$ and $c_2 = (0, 1.48)$.

We include four extra instruments, z_3, z_4, z_5, z_6, independent of each other and of z_1, z_2, U and V with common distribution $N(0, I_T)$ and proceed to select the best set of instruments using RMSC. The RMSC of all the 63 (57) combinations of IV has been assessed in the case of the models with 1 (2) endogenous variable(s) and the best model is the one with lowest RMSC. For a given candidate set of k instruments Z, the RMSC is:

$$RMSC = \ln \left| \hat{\sigma}_u^2 \left(\frac{X' P_Z X}{T} \right)^{-1} \right| + (k - p) \frac{\ln \sqrt{T}}{\sqrt{T}}.$$

In the case of one endogenous variable, if $\delta_1 < \delta_2$ only z_1 is relevant while all the other IV are redundant and if $\delta_1 = \delta_2$ both z_1 and z_2 determine the best set of IV while all the others are redundant. In the case of two endogenous variables, z_1 and z_2 constitute the best set of IV regardless of the values of δ_1 and δ_2.

We consider sample sizes T = 100; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. We include such large sample sizes because of possibilities of slow rate of convergence. Figures 1 below plot the proportion of correct model selection (hit rate) by sample size. The number of Monte Carlo replications is 10,000 throughout.

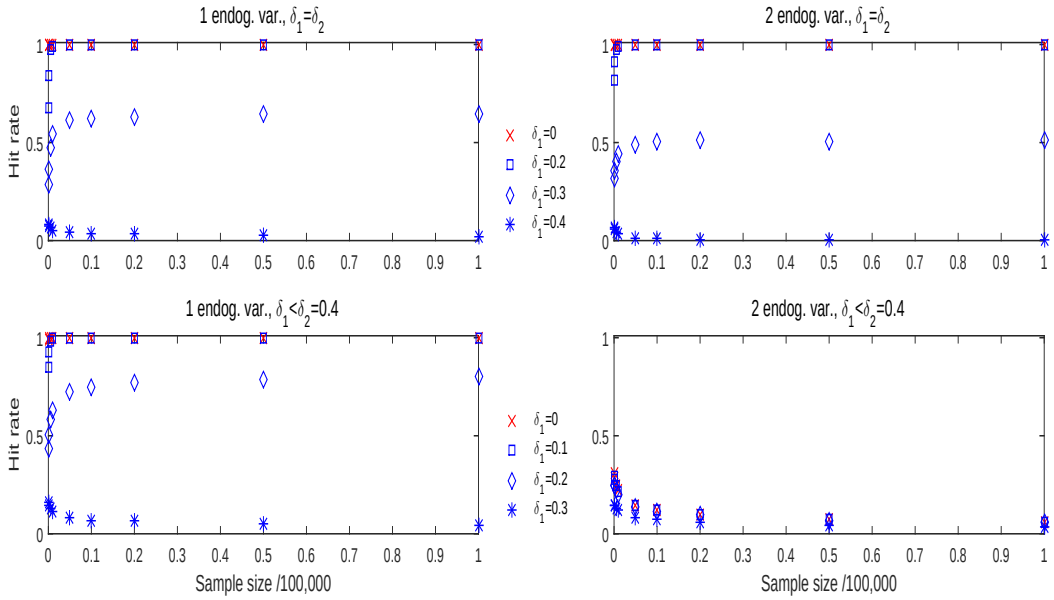
The results suggest that RMSC consistently selects the best model as the sample size increases in cases where the instruments are relatively strong (low δ_i). However, the failure of RMSC is striking in models with moderately large to large values of δ_i. The probability of selecting the best model does not seem to converge to 1 as the sample size grows. Specifically, for cases of $\delta_1 = \delta_2 = 0.3$, the best model is selected about 50% of the times for sample sizes as large as 50,000 or above. The selection procedure also fails to converge for $(\delta_1, \delta_2) = (0.2, 0.4)$ in models with one endogenous variable even though the sole relevant instrument in this configuration seems relatively strong. Also striking is the fact that the hit rate seem to decrease with sample size in many instances of nearly weak instruments. This is the case when $\delta_1, \delta_2 \geq 0.3$. Finally, the case of two endogenous variables and $\delta_1 < \delta_2$ appears to be the most difficult for RMSC to handle since the hit rate drops with the sample size for all combinations of instruments' strength including when a strong IV ($\delta_1 = 0$) is present.

The failure of RMSC can be related to the fact that the information part of the criterion diverges to infinity under nearly weak identification as we can see from Proposition 3.1(iii). This makes the penalty term inappropriate to balance out effectively the noise associated to the selection procedure. Also of importance is the fact the entropy or the asymptotic variance has to be estimated at a rate at least as fast as $\sqrt{T}$ for consistency to be guaranteed. (See Assumption 4 of Hall et al. (2007).) This is not guaranteed at all in this simulation exercise. We are rather certain that the entropy cannot be estimated at such a fast rate and can even have different rate of convergence depending on the set of instruments being assessed.

Accounting for these shortcomings of RMSC, we further analyze its properties in moment condition models with mixed identification strength. We then propose a modified version of this criterion which

robustly and consistently selects the best model regardless of the identification strength.

Figure 1: Proportion of best model selection (Hit rate) by RMSC for models with one and two endogenous variable. Sample size $T = 100$; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. Number of replications: 10,000.



4 A robust relevant moment selection procedure

In this section, we propose a moment selection method to consistently select the smallest (in terms of number of moment restrictions) most relevant model while accounting for the possibility of mixed identification strength of the moment restrictions. We first motivate and introduce a new criterion which is a modified version of RMSC with some robustness properties. We then outline the conditions under which this criterion delivers consistent selection of the best model. The section ends with a discussion on the robustness of the modified RMSC criterion.

4.1 The selection criterion

The problem that we address is one where we have a finite but possibly large number of moment candidate restrictions available to carry out inference about a p-vector parameter θ_0. These restrictions possibly do not have the same identification strength and our goal is to propose a criterion useful to select the best and most relevant moment condition model. As in Hall et al. (2007), we define this model as one from which it is impossible to improve the inference about θ_0 by adding other moment restrictions. Adding to the difficulty of the problem, we do not know what are the strength of the moment restrictions a priori and could not even provide a systematic ranking of them.

To simplify, we assume that the available moment restrictions fit into two categories of strength and that all the candidate models can be expressed as (3) with $0 \leq \delta_1 \leq \delta_2 < 1/2$. As in the previous section, we refer a generic candidate model by ϕ, the vector of the estimating functions that it contains. We shall focus on candidate models ϕ with partition $(\phi_1', \phi_2')'$ satisfying the conditions of Theorem 2.1. Note that ϕ_2 may be empty if all components of ϕ have the same strength. The most restrictive of these assumptions may be Assumption 1(i). However, we will show that the candidate models for which this condition fails are ruled out by the proposed selection procedure and as a result it makes sense to consider that this condition holds without loss of generality.

As established by Dovonon et al. (2019), the efficiency bound on the estimation of θ_0 by ϕ is:

$$V_\theta(\phi) = \left(J(\phi)'\Sigma(\phi)^{-1}J(\phi)\right)^{-1},$$

where

$$J(\phi) = \begin{pmatrix} \frac{\partial \rho_1}{\partial \theta'}(\theta_0)R_1(\phi) & 0 \\ 0 & \frac{\partial \rho_2}{\partial \theta'}(\theta_0)R_2(\phi) \end{pmatrix} \quad \text{and} \quad \Lambda_T(\phi) = \begin{pmatrix} T^{\frac{1}{2}-\delta_1(\phi)}I_{s_1(\phi)} & 0 \\ 0 & T^{\frac{1}{2}-\delta_2(\phi)}I_{s_2(\phi)} \end{pmatrix},$$

$\rho_i(\theta)/T^{\delta_i(\phi)} = E(\phi_i(\theta))$, $(i = 1, 2)$ $\quad s_1(\phi) = Rank\left(\partial \rho_1(\theta_0)/\partial \theta'\right)$, $\quad R(\phi) \equiv \left(R_1(\phi) \vdots R_2(\phi)\right)$ is a (p, p)-rotation matrix satisfying $R(\phi)'R(\phi) = I_p$ and $R_2(\phi)$ is a $(p, s_2(\phi))$-matrix with column vectors in the null space of $\frac{\partial \rho_1}{\partial \theta'}(\theta_0)$. (See (3) for more details.)

This bounds happens to be the asymptotic variance of the efficient GMM estimator

$$\hat{\theta}(\phi) \in \arg\min_{\theta \in \Theta} \bar{\phi}_T(\theta)'\hat{\Sigma}(\phi)^{-1}\bar{\phi}_T(\theta),$$

where $\hat{\Sigma}(\phi)$ is a consistent estimator $\lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_T(\theta_0)) \equiv \Sigma(\phi)$ and as previously, $\bar{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\phi(Y_t, \theta)$.

Recall also from Theorem 2.1 that different candidate models may lead to different rates of convergence of the GMM estimator or equivalently to different rates of accumulation of information. In that respect, letting $\phi^{(j)}$ $(j = 1, 2)$ be two candidate models, $\hat{\theta}_T(\phi^{(1)})$ may converge faster than $\hat{\theta}_T(\phi^{(2)})$ but with a larger information bound. In such a case, it is natural to prefer $\phi^{(1)}$ over $\phi^{(2)}$.

Hence, as a matter of fact, any relevant criterion in the current framework shall account for (i) the amount of information and (ii) the speed of information gathering which should be of first-order importance.

To account for the efficiency bound, we will follow Hall et al. (2007) who consider the entropy of the asymptotic distribution of the efficient GMM estimator. This distribution being Gaussian, the entropy is given by:

$$ent_\theta(\phi) = \frac{1}{2}p(1 + \ln(2\pi)) - \frac{1}{2}\ln\left[\left|J(\phi)'\Sigma(\phi)^{-1}J(\phi)\right|\right].$$

However, the dependence of $J(\phi)$ on the choice of parameter rotation matrix $R(\phi)$ raises the question of invariance of the entropy. The following proposition shows that regardless of the rotation matrix chosen, $ent_\theta(\phi)$ is unchanged. Hence, even though the asymptotic variance may depend on the choice of rotation, the entropy is rotation-invariant.

Proposition 4.1 *Let* $D = \begin{pmatrix} D_1 \\ D_2 \end{pmatrix}$ *be a* (k,p)*-matrix of rank* p *and let* s_1 *denote the rank of* D_1*. Assume that* $0 < s_1 < p$ *and let*

$$\mathcal{R} = \left\{ R = (R_1 \vdots R_2) \in \mathbb{R}^{p \times s_1} \times \mathbb{R}^{p \times p - s_1} : R'R = I_p, \text{ and } D_1 R_2 = 0 \right\}$$

and, for each $R \in \mathcal{R}$*, let*

$$J(R) = \begin{pmatrix} D_1 R_1 & 0 \\ 0 & D_2 R_2 \end{pmatrix}.$$

Then, for any $R, S \in \mathcal{R}$ *and any arbitrary* (k,k)*-matrix* V*, we have:*

$$\left| J(R)' V J(R) \right| = \left| J(S)' V J(S) \right|.$$

Proof. Let $\delta_1, \delta_2 \in \mathbb{R}$ such that $\delta_1 < \delta_2$ and $R \in \mathcal{R}$. Let

$$D_T = \begin{pmatrix} T^{-\delta_1} D_1 \\ T^{-\delta_2} D_2 \end{pmatrix}, \quad \text{and} \quad \ell_T = \begin{pmatrix} T^{\delta_1} I_{s_1} & 0 \\ 0 & T^{\delta_2} I_{p-s_1} \end{pmatrix}.$$

It is not hard to see that the sequence $D_T R \ell_T \to J(R)$ as $T \to \infty$. Hence, by continuity of the determinant function of a matrix, $|\ell_T' R' D_T' V D_T R \ell_T| \to |J(R)' V J(R)|$. Note that $|\ell_T' R' D_T' V D_T R \ell_T| = |\ell_T|^2 \cdot |R|^2 \cdot |D_T' V D_T| = |\ell_T|^2 \cdot |D_T' V D_T|$ and therefore the sequence of determinants does not depend on $R \in \mathcal{R}$. As a consequence, the limit $|J(R)' V J(R)|$ is also unrelated to $R \in \mathcal{R}$ and this concludes the proof. ■

The information measure $ent_\theta(\phi)$ has the following additional properties that are worth highlighting. If two candidate models $\phi^{(1)}$ and $\phi^{(2)}$ are such that $V_\theta(\phi^{(2)}) - V_\theta(\phi^{(1)})$ is nonzero and positive semidefinite, then $ent_\theta(\phi^{(1)}) < ent_\theta(\phi^{(2)})$. This follows readily by using Magnus and Neudecker (2002, Theorem 22). In addition, following the definition of Hall et al. (2007), we say that an estimating function $\phi^{(2)}$ is irrelevant (or redundant) given the estimating function $\phi^{(1)}$ if $V_\theta(\phi) = V_\theta(\phi^{(1)})$, with $\phi = (\phi^{(1)'}, \phi^{(2)'})'$. Hence, by definition, adding irrelevant (or redundant) moment restrictions does not change the level of entropy.

Thanks to these properties, the quest for the optimal model is consistent with the minimization of entropy as one should expect. However, if the limit amount of information about the true parameter value θ_0 plays an important role in the determination of the optimal model, this information is as mentioned only of second-order importance to the rate at which this information is gathered.

The setting of Hall et al. (2007) accounts only for cases where, for the best model, that rate is not heterogenous in the sense that all directions of the parameter space is estimated at the same standard rate $\sqrt{T}$. In this case, the effect of the rate can be ignored in the selection process and, as they point out in their Corollary 1(iii), any model yielding estimators that converge more slowly than the standard rate would have entropy equal to infinity and therefore would not be selected. Our framework departs from theirs by the fact that the best model may actually not only yield estimators converging at slower rate than standard but there are also possibilities of having estimators converging at different rate in various directions.

For our purposes, the rate of convergence needs to be accounted for in the definition of a meaningful selection criterion. A natural summary indicator for the rates of convergence from ϕ is the weighted average of those rates of convergence with weights given by the number of directions in the parameter space that they characterize. That is:

$$a(\phi) \equiv \frac{1}{p}\left(s_1(\phi)\left[\frac{1}{2} - \delta_1(\phi)\right] + s_2(\phi)\left[\frac{1}{2} - \delta_2(\phi)\right]\right).$$

(The rates of convergence are given by the scaling matrix $\Lambda_T(\phi)$ defined above.)

In the context of only two possible rates of convergence - say $\delta_i(\phi) = \delta_i$ $(i = 1, 2)$ for all ϕ - two models $\phi^{(1)}$ and $\phi^{(2)}$ can be compared along the number of fast converging directions that they estimate and the best model would be the one with the largest s_1. Since, in this case $s_2(\phi) = p - s_1(\phi)$, it is not hard, to see that

$$s_1(\phi^{(1)}) \geq s_1(\phi^{(2)}) \Leftrightarrow a(\phi^{(1)}) \geq a(\phi^{(2)}).$$

This further validates the choice of $a(\phi)$ as summary measure of the rates.

