Econometric Theory, **40**, 2024, 1003–1064.
doi:10.1017/S0266466622000640

RELEVANT MOMENT SELECTION UNDER MIXED IDENTIFICATION STRENGTH

PROSPER DOVONON
Department of Economics, Concordia University

FIRMIN DOKO TCHATOKA
School of Economics and Public Policy, The University of Adelaide

MICHAEL AGUESSY
Department of Economics, Concordia University

This paper proposes a robust moment selection method aiming to pick the best model even if this is a moment condition model 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, *Journal of Econometrics* 138, 488–512) 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 automatically accounts for both the convergence rate 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 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 the strong/point identification property. 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

The paper has benefited from many comments of the Co-Editor (Michael Jansson), the Editor (Peter Phillips), and two anonymous referees. We thank Jean-Jacques Forneron, Christian Gouriéroux, Zhongjun Qu, Eric Renault, Rami Tabri, Brendan K. Beare, and Ye Lu for helpful comments. We also thank the participants of the 2018 Africa Meeting of the Econometric Society in Benin and the 2018 Canadian Econometric Study Group meeting in Ottawa, and seminar participants at Boston University, York University, and the University of Sydney for helpful discussions. This research is supported by the Social Sciences and Humanities Research Council of Canada and by the Australian Research Council grant DP200101498. Address correspondence to Prosper Dovonon, Department of Economics, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC H3G 1M8, Canada; e-mail: prosper.dovonon@concordia.ca.

estimators that are all consistent and asymptotically normal under further regularity conditions. Moment selection methods have also been developed under standard identification settings.

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 literature includes Belloni et al. (2012), Caner and Fan (2015), Cheng and Liao (2015), and Windmeijer et al. (2019).

The second strand of literature on moment selection adopts a more classical methodology for model selection by relying on information criteria. This approach includes Andrews (1999), Andrews and Lu (2001), Donald and Newey (2001; henceforth DN), Hall and Peixe (2003), and 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. DN rely on the mean squared error (MSE) of some estimators including the two-stage least-squares 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, the relevant moment selection criterion (RMSC) of Hall et al. (2007) and the J-statistic selection criterion of Andrews (1999) are complementary. The *best* model in terms of RMSC is the smallest model (in number of moment restrictions) among those that are correct and yielding the maximum information about the parameter of interest.

Common to all the papers cited above is the requirement of strong identification for 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 failing this property. Identification properties are outlined by considering on the one hand strong identification and on the other hand the extreme lack of 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. Since the seminal work of Staiger and Stock (1997), the strength of a moment restriction has been captured by the possibility that the moment function vanishes uniformly over the whole parameter space as the sample size grows. The faster the moment function of the restriction vanishes, the weaker is the restriction.

Weak moment restrictions are those vanishing at least at the rate $T^{-\frac{1}{2}}$; strong ones are those vanishing only at the true value and do not drift to 0 at any other value, whereas those vanishing over the parameter set at rate $T^{-\alpha}$, $\alpha \in (0, \frac{1}{2})$, 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. Examples of such models can be found in Section 2. They include the classical linear instrumental variable (IV) model with nearly weak instruments and GMM inference on conditional moment restrictions with finite-support conditioning variables. We refer to Caner (2009), Andrews and Cheng (2012), Antoine and Renault (2012), and Han and McCloskey (2019) for further account of such models.

Even though point identification fails in the limit, consistent estimation is possible due to the fact that (by the central limit theorem) these models gather information about the parameter of interest at a faster rate than they lose their potential for identification. This feature has first been pointed out by Hahn and Kuersteiner (2002) and subsequently by Antoine and Renault (2009), who also show—in this setting—that consistent estimators may converge at faster rates in some directions of the parameters space. Interestingly, in this context, standard optimal GMM inference is valid without a specific characterization of the directions of faster convergence (see Antoine and Renault, 2009, 2012). More recently, Antoine and Renault (2020) have proposed a test to investigate whether a moment condition model is strong enough to warrant the validity of the standard inference.

This paper proposes a robust 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 two-step GMM estimator, which has linear reparameterizations shown to be asymptotically efficient in this framework by Dovonon, Atchadé, and Doko Tchatoka (2022). 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. More precisely, mRMSC first rewards the rate of estimation and then, for models with the same rate, it rewards (negative) 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 the RMSC and accounting for the scaling factor, mRMSC penalizes more strongly larger models. Indeed, the penalty term of mRMSC is proportional to $(1/\ln T)^{\alpha}$, $\alpha > 0$, whereas that of the Bayesian information criterion (BIC)-RMSC—identified as the best-performing version of RMSC—is $\ln\sqrt{T}/\sqrt{T}$. The choice of penalty for mRMSC is guided both by robustness to unknown model identification strength and 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 of our theory, the simulations reveal that, irrespective of the Monte Carlo design considered and except for the cases where all candidate models are weak, mRMSC selects the best model or set of instruments with probability (hit rate) growing to 1 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 hit rate 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. We also consider the MSE-based criterion of DN and found it bested by the two entropy-based criteria in terms of hit rate.

The post-selection performance of these methods has been analyzed. For this purpose, we also consider the moment condition model including all the available instruments. According to Chao and Swanson (2005), this is a recommended practice in settings where many weak instruments are available. We find that models selected by mRMSC dominate the other models in terms of coverage probability of confidence intervals except for the configuration where all relevant instruments are weak. In these cases, all the models in competition perform very poorly although the model with all instruments seems marginally better in relative terms. We also consider the bias and MSE of estimation. For both measures, mRMSC outperforms the other criteria except for the cases of weak identification where, again, all of them perform poorly with the model with all instruments having a slight edge followed by models selected by the Donald and Newey's criterion.

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 article is organized as follows: Section 2 introduces the setup 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 and post-selection properties. Relevant choices of penalty functions are also discussed. Simulation results are reported in Section 5, whereas Section 6 concludes. Lengthy proofs are relegated to Appendix B. The Supplementary Material of this article provides further simulation results.

Throughout the article, $|a|$ denotes the number of nonzero entries or the determinant of a if a is a vector or a square matrix; $\|a\|$ denotes the Frobenius norm of the matrix a, i.e., $\|a\| = \sqrt{\text{trace}(aa')}$; $a \vee b$ denotes $\max(a, b)$ and $a \wedge b$ denotes $\min(a, b)$.

2. SETUP, EXAMPLES, AND EXISTING RESULTS

Let us consider the sample $\{Y_{tT} : t = 1, \ldots, T\}$, $(T \geq 1)$, a triangular array with common distribution $\mathbb{P}_T$, described by the population moment condition

$$\mathbb{E}(\phi(Y_{tT}, \theta_0)) = 0, \tag{1}$$

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 $\mathbb{R}^p$, and $\mathbb{E}(\cdot)$ denotes expectation taken under $\mathbb{P}_T$—we do not explicitly mention its dependence on T for simplicity.

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

$$\mathbb{E}(\phi(Y_{tT}, \theta)) = 0, \quad \theta \in \Theta \quad \Leftrightarrow \theta = \theta_0. \tag{2}$$

This property plays an important role in the standard theory of GMM of Hansen (1982) to claim the 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 levels of identification strengths may be expected from the components of the estimating function as stressed by Hahn and Kuersteiner (2002), Antoine and Renault (2009, 2012), Caner (2009), Andrews and Cheng (2012), and Han and McCloskey (2019), among others.

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

$$\mathbb{E}(\phi(Y_{tT}, \theta)) = 0, \quad \forall \theta \in \Theta,$$

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, $\mathbb{E}(\phi(Y_{tT}, \theta))$ is local to 0:

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

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

When $0 < \delta < \frac{1}{2}$, the moment condition model is referred to as nearly weak and as weak when $\delta = \frac{1}{2}$. 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 levels of information about the parameters of interest.

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

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

In this representation, ϕ_1 has the potential to more strongly identify θ_0—or some of its components—than ϕ_2. Although 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 mild conditions. Standard identification features of moment condition models pertain to the case $\delta_1 = \delta_2 = 0$.

For simplicity of exposition, we maintain throughout that the moment functions exhibit either one of two identification strengths. Nevertheless, our results extend to more general settings as addressed by Appendix C (see also Remark 3).

Two examples of moment condition models where various levels of identification strengths may be expected from the components of the estimating function are presented below. Further examples of such models are detailed in Antoine and Renault (2012) and Han and McCloskey (2019).

Example 1. (Linear IV model with nearly weak instruments). Consider the classical linear IV model:

$$Y = X\theta + U, \tag{4}$$

$$X = Z\Pi_T + V, \tag{5}$$

with $Y = (y_1, \ldots, y_T)'$ the T-vector of realizations of the dependent variable; $X = [x_1, \ldots, x_T]'$ the (T,p)-matrix of p explanatory variables, some of which may be endogenous; $Z = [z_1, \ldots, z_T]'$ the (T,k)-matrix of IVs; $U = (u_1, \ldots, u_T)'$ and $V = [v_1, \ldots, v_T]'$ are T-vector and (T,p)-matrix of errors, respectively; and θ and Π_T the p-vector and (k,p)-matrix of parameters, respectively. Suppose that

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

for some $0 \leq \delta_1 \leq \delta_2 < \frac{1}{2}$, 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_T, i.e., Z_i, (T,k_i)-matrix for $i = 1,2$. Thus, we can write the system (4) and (5) as

$$Y = X\theta + U, \tag{6}$$

$$X = Z_1 \frac{C_1}{T^{\delta_1}} + Z_2 \frac{C_2}{T^{\delta_2}} + V \equiv Z_1 \Pi_{1T} + Z_2 \Pi_{2T} + V. \tag{7}$$

When $\delta_1 = \delta_2$, the instruments in Z_1 and Z_2 have equal strength, whereas those in Z_1 are stronger than those in Z_2 if $\delta_1 < \delta_2$. Suppose that $\{w_t \equiv (y_t, x_t, z_t) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^k : t = 1, \ldots, T\}$ is a sample of independent and identically distributed random vectors with finite second moments, and $\mathbb{E}(z_t u_t) = 0$, $\mathbb{E}(z_t v_t') = 0$ for all t. Then, the true parameter θ_0 solves the moment condition

$$\mathbb{E}\left(z_t(y_t - x_t'\theta)\right) = 0. \tag{8}$$

Specifically, letting

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

where $z_t \equiv (z_{1t}', z_{2t}')' \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$, we can write

$$\mathbb{E}[z_t(y_t - x_t'\theta)] = \begin{pmatrix} \mathbb{E}[z_{1t}(y_t - x_t'\theta)] \\ \mathbb{E}[z_{2t}(y_t - x_t'\theta)] \end{pmatrix} = \begin{pmatrix} T^{-\delta_1}\rho_1(\theta) + T^{-\delta_2}\nu_1(\theta) \\ T^{-\delta_2}\rho_2(\theta) + T^{-\delta_1}\nu_2(\theta) \end{pmatrix} \text{ with} \tag{9}$$

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

As indicated in Antoine and Renault (2009), we see that if the instruments z_{1t} and z_{2t} are orthogonal, i.e., if $\Delta_{12} = \Delta_{21}' = 0$, (9) becomes

$$\mathbb{E}[\phi_i(w_t, \theta)] = T^{-\delta_i}\rho_i(\theta),\ t = 1, \ldots, T,\ i = 1, 2, \tag{10}$$

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 (9), $i = 1, 2$.