Remark 2 *In the occurrence of mixed rate estimation involving more than two directions (see Antoine and Renault (2012)), direct comparison of two models using $a(\cdot)$ becomes problematic as this function no longer provides a natural ordering of the models. Nonetheless, $a(\phi)$ is maximized at $\phi = \phi_{max}$, the largest model available which also yields the best estimation rates. Hence, so long as $a(\phi)$ is the dominant term of the selection criterion, the best model selected shall be one that matches $a(\phi_{max})$ and it can be shown that $a(\phi)$ cannot be maximum without yielding the best estimation rates as well. The intuition is that estimation rates from ϕ_{max} are determined by its strongest elements. As a result, $a(\phi)$ cannot have maximum value if, for instance, the number of fastest estimation directions by ϕ does not match that of ϕ_{max}. One can proceed iteratively to claim that the map of rates for the estimator from ϕ_{max} is the same as that of any ϕ such that $a(\phi) = a(\phi_{max})$.*

These points make $a(\phi)$ a compelling summary of rates of convergence as far as model selection is concerned. As a result, the information-related part of the selection criterion that we shall consider is:

$$\iota_\theta(\phi) = -a(\phi) + \eta_T \cdot ent_\theta(\phi). \tag{21}$$

The sequence η_T depends on the sample size T and shall converge to 0 as T grows to infinity so that the rate component dominates the entropy component as one should expect. Nevertheless, η_T shall not converge too fast as this would destroy the valuable information encapsulated in the entropy function. In fact, $ent_\theta(\phi)$ is the component that ranks candidate models with the same rate component $a(\phi)$. For example, recall that candidates ϕ that estimate the whole parameter vector $\theta_0 \in \mathbb{R}^p$ at rate $\sqrt{T}$ are those with $s_1(\phi) = p$ and $\delta_1(\phi) = 0$. For them, $s_2(\phi) = 0$ and the leading term reaches its minimum value possible. The comparison of such candidate models is solely based on their entropies.

The natural question now is about the sample evaluation of $\iota_\theta(\phi)$. This question is of particular importance since, for a given model ϕ, $s_i(\phi)$ and $\delta_i(\phi)$ $(i = 1, 2)$ are unknown. Interestingly, $\iota_\theta(\phi)$ can

be mimicked by starting off with a naive estimator of the asymptotic variance $V_\theta(\phi)$. Recall that, as claimed by (7), under some regularity conditions,

$$\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}$$

converges in probability to $J(\phi)$. Hence,

$$\hat{V}_\theta(\phi) \equiv \left(\left(\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}\right)' \hat{\Sigma}(\phi)^{-1}\left(\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}\right)\right)^{-1} \tag{22}$$

consistently estimates the asymptotic variance $V_\theta(\phi) = (J(\phi)'\Sigma(\phi)^{-1}J(\phi))^{-1}$. Then, taking the determinant of $\hat{V}_\theta(\phi)$, $ent_\theta(\phi)$ can be estimated by

$$\widehat{ent}_\theta(\phi) = \frac{1}{2}p(1+\ln(2\pi)) - (s_1(\phi)\delta_1(\phi) + s_2(\phi)\delta_2(\phi))\ln T - \frac{1}{2}\ln\left|\frac{\partial \bar{\phi}_T'}{\partial \theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))\right|.$$

The choice of $\eta_T = 1/(p\ln T)$ arises naturally for the definition of $\iota_\theta(\phi)$ which then can be estimated by $\hat{\iota}_\theta(\phi)$ given by:

$$\hat{\iota}_\theta(\phi) \equiv -a(\phi) + \eta_T \cdot \widehat{ent}_\theta(\phi) = -\frac{1}{2} + \frac{1+\ln(2\pi)}{2\ln T} - \frac{1}{2p\ln T}\ln\left|\frac{\partial \bar{\phi}_T'}{\partial \theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))\right|.$$

The information-related part of the selection criterion can therefore effectively be considered as:

$$-\frac{1}{\ln T}\ln\left|\frac{\partial \bar{\phi}_T'}{\partial \theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))\right|.$$

The resulting family of information criterion for model selection that we label *Modified Relevant Moment Selection Criterion* (mRMSC) takes the form:

$$mRMSC(\phi) = -\frac{1}{\ln T}\ln\left|\hat{I}_{\theta,T}(\phi)\right| + \kappa_T, \quad \text{with} \quad \hat{I}_{\theta,T}(\phi) = \frac{\partial \bar{\phi}_T'}{\partial \theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi)), \tag{23}$$

where κ_T is the usual penalty term aiming to filter out noise without impacting consistent selection of the correct model. The choice of κ_T will be discussed in the next section. Despite the similarities, there are some key differences between mRMSC and RMSC. (a) The term appearing in the logarithm is not an estimator of the asymptotic variance of the efficient GMM estimator in general. This is the case only when estimation is done at the standard rate $\sqrt{T}$. (b) The information-related part is scaled down by the inverse of $\ln T$. This makes the rate component useful for moment selection in situations of interest where convergence is slower. Without scaling, this information-related term in mRMSC would explode and standard penalization components would not be as effective at excluding redundant moment restrictions as illustrated in Section 3.

4.2 Consistency

We now show that the proposed criterion leads to consistent selection of the relevant model. We follow Andrews (1999) and Hall et al. (2007) by relying on the following notation. Let $\phi_{max}(\cdot) \in \mathbb{R}^{k_{max}}$ be the vector of all available candidate moment restrictions. Let the selection vector $c \in \mathbb{R}^{k_{max}}$ with entries

0's and 1's denote the components of $\phi_{max}(\cdot)$ included in a particular moment condition model. Any subvector $\phi(\cdot)$ of the set of candidates $\phi_{max}(\cdot)$ is identified by a unique selection vector c with $c_j = 1$ if and only if $\phi(\cdot)$ contains the jth element of $\phi_{max}(\cdot)$. $|c| = c'c$ represents the number of moment restrictions in $\phi(\cdot)$ and write $\phi(\cdot) = \phi_{max}(\cdot, c)$. We, sometimes, omit the subscript 'max' for simplicity of presentation. The set of all possible selection vectors is denoted $\mathscr{C}$ and defined as:

$$\mathscr{C} = \left\{ c = (c_1, \ldots, c_{k_{max}})' \in \mathbb{R}^{k_{max}} : c_j = 0, 1 \text{ for } j = 1, \ldots, k_{max} \text{ and } |c| \geq p \right\}.$$

For notational simplicity, the statistics of interest are now indexed by c and so, $\hat{\theta}_T(c)$ denotes the GMM estimator based on $\phi \equiv \phi_{max}(\cdot, c)$; $V_\theta(c)$ its asymptotic variance and $R(c)$ the rotation matrix in which it is expressed; $\hat{I}_{\theta,T}(c)$ the estimated information matrix (see (23)).

We maintain the following assumption on ϕ_{max}.

Assumption 5 *(i) $\phi_{max}(\cdot)$ satisfies (3) that is: $\phi_{max} \equiv (\phi'_{max,1}, \phi'_{max,2})' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$:*

$$E\left(\phi_{max,i}(Y_t, \theta)\right) = \frac{\rho_{max}(\theta)}{T^{\delta_i}},$$

$i = 1, 2$, $0 \leq \delta_1 \leq \delta_2 < 1/2$ and $\rho_{max}(\cdot)$ is an $\mathbb{R}^{k_{max}}$-valued function defined on the parameter set $\Theta \subset \mathbb{R}^p$.

(ii) $\rho_{max} \equiv (\rho'_{max,1}, \rho'_{max,2})' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$ is continuous on Θ and satisfies over Θ: $[\rho_{max}(\theta) = 0 \Leftrightarrow \theta = \theta_0]$.

(iii) $\sup_{\theta \in \Theta} \sqrt{T} \left\| \bar{\phi}_{max,T}(\theta) - E\left(\phi_{max}(Y_t, \theta)\right) \right\| = O_P(1)$, where $\bar{\phi}_{max,T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{max}(Y_t, \theta)$.

(iv) θ_0 belongs to the interior of Θ and $\theta \mapsto \phi_{max}(Y_t, \theta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}_{\theta_0}$ of θ_0.

(v) $D_{max} = \frac{\partial \rho_{max}}{\partial \theta'}(\theta_0)$ is full column-rank and, letting $D_{max,i} = \frac{\partial \rho_{max,i}}{\partial \theta'}(\theta_0)$ $(i = 1, 2)$, we have $E\left(\frac{\partial \phi_{max,i}(Y_t, \theta_0)}{\partial \theta'}\right) = \frac{D_{max,i}}{T^{\delta_i}} + o(T^{-\delta_i})$ and $\sqrt{T} \sup_{\theta \in \mathcal{N}_{\theta_0}} \left\| \frac{\bar{\phi}_{max,T}(\theta)}{\partial \theta'} - E\left(\frac{\partial \phi_{max}(Y_t, \theta)}{\partial \theta'}\right) \right\| = O_P(1)$, where $\bar{\phi}_{max,i,T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{max,i}(Y_t, \theta)$.

(vi) $\theta \mapsto \phi_{max,1}(Y_t, \theta)$ is either linear or $\delta_2 < \frac{1}{4} + \frac{\delta_1}{2}$.

(vii) For all k: $1 \leq k \leq k_i$ $(i = 1, 2)$,

$$T^{\delta_i} \frac{\partial^2 \bar{\phi}^k_{max,i,T}(\theta)}{\partial \theta \partial \theta'} \xrightarrow{P} H_{max,i,k}(\theta),$$

uniformly over $\mathcal{N}_{\theta_0}$, where $H_{max,i,k}$ is a (p, p)-matrix function of θ and $\bar{\phi}^k_{max,i,T}(\theta)$ is the k-th component of $\bar{\phi}_{max,i,T}(\theta)$.

(viii) $\Sigma(\phi_{max}) = \lim_T Var(\sqrt{T} \bar{\phi}_{max,T}(\theta_0))$ is positive definite.

Assumption 5 is a partial collection of Assumptions 1, 2 and 3 omitting Assumption 2(ii). Note that this latter is useful to establish asymptotic normality of the GMM estimator but not crucial to obtain consistent selection of moments. The parts of Assumptions 1-3 highlighted by Assumption 5 are those

useful to establish consistency of the GMM estimator and the Jacobian matrix of the sample mean of the estimating function.

Since all the components of $\phi_{max}(\cdot)$ are valid estimating functions, inference based on the whole vector $\phi_{max}(\cdot)$ would lead to asymptotic efficiency. However, a plurality of moment restrictions has an adverse consequence of damaging finite sample properties of GMM inference. Simulation cases have been reported by Hall and Peixe (2003) showing the negative effect of redundant moment restrictions on inference. Formal analysis have also been carried out by Newey and Smith (2004) showing that larger moment condition models inflate finite sample bias. In this regard, researchers are motivated to select from $\phi_{max}(\cdot)$, the minimal set of relevant moments that achieves the same asymptotic efficiency as ϕ_{max}. We next introduce a formal definition of relevance that accounts for the possibility of mixed rate of convergence.

Letting c be a selection vector, we write $c = (c_1', c_2')' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$ and let $s_j(c)$ be the rank of the Jacobian matrix of $\rho_{max,j}(c_j)$ at θ_0.

Definition 1 *A subset of moment restriction characterized by $c_r \in \mathscr{C}$ is said to be relevant if the following two properties hold:*

(i) $s_1(c_r)\delta_1 + s_2(c_r)\delta_2 = s_1(\iota_{max})\delta_1 + s_2(\iota_{max})\delta_2$ *and* $V_\theta(\iota_{max}) = V_\theta(c_r)$, *where* ι_{max} *is a* k_{max}*-vector of 1's.*

(ii) For any decomposition $c_r = c_{r,1} + c_{r,2}$ *of* c_r *with* $c_{r,1}, c_{r,2} \in \mathscr{C}$, *either one of the following holds:*

(ii.a) $s_1(c_r)\delta_1 + s_2(c_r)\delta_2 < s_1(c_{r,1})\delta_1 + s_2(c_{r,1})\delta_2$,

(ii.b) $s_1(c_r)\delta_1 + s_2(c_r)\delta_2 = s_1(c_{r,1})\delta_1 + s_2(c_{r,1})\delta_2$ *and* $V_\theta(c_{r,1}) - V_\theta(c_r)$ *is positive semidefinite.*

This definition is of the same flavor as Definition 2 of Hall et al. (2007) while accounting explicitly for the rate of convergence. In particular, asymptotic variances can be compared only when rates of convergence are of the same magnitude. Consistent with our presentation so far, the definition implicitly assumes that the moment function $E(\phi_{max}(Y_t, \theta))$ partitions at most into two components with specific rate of convergence to 0 that are $T^{-\delta_1}$ and $T^{-\delta_2}$, respectively. We can be more general by allowing for more possibilities of rates at the cost of notation burden without any substantial added value.

It is worth mentioning that, because of the dependence of $V_\theta(c)$ on the choice of rotation matrix $R(c)$, the statement $V_\theta(i_{max}) = V_\theta(c_r)$ requires some clarification. We recall that $R(\iota_{max}) \equiv (R_1(\iota_{max}) \vdots R_2(\iota_{max}))$ is such that $R(\iota_{max})R(\iota_{max})' = I_p$ with the columns of $R_2(\iota_{max})$ spanning the null space of $\partial\rho_{max,1}(\theta_0)/\partial\theta'$.

Under the condition: $s_1(c_r)\delta_1 + s_2(c_r)\delta_2 = s_1(\iota_{max})\delta_1 + s_2(\iota_{max})\delta_2$, which is actually equivalent to $s_1(c_r) = s_1(\iota_{max})$ so long as $s_1 + s_2 = p$, Lemma B.1 in Appendix B claims that $R_2(\iota_{max})$ also span the null space of $\partial\rho_{max,1}(\theta_0, c)/\partial\theta'$. Hence the asymptotic distributions of $\hat{\theta}_T(c)$ and $\hat{\theta}_T(\iota_{max})$ can be explored in terms of the same rotation and their asymptotic variances shall be compared under this rotation. $V_\theta(\iota_{max})$ and $V_\theta(c_r)$ in Definition 1(i) are expressed in terms of that common rotation. Similar arguments can be made about the variance comparison in Definition 1(ii.b) as well.

We base the determination of c_r, the selection vector corresponding to the relevant set of moment condition on the *modified relevant moment selection criterion* mRMSC introduced by (23) with a penalization term κ_T, a function of sample size and the size of the estimating function. Note that parsimony is sought relatively to the number of moment restriction and not the number of parameter estimates which is always p. Specifically, we write:

$$mRMSC(c) = -\frac{1}{\ln T}\ln\left|\hat{I}_{\theta,T}(c)\right| + \kappa(|c|, T),$$

where $\hat{I}_{\theta,T}(c)$ is given by (23) with $\phi(\cdot) = \phi_{max}(\cdot, c)$. To estimate c_r, consider the value $\hat{c}_T$ of c minimizing $mRMSC(c)$ over $\mathscr{C}$:

$$\hat{c}_T = arg\min_{c\in\mathscr{C}} mRMSC(c).$$

Our next assumption pertains to the set of selection vectors. Let

$$\mathscr{C}_{\text{eff}} = \{c \in \mathscr{C} : s_1(c)\delta_1 + s_2(c)\delta_2 = s_1(\iota_{max})\delta_1 + s_2(\iota_{max})\delta_2 \wedge V_\theta(c) = V_\theta(\iota_{max})\}$$

and

$$\mathscr{C}_{\min} = \{c \in \mathscr{C}_{\text{eff}} : |c| \leq |\bar{c}| \quad \text{for all } \bar{c} \in \mathscr{C}_{\text{eff}}\}.$$

Assumption 6 *(i) c_r satisfies Definition 1 and $\mathscr{C}_{min} = \{c_r\}$; (ii) $\forall c \in \mathscr{C}$, $\rho_{max}(\theta, c) = 0 \Leftrightarrow \theta = \theta_0$, and $Rank(\partial\rho_{max}(\theta_0, c)/\partial\theta') = p$; (iii) $\hat{\Sigma}(c)$ converges in probability to $\lim_T Var(\sqrt{T}\bar{\phi}_{max,T}(\theta_0, c)) \equiv \Sigma(c)$, positive definite. (iv) $\hat{V}_\theta(c) = V_\theta(c) + O_P(\tau_{T,c}^{-1})$, where $\tau_{T,c} \to \infty$ as $T \to \infty$; (v) $\forall\, c \in \mathscr{C}$ and $\tilde{c}, \bar{c} \in \mathscr{C}$: $|\bar{c}| > |\tilde{c}|$, $\min(\tau_{T,\tilde{c}}, \tau_{T,\bar{c}}) \cdot \ln(T) \cdot (\kappa(|\bar{c}|, T) - \kappa(|\tilde{c}|, T)) \to \infty$ and $\ln T \cdot \kappa(|c|, T) \to 0$ as $T \to \infty$.*

This assumption is similar to Assumption 4 of Hall et al. (2007) that we adapt to our configuration. Part (i) is an identification condition for c_r allowing for its consistent estimation. Parts (iv) and (v) relate the rate of accumulation of information about θ_0 to the penalty term. These conditions allow the selection mechanism to favor, with large probability as the sample size grows, the less sophisticated model of two with comparable levels of information about θ_0. The convergence rate $\tau_{T,c}$ is tagged to the model choice c to stress the dependence of rate of estimation on the model under consideration.

In standard problems, the asymptotic variance is estimated at the rate $\tau_T = \sqrt{T}$ in the presence of cross sectional data whereas for weakly dependent data, this rate is slower ($\tau_T = \sqrt{T/\ell_T}$, where ℓ_T is the bandwidth parameter) due to the use of heteroskedasticity and autocorrelation consistent estimator of the asymptotic variance (see Andrews (1991)). These rates arise when the parameter itself is estimated at the rate $\sqrt{T}$ which is not the case in our setting. Proposition 4.4 derives the order of magnitude of $\hat{V}_\theta(c) - V_\theta(c)$ when the parameter is nearly weakly identified. Typically, $\tau_T = o(\sqrt{T})$ with cross-sectional data and $\tau_T = o(\sqrt{T/\ell_T})$ for weakly dependent data. The choice of penalty terms will be discussed after the following consistency result.

Assumption 6(ii) looks restrictive by imposing that all candidate models extracted from ϕ_{max} must globally identify θ_0 and must also identify θ_0 locally at first-order. This, indeed, needs not be the case. We will show next that candidate models that violate this assumption cannot score mRMSC as low as c_r asymptotically and, as a result, would not be selected.

Theorem 4.2 *If Assumptions 5 and 6 hold, then $\hat{c}_T$ converges in probability to c_r as $T \to \infty$.*

For completeness, we now analyze mRMSC when $\mathscr{C}$ contains candidate models that violate Assumption 6(ii). This is the case when point identification fails or when the Jacobian matrix of the moment function is rank deficient.

For a candidate model c, failure of point identification implies that $\hat{\theta}_T(c)$ is not consistent. If $\rho_{max}(\theta, c) = 0$ is solved by a continuum of values around θ_0, then the Jacobian matrix of the moment function is necessarily rank-deficient at θ_0.

Besides, point identification may hold while the Jacobian matrix is rank deficient at θ_0. In this case, $\hat{\theta}_T(c)$ is consistent but the first-order local approximation of the moment function fails to identify θ_0. Dovonon and Renault (2013, 2019), Dovonon and Hall (2018), Dovonon and Atchadé (2019) and Lee and Liao (2018) among others have studied the behaviour of GMM estimator in this condition. The expected outcome in this setting is that, overall, $\hat{\theta}_T(c)$ converges at a slower rate than $T^{\frac{1}{2}-\delta_2}$.

We shall examine rank deficiency in these two scenarios. Common to both is that $s_i(c)$ directions of the parameter are estimated at the rate $T^{\frac{1}{2}-\delta_i}$ $(i = 1, 2)$ with $s_i(c) = Rank\left(\frac{\partial \rho_{max,i}(\theta_0, c_i)}{\partial \theta'}\right)$ and $s_1(c) + s_2(c) = Rank\left(\frac{\partial \rho_{max}(\theta_0, c)}{\partial \theta'}\right) < p$. The remaining directions are estimated at a slower rate in the latter scenario while inconsistent in the former.

Another possibility is that the moment function is solved at isolated points including θ_0. In this case, we can claim that there is point identification relative to a smaller parameter set around θ_0. Full-rank Jacobian matrix of the moment function at θ_0 then fits into Theorem 4.2 while rank deficient Jacobian matrix at θ_0 fits into the second scenario discussed above. The following result extends Theorem 4.2 and shows that $\hat{c}$ is consistent for c_r even if $\mathscr{C}$ includes candidate models with identification issues.

Assumption 7 *Let $c = (c_1', c_2')' \in \mathbb{R}^{k_2} \times \mathbb{R}^{k_2}$ be a vector of 1's and 0's such that:*
(i-a) $[\rho_{max}(\theta, c) = 0 \Leftrightarrow \theta = \theta_0]$ and $Rank\left(\partial \rho_{max}(\theta_0, c)/\partial \theta'\right) < p$, or (i-b) $\rho_{max}(\theta, c) = 0$ on a continuum set containing θ_0 and, as $T \to \infty$, $\partial \rho_{max}(\hat{\theta}_T(c), c)/\partial \theta'$ converges in probability to M with rank $q < p$. *(ii) For any vector r in the null space of $\partial \rho_{max,1}(\theta_0, c_1)/\partial \theta'$ (in the setting of (i-a)) or the null space of M (in the setting of (i-b)), $[\partial \rho_{max,1}(\hat{\theta}_T(c), c_1)/\partial \theta']r = o_P(T^{\delta_1 - \delta_2})$.* *(iii) $\hat{\Sigma}(c)^{-1} = O_P(1)$.* *(iv) $\sup_{\theta \in \Theta} \left\| \partial \bar{\phi}_T(\theta, c)/\partial \theta' - \mathbb{L}_T^{-1} \partial \rho_{max}(\theta, c)/\partial \theta' \right\| = O_P(T^{-1/2})$, with $\mathbb{L}_T = \begin{pmatrix} T^{\delta_1} I_{k_1(c)} & 0 \\ 0 & T^{\delta_2} I_{k_2(c)} \end{pmatrix}$.*

Under Assumption 7(i-a), θ_0 is consistently estimated by $\hat{\theta}_T(c)$ and $\partial \rho_{max}(\hat{\theta}_T(c), c)/\partial \theta'$ converges in probability to $\partial \rho_{max}(\theta_0, c)/\partial \theta'$. The rank deficiency of the latter implies that of the former in the limit. The second part of Assumption 7(i-b) is not particularly restrictive, even though under its first part, θ_0 is not consistently estimable. Indeed, thanks to Lemma A.4 of Antoine and Renault (2009), $\rho_{max}(\hat{\theta}_T(c), c)$ converges to 0 in probability so that $\hat{\theta}_T(c)$ solves $\rho_{max}(\theta, c) = 0$ in the limit. Under a mere differentiability assumption, the Jacobian matrix of $\rho_{max}(\theta, c)$ at any accumulation point $\theta_* \in N$, the set of solutions of this equation, is rank deficient. Under the first part of Assumption 7(i-b), N is a continuum set and, the fact that $\hat{\theta}_T(c)$ lies on the closure of N in the limit implies that the Jacobian matrix at $\hat{\theta}_T(c)$ is rank deficient in the limit. This provides a motivation to the second part of the assumption. Of course, if $\rho_{max}(\theta, c)$ is linear in θ, the first and second parts of Assumption 7(i-b)

are trivially redundant. Assumption 7(ii) is useful to control the remainder of the expansion of the estimated Jacobian matrix. Note that if $\rho_{max}(\theta, c)$ is linear in θ, $[\partial\rho_{max,1}(\hat{\theta}_T(c), c_1)/\partial\theta']r = O_P(T^{-1/2})$ in both (i-a) and (i-b). Assumption 7(iii) is standard whereas Assumption 7(iv) is guaranteed by the functional central limit theorem.

Theorem 4.3 *Let $c = (c_1', c_2')' \in \mathscr{C}$ satisfying the identification features in Assumption 7. If Assumption 5 holds; c_r satisfies Definition 1; and for $a = c, c_r$, $\ln T \cdot \kappa(|a|, T) = o(1)$, then:*

$$mRMSC(c_r) < mRMSC(c)$$

with probability approaching 1 as $T \to \infty$.

4.3 Choice of penalty function and robustness

The conditions in Assumptions 6(iii) and (iv) are particularly crucial for the consistency of the model selection procedure and provide some guidelines for the choice of penalty function. It appears important to know the rate of convergence of the estimator of asymptotic variance used and then select the penalty function $\kappa(\cdot, T)$ in such a way that Assumptions 6(iv) holds. The following proposition gives the rate of convergence of the asymptotic variance estimator $\hat{V}_\theta(\phi)$ given by (22) for a model candidate ϕ. We consider the case where cross-section independent and identically distributed data are involved and the case of weakly dependent time series data.

In the case of cross-section data, the estimator of the long run variance is the sample variance given by:

$$\hat{\Sigma}_{iid}(\phi) = \frac{1}{T}\sum_{t=1}^{T} \phi_t(\hat{\theta}_T(\phi))\phi_t(\hat{\theta}_T(\phi))'$$

whereas in the case of time series data, one shall rely on $\hat{\Sigma}_{hac}(\phi)$, any heteroskedasticity and autocorrelation consistent estimator of the long run variance. See e.g. Andrews (1991). We let ℓ_T denote the kernel bandwidth of this estimator,

$$c_t(\theta) = Vec\left(\frac{\partial\phi}{\partial\theta'}(\theta)\right)\phi_t(\theta)', \quad \text{and} \quad m_t = \phi_t(\theta_0)\phi_t(\theta_0)',$$

where $Vec(\cdot)$ is the standard matrix vectorization operator. We have the following result.

Proposition 4.4 *Assume that the model ϕ satisfies* (3) *and that Assumptions 1 to 3 hold.*

(i) If $\{Y_t : t = 1, \ldots, T\}$ are independent and identically distributed; $\frac{1}{T}\sum_{t=1}^{T} c_t(\theta)$ converges uniformly to a function $c(\theta)$ in a neighborhood of θ_0; and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(m_t - E(m_t)) = O_P(1)$, then

$$\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P\left(T^{(-\frac{1}{2}-\delta_1+2\delta_2)\vee(\delta_1-\delta_2)}\right).$$

If in addition the model is linear in θ, $\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P(T^{(-\frac{1}{2}+\delta_2)\vee(-\delta_2+\delta_1)})$.

(ii) If $\{Y_t : t = 1, \ldots, T\}$ is a weakly dependent time series process, $\delta_2 < \frac{1}{6}$; $\ell_T \sim T^a$, with $a \in (2\delta_2, \frac{1}{2} - \delta_2)$ such that the condition (ii) of Proposition A.1 in Appendix A is satisfied and, in addition, all the conditions of that proposition hold with $\delta = \delta_2$, then:

$$\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P\left(T^{(\delta_1-\delta_2)\vee(-\frac{1}{2}(1-a))}\right).$$

This proposition shows that the rate of convergence of $\hat{V}_\theta(\phi)$ depends on the identification strength of the model under consideration. In the case of cross-sectional data, the requirement in Assumption 6(iv) translates into:

$$(\ln T)T^{(\frac{1}{2}-\delta_2)\vee(\delta_2-\delta_1)}(\kappa(|\tilde{c}|,T)-\kappa(|\bar{c}|,T))\to\infty,$$

as $T\to\infty$ and $\tilde{c},\bar{c}\in\mathscr{C}$ such that $|\bar{c}|>|\tilde{c}|$. Since δ_1,δ_2 can take any arbitrary value in $[0,1/2)$, the commonly used penalty functions such as the BIC-type information criterion ($\kappa(|c|,T)=(|c|-p)\ln\sqrt{T}/\sqrt{T}$) and the Hannan-Quinn-type of criterion ($\kappa(|c|,T)=(|c|-p)b\ln(\ln\sqrt{T})/\sqrt{T},\ b>2$) would not fulfill this requirement since we can always find some values of δ_1 and δ_2 in $[0,1/2)$ that make these criteria violate the condition.

A natural choice of penalty function to consider is:

$$\kappa(|c|,T)=\frac{h(|c|,p)}{(\ln T)^{1+\alpha}},\quad \text{for some}\quad \alpha>0 \tag{24}$$

and $h(|c|,p)$ a positive and strictly increasing function of $|c|$ for all value of p. Examples of function h include:

$$h(|c|,p)=1-\frac{p}{|c|}\quad\text{and}\quad h(|c|,p)=|c|-p.$$

Thanks to (23), the modified relevant moment selection criterion is given by:

$$mRMSC(\phi)=\frac{1}{\ln T}\ln\left|\left(\frac{\partial\bar{\phi}_T'}{\partial\theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))\right)^{-1}\right|+\frac{h(|c|,p)}{(\ln T)^{1+\alpha}}.$$

Obviously, since T is the same across the models under assessment in selection procedure, we can simply write:

$$mRMSC(\phi)=\ln\left|\left(\frac{\partial\bar{\phi}_T'}{\partial\theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))\right)^{-1}\right|+\frac{h(|c|,p)}{(\ln T)^{\alpha}}, \tag{25}$$

for some $\alpha>0$ and $h(|c|,p)$ is as introduced above.

It is not hard to see that such a penalty function satisfies the requirements in Assumption 6(iv) regardless of the values of δ_1 and δ_2 and therefore leads to consistent selection of the best model. This penalty function also works when the data are time series as this can be seen from the order of magnitude derived in Proposition 4.4(ii) for the asymptotic variance estimator.

While the best choice of α in (24) may be of independent interest that we shall pursue in future work, it is of interest to mention that $\alpha>0$ is important a condition to ensure that the penalty function is smaller than the information component. Also, note that the higher α, the less "bad" models are penalized. Since the mixed identification framework is one where signals are by definition weak, it is even more important to exercise higher penalty on "bad" models to obtain consistent selection. In the simulation results reported in next section, we have set $\alpha=0.1$ and use $h(|c|,p)=1-\frac{p}{|c|}$.

5 Simulations results

In this section, we study the finite sample performance of the proposed selection criterion (mRMSC) through a Monte Carlo experiment. For this purpose, we use the same simulation setup of Section 3,

and present the results for both the efficient GMM (2SLS in this case) and the limited information maximum likelihood (LIML) estimators. Since we focus on the classical linear IV model in this experiment, the inclusion of the results with LIML allows us to explore how our selection criterion behaves with an alternative k-class estimator other than 2SLS. LIML has been known to have a good finite-sample performance in IV regressions when identification is strong (see e.g. Bekker (1994)), and its properties have been contrasted with other k-class estimators under weak instruments, from the viewpoint of both estimation and testing (see e.g. Nelson and Startz (1990), Staiger and Stock (1997), Blomquist and Dahlberg (1999), and Wang and Doko Tchatoka (2018)). For the purpose of comparison results are shown for both mRMSC and RMSC.