Example 2. (Kernel Smoothing; Antoine and Renault, 2012). Let (X_t, Y_t), $t = 1, \ldots, T$, be the observed sample on a stationary process with stationary distribution (X, Y). Let $m_T(\theta)$ be a Nadaraya–Watson estimator of the conditional expectation $\mathbb{E}[g(Y, \theta)|X = x]$, where g is a known function of an unknown parameter θ of interest. Letting $h_T = T^{-2\delta_1}$ for a suitably chosen $0 < \delta_1 < \frac{1}{2}$, denote the bandwidth sequence, $\sqrt{Th_T}\{m_T(\theta) - - - \mathbb{E}[g(Y, \theta)|X = x]\}$ is pointwise asymptotically Gaussian with zero mean and under further mild conditions may, as a function of θ, converge to a Gaussian process. Suppose now that for inference about the true value θ_0 of θ, the estimating equation $\mathbb{E}[g(Y, \theta_0)|X = x] = 0$ is valid for a given value x but may not be uniformly valid over all the support of X.[1] Then,

$$\sqrt{T}\left[\bar{\phi}_T(\theta) - \frac{\rho(\theta)}{T^{\delta_1}}\right]$$

[1] Such models appear in the applied finance literature. We refer to Gagliardini, Gourieroux, and Renault (2011) and Antoine and Renault (2012, Exam. 1, p. 351) for more details about this example.

is asymptotically Gaussian, where $\rho(\theta) = \mathbb{E}[g(Y,\theta)|X = x]$, and $\bar{\phi}_T(\theta) \equiv \sqrt{h_T} m_T(\theta)$. This property of $\bar{\phi}_T(\theta)$ gives rise to the moment condition:

$$\mathbb{E}\left(\bar{\phi}_T(\theta)\right) = \frac{\rho(\theta)}{T^{\delta_1}}, \tag{11}$$

where $\bar{\phi}_T(\theta)$ is a sample mean of a double array given by

$$\bar{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \phi_{t,T}(\theta) \text{ where } \phi_{t,T}(\theta) = \sqrt{h_T} K\left(\frac{x_t - x}{h_T}\right) g(y_t,\theta) \Big/ \sum_{s=1}^{T} K\left(\frac{x_s - x}{h_T}\right)$$

and K is a kernel function. Euler optimality conditions are fulfilled for the true unknown value θ_0 of θ ensuring that $\rho(\theta_0) = 0$.

Suppose now that another conditional expectation is informative about θ. The same reasoning as above leads to additional moment restrictions similar to (11) with a possibly different degree of smoothness δ_2. We then end up with vectorial functions $\phi_T(\theta)$ and $\rho(\theta)$ such that

$$\mathbb{E}\left(\bar{\phi}_{T,i}(\theta)\right) = \frac{\rho_i(\theta)}{T^{\delta_i}}, \quad i = 1,2,$$

which has the form in (3).

Returning to the general framework, the following assumption is made to obtain consistent estimators for θ_0 from model (1) under the mixed identification framework in (3). We let $\bar{\phi}_T(\theta) = T^{-1}\sum_{t=1}^{T} \phi(Y_{tT},\theta)$.

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) $\sup_{\theta\in\Theta} \sqrt{T} \left\| \bar{\phi}_T(\theta) - \mathbb{E}(\phi(Y_{tT},\theta)) \right\| = O_P(1)$, under $\mathbb{P}_T$.

Assumption 1(i) imposes global identification of θ_0 by the suitably inflated estimating moment function, whereas 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 for triangular arrays. See, e.g., Ziegler (1997). 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

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \bar{\phi}_T(\theta)' W_T \bar{\phi}_T(\theta), \tag{12}$$

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

$$\rho(\hat{\theta}_T) = O_P\left(T^{\delta_2 - \frac{1}{2}}\right). \tag{13}$$

Hence, $\delta_2 < \frac{1}{2}$ 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 $\delta_2 < \frac{1}{2}$ is not a 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 $\delta_1 < \frac{1}{2}$ 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 let $R = (R_1 \vdots R_2)$ be a (p,p)-nonsingular 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

$$J = \begin{pmatrix} \frac{\partial \rho_1}{\partial \theta'}(\theta_0)R_1 & 0 \\ 0 & \frac{\partial \rho_2}{\partial \theta'}(\theta_0)R_2 \end{pmatrix} \quad \text{and} \quad \Lambda_T = \begin{pmatrix} T^{\frac{1}{2}-\delta_1} I_{s_1} & 0 \\ 0 & T^{\frac{1}{2}-\delta_2} I_{s_2} \end{pmatrix}. \tag{14}$$

The following assumptions are made.

Assumption 2. (i) θ_0 is interior to Θ and $\phi(Y_{tT}, \theta)$ is continuously differentiable on Θ.

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

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

$$\mathbb{E}\left(\frac{\partial \phi_i(Y_{tT}, \theta_0)}{\partial \theta'}\right) = \frac{C_i}{T^{\delta_i}} + o(T^{-\delta_i}), \quad \text{and}$$

$$\sqrt{T} \sup_{\theta \in \mathcal{N}_{\theta_0}} \left\| \frac{\partial \bar{\phi}_{iT}(\theta)}{\partial \theta'} - \mathbb{E}\left(\frac{\partial \phi_i(Y_{tT}, \theta)}{\partial \theta'}\right) \right\| = O_P(1),$$

under $\mathbb{P}_T$, where $\mathcal{N}_{\theta_0}$ is a neighborhood of θ_0.

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

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

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

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 1–3 and $0 < s_1 < p$, then:*

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

$$\sqrt{T}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\tilde{\theta}_T)R\Lambda_T^{-1} \xrightarrow{P} J, \quad \text{under } \mathbb{P}_T, \tag{15}$$

(ii)

$$\Lambda_T R^{-1}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, (J'WJ)^{-1}J'W\Sigma WJ(J'WJ)^{-1}), \quad \text{under } \mathbb{P}_T, \tag{16}$$

where $\hat{\theta}_T$ is the GMM estimator defined by (12) 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 $T^{\frac{1}{2}-\delta_1}$, and 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. Let W_{11} be the limit weighting matrix for estimation based only on ϕ_1. If W_{11} matches the upper-left (k_1, k_1)-submatrix of W, 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 where $s_1 = 0$, it is ϕ_2 that ensures local identification and ϕ_1 may turn out to be the irrelevant set of moment restrictions.

It is not hard to see that the asymptotic variance in (16) is smallest for the choice of $W = \Sigma^{-1}$ where it is equal to $V_* = \left(J'\Sigma^{-1}J\right)^{-1}$. Dovonon et al. (2022) 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. We next revisit our comments in Remark 1 in the setting of efficient GMM estimation.

Remark 2. In Remark 1, consider again the case $s_1 = p$. We observe that if estimation involves optimal weighting matrices, W_{11} does not always match the upper-left block of $W = \Sigma^{-1}$ and efficiency gain becomes possible when the weaker estimating function ϕ_2 is added. To see this, consider the necessary and sufficient condition for efficiency gain derived by Breusch et al. (1999, Thm. 1) in a general GMM inference setting:

$$\Gamma_2 - \Sigma_{21}\Sigma_{11}^{-1}\Gamma_1 \neq 0, \tag{17}$$

with $\Gamma_i = \mathbb{E}(\partial\phi_i(Y_{tT}, \theta_0)/\partial\theta')$ and Σ_{11} and Σ_{21} are suitable upper-left and lower-left submatrices of Σ.

When $\Gamma_2 = 0$, this condition boils down to[2]

$$\Sigma_{21}\Sigma_{11}^{-1}\Gamma_1 \neq 0.$$

In particular, when $k_1 = p$ (just-identification by ϕ_1) and $\Gamma_2 = 0$, equation (17) is equivalent to $\Sigma_{21} \neq 0$, which is the condition derived by Antoine and Renault (2017). In general, if ϕ_1 is overidentifying, this condition is necessary but not sufficient for efficiency gain.

In the context of linear IV model (Example 1), given the moment restriction $\mathbb{E}(z_{1t}u_t) = 0$, the necessary and sufficient condition that $\mathbb{E}(z_{2t}u_t) = 0$ induces efficiency gain as per (17) can be written:

$$\mathbb{E}(z_{2t}x_t') - \Delta_{21}\Delta_{11}^{-1}\mathbb{E}(z_{1t}x_t') \neq 0.$$

We can show—see also Hall, Inoue, and Shin, 2008, p. 499—that this condition is equivalent to $\Pi_{2T} \neq 0$ and this, *regardless of the canonical correlations between* z_{1t} *and* z_{2t}. In particular, if Π_{2T} vanishes faster than Π_{1T} and z_{1t} identifies θ_0, then including z_{2t} would not improve efficiency. This claim is confirmed by Proposition 3.1(ii) in Section 3, which shows that the asymptotic distribution of the efficient GMM estimator is unchanged whether or not the second (weaker) set of instruments is included for inference.

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 the moment function's partition in (3) is essential. Interestingly, however, Antoine and Renault (2009, 2012) have shown that such knowledge is not required. In particular, inference about θ_0

[2] In our framework of mixed identification strength, since Γ_2 is of smaller magnitude than Γ_1, condition (17) amounts to $\Sigma_{21}\Sigma_{11}^{-1}\frac{\partial\rho_1(\theta_0)}{\partial\theta'} \neq 0$.

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 an mRMSC that robustly selects 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 IV models, and we provide some intuition about potential shortcomings that paves the way for an *mRMSC* that we introduce in the next section. Before introducing the simulation setup, let us first introduce the RMSC.

RMSC is a penalized entropy measure that is minimized over candidate models to obtain the most relevant one. Let ϕ denote the estimating function of the moment condition model in (1) 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 (1) which, up to a constant, is

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

where $V(\phi) = \left(G(\phi)'\Sigma(\phi)^{-1}G(\phi)\right)^{-1}$, $G(\phi) = \mathbb{E}(\partial\phi(Y_{tT},\theta_0)/\partial\theta')$, $\Sigma(\phi) = \lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_T(\theta_0))$, with variance under $\mathbb{P}_T$. The sample estimate of $ent_\theta(\phi)$ yields RMSC:

$$RMSC(\phi) = -\ln\left|\hat{G}_T(\phi)'\hat{\Sigma}(\phi)^{-1}\hat{G}_T(\phi)\right| + \kappa(|\phi|,T) = \frac{1}{2}\ln\left|\hat{V}_T(\phi)\right| + \kappa(|\phi|,T), \tag{18}$$

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:

$$\kappa(k,T) = (k-p)\frac{\ln\sqrt{\tau_T}}{\sqrt{\tau_T}} \tag{19}$$

which has been identified by Hall et al. (2007) as the best performing one compared to other alternatives including the Hannan-Quinn penalty. In (19), 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 $\{\phi(Y_{tT}, \theta_0) : t = 1, \ldots, T\}$ 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).

Next, we shed some light on the performance of the RMSC procedure in the presence of moment restrictions with mixed strength. We achieve this in the context of the classical linear IV model of Example 1. For a larger perspective, it is worth deriving the asymptotic properties of the two-stage least-squares estimator. 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

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

is nonsingular.

(iii) $\mathbb{E}(z_t u_t) = 0, \quad \mathbb{E}(z_t v_t') = 0,$

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

where $\sigma_u^2 = \mathbb{E}(u_t^2)$.

Assumption 4(i) restricts 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 $\mathbb{E}(z_t x_t')$. 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 the presence of endogenous regressors.

The efficient GMM estimator of θ_0 from the moment condition (8) is the two-stage least-squares 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 $s_1 \equiv \text{Rank}(C_1)$, 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 (DGP).

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

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 one endogenous variable and set $\theta_0 = 0.1$ and $c_1 = c_2 = 1.48$ 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 each of the 63 (57) combinations of IV has been assessed in the case of the models with 1 (2) endogenous variable(s)

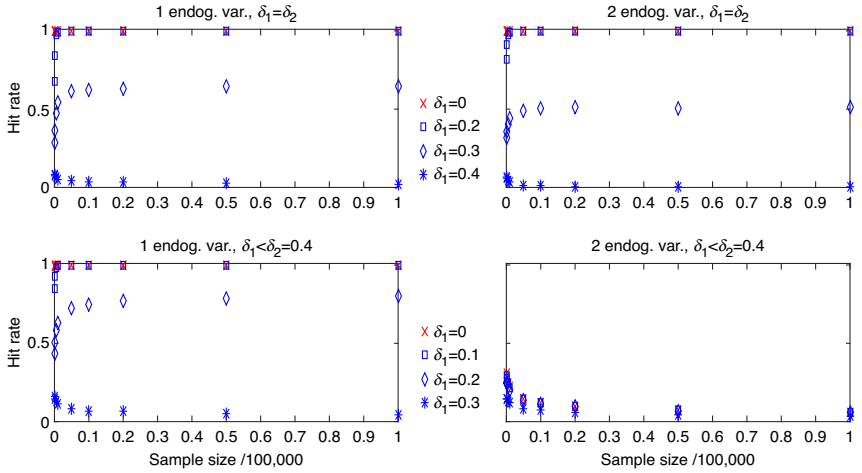


FIGURE 1. Proportion of best model selection (Hit rate) by RMSC for models with one and two endogenous variables. Sample size $T = 100, 200, 500, 1{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000, 100{,}000$; number of replications: 10,000.

and the best model is the one with the lowest RMSC. For a given candidate set of k instruments Z, the RMSC is:

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

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. Figure 1 plots 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 time 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

seems 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 with the selection procedure. Also of importance is the fact that 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 a 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.

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 mRMSC.

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 strengths 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 strengths and that all the candidate models can be expressed as (3) with $0 \leq \delta_1 \leq \delta_2 < \frac{1}{2}$. As in the previous section, we refer to 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 the 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. (2022), the efficiency bound on the estimation of $R(\phi)^{-1}\theta_0$ by ϕ is

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

where

$$J(\phi) = \begin{pmatrix} \frac{\partial \rho_1}{\partial \theta'}(\theta_0)R_1(\phi) & 0 \\ 0 & \frac{\partial \rho_2}{\partial \theta'}(\theta_0)R_2(\phi) \end{pmatrix} \quad \text{and}$$

$$\Lambda_T(\phi) = \begin{pmatrix} T^{\frac{1}{2}-\delta_1(\phi)}I_{s_1(\phi)} & 0 \\ 0 & T^{\frac{1}{2}-\delta_2(\phi)}I_{s_2(\phi)} \end{pmatrix},$$

$\rho_i(\theta)/T^{\delta_i(\phi)} = \mathbb{E}(\phi_i(\theta))$, $(i = 1,2)$, $s_1(\phi) = \text{Rank}\left(\partial\rho_1(\theta_0)/\partial\theta'\right)$, $R(\phi) \equiv \left(R_1(\phi) \vdots R_2(\phi)\right)$ 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 bound happens to be the asymptotic variance of the efficient GMM estimator

$$\hat{\theta}(\phi) \in \arg\min_{\theta\in\Theta} \bar{\phi}_T(\theta)'\hat{\Sigma}(\phi)^{-1}\bar{\phi}_T(\theta), \tag{21}$$

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, $\bar{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^T \phi(Y_{tT}, \theta)$.

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*

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

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, $|\ell_T' R' D_T' V D_T R \ell_T| \to |J(R)'VJ(R)|$. 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 $|J(R)'VJ(R)|$ 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, Thm. 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 that rate is not heterogeneous for the best model in the sense that all directions of the parameter space are 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 a slower rate than standard, but there are also possibilities of having estimators converging at different rates in various directions.

For our purpose, 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 the summary measure of the rates.

Remark 3. In the occurrence of mixed rate estimation involving more than two directions (see Theorem 2.1), direct comparison of two models using the analog of $a(\cdot)$ may look problematic as this function no longer provides a natural ordering of the models. Nonetheless, this analog $a(\phi)$ 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 $a(\phi_{\max})$. Lemma C.1 in Appendix C establishes that $a(\phi)$ cannot be maximum without yielding the best estimation rates as well. The intuition is that estimation rates from $\phi_{\max}$ 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 $\phi_{\max}$. One can proceed iteratively to claim that the map of rates for the estimator from $\phi_{\max}$ is the same as that of any ϕ such that $a(\phi) = a(\phi_{\max})$. This general case is formally studied in Appendix C.

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

$$\iota_\theta(\phi) = -a(\phi) + v_T \cdot ent_\theta(\phi). \tag{22}$$

The sequence v_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, v_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 (15), 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,

$$\hat{V}_\theta(\phi) \equiv \left(\left(\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}\right)' \hat{\Sigma}(\phi)^{-1}\left(\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi))R(\phi)\Lambda_T(\phi)^{-1}\right)\right)^{-1} \tag{23}$$

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

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

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

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

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 *mRMSC* takes the form:

$$mRMSC(\phi) = -\frac{1}{\ln T}\ln\left|\hat{I}_{\theta,T}(\phi)\right| + \kappa_T, \quad \text{with} \quad \hat{I}_{\theta,T}(\phi) = \frac{\partial \bar{\phi}_T'}{\partial \theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(\phi)), \tag{24}$$

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 $\phi(\cdot) = \phi_{\max}(\cdot, c)$. 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, and $\hat{I}_{\theta,T}(c)$ the estimated information matrix (see (24)).

We maintain the following assumption on $\phi_{\max}$.

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

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

$i = 1, 2$, $0 \leq \delta_1 \leq \delta_2 < \frac{1}{2}$ and $\rho_{\max}(\cdot)$ is an $\mathbb{R}^{k_{\max}}$-valued function defined on the compact 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) $\sup_{\theta \in \Theta} \sqrt{T} \left\| \bar{\phi}_{\max,T}(\theta) - \mathbb{E}\left(\phi_{\max}(Y_{tT}, \theta)\right) \right\| = O_P(1)$ under $\mathbb{P}_T$, with $\bar{\phi}_{\max,T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{\max}(Y_{tT}, \theta)$.

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

(v) $\frac{\partial \rho_{\max}}{\partial \theta'}(\theta_0)$ is full column rank, and, for $i = 1,2$, $\mathbb{E}\left(\frac{\partial \phi_{\max,i}(Y_{tT},\theta_0)}{\partial \theta'}\right) = T^{-\delta_i}\frac{\partial \rho_{\max,i}}{\partial \theta'} + o(T^{-\delta_i})$ and $\sqrt{T}\sup_{\theta\in\mathcal{N}_{\theta_0}}\left\|\frac{\partial \bar{\phi}_{\max,T}(\theta)}{\partial \theta'} - \mathbb{E}\left(\frac{\partial \phi_{\max}(Y_{tT},\theta)}{\partial \theta'}\right)\right\| = O_P(1)$ under $\mathbb{P}_T$, with $\bar{\phi}_{\max,i,T}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\phi_{\max,i}(Y_{tT},\theta)$.

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

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

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

under $\mathbb{P}_T$, 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 kth component of $\bar{\phi}_{\max,i,T}(\theta)$.

(viii) $\Sigma(\phi_{\max}) = \lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_{\max,T}(\theta_0))$ is positive definite, with variance taken under $\mathbb{P}_T$.

Assumption 5 is a partial collection of Assumptions 1–3 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 the 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 $\phi_{\max}$. 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 $s_1(c)$ be the rank of the Jacobian matrix of $\rho_{\max,1}(c_1)$ at θ_0 and $s_2(c) = p - s_1(c)$.

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* $\iota_{\max}$ *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 $\mathbb{E}(\phi_{\max}(Y_{tT}, \theta))$ partitions at most into two components with specific rate of convergence to 0 that are $T^{-\delta_1}$ and $T^{-\delta_2}$, respectively. The general case is studied in Appendix C.

Nevertheless, because of the dependence of $V_\theta(c)$ on the choice of rotation matrix $R(c)$, the statement $V_\theta(\iota_{\max}) = V_\theta(c_r)$ 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 $s_1(c_r) = s_1(\iota_{\max})$, 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 $\hat{\theta}_T(c_r)$ 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 conditions on the *mRMSC* introduced by (24) with a penalization term κ_T, a function of sample size, and the size of the estimating function. Note that parsimony is sought relative to the number of moment restrictions 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 (24) 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

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

and

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

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 $\text{Rank}(\partial\rho_{\max}(\theta_0, c)/\partial\theta') = p$; (iii) $\hat{\Sigma}(c)$ converges in

probability, under $\mathbb{P}_T$, to $\Sigma(c) \equiv \lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_{\max,T}(\theta_0, c))$, positive definite, with variance taken under $\mathbb{P}_T$; (iv) $\hat{V}_\theta(c) = V_\theta(c) + O_P(\tau_{T,c}^{-1})$ under $\mathbb{P}_T$, where $\tau_{T,c} \to \infty$ as $T \to \infty$; and (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. Part (ii) may look restrictive by imposing that all candidate models extracted from $\phi_{\max}$ must globally identify θ_0 and must also identify θ_0 locally at first order. This, indeed, need not be the case. However, it turns out that consistency of model selection within that category of models is the most relevant. As we show in Theorem 4.3, potential nonidentifying and/or rank deficient candidate models are strongly outscored by c_r in terms of minimum mRMSC, at the limit.

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 a bandwidth parameter; 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 4.5 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.

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

An extension of this result to the case of more than two identification strengths is given by Theorem C.1 in Appendix C. 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.

In addition, 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, 2020), Dovonon and Hall (2018), Lee and Liao (2018), Han and McCloskey (2019), and Dovonon and Atchadé (2020), among others, have studied the behavior of the

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 $s_1(c)=\text{Rank}\left(\frac{\partial \rho_{\max,1}(\theta_0,c_1)}{\partial \theta'}\right)$ and $s_1(c)+s_2(c)=\text{Rank}\left(\frac{\partial \rho_{\max}(\theta_0,c)}{\partial \theta'}\right)<p$. The remaining directions are estimated at a slower rate in the latter scenario, whereas 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. The full-rank Jacobian matrix of the moment function at θ_0 then fits into Theorem 4.2, whereas the 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 $c=(c_1',c_2')'\in\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}$ be a vector of 1's and 0's such that:

(i.a) $[\rho_{\max}(\theta,c)=0\Leftrightarrow\theta=\theta_0]$ and $\text{Rank}\left(\partial\rho_{\max}(\theta_0,c)/\partial\theta'\right)<p$, 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 $(\mathbb{P}_T)$ 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)), $[\partial\rho_{\max,1}(\hat{\theta}_T(c),c_1)/\partial\theta']r=o_P(T^{\delta_1-\delta_2})$, under $\mathbb{P}_T$.

(iii) $\hat{\Sigma}(c)^{-1}=O_P(1)$, under $\mathbb{P}_T$.