Figures 2-5 contain the results. The results for 2SLS estimator are reported in Figures 2 and 3 with one ($p = 1$) and two ($p = 2$) endogenous regressor(s), respectively. The results for LIML are presented in Figures 4-5 and are qualitatively similar to those of 2SLS in most cases presented here. The subfigures in each case show, for a combination of identification strength (i.e. the values of δ_i, $i = 1, 2$), the plots of the proportion of correct model selection (*hit rate*) by sample size. Specifically, the top four subfigures in each case report the results where both instruments z_1 and z_2 have equal identification strength (i.e. $\delta_1 = \delta_2$), while the remaining four subfigures plot the hit rates where z_1 and z_2 have different identification strength ($\delta_1 < \delta_2$). The red curve with pentagram-markers in each subfigure represents the performance of the modified relevant moment selection criterion (mRMSC), while the blue curve with circle-markers represents the RMSC of Hall et al. (2007).

Three main results stand out from this exercise. First, the *hit rate* of mRMSC seems to increase to one as the sample size grows in all simulation designs regardless the estimator used (2SLS or LIML). This confirms the consistency result for mRMSC established by Theorems 4.2 and 4.3. While mRMSC displays evidence of consistency throughout, there are many instances where the *hit rate* of RMSC drops to 0 or plateaus way below 1 highlighting the limitation of this criterion to consistently select the correct model when operating on models with poor identification strength. For $p = 1$, these cases are contained in Subfigures '$\delta_1 = \delta_2 = 0.3$, $\delta_1 = 0.2, \delta_2 = 0.4$' and $\delta_1 = 0.3$, $\delta_2 = 0.4$ of Figures 2 and 4. The lacklustre performance of RMSC is more pronounced for $p = 2$ (two endogenous variables). In this case, RMSC seems to be consistent only when the model is strongly identified or close to being so: '$\delta_1 = \delta_2 = 0$' and '$\delta_1 = \delta_2 = 0.1$.' See Figures 3 and 5.

Second, when θ is strongly identified or close to being so, whether 2SLS or LIML estimator is employed, RMSC performs better than mRMSC for small sample sizes ($T = 100$, 200) when $p = 1$ but this gap vanishes in case of two endogenous variables ($p = 2$). As the sample size grows, the proportion of correct model selection of both mRMSC and RMSC converges quickly to 1. See e.g. the Subfigures '$\delta_1 = \delta_2 = 0$' and '$\delta_1 = \delta_2 = 0.1$' in Figures 2 and 5).

Third, as the identification strength deteriorates, that is $\delta_1 = \delta_2 \geq 0.2$ or $\delta_2 > \delta_1 \geq 0.2$, mRMSC expectedly outperforms RMSC. This dominance of the mRMSC is even more pronounced and systematic in models with 2 endogenous variables ($p = 2$), regardless of the estimator used. In this case, mRMSC largely dominates RMSC.

Overall, this simulation exercise illustrates that our modified relevant moment selection criterion perform well even with moderately large to large values of δ_i ($i = 1, 2$), while the RMSC fails to handle these cases, as per its decreasing *hit rate* as the sample size increases for high values of δ_i.

Moreover, Tables 1-6 in the appendix show in detail the empirical selection probabilities of the models that are aggregated to plot the *hit rates* in Figures 2-5. The tables contain the results of both RMSC and mRMSC for sample sizes $T = 100$; 500; 1,000; 10,000; 50,000; 100,000. The results with one endogenous regressor ($p = 1$) are presented in Tables 1-3, while those with two endogenous regressors ($p = 2$) are shown in Tables 4-6. More specifically, each table indicates, for each sample size and each estimator (2SLS and LIML), the empirical selection rates of all possible models for a given criterion (RMSC or mRMSC) and given values of δ_i ($i = 1, 2$). The results support our previous analysis in Figures 2-5.

Considering first the case with one endogenous regressor (Tables 1-3), we see that when $\delta_1 < \delta_2 = 0.4$ (first part of the tables for each sample size), mRMSC outperforms RMSC even for relatively small sample sizes. For example, when $T = 100$ and '$\delta_1 = 0.2 < \delta_2 = 0.4$', RMSC only selects the relevant model (i.e. columns 'z_1' in Table 1 for $T = 100$) 44% of the time with 2SLS and 59% with LIML, while mRMSC selects this model 68% of the time with 2SLS and 79% with LIML. As the sample size increases to $T = 50{,}000$, these empirical selection probabilities bounced to 77% with 2SLS and 79% with LIML for RMSC, and to 100% with both 2SLS and LIML for our mRMSC. Furthermore, looking at columns 'z_1' in Tables 1-3 for $\delta_1 < \delta_2 = 0.4$ (first part of the tables), it is obvious that the dominance of mRMSC is even more pronounced when '$\delta_1 = 0.3 < \delta_2 = 0.4$' regardless of the estimators or the sample size. The dominance of mRMSC when $\delta_1 < \delta_2 = 0.4$ becomes even more visible as the sample size increases. For example, when '$\delta_1 = 0.3 < \delta_2 = 0.4$' and $T = 100{,}000$ (Table 3), the empirical selection rate of the more relevant model (columns 'z_1' of Table 3 for $T = 100{,}000$) is 4% with 2SLS and 9% LIML for RMSC. Meanwhile, the empirical selection rate of this model is 67% with 2SLS and 69% with LIML for our mRMSC. Clearly, we see that as identification weakens, RMSC has a tendency to often select less relevant models (see e.g. the selection probabilities in columns 'all I' in the tables), while mRMSC still has an overall good performance at selecting the more relevant model. Now, when $\delta_1 = \delta_2$ (second part of the Tables 1-3 for each sample size), both RMSC and mRMSC perform relatively well in selecting the correct model (i.e. column '$z_1 + z_2$' of the tables) even with moderate identification strength ($\delta_1 = \delta_2 \leq 0.2$). As identification deteriorates, the two criteria perform less in selecting the correct model. We note that when $p = 1$ and $\delta_1 = \delta_2$, there are instances where RMSC has a slight edge on mRMSC for sample sizes less than 50,000. However, this edge vanishes when $p = 2$ as discussed in the next paragraph.

Now, let us consider the case with two endogenous regressors (Tables 4-6), where the more relevant model is column '$z_1 + z_2$'. We see that for both '$\delta_1 < \delta_2 = 04$' and '$\delta_1 = \delta_2$', mRMSC outperforms RMSC in all combination of identification strength δ_i ($i = 1, 2$), regardless the sample size or the estimator utilized. As in the case of one endogenous regressor, the dominance of mRMSC increases with the sample size. Again, RMSC shows a tendency to often select less relevant models when identification deteriorates (i.e. high values of δ_i, $i = 1, 2$). In addition, the empirical selection probabilities of the relevant model increase with the sample size for mRMSC for all combinations of identification strength used, while that of RMSC often decrease as the sample size increases for high values of δ_i ($i = 1, 2$). This illustrates why the aggregate *hit rate* of RMSC decreases as the sample size increases for high values of δ_i, as shown in Figures 2-5.

Figure 2: Hit rate of mRMSC and RMSC with 2SLS: model with one endogenous variable ($p = 1$). Sample size $T = 100$; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. Number of replications: 10,000.

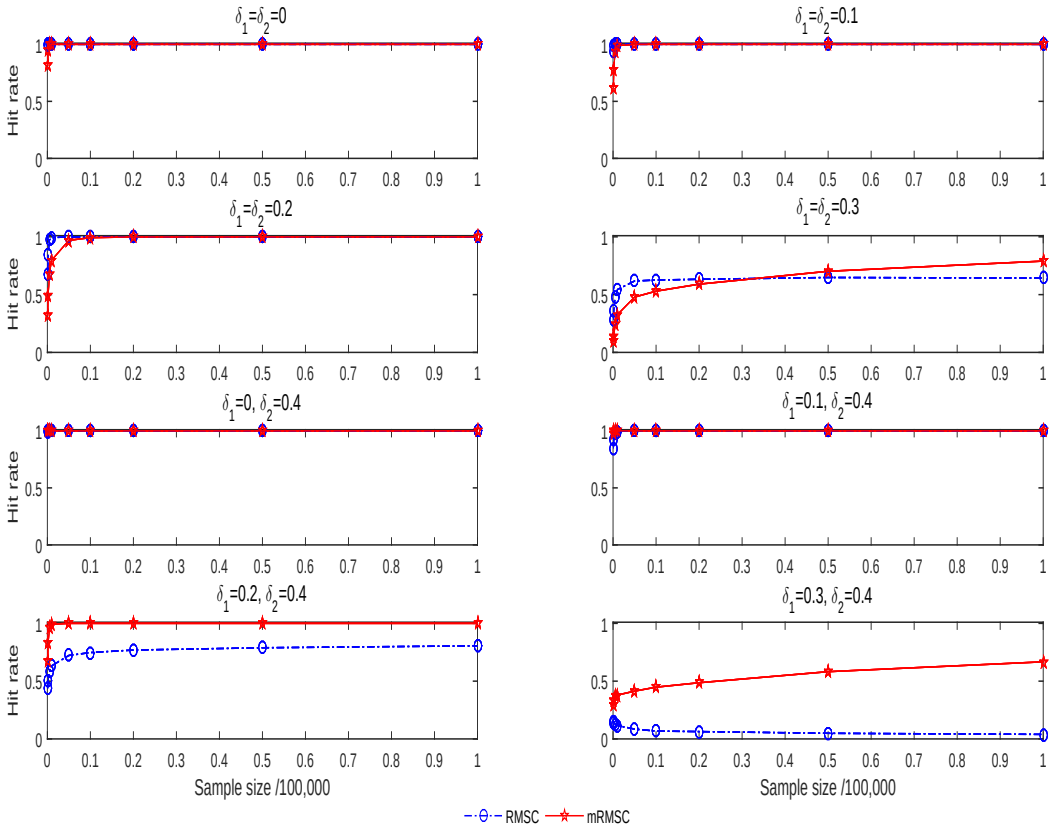


Figure 3: Hit rate by mRMSC and RMSC using 2SLS: model with two endogenous variables ($p = 2$). Sample size $T = 100$; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. Number of replications: 10,000.

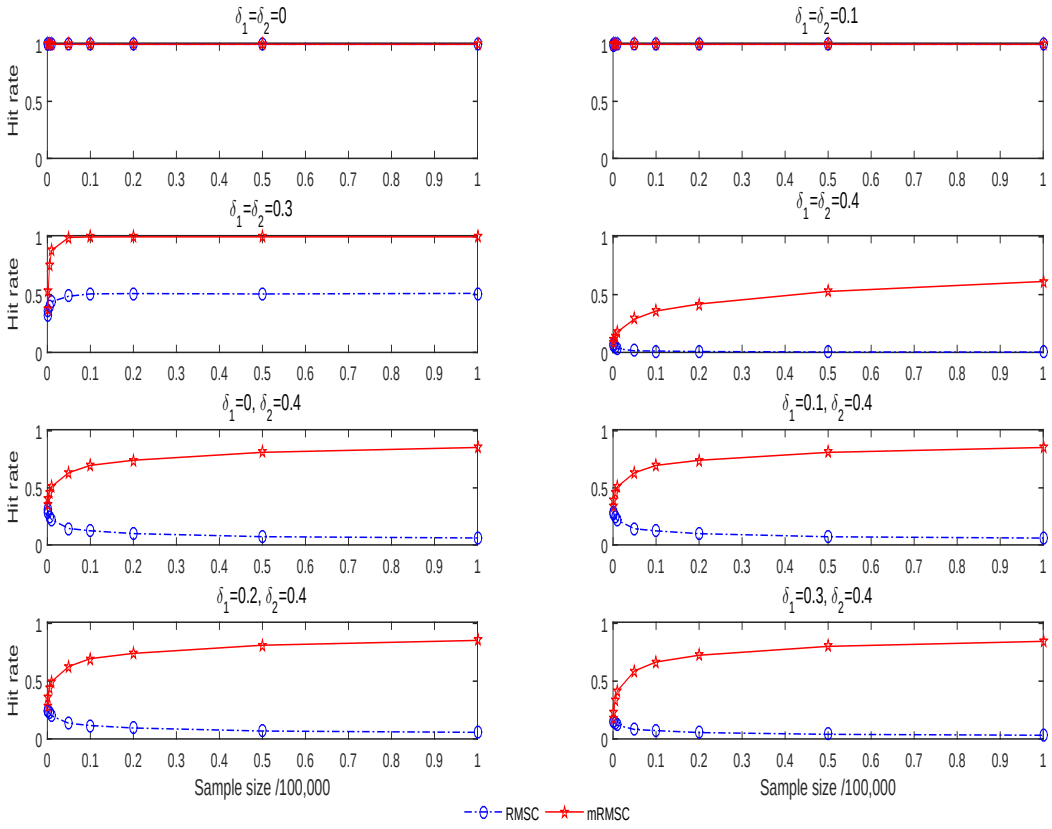


Figure 4: Hit rate of mRMSC and RMSC using LIML: model with one endogenous variable ($p = 1$). Sample size $T = 100$; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. Number of replications: 10,000.