(iv) $\sup_{\theta\in\Theta}\left\|\frac{\partial\bar{\phi}_T}{\partial\theta'}(\theta,c)-\mathbb{L}_T^{-1}\frac{\partial\rho_{\max}}{\partial\theta'}(\theta,c)\right\|=O_P(T^{-\frac{1}{2}})$ under $\mathbb{P}_T$; $\mathbb{L}_T=\begin{pmatrix}T^{\delta_1}I_{k_1(c)} & 0\\ 0 & T^{\delta_2}I_{k_2(c)}\end{pmatrix}$.

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, $[\partial\rho_{\max,1}(\hat{\theta}_T(c),c_1)/\partial\theta']r=O_P(T^{-\frac{1}{2}})$ 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 $c = (c_1', c_2')' \in \mathscr{C}$ failing identification as in Assumption 7. If (i) Assumption 5 holds, (ii) there exists c_r satisfying Definition 1, and (iii) for $\gamma = c, c_r$, $\ln T \cdot \kappa(|\gamma|, T) = o(1)$, then*

$$(mRMSC(c_r) - mRMSC(c)) \ln T = a_0 + a_1 \ln T + a_{1T} + o_P(1),$$

where a_0 is a constant, a_1 is a nonpositive constant, and a_{1T} is a random sequence diverging to $-\infty$ as $T \to \infty$. Consequently,

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

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

Remark 4. This result shows that candidate models that fail identification as indicated by Assumption 7 are dominated by the relevant model c_r in terms of mRMSC and cannot be picked. Meanwhile, Theorem 4.2 shows that any candidate model that satisfies identification properties as in Assumption 6(ii) is also dominated by c_r. These two results show that mRMSC is consistent in a wide range of candidate model configurations. It is not hard to see that Theorem 4.3 covers models that are completely uninformative about the true parameter value θ_0. This is the case if $\rho_{\max}(\theta, c) = 0$ for all θ in the parameter set.

Remark 5. The consistency of mRMSC is established under the condition that the GMM estimator from the estimating function $\phi_{\max}$ is consistent and asymptotically normal. If this condition fails—which is the case if $\phi_{\max}$ is uninformative about θ_0 or $\delta_2 \geq \delta_1 \geq \frac{1}{2}$—mRMSC is not guaranteed to behave well. Actually, simulation results in Section 5 show that it behaves very poorly as does the RMSC and the MSE criterion of DN. Hall et al. (2008) have studied the behavior of RMSC in the condition where all candidate models are weak. They advocate a two-step procedure in which standard identification of $\phi_{\max}$ is first tested following, e.g., the approach of Stock and Yogo (2005) and, only if identification is maintained, can the researchers proceed with RMSC for relevant model selection. Following them, we shall advocate a two-step procedure as well. Antoine and Renault (2020) have recently proposed a test to investigate whether a moment restriction is strong enough to warrant consistent and asymptotically normal GMM estimation. We recommend to first apply this test to $\phi_{\max}$ and only when there is indication of consistency and asymptotic normality that one can apply mRMSC for model selection.

The next result addresses the efficiency of inference post-selection. For this, let us consider the partition $\mathscr{C} = \mathscr{C}_0 \bigcup \mathscr{C}_1$, where $\mathscr{C}_0$ and $\mathscr{C}_1$ are the subsets of elements $c \in \mathscr{C}$ such that $\text{Rank}(\partial \rho_{\max}(\theta_0, c)/\partial \theta') = p$ and $\text{Rank}(\partial \rho_{\max}(\theta_0, c)/\partial \theta') < p$, respectively. We have the following result.

PROPOSITION 4.4. *If: (i) Assumption 5 holds and $\sqrt{T}\bar{\phi}_{\max,T}(\theta_0) \xrightarrow{d} N(0, \Sigma(\phi_{\max}))$, under $\mathbb{P}_T$; (ii) Assumption 6 holds with $\mathscr{C}$ replaced by $\mathscr{C}_0$; and (iii) any $c \in \mathscr{C}_1$*

satisfies Assumption 7. Then, $\hat{c}_T$ converges in probability to c_r and, letting $\hat{\phi} = \phi_{\max}(\cdot, \hat{c}_T)$, we have

$$\Lambda_T(\hat{\phi})R(\hat{\phi})^{-1}\left(\hat{\theta}_T(\hat{\phi}) - \theta_0\right) \xrightarrow{d} N(0, V_\theta(c_r))$$

under $\mathbb{P}_T$, where $\hat{\theta}_T(\hat{\phi})$ is defined by (21).

4.3. Choice of Penalty Function and Robustness

The conditions in Assumption 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 Assumption 6(iv) holds. The following proposition gives the rate of convergence of the asymptotic variance estimator $\hat{V}_\theta(\phi)$ given by (23) for a model candidate ϕ. We consider the case where cross-sectional independent and identically distributed data are involved and the case of weakly dependent time series data.

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

$$\hat{\Sigma}_{iid}(\phi) = \frac{1}{T}\sum_{t=1}^{T} \phi_{tT}(\hat{\theta}_T(\phi))\phi_{tT}(\hat{\theta}_T(\phi))', \quad \phi_{tT}(\theta) = \phi(Y_{tT}, \theta),$$

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,

$$n_{tT}(\theta) = vec\left(\frac{\partial \phi_{tT}}{\partial \theta'}(\theta)\right)\phi_{tT}(\theta)', \quad \text{and} \quad m_{tT} = \phi_{tT}(\theta_0)\phi_{tT}(\theta_0)',$$

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

PROPOSITION 4.5. *Assume that the model ϕ satisfies (3) and that Assumptions 1–3 hold.*

(i) *If $\{Y_{tT} : t = 1, \ldots, T\}$ are independent and identically distributed, $\frac{1}{T}\sum_{t=1}^{T} n_{tT}(\theta)$ converges uniformly in probability to a function $n(\theta)$ in a neighborhood of θ_0, and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(m_{tT} - \mathbb{E}(m_{tT})) = O_P(1)$ under $\mathbb{P}_T$, then*

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

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

(ii) *If* $\{Y_{tT} : t = 1, \ldots, T\}$ *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*

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

under $\mathbb{P}_T$.

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(v) 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 $[0, \frac{1}{2})$, 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 $[0, \frac{1}{2})$ that make these criteria violate the condition.

A natural choice of penalty function to consider is

$$\kappa(|c|,T) = \frac{h(|c|,p)}{(\ln T)^{1+\alpha}}, \quad \text{for some} \quad \alpha > 0, \tag{25}$$

and $h(|c|,p)$ a nonnegative and strictly increasing function of $|c|$ for all values 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 (24), the mRMSC 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 the selection procedure, we can simply write

$$mRMSC(\phi) = \ln\left|\left(\frac{\partial\bar{\phi}_T'}{\partial\theta}(\hat{\theta}_T(\phi))\hat{\Sigma}(\phi)^{-1}\frac{\partial\bar{\phi}_T}{\partial\theta'}(\hat{\theta}_T(\phi))\right)^{-1}\right| + \frac{h(|c|,p)}{(\ln T)^{\alpha}}, \tag{26}$$

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 4.5(ii) for the asymptotic variance estimator.

While the best choice of α in (25) may be of independent interest that we shall pursue in future work, it is of interest to mention that $\alpha > 0$ is important as a condition to ensure that the penalty function is smaller than the information component. Moreover, 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 the next section, we have set $\alpha = 0.1$ and use $h(|c|, p) = 1 - \frac{p}{|c|}$.[3]

5. SIMULATION RESULTS

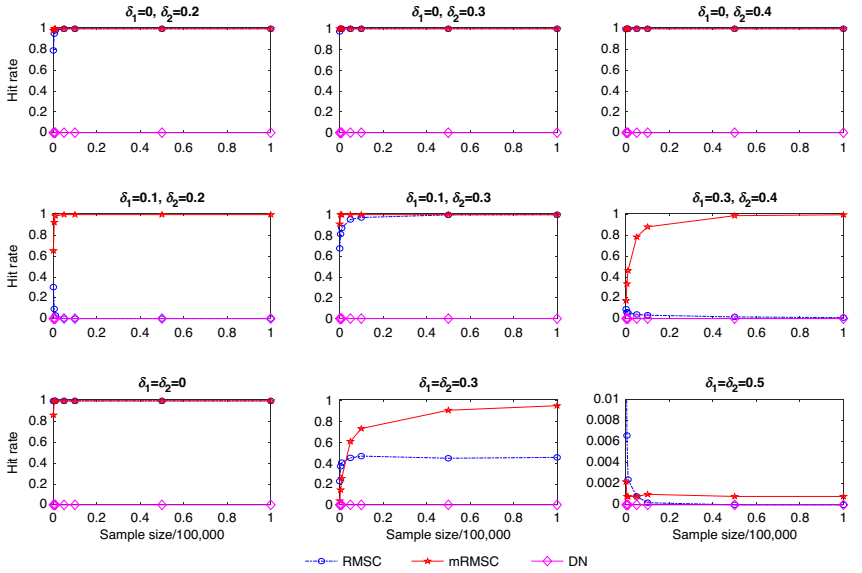
In this section, we study the finite-sample performance of the proposed selection criterion (mRMSC) and the post-selection properties of the GMM estimator (bias, MSE, and coverage rate of confidence sets) through a Monte Carlo experiment. For this purpose, we use the same simulation setup of Section 3 but increase the set of candidate instruments to 12 (i.e., $z_1, z_2, \ldots, z_{12}$). For clarity, the analysis on the performance of the selection criterion (mRMSC) is separated from that on the post-selection properties of the GMM estimator.

5.1. Performance of the mRMSC

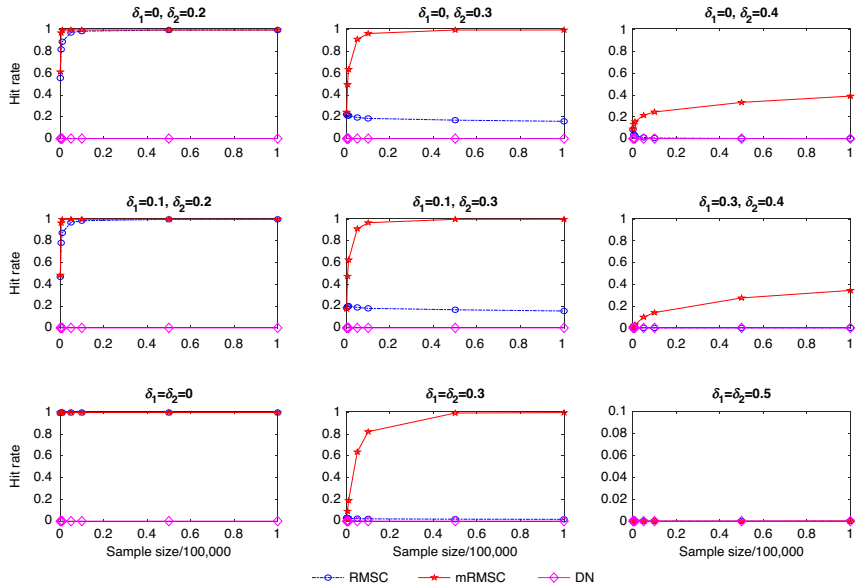
In this section, we compare the finite-sample performance of the proposed mRMSC with existing methods in the literature, namely, the RMSC of Hall et al. (2007) and the MSE-based criterion of DN. Since entropy-based selection criteria (mRMSC and RMSC) are conceptually different from the MSE type-selection criteria, the inclusion of the DN criterion is useful to determine which types of criteria perform the best, at least from the finite-sample perspective.

Figure 2 contains the results for both the model with one endogenous regressor ($p = 1$; Figure 2a) and the model with two endogenous regressors ($p = 2$; Figure 2b). Each subfigure shows, for a combination of identification strengths (i.e., the values of $\delta_i, i = 1, 2$), the plots of the proportion of best model selection (*hit rate*) by sample size. Specifically, the first two rows in Figure 2a,b report the results where the two instruments z_1 and z_2 have different identification strength ($\delta_1 < \delta_2$), whereas the last row contains the plots of the hit rates where both instruments z_1 and z_2 have equal identification strength (i.e., $\delta_1 = \delta_2$). As part of this, we include the case $\delta_1 = \delta_2 = 0.5$ to assess selection performance when the sample moment of the estimation function does not accumulate sufficient information to allow for consistent point estimation as the sample size grows (see, e.g., Staiger and Stock, 1997). Three main results stand out from this exercise.

[3]The post-selection cross-validation prediction performance of mRMSC is assessed for different values of α in Appendix S2 of the Supplementary Material. The best performance is reached for $\alpha \in [0.1, 1.0]$, which motivates our choice for $\alpha = 0.1$.



(a) Model with $p = 1$



(b) Model with $p = 2$

FIGURE 2. Hit rate of mRMSC, RMSC, and DN. Sample size $T = 100$, 200, 500, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000; number of replications: 5,000.

First, with the exception of the Staiger and Stock (1997) weak identification setup ($\delta_1 = \delta_2 = 0.5$), the *hit rate* of mRMSC increases to 1 in all cases as the sample size grows except for very few cases where its convergence is expectedly very slow. 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 and DN drops to 0 or plateaus way below 1.0, highlighting the limitation of these criteria at consistently selecting the correct model when operating on models with poor identification strength. Looking specifically at the DN criterion, the *hit rate* is almost flat at 0 in all cases considered, including when standard strong identification holds ($\delta_1 = \delta_2 = 0$). This is because DN is not based on maximizing the entropy of GMM estimator asymptotic distribution and does not penalize larger models either. As a result, the DN criterion always tends to select models that include irrelevant instruments (see Tables 1 and 2). Regarding RMSC, the cases where the *hit rate* drops to 0 or plateaus way below 1.0 are seen clearly in the subfigures "$\delta_1 = 0.3, \delta_2 = 0.4$" and "$\delta_1 = \delta_2 = 0.3$" of Figure 2a. The lackluster performance of RMSC and DN is more pronounced in models with two endogenous regressors (Figure 2b). In this case, RMSC seems to be consistent only when the model is strongly identified (i.e., "$\delta_1 = \delta_2 = 0$") or close to being so (e.g., "$\delta_1 = 0, \delta_2 = 0.2$" or "$\delta_1 = 0.1, \delta_2 = 0.2$"), whereas DN almost never selects the correct model even when identification is strong. Clearly, DN seems to be the least-performing criterion in selecting the correct model. In both Figure 2a (model with one endogenous regressor) and Figure 2b (model with two endogenous regressors), the *hit rate* of all selection criteria is almost flat at 0 when $\delta_1 = \delta_2 = 0.5$, the weak identification framework.

Second, when θ is strongly identified, RMSC performs slightly better than mRMSC for small sample sizes $T = 100, 200$ when $p = 1$ (subfigure "$\delta_1 = \delta_2 = 0$" in Figure 2a), but this gap vanishes in models with two endogenous regressors (subfigure "$\delta_1 = \delta_2 = 0$" in Figure 2b). As the sample size increases, the proportion of correct model selection of both mRMSC and RMSC approaches quickly 1.0. See, e.g., the subfigures "$\delta_1 = \delta_2 = 0$" in Figure 2a,b.

Third, as the identification strength deteriorates, that is, $\delta_1 = \delta_2 = 0.3$ or $\delta_2 > \delta_1 \geq 0.3$ in Figure 2a, mRMSC expectedly outperforms RMSC. This dominance of mRMSC is even more pronounced and systematic in models with two endogenous variables (see all subfigures "$\delta_1 = 0, \delta_2 \geq 0.3$," "$\delta_1 = 0.1, \delta_2 \geq 0.3$," and "$\delta_1 = \delta_2 = 0.3$" in Figure 2b). Overall, this simulation exercise illustrates that our mRMSC performs well even with moderately large to large values of δ_i $(i = 1, 2)$, whereas the RMSC and the DN fail to handle these cases, as per their declining *hit rate* as T increases for high values of δ_i.

While Figure 2 focuses only on *hit rates*, more results are presented by Tables 1 and 2, which show in detail the empirical selection probabilities of the candidate models. These tables contain the results of RMSC, DN, and mRMSC for sample sizes $T = 100$, 1,000, 5,000, 50,000. The results with one endogenous regressor ($p = 1$) are presented in Table 1, whereas those with two endoge-

TABLE 1. Empirical selection probabilities: One endogenous regressor ($p = 1$), $T = 100$, 1,000, 5,000, 50,000.

			$T = 100$										$T = 1{,}000$									
	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	0.2	0.79	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.2	0.30	0.00	0.67	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0	0.3	0.97	0.00	0.03	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	0.3	0.68	0.00	0.28	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.87	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	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
	0.3	0.4	0.09	0.02	0.12	0.22	0.09	0.08	0.01	0.09	0.28	0.00	0.05	0.00	0.13	0.13	0.01	0.21	0.12	0.03	0.31	0.00
mRMSC	0	0.2	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
	0.1	0.2	0.65	0.02	0.27	0.00	0.00	0.02	0.01	0.00	0.03	0.00	0.99	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0	0.3	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	0.3	0.91	0.00	0.05	0.00	0.00	0.01	0.00	0.00	0.03	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	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
	0.3	0.4	0.17	0.04	0.02	0.02	0.01	0.01	0.01	0.04	0.68	0.01	0.46	0.03	0.04	0.01	0.00	0.02	0.02	0.02	0.39	0.01
DN	0	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33
	0.1	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33
	0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33

(Continues)

TABLE 1. Continued

			$T = 100$										$T = 1,000$									
	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
	0.1	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32
	0	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.28
	0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.84	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.18
	$\delta_1 = \delta_2$																					
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
	0.3	0.3	0.07	0.06	0.23	0.15	0.14	0.11	0.01	0.04	0.20	0.00	0.01	0.01	0.41	0.02	0.02	0.33	0.10	0.01	0.09	0.00
	0.5	0.5	0.01	0.01	0.02	0.10	0.10	0.04	0.01	0.32	0.38	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.06	0.20	0.69	0.00
mRMSC	0	0	0.07	0.07	0.86	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	0.13	0.13	0.04	0.01	0.01	0.02	0.02	0.02	0.61	0.01	0.18	0.18	0.26	0.00	0.00	0.05	0.03	0.00	0.30	0.00
	0.5	0.5	0.03	0.02	0.00	0.01	0.01	0.00	0.01	0.10	0.80	0.02	0.02	0.03	0.00	0.01	0.01	0.00	0.01	0.12	0.80	0.02
DN	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33
	0.3	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.80	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.74	0.26
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.76	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.79	0.16

(Continues)

TABLE 1. Continued

			$T = 5,000$										$T = 50,000$									
	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	0.2	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	0.2	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	0.3	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	0.3	0.95	0.00	0.05	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	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	0.04	0.00	0.13	0.08	0.00	0.25	0.19	0.00	0.30	0.00	0.02	0.00	0.13	0.04	0.00	0.27	0.25	0.00	0.30	0.00
mRMSC	0	0.2	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	0.2	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	0.3	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	0.3	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	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	0.79	0.01	0.06	0.00	0.00	0.01	0.01	0.00	0.11	0.00	0.99	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DN	0	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	0.1	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	0.1	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	0	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.30
	0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.28

(Continues)

TABLE 1. Continued

			$T = 5,000$										$T = 50,000$									
	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	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	0.45	0.00	0.00	0.38	0.13	0.00	0.04	0.00	0.00	0.00	0.45	0.00	0.00	0.39	0.13	0.00	0.03	0.00
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.15	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.90	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
	0.3	0.3	0.14	0.14	0.61	0.00	0.00	0.03	0.01	0.00	0.06	0.00	0.05	0.04	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	0.5	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.11	0.81	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.10	0.82	0.03
DN	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.32
	0.3	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.78	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.79	0.16

Notes: "z1 + z2" denotes models with the two 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 two instruments z1 and z2 + 1 irrelevant instrument; "z1 + z2 + 2I" denotes models with the two 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 one 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: Two endogenous regressors ($p = 2$), $T = 100$, 1,000, 5,000, 50,000.

			$T = 100$								$T = 1{,}000$							
	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	0.2	0.56	0.00	0.00	0.34	0.08	0.00	0.02	0.00	0.89	0.00	0.00	0.10	0.01	0.00	0.00	0.00
	0.1	0.2	0.47	0.00	0.00	0.38	0.13	0.00	0.03	0.00	0.87	0.00	0.00	0.12	0.01	0.00	0.00	0.00
	0	0.3	0.22	0.01	0.00	0.40	0.24	0.00	0.12	0.00	0.21	0.00	0.00	0.37	0.28	0.00	0.14	0.00
	0.1	0.3	0.18	0.01	0.00	0.38	0.28	0.00	0.15	0.00	0.20	0.00	0.00	0.36	0.28	0.00	0.16	0.00
	0	0.4	0.08	0.01	0.00	0.30	0.29	0.00	0.31	0.00	0.03	0.00	0.00	0.15	0.30	0.00	0.53	0.00
	0.3	0.4	0.01	0.00	0.00	0.08	0.22	0.02	0.67	0.00	0.00	0.00	0.00	0.02	0.10	0.00	0.87	0.00
mRMSC	0	0.2	0.61	0.00	0.00	0.10	0.05	0.00	0.24	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.2	0.49	0.00	0.00	0.07	0.04	0.00	0.38	0.02	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0	0.3	0.24	0.00	0.00	0.08	0.07	0.00	0.58	0.02	0.64	0.00	0.00	0.09	0.06	0.00	0.22	0.00
	0.1	0.3	0.18	0.00	0.00	0.05	0.04	0.00	0.68	0.05	0.63	0.00	0.00	0.08	0.06	0.00	0.23	0.00
	0	0.4	0.09	0.01	0.00	0.05	0.05	0.00	0.76	0.03	0.16	0.01	0.00	0.10	0.08	0.00	0.66	0.01
	0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.15	0.03	0.00	0.00	0.01	0.01	0.00	0.83	0.12
DN	0	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.18
	0.1	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.74	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.81	0.19
	0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.78	0.22

(Continues)