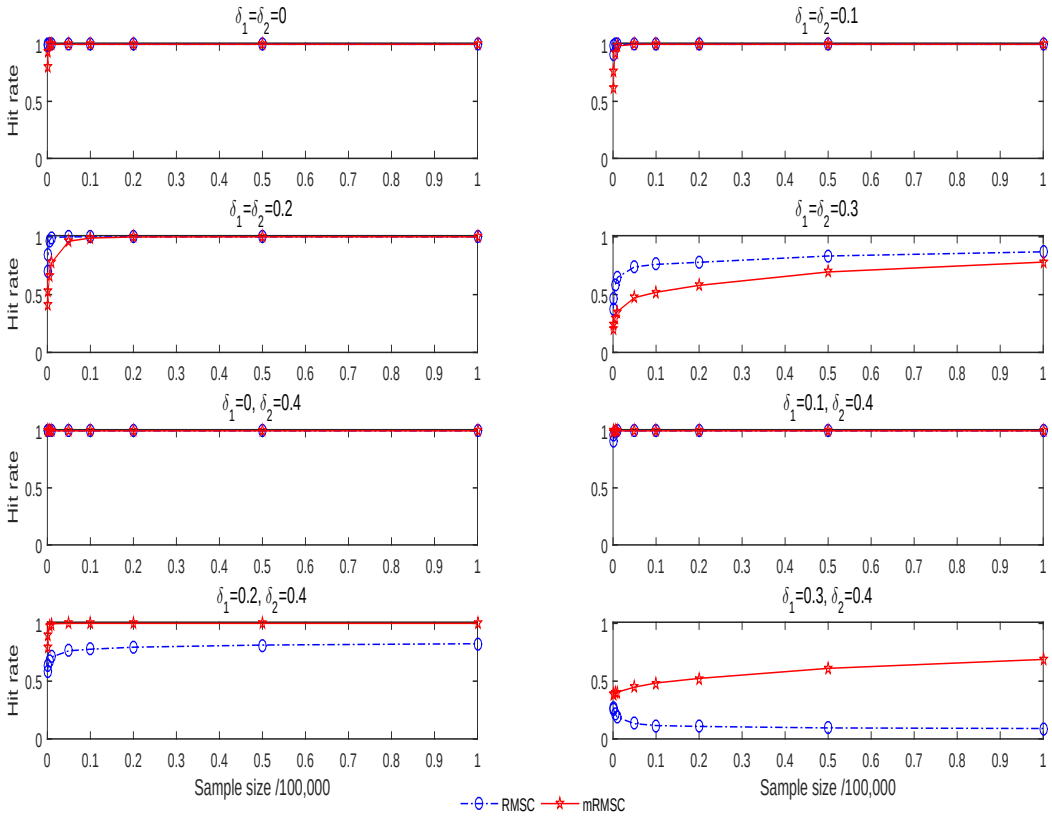
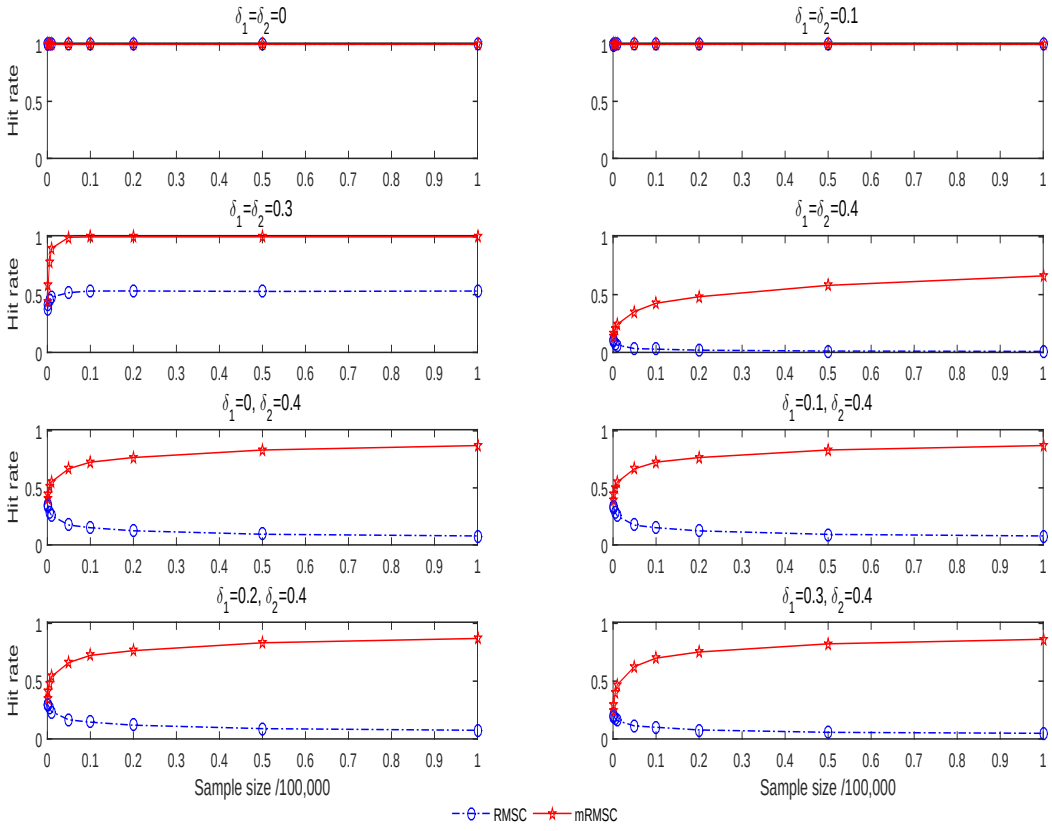


Figure 5: Hit rate of mRMSC and RMSC using LIML: model with two endogenous variables ($p = 2$). Sample size $T = 100$; 200; 500; 1,000; 5,000; 10,000; 20,000; 50,000; 100,000. Number of replications: 10,000.



6 Conclusion

In this paper, we study model selection in moment condition models with mixed identification strength. Our investigation reveals that standard model selection procedures, such as the relevant model selection criterion (RMSC) of Hall et al. (2007), are inconsistent in this setting as they do not explicitly account for the rate of convergence of parameter estimation of candidate models which may vary. We thus propose new entropy-based relevant moment selection criteria and establish their consistency properties in settings that include moment restrictions with mixed strength. The benchmark estimator that we consider is the two-step efficient generalized method of moments (GMM) estimator which is known to be efficient in this framework as well (see Dovonon et al. (2019)). A family of penalization functions is introduced that guarantees the consistency of the selection procedure. We illustrate the finite sample performance of the proposed method through Monte Carlo simulations.

A Convergence rate of HAC using slow parameter estimate

As we have seen, the under mixed strength identifying moment restrictions, the resulting parameter estimator has a slow rate of convergence ($O_P(T^{\frac{1}{2}-\delta})$). Standard theories for HAC estimators of the long-run variance apply to $\sqrt{T}$-consistent parameter estimators. The next proposition gives the rate of convergence of HAC estimators of the long-run variance, Σ, of $\phi(Y_t,\theta_0)$: $\Sigma = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi(Y_t,\theta_0)\right)$ when estimators of θ_0 based on the moment condition $E(\phi(Y_t,\theta_0))=0$ are available and the components of ϕ have mixed identification strength for θ_0. We know in this case that standard estimators $\hat{\theta}_T$ are such that $T^{\frac{1}{2}-\delta}(\hat{\theta}_T-\theta_0)=O_P(1)$ for some $\delta\geq 0$.

Let $\hat{\Sigma}_{hac}$ be the HAC estimator of Σ using the kernel function $k(x)$ and bandwidth parameter ℓ_T. (See Andrews (1991) for more explicit definitions.) We shall assume that $k(\cdot)$ belong to the class $\mathscr{K}_1$, i.e. $k(\cdot)$ is symmetric, continuous at 0 and at all but a finite number of other points, square integrable and takes values in $[-1,1]$, with $k(0)=1$.

Let $k_q=\lim_{x\to 0}\frac{1-k(x)}{|x|^q}$ (see Andrews (1991, p 824)) and $R(j)=E(\phi(Y_t,\theta_0)\phi(Y_{t-j},\theta_0)')$, $j\in\mathbb{Z}$ the autocovariance function of $\phi(Y_t,\theta_0)$ which is assumed to be covariance stationary. We have the following.

Proposition A.1 *Assume that*

(i) Assumptions A, B and C of Andrews (1991) hold with V replaced by ϕ and B(i) replaced by

$$T^{\frac{1}{2}-\delta}(\hat{\theta}_T-\theta_0)=O_P(1)$$

for some $\delta\geq 0$.

(ii) $0\leq\delta\leq\frac{1}{6}$ and there exists $a\in(2\delta,\frac{1}{2}-\delta)$ such that

$$k_q<\infty,\quad \textit{and}\quad \sum_{j=-\infty}^{+\infty}|j|^q\|R(j)\|<\infty,$$

with some $q\geq\frac{1-a}{2a}$.

(iii) $\ell_T\sim T^a$.

Then,

$$\sqrt{\frac{T}{\ell_T}}\left(\hat{\Sigma}_{hac}-\Sigma\right)=O_P(1).$$

Proof of Proposition A.1: Let

$$\Sigma_T(\theta_0)=\sum_{j=-T+1}^{T-1}\left(1-\frac{|j|}{T}\right)R(j),\quad \Sigma=\sum_{j=-\infty}^{+\infty}R(j)\quad\text{with}\quad R(j)=E\left(\phi(Y_t,\theta_0)\phi(Y_{t-j},\theta_0)'\right).$$

Note that Σ is the limit of Σ_T as $T\to\infty$. We have

$$\begin{aligned}\|\Sigma_T(\theta_0)-\Sigma\| &= \left\|\sum_{|j|\geq T}R(j)-\tfrac{1}{T}\sum_{|j|\leq T-1}|j|R(j)\right\|\leq\sum_{|j|\geq T}\|R(j)\|+\tfrac{1}{T}\sum_{|j|\leq T-1}|j|\|R(j)\|\\ &\leq \sum_{|j|\geq T}\|R(j)\|+\tfrac{1}{\sqrt{T}}\sum_{|j|\leq T-1}|j|^{1/2}\|R(j)\|.\end{aligned}$$

Under the condition (ii) of the proposition, $q\geq 1/2$ and as a result, we also have

$$\sum_{j=-\infty}^{+\infty}|j|^{1/2}\|R(j)\|<\infty.$$

Thus, from Lemma 4 of Parzen (1957), we have, as $T \to \infty$,

$$\sqrt{T} \sum_{|j| \geq T} \|R(j)\| \to 0.$$

Hence, $\sqrt{T}\,\|\Sigma_T(\theta_0) - \Sigma\| \leq C$ for some C positive and for T large enough. As a result,

$$\sqrt{\frac{T}{\ell_T}}\,\|\Sigma_T(\theta_0) - \Sigma\| \to 0.$$

Therefore, to complete the proof, it suffices to show that

$$\sqrt{\frac{T}{\ell_T}}\left(\hat{\Sigma}_{hac} - \Sigma_T(\theta_0)\right) = O_P(1). \tag{A.1}$$

This is done by adapting the proof of Andrews (1991, Th. 1(a),(b)) to our setting where $T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0) = O_P(1)$. Following him and without loss of generality, we assume that Σ's are scalars.

Define $\tilde{\Sigma}(\theta)$ similarly to $\hat{\Sigma}_{hac}$ but with $\hat{\theta}_T$ replaced by θ and let $\tilde{\Sigma} \equiv \tilde{\Sigma}(\theta_0)$. By definition, $\hat{\Sigma}_{hac} = \tilde{\Sigma}(\hat{\theta}_T)$. Under our maintained assumptions, the conditions of Andrews (1991, Th. 1(a),(b)) hold and a close consideration of his proof reveals that we only need to show that

$$\sqrt{\frac{T}{\ell_T}}(\hat{\Sigma}_{hac} - \tilde{\Sigma}) = o_P(1)$$

to conclude (A.1). Similar to Andrews (1991, Eq. (A.11)), a two term Taylor expansion gives:

$$\begin{aligned}
\sqrt{\tfrac{T}{\ell_T}}(\hat{\Sigma}_{hac} - \tilde{\Sigma}) &= \left(\tfrac{T^\delta}{\sqrt{\ell_T}}\tfrac{\partial}{\partial\theta'}\tilde{\Sigma}(\theta_0)\right) T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0) + \tfrac{1}{2}T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0)' \left[\tfrac{T^{2\delta-\frac{1}{2}}}{\sqrt{\ell_T}}\tfrac{\partial^2}{\partial\theta\partial\theta'}\tilde{\Sigma}(\bar{\theta})\right] T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0) \\
&\equiv L_{1T}' T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0) + \tfrac{1}{2}T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0)' L_{2T} T^{\frac{1}{2}-\delta}(\hat{\theta}_T - \theta_0),
\end{aligned}$$

where $\bar{\theta} \in (\theta_0, \hat{\theta}_T)$. Similar treatments leading to Andrews (1991, Eq. (A.12)) yield:

$$\begin{aligned}
\|L_{2T}\| &\leq \tfrac{T^{2\delta-\frac{1}{2}}}{\sqrt{\ell_T}} \sum_{j=-T+1}^{T-1} |k(j/\ell_T)| \tfrac{1}{T} \sum_{t=|j|+1}^{T} \sup_{\theta\in\Theta} \left\| \tfrac{\partial^2}{\partial\theta\partial\theta'}\phi(Y_t,\theta)\phi(Y_{t-|j|},\theta) \right\| \\
&= T^{2\delta-\frac{1-a}{2}} \left(\tfrac{1}{\ell_T} \sum_{j=-T+1}^{T-1} |k(j/\ell_T)| \right) O_P(1) = O_P\left(T^{2\delta-\frac{1-a}{2}}\right).
\end{aligned}$$

Also, we have:

$$\begin{aligned}
L_{1T} &= \tfrac{T^\delta}{\sqrt{\ell_T}} \sum_{j=-T+1}^{T-1} k\left(\tfrac{j}{\ell_T}\right) \tfrac{1}{T} \sum_{t=|j|+1}^{T} \phi(Y_t,\theta_0) \left(\tfrac{\partial}{\partial\theta}\phi(Y_{t-|j|},\theta_0) - \lambda\right) \\
&\quad + \tfrac{T^\delta}{\sqrt{\ell_T}} \sum_{j=-T+1}^{T-1} k\left(\tfrac{j}{\ell_T}\right) \tfrac{1}{T} \sum_{t=|j|+1}^{T} \left(\tfrac{\partial}{\partial\theta}\phi(Y_t,\theta_0) - \lambda\right) \phi(Y_{t-|j|},\theta_0) \\
&\quad + T^\delta D_T \lambda,
\end{aligned}$$

with $\lambda = E(\partial/\partial\theta)\phi(Y_t,\theta_0)$ and

$$D_T = \frac{1}{\sqrt{\ell_T}} \sum_{j=-T+1}^{T-1} k\left(\frac{j}{\ell_T}\right) \frac{1}{T} \sum_{t=|j|+1}^{T} (\phi(Y_t,\theta_0) + \phi(Y_{t-|j|},\theta_0)).$$

Clearly, the first two terms in the expansion of L_{1T} are of order $O_P(T^\delta/\sqrt{\ell_T})$. Also, from Andrews (1991, Eq. (A.15)), we can claim that $D_T = O_P(\sqrt{\ell_T/T})$. As a result,

$$L_{1T} = O_P\left(T^{\delta-\frac{a}{2}}\right) + O_P\left(T^{\delta+\frac{a-1}{2}}\right).$$

Since $a \in (2\delta, \frac{1}{2}-\delta)$, $L_{2T} = o_P(1)$. Also, since $\delta < 1/6$, $a < 1-4\delta$ and $L_{1T} = o_P(1)$ and this completes the proof. □

B Auxiliary results and proofs

Lemma B.1 *Let* $s_1(\iota_{max}) = Rank\left(\frac{\partial \rho_{max,1}(\theta_0)}{\partial \theta'}\right)$ *and* $R(\iota_{max}) = \left(R_1(\iota_{max}) \vdots R_2(\iota_{max})\right)$ *such that* $R(\iota_{max})R(\iota_{max})' = I_p$ *and* $\frac{\partial \rho_{max,1}(\theta_0)}{\partial \theta'} R_2(\iota_{max}) = 0$.

Let $c = (c_1', c_2')' \in \mathscr{C}$. *If* $s_1(c) = s_1(\iota_{max})$, *that is* $Rank\left(\frac{\partial \rho_{max,1}(\theta_0)}{\partial \theta'}\right) = Rank\left(\frac{\partial \rho_{max,1}(\theta_0,c)}{\partial \theta'}\right)$, *then* $\frac{\partial \rho_{max,1}(\theta_0,c)}{\partial \theta'} R_2(\iota_{max}) = 0$.

Proof of Lemma B.1. Omitted. □

Proof of Proposition 3.1. We have $\hat{\theta}_T - \theta_0 = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'U$.