TABLE 2. Continued

			$T = 100$								$T = 1{,}000$							
	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
	0.1	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.25
	0	0.4	0.00	0.00	0.00	0.00	0.00	0.01	0.59	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.36
	0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.01	0.77	0.22	0.00	0.00	0.00	0.00	0.00	0.01	0.75	0.25
	$\delta_1 = \delta_2$																	
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
	0.3	0.3	0.03	0.00	0.00	0.16	0.31	0.01	0.49	0.00	0.03	0.00	0.00	0.13	0.26	0.00	0.59	0.00
	0.5	0.5	0.00	0.00	0.00	0.01	0.06	0.12	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.94	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
	0.3	0.3	0.02	0.00	0.00	0.01	0.01	0.00	0.83	0.13	0.19	0.00	0.00	0.04	0.02	0.00	0.68	0.07
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.18	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.18
DN	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.21
	0.3	0.3	0.00	0.00	0.00	0.00	0.00	0.01	0.78	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.19
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.02	0.79	0.19	0.00	0.00	0.00	0.00	0.00	0.02	0.80	0.18

(Continues)

TABLE 2. Continued

			$T = 5,000$								$T = 50,000$							
	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	0.2	0.98	0.00	0.00	0.02	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	0.2	0.97	0.00	0.00	0.03	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	0.3	0.19	0.00	0.00	0.36	0.27	0.00	0.17	0.00	0.17	0.00	0.00	0.34	0.28	0.00	0.21	0.00
	0.1	0.3	0.19	0.00	0.00	0.35	0.28	0.00	0.18	0.00	0.17	0.00	0.00	0.34	0.28	0.00	0.21	0.00
	0	0.4	0.01	0.00	0.00	0.06	0.18	0.00	0.75	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.94	0.00
	0.3	0.4	0.00	0.00	0.00	0.01	0.05	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.98	0.00
mRMSC	0	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
	0.1	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	0.3	0.92	0.00	0.00	0.04	0.01	0.00	0.03	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.1	0.3	0.91	0.00	0.00	0.04	0.01	0.00	0.04	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0	0.4	0.21	0.00	0.00	0.10	0.09	0.00	0.59	0.00	0.33	0.00	0.00	0.13	0.08	0.00	0.45	0.00
	0.3	0.4	0.10	0.00	0.00	0.03	0.03	0.00	0.77	0.08	0.28	0.00	0.00	0.09	0.06	0.00	0.55	0.02
DN	0	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.16
	0.1	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.16
	0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.16
	0.1	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.17
	0	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.74	0.26
	0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.30

(Continues)

TABLE 2. Continued

			$T = 5{,}000$							$T = 50{,}000$							
	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	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	0.02	0.00	0.00	0.11	0.24	0.00	0.63	0.00	0.02	0.00	0.00	0.09	0.21	0.00	0.68	0.00
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.03	0.96	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.90	0.09
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
	0.3	0.3	0.64	0.00	0.00	0.06	0.03	0.00	0.26	0.02	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.02	0.80	0.18	0.00	0.00	0.00	0.00	0.00	0.02	0.80	0.19
DN	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.20
	0.3	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.21
	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.02	0.80	0.18	0.00	0.00	0.00	0.00	0.00	0.02	0.80	0.18

Notes: "z1 + z2" denotes models with the two 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 two instruments z1 and z2 + 1 irrelevant instrument; "z1 + z2 + 2I" denotes models with the two 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 one 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.

nous regressors ($p = 2$) are presented in Table 2. More specifically, each table indicates, for each sample size, the empirical selection rates of all possible models for a given criterion (RMSC, DN, or mRMSC) and given values of δ_i ($i = 1, 2$).

Considering first the *case with one endogenous regressor* (Table 1), we see that when $\delta_1 < \delta_2$ (first part of the table for each sample size), mRMSC outperforms RMSC even for relatively small sample sizes, and the latter dominates DN, which almost never selects the correct model. For example, when $T = 100$ and "$\delta_1 = 0.1 < \delta_2 = 0.3$," RMSC only selects the relevant model (i.e., columns "z1" in Table 1 for $T = 100$) 68% of the time, whereas mRMSC selects this model 91% of the time. As the sample size increases to $T = 50{,}000$, these empirical selection probabilities bounce to 100% for both RMSC and mRMSC. The empirical selection probabilities of DN remain flat at 0 along T.

Furthermore, looking at columns "z1" in Table 1 for $\delta_1 < \delta_2$ (first part of the table), it is obvious that the dominance of mRMSC is even more pronounced when "$\delta_1 = 0.3 < \delta_2 = 0.4$" (i.e., when identification strength deteriorates) regardless of the sample size, with this dominance becoming even more visible as the sample size increases. For example, when "$\delta_1 = 0.3 < \delta_2 = 0.4$," RMSC only selects the relevant model 9% of the time when $T = 100$, whereas mRMSC selects this model 17% of the time. As the sample size increases to $T = 50{,}000$, the empirical selection probability for RMSC decreases drastically to 2%, whereas that of mRMSC bounces to 99%. Clearly, we see that as identification weakens, RMSC has a tendency to often select less relevant and less sparse models for small samples (see the selection probabilities in columns "All I," "z1+z2," "z1+I," "z2+I," and "zj+more" in Table 1 for $T = 100$) or less sparse models with at least one of both instruments z_1 and z_2 (see, e.g., the selection probabilities in columns "z1+z2" in Table 1 for $T = 1{,}000$, 5,000, 50,000). Meanwhile, mRMSC still has an overall good performance in selecting the more relevant model. Now, when $\delta_1 = \delta_2$ (second part of Table 1 for each sample size), both RMSC and mRMSC perform relatively well in selecting the correct model (i.e., columns "z1+z2" of the tables) even with moderate identification strength ($\delta_1 = \delta_2 \leq 0.3$), but RMSC performs slightly better for sample sizes $T = 100$, 1,000, whereas this dominance is reversed for larger sample sizes ($T = 5{,}000$, 50,000). As identification deteriorates (see the selection probabilities in columns "$\delta_1 = \delta_2 = 0.3$" in Table 1 for $T = 5{,}000$, 50,000), mRMSC improves substantially over RMSC when $T = 5{,}000$, 50,000. However, all criteria become weaker in selecting the correct model when identification is weak (see columns "$\delta_1 = \delta_2 = 0.5$" in Table 1).

We now consider the *case with two endogenous regressors* (Table 2) where the most relevant model is column "z1+z2." We see that for both "$\delta_1 < \delta_2$" and "$\delta_1 = \delta_2$," mRMSC outperforms RMSC in most combinations of identification strength δ_i ($i = 1, 2$), especially when the sample size increases ($T = 1{,}000$, 5,000, 50,000). Again, RMSC shows a tendency to often select less relevant models when identification deteriorates (i.e., high values of $\delta_i, i = 1, 2$). In addition, the empirical selection probabilities of the relevant model increase with the sample size for

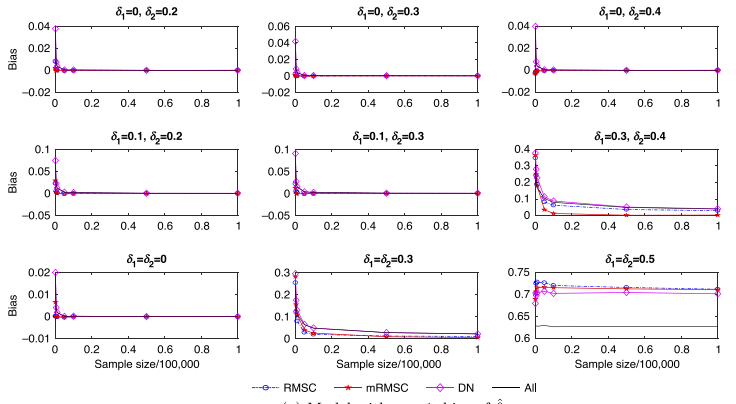
mRMSC for all combinations of identification strength used, whereas those of RMSC often decrease as the sample size increases for high values of δ_i ($i = 1, 2$). This illustrates why the aggregate *hit rate* of RMSC decreases as the sample size increases for high values of δ_i, as shown in Figure 2. Remarkably, DN never selects the correct model even when identification is strong ($\delta_1 = \delta_2 = 0$).

5.2. Performance of Ppost-Selection Inference

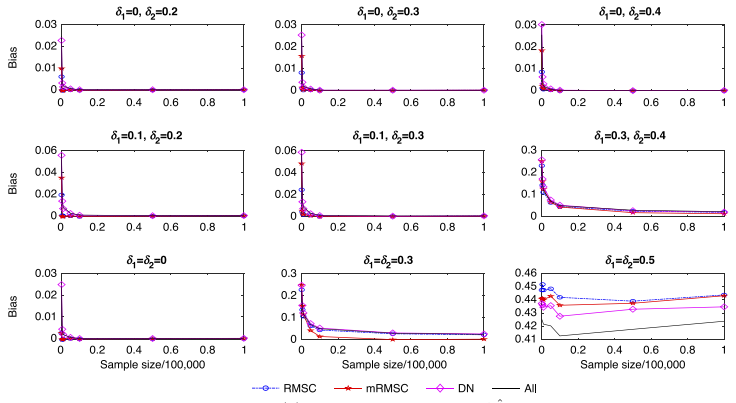
We now investigate the bias, MSE, and coverage rate of confidence sets of the GMM estimator post selection. In addition to the criteria under consideration, namely, mRMSC, RMSC, and DN, we also analyze the performance of the (naive) GMM estimator that uses all available instruments. This naive GMM model corresponds to the estimating function $\phi_{\max}$ in Section 4.2. Its inclusion allows us to illustrate the importance of moment selection in GMM models with relatively poor identification strength, even when the set of moment conditions available is not large.

For clarity, let us focus first on the bias and MSE of the post-selection GMM estimator $\hat{\theta}$. We consider both models with one endogenous regressor ($p = 1$) and two endogenous regressors ($p = 2$). In the latter case, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)'$ has two components, so we shall show the bias and MSE results of both components. Figures 3 and 4 report the results where the bias and MSE of the post-selection estimators $\hat{\theta}$ (for $p = 1$) and $\hat{\theta}_1, \hat{\theta}_2$ (for $p = 2$) are plotted against the sample size/100,000 with the various selection criteria for sample sizes $T =$ 100, 200, 500, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000.

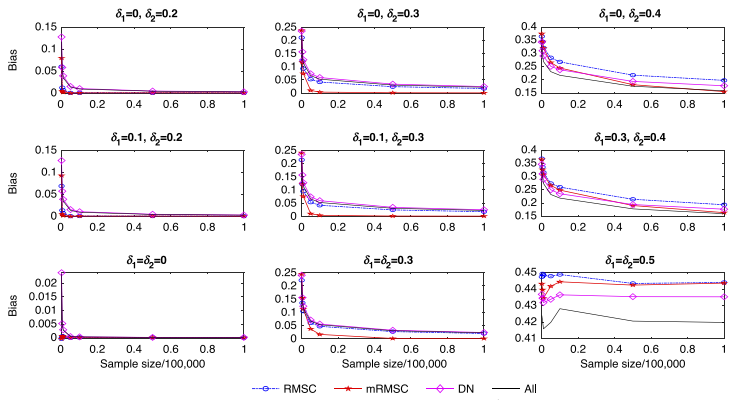
Looking first at the plots of the bias in Figure 3, we see similar patterns in the behavior of the GMM estimator for both $p = 1$ (Figure 3a) and $p = 2$ (Figure 3b,c). Considering the case where $p = 1$ (Figure 3a), with the exception of the Staiger and Stock (1997) weak identification setup ("$\delta_1 = \delta_2 = 0.5$"), the post-selection GMM estimator with DN has the highest bias for most combinations of the identification strength δ_i ($i = 1, 2$). Both mRMSC and RMSC outperform the naive GMM estimator except when identification is weak ("$\delta_1 = \delta_2 = 0.5$"). The dominance of the naive estimator under weak identification is in line with the widely documented weak IV literature (see, e.g., Chao and Swanson, 2005; Andrews and Stock, 2007). For all the combinations of the identification strength $\delta_1 = \delta_2 = 0.3$ and $\delta_1 < \delta_2 \leq 0.4$ shown in Figure 3a, mRMSC dominates or performs as well as RMSC. The dominance of the mRMSC is especially pronounced when $\delta_1 = 0.3, \delta_2 = 0.4$, where the bias of the post-selection GMM estimator $\hat{\theta}$ with mRMSC vanishes as the sample size increases, whereas the bias of that of the RMSC plateaus far away from zero. Second, for the models with $p = 2$ (Figure 3b,c), the bias of the GMM estimator $\hat{\theta}_1$ (the strongest identified component of θ) across various selection criteria is quite similar to the results with $p = 1$ depicted in Figure 3a, with the exception of subfigure "$\delta_1 = \delta_2 = 0.3$" where mRMSC's dominance is even clearer. Looking at the bias of $\hat{\theta}_2$—the least identified component of θ (Figure 3c)—we observe that in most cases the bias of the post-selection GMM estimator $\hat{\theta}_2$ with mRMSC is smaller than the ones resulting from both RMSC and DN. While



(a) Model with $p = 1$: bias of $\hat{\theta}$

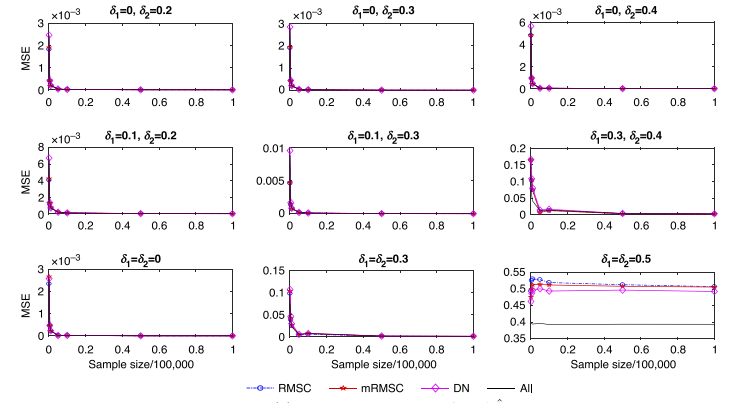


(b) Model with $p = 2$: bias of $\hat{\theta}_1$

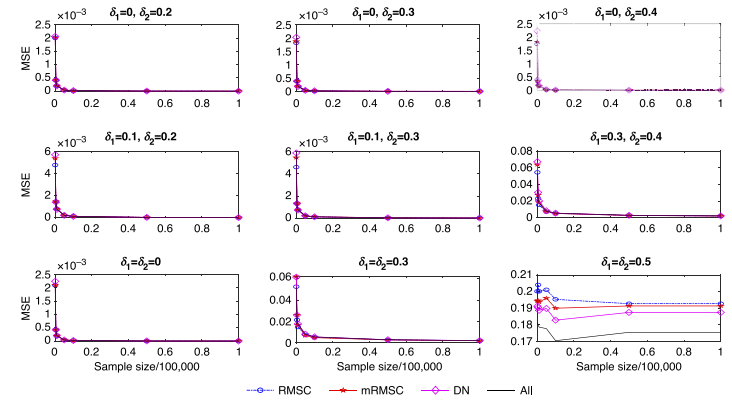


(c) Model with $p = 2$: bias of $\hat{\theta}_2$

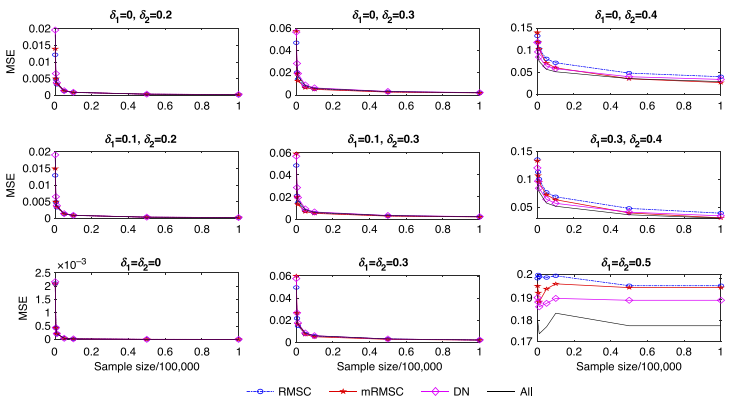
FIGURE 3. Bias of GMM post selection with mRMSC, RMSC, and DN, and GMM with full set of IVs. Sample size $T = 100, 200, 500, 1{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000, 100{,}000$; number of replications: 5,000.



(a) Model with $p = 1$: MSE of $\hat{\theta}$



(b) Model with $p = 2$: MSE of $\hat{\theta}_1$



(c) Model with $p = 2$: MSE of $\hat{\theta}_2$

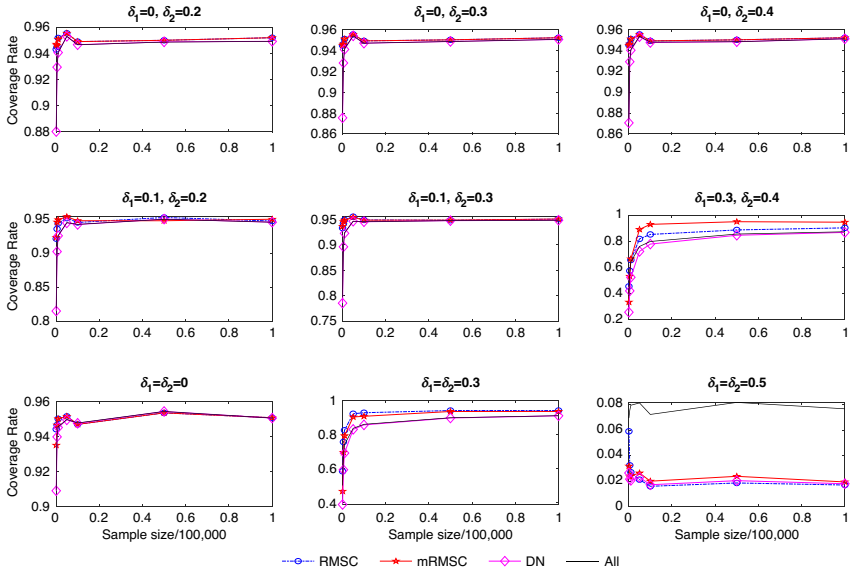
FIGURE 4. MSE of GMM post selection with mRMSC, RMSC, and DN, and GMM with full set of IVs. Sample size $T = 100$, 200, 500, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000; number of replications: 5,000.

the naive GMM estimation appears competitive at mitigating the bias of $\hat{\theta}_2$ when $\delta_1 = \delta_2 = 0.5$, it does not perform as well as mRMSC when $0.1 < \delta_1 < \delta_2 \leq 0.4$. Interestingly, the post-selection GMM estimator with mRMSC performs as well as the naive GMM estimator as the sample size increases in the rare cases where the latter is competitive.

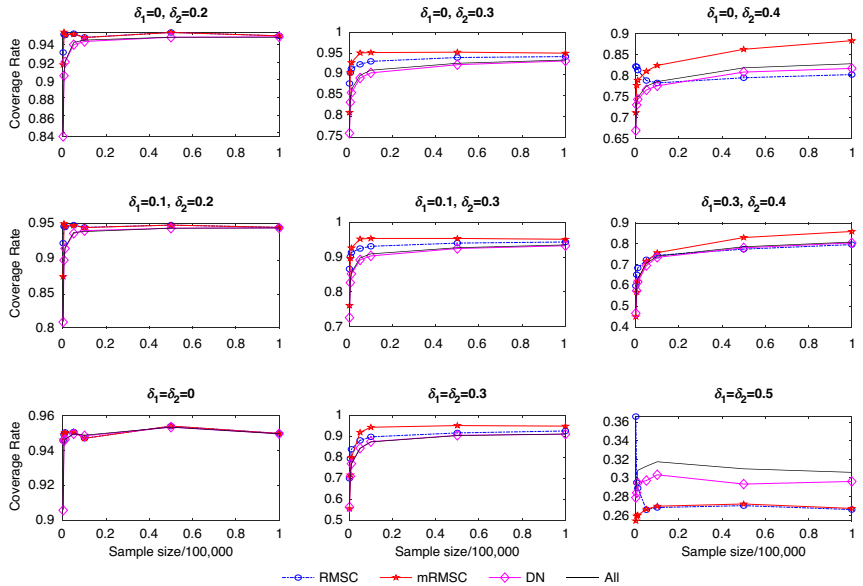
Let us now focus on the MSE results in Figure 4. Considering the model with $p = 1$ (Figure 4a), we see that all selection criteria (mRMSC, RMSC, and DN) perform quite similarly. In particular, when $\delta_1 < \delta_2 \leq 0.4$ or $\delta_1 = \delta_2 < 0.5$, mRMSC and RMSC have a slight edge on DN when the sample size is small. However, this advantage of mRMSC and RMSC disappears as the sample size increases for a similar level of performance. When $\delta_1 = \delta_2 = 0.5$ (i.e., under weak identification), all the models perform poorly with the naive GMM displaying the smallest MSE, followed by the post-selection GMM estimators with DN, mRMSC, and RMSC, respectively. Considering now the model with $p = 2$ (Figure 4b,c), we observe again that the MSE results of $\hat{\theta}_1$ are quite similar to those in Figure 4a for $p = 1$. However, the MSE of the estimator $\hat{\theta}_2$ (the estimator of the weakest identified component) depicts a different picture. Indeed, there are many instances in Figure 4c where the MSEs of the post-selection GMM estimators with mRMSC and RMSC are smaller than that with DN, especially for small samples. As identification deteriorates (see subfigures "$\delta_1 = 0, \delta_2 = 0.4$" and "$\delta_1 = 0.3$, $\delta_2 = 0.4$"), post-selection GMM-RMSC is dominated by GMM-mRMSC, which also matches both the naive GMM and GMM-DN as the sample size increases.

Aside the bias and MSE, one of the important properties of post-selection inference is whether a given selection criterion leads to confidence sets with correct coverage post selection. It is well known that standard selection methods based on information criteria, such as Akaike information criterion (AIC) and BIC, do not enjoy this property (see, e.g., Kabaila and Leeb, 2006). To investigate this further, we consider both the models with $p = 1$ and $p = 2$ and explore the coverage rate of Wald-type confidence intervals based on the post-selection or naive GMM estimators $\hat{\theta}$. As θ is a scalar when $p = 1$, we consider t-type confidence intervals based on the post-selection or naive GMM estimator in that case. However, for $p = 2$, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)' \in \mathbb{R}^2$, so we shall consider Wald-type joint confidence sets based on the post-selection or naive GMM estimator $\hat{\theta}$. As in the previous sections, we report the results across various selection criteria (mRMSC, RMSC, and DN), along with the naive GMM estimator that utilizes all 12 available instruments. Figure 5 shows the results, where the coverage rates are plotted against the sample size. The nominal confidence level is set to 95%, but the results are not sensitive to alternative choices of this significance level.