(i) Note that

$$X'Z = RR'X'Z = R\left(R'C'\mathbb{L}_T^{-1}Z'Z + R'V'Z\right)$$

and

$$\mathbb{L}_T^{-1}CR = \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix}\ell_T^{-1}, \quad \text{with} \quad \ell_T = \begin{pmatrix} T^{\delta_1}I_{s_1} & 0 \\ 0 & T^{\delta_2}I_{p-s_1} \end{pmatrix}.$$

Hence, $X'Z = R\ell_T^{-1}A_T$, with $A_T = \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix}' Z'Z + \ell_T R'V'Z$ and

$$\sqrt{T}\ell_T^{-1}R'(\hat{\theta}_T - \theta_0) = \left(\frac{A_T(Z'Z)^{-1}A_T'}{T}\right)^{-1} A_T(Z'Z)^{-1}\frac{Z'U}{\sqrt{T}}. \tag{B.1}$$

We have:

$$\begin{aligned}
\tfrac{A_T(Z'Z)^{-1}A_T'}{T} &= \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix}' \tfrac{Z'Z}{T} \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix} \\
&\quad + \tfrac{\ell_T}{\sqrt{T}}R'\tfrac{V'Z}{\sqrt{T}} \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix} \\
&\quad + \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix}' \tfrac{Z'V}{\sqrt{T}}R\tfrac{\ell_T}{\sqrt{T}} + \tfrac{\ell_T}{\sqrt{T}}R'\tfrac{V'Z}{\sqrt{T}}\left(\tfrac{Z'Z}{T}\right)^{-1}\tfrac{Z'V}{\sqrt{T}}R\tfrac{\ell_T}{\sqrt{T}} \\
&= \begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix}' \Delta \begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix} + o_P(1).
\end{aligned}$$

Thus,

$$\left(\frac{A_T(Z'Z)^{-1}A_T'}{T}\right)^{-1} = \left[\begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix}' \Delta \begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix}\right]^{-1} + o_P(1). \tag{B.2}$$

Also,

$$\begin{aligned}
A_T(Z'Z)^{-1}\tfrac{Z'U}{\sqrt{T}} &= \begin{pmatrix} C_1R_1 & 0 \\ C_2R_1T^{\delta_1-\delta_2} & C_2R_2 \end{pmatrix}' \tfrac{Z'U}{\sqrt{T}} + \tfrac{\ell_T}{\sqrt{T}}R'\tfrac{V'Z}{\sqrt{T}}\left(\tfrac{Z'Z}{T}\right)^{-1}\tfrac{Z'U}{\sqrt{T}} \\
&= \begin{pmatrix} C_1R_1 & 0 \\ 0 & C_2R_2 \end{pmatrix}' \tfrac{Z'U}{\sqrt{T}} + o_P(1).
\end{aligned} \tag{B.3}$$

(i) follows from (B.1), (B.2), (B.3) and Assumption 4(iii).

(ii) In the case $s_1 = p$, we use the fact that $T^{\delta_1 - 1} X'Z = \begin{pmatrix} C_1' & 0 \end{pmatrix} \Delta + o_P(1)$. Therefore,

$$X'Z(Z'Z)^{-1}Z'X = T^{1-2\delta_1}\left(C_1'\Delta_{11}C_1 + o_P(1)\right) \tag{B.4}$$

and since C_1 is of rank p, $\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} = T^{-1+2\delta_1}\left[(C_1'\Delta_{11}C_1)^{-1} + o_P(1)\right]$. Also,

$$X'Z(Z'Z)^{-1}Z'U = T^{\frac{1}{2}-\delta_1}\left[\begin{pmatrix} C_1' & 0 \end{pmatrix}\Delta + o_P(1)\right](\Delta^{-1} + o_P(1))\frac{Z'U}{\sqrt{T}}$$

As a result,

$$T^{\frac{1}{2}-\delta_1}(\hat{\theta}_T - \theta_0) = (C_1'\Delta_{11}C_1)^{-1}C_1'\frac{Z_1'U}{\sqrt{T}} + o_P(1)$$

and (ii) follows from Assumption 4(iii).

(iii) $\hat{\sigma}_u^2$ converges in probability to σ_u^2 by the law of large numbers. For the case $0 < s_1 < p$, we have

$$\left(\Lambda_T^{-1}R'X'P_ZXR\Lambda_T^{-1}\right)^{-1} = \left(\frac{\ell_T R'X'P_ZXR\ell_T}{T}\right)^{-1} = \left(\frac{A_T(Z'Z)^{-1}A_T'}{T}\right)^{-1}$$

and using (B.2), we have the expected result. For the case $s_1 = p$, the expected result follows from (B.4). □

Proof of Theorem 4.2: Analogue to previous notation, let

$$\hat{V}_\theta(c) = \left(\left(\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(c))R(c)\Lambda_T(c)^{-1}\right)'\hat{\Sigma}(c)^{-1}\left(\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(c))R(c)\Lambda_T(c)^{-1}\right)\right)^{-1},$$

where we use ϕ in this definition includes only the components of ϕ_{max} selected by c. Under Assumptions 5 and 6(ii), $\|\hat{\theta}_T(c) - \theta_0\| = O_P(T^{-\frac{1}{2}+\delta_2})$. Thanks to Lemma A.5 of Antoine and Renault (2009), we can claim that $\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(c))R(c)\Lambda_T(c)^{-1}$ converges in probability to $J(c)$ and as a result, $\hat{V}_\theta(c)$ converges in probability to $(J(c)'\Sigma(c)^{-1}J(c))^{-1}$.

Note that

$$\hat{V}_\theta(c) = \frac{1}{T}\Lambda_T(c)R(c)'\left(\hat{I}_{\theta,T}\right)^{-1}R(c)\Lambda_T(c).$$

Hence,

$$\ln\left|\hat{V}_\theta(c)\right| = -2(s_1(c)\delta_1 + s_2(c)\delta_2)\ln T - \ln\left|\hat{I}_{\theta,T}\right|.$$

Thus,

$$-\frac{\ln\left|\hat{I}_{\theta,T}\right|}{\ln T} = 2(s_1(c)\delta_1 + s_2(c)\delta_2) + \frac{\ln\left|\hat{V}_\theta(c)\right|}{\ln T} = 2(s_1(c)(\delta_1 - \delta_2) + p\delta_2) + \frac{\ln\left|\hat{V}_\theta(c)\right|}{\ln T}$$

and

$$mRMSC(c) = 2(s_1(c)(\delta_1 - \delta_2) + p\delta_2) + \frac{\ln\left|\hat{V}_\theta(c)\right|}{\ln T} + \kappa(|c|, T). \tag{B.5}$$

Let

$$\Delta_T(c, c_r) = mRMSC(c) - mRMSC(c_r).$$

Thanks to Assumptions 1(i) and (ii), $s_1(c_r) = s_1(\iota_{max})$. This rules out $s_1(c) > s_1(c_r)$ and we shall distinguish the following two cases: (1) $s_1(c) < s_1(c_r)$ and (2) $s_1(c) = s_1(c_r)$.

Case (1): $s_1(c) < s_1(c_r)$. Since $\delta_1 - \delta_2 < 0$, we have: $s_1(c_r)(\delta_1 - \delta_2) + p\delta_2 < s_1(c)(\delta_1 - \delta_2) + p\delta_2$.

Moreover, since $\hat{V}_\theta(c) \xrightarrow{P} V_\theta(c)$, and $\hat{V}_\theta(c_r) \xrightarrow{P} V_\theta(c_r)$ (with both limits finite) and $\kappa(|c|, T) \to 0$ as $T \to \infty$ for all c, we can claim that $\Delta_T(c, c_r) \xrightarrow{P} 2(s_1(c_r) - s_1(c))(\delta_1 - \delta_2) < 0$ meaning that c_r will be chosen over c as

T gets large with probability approaching 1.

Case (2): $s_1(c) = s_1(c_r)$. Lemma B.1 ensures that $V_\theta(c)$, $V_\theta(c_r)$ and $V_\theta(\iota_{max})$ can be expressed in terms of the same rotation matrix $R(\iota_{max})$. By definition, $V_\theta(c_r) = V_\theta(\iota_{max})$ and, considering $V_\theta(c)$ as expressed in terms of $R(\iota_{max})$ as well, standard results of GMM theory ensure that we either have $V_\theta(c) = V_\theta(c_r)$ or $V_\theta(c) - V_\theta(c_r)$ is positive semi definite. We further consider these two cases.

Case (2-i): $V_\theta(c) = V_\theta(c_r)$. We have

$$\begin{aligned} & \min(\tau_{T,c}, \tau_{T,c_r}) \ln(T) \Delta_T(c, c_r) \\ = \; & \min(\tau_{T,c}, \tau_{T,c_r}) \left(\ln |\hat{V}_\theta(c)| - \ln |V_\theta(c)| \right) - \min(\tau_{T,c}, \tau_{T,c_r}) \left(\ln |\hat{V}_\theta(c_r)| - \ln |V_\theta(c_r)| \right) \\ & + \min(\tau_{T,c}, \tau_{T,c_r}) \ln(T) \left(\kappa(|c|, T) - \kappa(|c_r|, T) \right) \\ = \; & O_P(1) + \min(\tau_{T,c}, \tau_{T,c_r}) \ln(T) \left(\kappa(|c|, T) - \kappa(|c_r|, T) \right). \end{aligned}$$

Thanks to Assumption 6(iv), this quantity tends to $+\infty$ with probability 1 as T grows and we can deduce that $\Delta_T(c, c_r)$ is positive with probability 1 as T grows. This means that c_r is eventually selected over c.

Case (2-ii): $V_\theta(c) - V_\theta(c_r)$ is positive semi definite and different from 0. From Theorem 22 of Magnus and Neudecker (1999), $|V_\theta(c)| > |V_\theta(c_r)|$ and we have

$$\ln(T) \Delta_T(c, c_r) = \ln |\hat{V}_\theta(c)| - \ln |\hat{V}_\theta(c_r)| + \ln(T) \left(\kappa(|c|, T) - \kappa(|c_r|, T) \right) = \ln |V_\theta(c)| - \ln |V_\theta(c_r)| + o_P(1).$$

Therefore, $\Delta_T(c, c_r)$ is positive with probability 1 as T grows.

Taken together, Cases (1), (2-i) and (2-ii) establish that $\hat{c} \xrightarrow{P} c_r$ as $T \to \infty$. $\square$

Proof of Theorem 4.3: We have that

$$\frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)}{\partial \theta'} = \left(\frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)}{\partial \theta'} - \mathbb{L}_T^{-1} \frac{\partial \rho_{max}(\hat{\theta}_T(c), c)}{\partial \theta'} \right) + \mathbb{L}_T^{-1} \left(\frac{\partial \rho_{max}(\hat{\theta}_T(c), c)}{\partial \theta'} - M \right) + \mathbb{L}_T^{-1} M,$$

where M stands for either $\frac{\partial \rho_{max}(\theta_0, c)}{\partial \theta'}$ or M in Assumptions 7(i-a) and (i-b).

Let M_1 be the submatrix of M given by its first k_1 rows and M_2 the submatrix of M given by its last k_1 rows. Let $s_1(c) = Rank(M_1)$ and $R = \left(R_1 \vdots R_2 \vdots R_3 \right)$ the orthogonal matrix (i.e. $RR' = I_p$) such that $M_1 R_2 = 0$ and $M R_3 = 0$. Note R_1 is void if $s_1(c) = 0$ and R_2 is void if $s_2(c) = 0$ while R_3 has $p - q > 0$ columns corresponding to an orthogonal basis of the null space of M. $s_1(c) + s_2(c) = q$.

Let $\ell_T = \begin{pmatrix} T^{\delta_1} I_{s_1(c)} & 0 & 0 \\ 0 & T^{\delta_2} I_{s_2(c)} & 0 \\ 0 & 0 & T^{\delta_2} I_{p-q} \end{pmatrix}$. We have:

$$\frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)}{\partial \theta'} R \ell_T = \mathbb{L}_T^{-1} M R \ell_T + o_P(1) = \begin{pmatrix} M_1 R_1 & 0 & 0 \\ M_2 R_1 T^{\delta_1 - \delta_2} & M_2 R_2 & 0 \end{pmatrix} + o_P(1) = \begin{pmatrix} M_1 R_1 & 0 & 0 \\ 0 & M_2 R_2 & 0 \end{pmatrix} + o_P(1).$$

We have

$$mRMSC(c) = -\frac{\ln \left| \hat{I}_{\theta,T}(c) \right|}{\ln T} + \kappa(|c|, T),$$

with $\hat{I}_{\theta,T}(c) = \frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)'}{\partial \theta} \hat{\Sigma}(c)^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)}{\partial \theta'}$. We can write:

$$\hat{I}_{\theta,T}(c) = R \ell_T^{-1} \left(\ell_T R' \frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)'}{\partial \theta} \hat{\Sigma}(c)^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T(c), c)}{\partial \theta'} R \ell_T \right) \ell_T^{-1} R' \equiv R \ell_T^{-1} \hat{K}_{\theta,T} \ell_T^{-1} R'.$$

Thus,

$$\ln\left|\hat{I}_{\theta,T}(c)\right| = 2\ln\left|\ell_T^{-1}\right| + \ln|\hat{K}_{\theta,T}|$$

so that

$$mRMSC(c) = 2[s_1(c)(\delta_1-\delta_2)+p\delta_2] - \frac{\ln|\hat{K}_{\theta,T}|}{\ln T} + \kappa(|c|,T).$$

Using (B.5), we have

$$mRMSC(c_r) = 2(s_1(c_r)(\delta_1-\delta_2)+p\delta_2) + \frac{\ln|V_\theta(c_r)+o_P(1)|}{\ln T} + \kappa(|c_r|,T).$$

Note that $\ln T\cdot\kappa(|a|,T)\to 0$ as $T\to\infty$ for $a=c,c_r$. Also, since $V_\theta(c_r)$ is positive definite, $\ln|V_\theta(c_r)|\in\mathbb{R}$ whereas $-\ln|\hat{K}_{\theta,T}|\to+\infty$, since $\hat{K}_{\theta,T}$ is asymptotically degenerate. Furthermore, by definition of c_r,

$$s_1(c_r)(\delta_1-\delta_2)+p\delta_2 = s_1(\iota_{max})(\delta_1-\delta_2)+p\delta_2$$

and we have

$$[s_1(\iota_{max})(\delta_1-\delta_2)+p\delta_2] - [s_1(c)(\delta_1-\delta_2)+p\delta_2] = (s_1(\iota_{max})-s_1(c))(\delta_1-\delta_2)\le 0$$

since $s_1(\iota_{max})\ge s_1(c)$ and $p>q$. We can therefore conclude that $mRMSC(c_r)<mRMSC(c)$ with probability approaching 1 as $T\to\infty$. □

Proof of Proposition 4.4: Under Assumptions 1 and 2(ii), $\hat{\theta}_T(\phi)-\theta_0 = O_P(T^{-\frac{1}{2}+\delta_2})$. By a mean-value expansion, we have:

$$\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi)) = \frac{\partial\bar{\phi}_T}{\partial\theta'}(\theta_0) + \frac{\partial^2\bar{\phi}_T(\ddot{\theta})}{Vec(\partial\theta\partial\theta')'}[I_p\otimes(\hat{\theta}_T(\phi)-\theta_0)],$$

where $\ddot{\theta}\in(\hat{\theta}_T(\phi),\theta_0)$ and may vary with the entries of $\frac{\partial\bar{\phi}_T}{\partial\theta'}(\theta)$, '$\otimes$' is the Kronecker product and $Vec(A)$ transforms the matrix A into a vector by stacking its columns. By post-multiplying this equality by $\sqrt{T}R(\phi)\Lambda_T(\phi)^{-1}$, we have:

$$\begin{aligned}\sqrt{T}\tfrac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1} &= \sqrt{T}\left(\tfrac{\partial\bar{\phi}_T}{\partial\theta'}(\theta_0) - E\left(\tfrac{\partial\bar{\phi}_T}{\partial\theta'}(\theta_0)\right)\right)R(\phi)\Lambda_T(\phi)^{-1} + \sqrt{T}E\left(\tfrac{\partial\phi_t}{\partial\theta'}(\theta_0)\right)R(\phi)\Lambda_T(\phi)^{-1}\\ &\quad +\sqrt{T}\tfrac{\partial^2\bar{\phi}_T(\ddot{\theta})}{Vec(\partial\theta\partial\theta')'}[I_p\otimes(\hat{\theta}_T(\phi)-\theta_0)]R(\phi)\Lambda_T(\phi)^{-1}\\ &\equiv (1)+(2)+(3).\end{aligned}$$

By Assumption 2(ii), $(1) = O_P(1)O_P(T^{-\frac{1}{2}+\delta_2}) = O_P(T^{-\frac{1}{2}+\delta_2})$. By Assumption 3,

$$(3) = \sqrt{T}O_P(T^{-\delta_1})O_P(T^{-\frac{1}{2}+\delta_2})O_P(T^{-\frac{1}{2}+\delta_2}) = O_P(T^{-\frac{1}{2}-\delta_1+2\delta_2}).$$

Besides,

$$(2) = \begin{pmatrix} \frac{\partial\rho_1(\theta_0)}{\partial\theta'}R_1(\phi) & 0\\ \\ \frac{1}{T^{\delta_2-\delta_1}}\frac{\partial\rho_2(\theta_0)}{\partial\theta'}R_1(\phi) & \frac{\partial\rho_2(\theta_0)}{\partial\theta'}R_2(\phi)\end{pmatrix} = J(\phi)+O(T^{-\delta_2+\delta_1}),$$

where we consider the usual partition of the moment restriction, i.e.

$$E(\phi_j(\theta)) = \frac{\rho_j(\theta)}{T^{\delta_j}}\ (j=1,2),\quad \text{and}\quad R(\phi) = (R_1(\phi)\vdots R_2(\phi)),$$

with the columns of $R_2(\phi)$ spanning the null space of $\frac{\partial\rho_1}{\partial\theta'}(\theta_0)$. As a result,

$$\begin{aligned}\sqrt{T}\tfrac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1} &= J + O_P(T^{-\delta_2+\delta_1}) + O_P(T^{-\frac{1}{2}+\delta_2}) + O_P(T^{-\frac{1}{2}-\delta_1+2\delta_2})\\ &= J + O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}-\delta_1+2\delta_2)}).\end{aligned} \tag{B.6}$$

If the model is linear in θ, (3) is not involved and

$$\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1} = J + O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}-\delta_2)}).$$

To complete the proof for (i), we derive the asymptotic order of magnitude of $\hat{\Sigma}_{iid}(\phi) - \Sigma(\phi)$, where $\Sigma(\phi) = E(\phi_t(\theta_0)\phi_t(\theta_0)')$. By a mean-value expansion, we have

$$\begin{aligned} Vec\left(\tfrac{1}{T}\textstyle\sum_{t=1}^T \phi_t(\hat{\theta}_T(\phi))\phi_t(\hat{\theta}_T(\phi))'\right) &= Vec(\Sigma(\phi)) + Vec\left(\tfrac{1}{T}\textstyle\sum_{t=1}^T \phi_t(\theta_0)\phi_t(\theta_0)' - \Sigma(\phi)\right) \\ &\quad + \tfrac{1}{T}\textstyle\sum_{t=1}^T \frac{\partial}{\partial\theta'}Vec[\phi_t(\theta)\phi_t(\theta)']\Big|_{\theta=\ddot{\theta}}(\hat{\theta}_T(\phi) - \theta_0) \\ &= Vec(\Sigma(\phi)) + O_P(T^{-\frac{1}{2}+\delta_2}) = \Sigma(\phi) + O_P(T^{-\frac{1}{2}-\delta_1+2\delta_2}), \end{aligned}$$

where the last equality follows from the fact that $-\frac{1}{2}+\delta_2 \leq -\frac{1}{2}-\delta_1+2\delta_2$. Since $\hat{V}_\theta(\phi)$ is a smooth function of $\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}$ and $\hat{\Sigma}_{iid}(\phi)$, the claimed result follows by the delta method. If the model is linear in θ, we would have $\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}+\delta_2)})$.

To complete the proof for (ii), we rely on Proposition A.1 to obtain the asymptotic order of magnitude of $\hat{\Sigma}_{hac}(\phi) - \Sigma(\phi)$, where Σ is the long run variance of $\phi_t(\theta_0)$. Under the conditions in (ii), we can claim applying Proposition A.1 that

$$\hat{\Sigma}_{hac}(\phi) - \Sigma(\phi) = O_P(T^{-\frac{1}{2}+\frac{\alpha}{2}}).$$

Again, by the delta method, we can claim using (B.6) that

$$\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}-\delta_1+2\delta_2)\vee(-\frac{1}{2}(1-\alpha))}) = O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}(1-\alpha))}). \tag{B.7}$$

If the model is linear in θ, we have

$$\hat{V}_\theta(\phi) - V_\theta(\phi) = O_P(T^{(-\delta_2+\delta_1)\vee(-\frac{1}{2}+\delta_2)\vee(-\frac{1}{2}(1-\alpha))})$$

which, since $2\delta_2 < \alpha$, also implies (B.7). □

C Additional Monte Carlo results

Table 1: Empirical selection probabilities: one endogenous regressor ($p = 1$), $T = 100$; 500

									$T = 100$													
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	0.99	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	0.85	0.00	0.11	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.44	0.02	0.23	0.17	0.04	0.03	0.00	0.03	0.04	0.00	0.59	0.04	0.27	0.01	0.04	0.00	0.00	0.05	0.01	0.00
	0.3	0.4	0.16	0.05	0.16	0.15	0.12	0.03	0.00	0.23	0.10	0.00	0.27	0.10	0.23	0.02	0.05	0.00	0.00	0.33	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	0.97	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.68	0.05	0.06	0.02	0.01	0.03	0.03	0.03	0.07	0.01	0.79	0.07	0.05	0.00	0.01	0.00	0.00	0.05	0.02	0.00
	0.3	0.4	0.29	0.10	0.04	0.03	0.03	0.03	0.04	0.23	0.19	0.02	0.39	0.16	0.09	0.00	0.01	0.01	0.00	0.33	0.01	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.03	0.03	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.11	0.11	0.67	0.03	0.03	0.04	0.00	0.00	0.01	0.00	0.14	0.14	0.70	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.11	0.11	0.29	0.13	0.13	0.06	0.00	0.12	0.07	0.00	0.19	0.18	0.38	0.03	0.03	0.00	0.00	0.18	0.00	0.00
	0.4	0.4	0.06	0.06	0.08	0.10	0.11	0.02	0.00	0.47	0.10	0.00	0.12	0.13	0.13	0.03	0.03	0.00	0.00	0.57	0.00	0.00
mRMSC	0	0	0.09	0.09	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.09	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.19	0.18	0.62	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.20	0.19	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.26	0.26	0.33	0.00	0.00	0.07	0.04	0.00	0.03	0.01	0.28	0.29	0.42	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	0.3	0.3	0.22	0.22	0.10	0.02	0.03	0.06	0.06	0.11	0.15	0.02	0.29	0.30	0.20	0.00	0.00	0.01	0.00	0.19	0.01	0.00
	0.4	0.4	0.12	0.12	0.02	0.02	0.03	0.02	0.02	0.46	0.18	0.01	0.18	0.18	0.06	0.00	0.00	0.01	0.00	0.56	0.00	0.00
									$T = 500$													
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	0.98	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.58	0.00	0.28	0.09	0.00	0.03	0.00	0.01	0.01	0.00	0.68	0.00	0.31	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	0.3	0.4	0.13	0.02	0.21	0.09	0.04	0.06	0.00	0.43	0.04	0.00	0.22	0.05	0.28	0.00	0.01	0.00	0.00	0.44	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.96	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.98	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	0.3	0.4	0.37	0.07	0.07	0.00	0.01	0.02	0.01	0.42	0.02	0.00	0.40	0.10	0.08	0.00	0.00	0.00	0.00	0.42	0.00	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.01	0.01	0.97	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.02	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.05	0.05	0.47	0.05	0.05	0.12	0.01	0.17	0.02	0.00	0.10	0.11	0.58	0.01	0.01	0.00	0.00	0.19	0.00	0.00
	0.4	0.4	0.02	0.03	0.06	0.03	0.03	0.02	0.00	0.78	0.02	0.00	0.06	0.07	0.10	0.00	0.00	0.00	0.00	0.77	0.00	0.00
mRMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.03	0.03	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.15	0.16	0.68	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.16	0.17	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.23	0.23	0.24	0.00	0.00	0.05	0.03	0.19	0.02	0.00	0.25	0.25	0.30	0.00	0.00	0.00	0.00	0.20	0.00	0.00
	0.4	0.4	0.08	0.09	0.03	0.00	0.00	0.01	0.00	0.78	0.01	0.00	0.10	0.10	0.04	0.00	0.00	0.00	0.00	0.76	0.00	0.00

Note: 'z1 + z2' denotes models with the 2 instruments z1 and z2; 'zj+I' ($j = 1, 2$) denotes models with zj + 1 irrelevant (i.e. completely unrelated) instrument; 'z1 + z2 + I' denotes models with the 2 instruments z1 and z2 + 1 irrelevant instrument; 'z1 + z2 + 2I' denotes models with the 2 instruments z1 and z2 + 2 irrelevant instruments; 'all I' denotes models with irrelevant instruments only; 'zj + more I' denotes models with zj ($j = 1, 2$) + more than 1 irrelevant instrument; 'All' denotes model with all instruments. The highlighted columns correspond to the best subset of instruments. This subset depends on the combination of strengths (δ_1, δ_2) and the number p of estimated parameters.

Table 2: Empirical selection probabilities: one endogenous regressor ($p = 1$), T = 1,000; 10,000

			T = 1,000																			
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.63	0.00	0.28	0.06	0.00	0.02	0.00	0.00	0.00	0.00	0.71	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.12	0.01	0.23	0.06	0.02	0.06	0.00	0.50	0.02	0.00	0.19	0.03	0.29	0.00	0.01	0.00	0.00	0.49	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.38	0.05	0.08	0.00	0.00	0.01	0.00	0.47	0.01	0.00	0.41	0.07	0.07	0.00	0.00	0.00	0.00	0.46	0.00	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.03	0.03	0.54	0.03	0.03	0.13	0.01	0.21	0.01	0.00	0.07	0.07	0.65	0.00	0.00	0.00	0.00	0.21	0.00	0.00
	0.4	0.4	0.02	0.02	0.05	0.02	0.02	0.02	0.00	0.85	0.01	0.00	0.04	0.05	0.08	0.00	0.00	0.00	0.00	0.82	0.00	0.00
mRMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.10	0.10	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.21	0.21	0.33	0.00	0.00	0.03	0.01	0.22	0.00	0.00	0.22	0.22	0.35	0.00	0.00	0.00	0.00	0.21	0.00	0.00
	0.4	0.4	0.06	0.06	0.02	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.07	0.08	0.03	0.00	0.00	0.00	0.00	0.82	0.00	0.00
			T = 10,000																			
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.75	0.00	0.24	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.07	0.00	0.26	0.02	0.00	0.05	0.00	0.59	0.00	0.00	0.12	0.00	0.34	0.00	0.00	0.00	0.00	0.54	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.45	0.01	0.03	0.00	0.00	0.00	0.00	0.51	0.00	0.00	0.48	0.01	0.02	0.00	0.00	0.00	0.00	0.48	0.00	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.00	0.00	0.63	0.00	0.00	0.12	0.00	0.24	0.00	0.00	0.01	0.01	0.76	0.00	0.00	0.00	0.00	0.23	0.00	0.00
	0.4	0.4	0.00	0.00	0.04	0.00	0.01	0.02	0.00	0.93	0.00	0.00	0.02	0.02	0.07	0.00	0.00	0.00	0.00	0.89	0.00	0.00
mRMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.11	0.11	0.53	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.12	0.12	0.52	0.00	0.00	0.00	0.00	0.24	0.00	0.00
	0.4	0.4	0.03	0.03	0.02	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.04	0.04	0.03	0.00	0.00	0.00	0.00	0.89	0.00	0.00

Note: See note at Table 1.

Table 3: Empirical selection probabilities: one endogenous regressor ($p = 1$), T = 50,000; 100,000

			T = 50,000																			
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.79	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.81	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.05	0.00	0.30	0.02	0.00	0.10	0.00	0.52	0.00	0.00	0.10	0.00	0.43	0.00	0.00	0.00	0.00	0.47	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.58	0.00	0.01	0.00	0.00	0.00	0.00	0.41	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.00	0.00	0.65	0.00	0.00	0.17	0.00	0.18	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.17	0.00	0.00
	0.4	0.4	0.00	0.00	0.03	0.00	0.00	0.03	0.01	0.92	0.01	0.00	0.01	0.01	0.08	0.00	0.00	0.01	0.00	0.88	0.00	0.00
mRMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.05	0.06	0.70	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.06	0.06	0.70	0.00	0.00	0.00	0.00	0.18	0.00	0.00
	0.4	0.4	0.03	0.03	0.03	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.88	0.00	0.00
			T = 100,000																			
			2SLS										LIML									
	δ_1	δ_2	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1	z2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																					
RMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	0.81	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.04	0.00	0.31	0.03	0.00	0.16	0.01	0.45	0.00	0.00	0.09	0.00	0.50	0.00	0.00	0.00	0.00	0.41	0.00	0.00
mRMSC	0	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.4	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.4	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00
	$\delta_1 = \delta_2$																					
RMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.00	0.00	0.64	0.00	0.00	0.21	0.01	0.14	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.13	0.00	0.00
	0.4	0.4	0.00	0.00	0.02	0.00	0.00	0.04	0.01	0.92	0.01	0.00	0.01	0.01	0.09	0.00	0.00	0.02	0.00	0.87	0.00	0.00
mRMSC	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.03	0.03	0.79	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.04	0.04	0.78	0.00	0.00	0.00	0.00	0.14	0.00	0.00
	0.4	0.4	0.03	0.03	0.04	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.04	0.04	0.05	0.00	0.00	0.00	0.00	0.87	0.00	0.00

Note: See note at Table 1.

Table 4: Empirical selection probabilities: two endogenous regressors ($p = 2$), $T = 100$; 500

			$T = 100$															
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.31	0.03	0.00	0.43	0.15	0.00	0.09	0.00	0.35	0.05	0.00	0.41	0.11	0.00	0.09	0.00
	0.1	0.4	0.29	0.03	0.00	0.43	0.16	0.00	0.09	0.00	0.34	0.05	0.00	0.41	0.11	0.00	0.09	0.00
	0.2	0.4	0.24	0.02	0.00	0.43	0.20	0.00	0.11	0.00	0.30	0.04	0.00	0.42	0.14	0.00	0.10	0.00
	0.3	0.4	0.15	0.01	0.00	0.39	0.29	0.00	0.16	0.00	0.20	0.03	0.00	0.41	0.21	0.00	0.14	0.00
mRMSC	0	0.4	0.36	0.03	0.00	0.23	0.18	0.00	0.16	0.04	0.40	0.06	0.00	0.23	0.15	0.00	0.13	0.02
	0.1	0.4	0.34	0.03	0.00	0.22	0.18	0.00	0.17	0.06	0.39	0.06	0.00	0.23	0.16	0.00	0.14	0.03
	0.2	0.4	0.29	0.03	0.00	0.19	0.18	0.00	0.21	0.10	0.34	0.05	0.00	0.21	0.17	0.00	0.17	0.05
	0.3	0.4	0.18	0.02	0.00	0.15	0.19	0.00	0.28	0.18	0.24	0.04	0.00	0.20	0.20	0.00	0.24	0.08
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.82	0.00	0.00	0.17	0.01	0.00	0.00	0.00	0.84	0.00	0.00	0.15	0.01	0.00	0.00	0.00
	0.3	0.3	0.32	0.00	0.00	0.44	0.19	0.00	0.05	0.00	0.38	0.00	0.00	0.43	0.14	0.00	0.05	0.00
	0.4	0.4	0.07	0.01	0.01	0.30	0.34	0.00	0.27	0.01	0.11	0.02	0.03	0.34	0.25	0.01	0.23	0.00
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.87	0.00	0.00	0.07	0.03	0.00	0.02	0.01	0.90	0.00	0.00	0.06	0.02	0.00	0.01	0.01
	0.3	0.3	0.37	0.00	0.00	0.18	0.16	0.00	0.17	0.11	0.44	0.01	0.01	0.20	0.16	0.00	0.14	0.06
	0.4	0.4	0.08	0.01	0.01	0.11	0.19	0.00	0.36	0.24	0.14	0.03	0.03	0.16	0.21	0.01	0.32	0.09
			$T = 500$															
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.25	0.01	0.00	0.44	0.23	0.00	0.07	0.00	0.29	0.01	0.00	0.44	0.19	0.00	0.07	0.00
	0.1	0.4	0.24	0.01	0.00	0.44	0.24	0.00	0.07	0.00	0.28	0.01	0.00	0.44	0.19	0.00	0.07	0.00
	0.2	0.4	0.22	0.01	0.00	0.42	0.26	0.00	0.08	0.00	0.27	0.01	0.00	0.43	0.21	0.00	0.07	0.00
	0.3	0.4	0.13	0.00	0.00	0.37	0.34	0.00	0.14	0.01	0.18	0.01	0.00	0.41	0.28	0.00	0.11	0.01
mRMSC	0	0.4	0.46	0.02	0.00	0.23	0.15	0.00	0.11	0.03	0.51	0.03	0.00	0.23	0.13	0.00	0.09	0.01
	0.1	0.4	0.46	0.02	0.00	0.23	0.16	0.00	0.11	0.03	0.50	0.03	0.00	0.23	0.13	0.00	0.09	0.02
	0.2	0.4	0.44	0.02	0.00	0.22	0.16	0.00	0.12	0.04	0.48	0.03	0.00	0.23	0.14	0.00	0.10	0.02
	0.3	0.4	0.34	0.01	0.00	0.18	0.18	0.00	0.19	0.10	0.40	0.03	0.00	0.21	0.17	0.00	0.14	0.05
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.98	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.02	0.00	0.00	0.00	0.00
	0.3	0.3	0.40	0.00	0.00	0.42	0.15	0.00	0.02	0.00	0.45	0.00	0.00	0.41	0.12	0.00	0.02	0.00
	0.4	0.4	0.05	0.00	0.00	0.22	0.39	0.00	0.29	0.04	0.08	0.01	0.01	0.30	0.35	0.00	0.23	0.02
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.75	0.00	0.00	0.11	0.06	0.00	0.05	0.03	0.78	0.00	0.00	0.11	0.06	0.00	0.04	0.02
	0.4	0.4	0.14	0.01	0.00	0.13	0.20	0.00	0.31	0.21	0.20	0.02	0.02	0.19	0.22	0.00	0.26	0.09

Note: See note at Table 1.

Table 5: Empirical selection probabilities: two endogenous regressors ($p = 2$), T = 1,000; 10,000

			T = 1,000															
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.22	0.00	0.00	0.43	0.27	0.00	0.08	0.00	0.26	0.01	0.00	0.44	0.23	0.00	0.07	0.00
	0.1	0.4	0.22	0.00	0.00	0.42	0.27	0.00	0.08	0.00	0.25	0.01	0.00	0.44	0.23	0.00	0.07	0.00
	0.2	0.4	0.20	0.00	0.00	0.42	0.29	0.00	0.09	0.01	0.24	0.01	0.00	0.43	0.25	0.00	0.08	0.00
	0.3	0.4	0.12	0.00	0.00	0.36	0.35	0.00	0.15	0.02	0.16	0.01	0.00	0.40	0.31	0.00	0.12	0.01
mRMSC	0	0.4	0.51	0.01	0.00	0.23	0.14	0.00	0.09	0.02	0.55	0.02	0.00	0.22	0.12	0.00	0.07	0.01
	0.1	0.4	0.51	0.01	0.00	0.22	0.14	0.00	0.09	0.02	0.55	0.02	0.00	0.22	0.12	0.00	0.07	0.01
	0.2	0.4	0.50	0.01	0.00	0.22	0.14	0.00	0.10	0.03	0.54	0.02	0.00	0.22	0.13	0.00	0.08	0.01
	0.3	0.4	0.42	0.01	0.00	0.19	0.15	0.00	0.15	0.08	0.47	0.02	0.00	0.20	0.15	0.00	0.12	0.04
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	0.3	0.3	0.44	0.00	0.00	0.41	0.13	0.00	0.02	0.00	0.48	0.00	0.00	0.40	0.11	0.00	0.01	0.00
	0.4	0.4	0.03	0.00	0.00	0.20	0.38	0.00	0.31	0.07	0.06	0.01	0.00	0.28	0.38	0.00	0.24	0.03
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.88	0.00	0.00	0.07	0.03	0.00	0.02	0.01	0.90	0.00	0.00	0.06	0.02	0.00	0.01	0.01
	0.4	0.4	0.18	0.00	0.00	0.15	0.20	0.00	0.28	0.19	0.24	0.01	0.01	0.19	0.22	0.00	0.24	0.09
	T = 10,000																	
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.12	0.00	0.00	0.34	0.36	0.00	0.15	0.02	0.15	0.00	0.00	0.38	0.33	0.00	0.12	0.01
	0.1	0.4	0.12	0.00	0.00	0.34	0.36	0.00	0.15	0.02	0.15	0.00	0.00	0.38	0.33	0.00	0.13	0.01
	0.2	0.4	0.12	0.00	0.00	0.34	0.36	0.00	0.16	0.02	0.15	0.00	0.00	0.37	0.33	0.00	0.13	0.01
	0.3	0.4	0.07	0.00	0.00	0.27	0.38	0.00	0.23	0.04	0.10	0.00	0.00	0.32	0.37	0.00	0.19	0.03
mRMSC	0	0.4	0.70	0.00	0.00	0.17	0.08	0.00	0.04	0.01	0.73	0.00	0.00	0.16	0.07	0.00	0.03	0.01
	0.1	0.4	0.70	0.00	0.00	0.17	0.08	0.00	0.04	0.01	0.73	0.00	0.00	0.16	0.07	0.00	0.03	0.01
	0.2	0.4	0.70	0.00	0.00	0.17	0.08	0.00	0.04	0.01	0.72	0.00	0.00	0.16	0.07	0.00	0.03	0.01
	0.3	0.4	0.67	0.00	0.00	0.17	0.09	0.00	0.06	0.02	0.70	0.00	0.00	0.16	0.07	0.00	0.05	0.01
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.51	0.00	0.00	0.37	0.11	0.00	0.01	0.00	0.53	0.00	0.00	0.36	0.09	0.00	0.01	0.00
	0.4	0.4	0.01	0.00	0.00	0.10	0.30	0.00	0.41	0.18	0.03	0.00	0.00	0.17	0.36	0.00	0.34	0.10
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.4	0.4	0.36	0.00	0.00	0.17	0.16	0.00	0.19	0.12	0.43	0.00	0.00	0.19	0.16	0.00	0.15	0.06

Note: See note at Table 1.

Table 6: Empirical selection probabilities: two endogenous regressors ($p = 2$), T = 50,000; 100,000

			T = 50,000															
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.07	0.00	0.00	0.28	0.38	0.00	0.22	0.05	0.09	0.00	0.00	0.32	0.36	0.00	0.19	0.03
	0.1	0.4	0.07	0.00	0.00	0.28	0.38	0.00	0.22	0.05	0.09	0.00	0.00	0.32	0.36	0.00	0.19	0.03
	0.2	0.4	0.07	0.00	0.00	0.28	0.38	0.00	0.23	0.05	0.09	0.00	0.00	0.31	0.37	0.00	0.20	0.04
	0.3	0.4	0.04	0.00	0.00	0.21	0.37	0.00	0.29	0.08	0.06	0.00	0.00	0.26	0.37	0.00	0.25	0.06
mRMSC	0	0.4	0.81	0.00	0.00	0.12	0.05	0.00	0.02	0.00	0.83	0.00	0.00	0.11	0.04	0.00	0.02	0.00
	0.1	0.4	0.81	0.00	0.00	0.12	0.05	0.00	0.02	0.00	0.83	0.00	0.00	0.11	0.04	0.00	0.02	0.00
	0.2	0.4	0.81	0.00	0.00	0.12	0.04	0.00	0.02	0.00	0.83	0.00	0.00	0.11	0.04	0.00	0.02	0.00
	0.3	0.4	0.80	0.00	0.00	0.12	0.05	0.00	0.02	0.01	0.82	0.00	0.00	0.11	0.04	0.00	0.02	0.00
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.51	0.00	0.00	0.38	0.10	0.00	0.01	0.00	0.53	0.00	0.00	0.37	0.09	0.00	0.01	0.00
	0.4	0.4	0.00	0.00	0.00	0.05	0.23	0.00	0.43	0.29	0.01	0.00	0.00	0.10	0.32	0.00	0.39	0.18
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.4	0.4	0.53	0.00	0.00	0.16	0.12	0.00	0.12	0.07	0.58	0.00	0.00	0.17	0.12	0.00	0.09	0.04
			T = 100,000															
			2SLS								LIML							
	δ_1	δ_2	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All	z1+z2	z1+I	z2+I	z1+z2+I	z1+z2+2I	all I	zj+more I	All
	$\delta_1 < \delta_2$																	
RMSC	0	0.4	0.06	0.00	0.00	0.26	0.37	0.00	0.25	0.06	0.08	0.00	0.00	0.29	0.37	0.00	0.21	0.05
	0.1	0.4	0.06	0.00	0.00	0.26	0.37	0.00	0.25	0.06	0.08	0.00	0.00	0.29	0.37	0.00	0.21	0.05
	0.2	0.4	0.06	0.00	0.00	0.25	0.37	0.00	0.25	0.06	0.07	0.00	0.00	0.29	0.38	0.00	0.22	0.05
	0.3	0.4	0.03	0.00	0.00	0.19	0.36	0.00	0.32	0.10	0.05	0.00	0.00	0.24	0.38	0.00	0.27	0.07
mRMSC	0	0.4	0.85	0.00	0.00	0.10	0.03	0.00	0.01	0.00	0.87	0.00	0.00	0.09	0.03	0.00	0.01	0.00
	0.1	0.4	0.85	0.00	0.00	0.10	0.03	0.00	0.01	0.00	0.87	0.00	0.00	0.09	0.03	0.00	0.01	0.00
	0.2	0.4	0.85	0.00	0.00	0.10	0.04	0.00	0.01	0.00	0.87	0.00	0.00	0.09	0.03	0.00	0.01	0.00
	0.3	0.4	0.84	0.00	0.00	0.10	0.04	0.00	0.02	0.00	0.86	0.00	0.00	0.09	0.03	0.00	0.01	0.00
	$\delta_1 = \delta_2$																	
RMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	0.51	0.00	0.00	0.37	0.10	0.00	0.01	0.00	0.53	0.00	0.00	0.36	0.09	0.00	0.01	0.00
	0.4	0.4	0.00	0.00	0.00	0.04	0.20	0.00	0.43	0.34	0.01	0.00	0.00	0.08	0.28	0.00	0.41	0.22
mRMSC	0	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.3	0.3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.4	0.4	0.61	0.00	0.00	0.15	0.10	0.00	0.09	0.05	0.66	0.00	0.00	0.15	0.09	0.00	0.07	0.03

Note: See note at Table 1.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica 59*(3), 817–858.

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica 67*(3), 543–563.

Andrews, D. W. K. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

Andrews, D. W. K. and B. Lu (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics 101*(1), 123–164.

Antoine, B. and E. Renault (2009). Efficient GMM with nearly-weak instruments. *The Econometrics Journal 12*, S135–S171.

Antoine, B. and E. Renault (2012). Efficient minimum distance estimation with multiple rates of convergence. *Journal of Econometrics 170*(2), 350–367.

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

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Blomquist, S. and M. Dahlberg (1999). Small sample properties of LIML and jackknife IV estimators: Experiments with weak instruments. *Journal of Applied Econometrics 14*(1), 69–88.

Caner, M. (2009). Testing, estimation in GMM and CUE with nearly-weak identification. *Econometric Reviews 29*(3), 330–363.

Caner, M. and Q. Fan (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive Lasso. *Journal of Econometrics 187*(1), 256–274.

Cheng, X. and Z. Liao (2015). Select the valid and relevant moments: An information-based Lasso for GMM with many moments. *Journal of Econometrics 186*(2), 443–464.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

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

Dovonon, P. and F. Y. Atchadé (2019). Efficiency bounds for semiparametric models with singular score functions. Technical report, Concordia University, Canada and Boston University, USA.

Dovonon, P., F. Y. Atchadé, and F. Doko Tchatoka (2019). Efficiency bounds for moment condition models with mixed strength. Technical report, Department of Economics, Concordia University, Canada.

Dovonon, P. and A. R. Hall (2018). The asymptotic properties of GMM and indirect inference under second-order identification. *Journal of econometrics 205*(1), 76–111.

Dovonon, P. and E. Renault (2013). Testing for common conditionally heteroskedastic factors. *Econometrica 81*(6), 2561–2586.

Dovonon, P. and E. Renault (2019). GMM overidentification test with first-order underidentification. Technical report, Department of Economics, Concordia University, Canada.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007). Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics 138*(2), 488–512.

Hall, A. R. and F. P. Peixe (2003). A consistent method for the selection of relevant instruments. *Econometric Reviews 22*(3), 269–287.

Inoue, A. and M. Shintani (2018). Quasi-bayesian model selection. *Quantitative Economics 9*(3), 1265–1297.

Lee, J. H. and Z. Liao (2018). On standard inference for GMM with local identification failure of known forms. *Econometric Theory 34*(4), 790–814.

Magnus, J. R. and H. Neudecker (2002). *Matrix differential calculus with applications in statistics and econometrics*. Wiley.

Nelson, C. R. and R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica: Journal of the Econometric Society*, 967–976.

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

Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *The Annals of Mathematical Statistics 28*(2), 329–348.

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

Wang, W. and F. Doko Tchatoka (2018). On bootstrap inconsistency and bonferroni-based size-correction for the subset Anderson-Rubin test under conditional homoskedasticity. *Journal of Econometrics 207*(1), 188–211.

Windmeijer, F., H. Farbmacher, N. Davies, and D. G. Smith (2018). On the use of the Lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*. DOI:10.1080/01621459.2018.1498346.