Considering first the model with $p = 1$ (Figure 5a), two main observations stand out. First, when $\delta_1 < \delta_2$, the post-selection GMM-mRMSC outperforms or performs as well as both the naive GMM and the GMM-RMSC and GMM-DN. As identification deteriorates, the dominance of post-selection GMM-mRMSC is visible (see, e.g., subfigures "$\delta_1 = 0.3, \delta_2 = 0.4$" in Figure 5a). In the latter case, the coverage rate of confidence intervals from post-selection GMM-mRMSC



(a) Model with $p = 1$



(b) Model with $p = 2$

FIGURE 5. Coverage rate of confidence sets at nominal level 95%: GMM post selection with mRMSC, RMSC, and DN, and GMM with full set of IVs. Sample size $T = 100$, 200, 500, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000; number of replications: 5,000.

approaches 95% as the sample size increases, whereas naive GMM, GMM-RMSC, and GMM-DN have coverage rates that plateau slightly above 85% even for a sample size as large as $T = 100,000$. Second, when $\delta_1 = \delta_2$, with the exception of the weak identification case ($\delta_1 = \delta_2 = 0.5$), GMM-mRMSC performs as well as GMM-RMSC and both dominate GMM-DN and naive GMM, with greater magnitude for small sample sizes (see, e.g., subfigure "$\delta_1 = \delta_2 = 0.3$" in Figure 5a).

Moving to models with two endogenous variables, $p = 2$ (Figure 5b), the dominance of post-selection GMM-mRMSC is even more noticeable. With the exception of the weak identification case ($\delta_1 = \delta_2 = 0.5$), the edge of GMM-mRMSC on GMM-RMSC, GMM-DN, and the naive GMM is remarkable. For example, when $\delta_1 = 0.3, \delta_2 = 0.4$, the coverage rate of the joint confidence sets with GMM-RMSC, GMM-DN, and the naive GMM plateaus below 80% as the sample size increases. Meanwhile, the coverage rate of joint confidence sets with GMM-mRMSC continue to increase with the sample size. At many instances in Figure 5b, GMM-RMSC outperforms both naive GMM and GMM-DN, although it is not as competitive as GMM-mRMSC. Such results are shown in all subfigures with $\delta_1 = 0.3$. Clearly, this simulation exercise illustrates the overall good performance of our post-selection GMM-mRMSC compared with post-selection GMM-RMSC and DN, as well as the naive GMM.

Note that while the DGP in Sections 5.1 and 5.2 imposes z_1 to be uncorrelated z_2, we report in the Supplementary Material experiment results for correlated relevant instruments. These results are qualitatively the same as those reported in Sections 5.1 and 5.2. (See Figures S1.1–S1.4 and Table S1.1 in the Supplementary Material.) Moreover, note that Figures 3 and 4 may not visually distinguish the methods because of their close performance magnitudes in many instances. To provide more readability, the Supplementary Material includes tables of the bias, MSE, and coverage rate of the joint confidence sets plotted in those figures. (See Table S1.2 for correlated instruments and Table S1.3 for uncorrelated instruments in the Supplementary Material.)

6. CONCLUSION

In this paper, we study model selection in moment condition models with mixed identification strength that allow for consistent and asymptotically normal parameter estimation. Our investigation reveals that standard model selection procedures, such as the relevant model selection criterion 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 introduce new entropy-based relevant moment selection criteria, the mRMSC. Similar to RMSC, mRMSC are evaluated using the two-step GMM estimator, which has linear reparameterizations known to be efficient in this framework as well (see Dovonon et al., 2022). In the case of the multivariate parameter, the asymptotic distribution of this estimator is, in general, characterized by directions of fast convergence and directions of slow convergence rate. The best or relevant model is the smallest

model (in terms of number of moment restrictions) that delivers the same rate of convergence and the same asymptotic variance as those obtained when all the moment restrictions are used.

By construction, mRMSC first rewards the rate of estimation and then, for models with the same rate, it rewards (negative) entropy. In addition, suitable penalty terms are introduced that guarantee the consistency of the selection procedure. Conditions under which mRMSC lead to consistent selection of the best model are outlined, and we show that this new selection procedure is robust to the presence of uninformative and weak models.

We illustrate the finite-sample performance of the proposed method through Monte Carlo simulations. In addition to mRMSC and RMSC, we also consider the MSE criterion of DN along with the moment condition model including all the available instruments serving as a benchmark. In almost all the considered Monte Carlo designs, mRMSC dominates the other selection criteria in terms of hit rate. The post-selection performance is also investigated, revealing that mRMSC-selected models produce confidence intervals with the best coverage probability in most of the Monte Carlo designs. These models are also among those with the smallest bias and MSE. Nevertheless, when all the available moment restrictions are weak, mRMSC performs quite poorly as the other selection methods. In this case, the model with all instruments performs marginally better than all the selected ones although unreliably.

APPENDIX A. Convergence Rate of HAC Using Slow Estimators

As we have seen, 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 $\phi(Y_{tT},\theta_0)$: $\Sigma = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi(Y_{tT},\theta_0)\right)$, with variance taken under $\mathbb{P}_T$, when estimators of θ_0 based on the moment condition $\mathbb{E}(\phi(Y_{tT},\theta_0)) = 0$ 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 $R(j) = \mathbb{E}(\phi(Y_{tT},\theta_0)\phi(Y_{t-j,T},\theta_0)')$, $j \in \mathbb{Z}$, the autocovariance function of $\phi(Y_{tT},\theta_0)$, which is assumed to be covariance stationary, where we recall that $\mathbb{E}(\cdot)$ stands for expectation taken under $\mathbb{P}_T$. We have the following.

PROPOSITION A.1. *Assume that*

(i) $z_{tT} = \left(\phi_{tT}(\theta_0)', vec[(\partial/\partial\theta')\phi_{tT}(\theta_0) - \mathbb{E}(\partial/\partial\theta')\phi_{tT}(\theta_0)]'\right)'$ *is fourth-order stationary with autocovariance function under* $\mathbb{P}_T$ *(i.e.,* $j \mapsto R(j) = \mathbb{E}(z_{tT}z_{t-j,T})$*) and fourth-order cumulant function under* $\mathbb{P}_T$ *not T dependent.*

(i) *Assumptions B and C of Andrews (1991) hold with* $V_t(\theta)$ *replaced by* $\phi_{tT}(\theta)$*, the supremum* $\sup_{t\geq 1}(\cdot)$ *in B(ii), B(iii), and C(ii) replaced by* $\sup_{T\geq 1}\sup_{1\leq t\leq T}(\cdot)$*, and B(i) replaced by*

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

under $\mathbb{P}_T$ *for some* $\delta\geq 0$.

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

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

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

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

Then,

$$\sqrt{\frac{T}{\ell_T}}\left(\hat{\Sigma}_{hac}-\Sigma\right)=O_P(1),\quad \text{under } \mathbb{P}_T.$$

Proof of Proposition A.1. Let

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

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

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

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

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

Thus, 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, Thm. 1(a) and (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, Thm. 1(a) and (b)) hold and a close consideration of his proof reveals that we only need to show that

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

under $\mathbb{P}_T$ to conclude (A.1). Similar to Andrews (1991, eqn. (A.11)), a two-term Taylor expansion gives

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

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

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

where the $O_P(\cdot)$ holds under $\mathbb{P}_T$. Furthermore, we have

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

with $\lambda = \mathbb{E}(\partial/\partial\theta)\phi(Y_{tT},\theta_0)$ and

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

Clearly, the first two terms in the expansion of L_{1T} are of order $O_P(T^{\delta}/\sqrt{\ell_T})$ under $\mathbb{P}_T$. Furthermore, from Andrews (1991, eqn. (A.15)), we can claim that $D_T = O_P(\sqrt{\ell_T/T})$

under $\mathbb{P}_T$. As a result,

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

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

APPENDIX B. Auxiliary Results and Proofs

LEMMA B.1. *Let* $s_1(\iota_{\max}) = Rank\left(\frac{\partial \rho_{\max,1}}{\partial \theta'}(\theta_0)\right)$ *and* $R = \left(R_1 \vdots R_2\right)$ *a nonsingular* (p,p)*-matrix such that* $RR' = I_p$*, with the columns of* R_1 *spanning the range of* $\frac{\partial \rho'_{\max,1}}{\partial \theta}(\theta_0)$ *and* $\frac{\partial \rho_{\max,1}}{\partial \theta'}(\theta_0)R_2 = 0$*. Let* $c = (c_1', c_2')' \in \mathscr{C}$*. If* $s_1(c) = s_1(\iota_{\max})$*, that is,* $Rank\left(\frac{\partial \rho_{\max,1}}{\partial \theta'}(\theta_0)\right) = Rank\left(\frac{\partial \rho_{\max,1}}{\partial \theta'}(\theta_0, c_1)\right)$*, then the columns of* R_1 *span the range of* $\frac{\partial \rho'_{\max,1}}{\partial \theta}(\theta_0, c_1)$ *and* $\frac{\partial \rho_{\max,1}}{\partial \theta'}(\theta_0, c_1)R_2 = 0$.

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

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

Thus,

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

Furthermore,

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

(i) follows from (B.1)–(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, $\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} = T^{-1+2\delta_1}\left[(C_1'\Delta_{11}C_1)^{-1} + o_P(1)\right]$. Moreover,

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

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

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

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. Analogous 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 ϕ in this definition includes only the components of $\phi_{\max}$ selected by c. Under Assumptions 5 and 6(ii), $\|\hat{\theta}_T(c) - \theta_0\| = O_P(T^{-\frac{1}{2}+\delta_2})$ under $\mathbb{P}_T$. 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 Assumption 6(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): $s_1(c) < s_1(c_r)$. Assume first that $\delta_1 - \delta_2 < 0$, and 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 $\hat{V}_\theta(c_r) \xrightarrow{P} V_\theta(c_r)$ (both under $\mathbb{P}_T$ with finite limits) 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.

If $\delta_1 = \delta_2$, then $\rho_{\max,1}(\theta,c) = \rho_{\max}(\theta,c)$, for all $c \in \mathscr{C}$. Hence, from Assumption 6(ii), we have $s_1(c) = p = s_1(c_r)$. This case is covered by Case (2).

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

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

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 Magnus and Neudecker (2002, Thm. 22), $|V_\theta(c)| > |V_\theta(c_r)|$ and we have

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

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 $\hat{c} \xrightarrow{P} c_r$ under $\mathbb{P}_T$ as $T \to \infty$. □

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 $\frac{\partial \rho_{\max}(\theta_0, c)}{\partial \theta'}$ under Assumption 7(i.a) or for the actual M in Assumption 7(i.b).

Let M_1 be the submatrix of M given by its first k_1 rows, and let M_2 be the submatrix of M given by its last k_2 rows. Let $s_1(c) = \text{Rank}(M_1)$ 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 $MR_3 = 0$. Note that R_1 is void if $s_1(c) = 0$ and R_2 is void if $s_2(c) = 0$, whereas 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 $\lambda_T = \begin{pmatrix} T^{\delta_1} I_{s_1(c)} & 0 & 0 \\ 0 & T^{\delta_2} I_{s_2(c)} & 0 \\ 0 & 0 & T^{\delta_2} I_{p-q} \end{pmatrix}$. We have

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

where the $o_P(1)$ terms are negligible under $\mathbb{P}_T$. 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

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

Thus,

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

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 $\ln T \cdot \kappa(|\gamma|, T) \to 0$ as $T \to \infty$ for $\gamma = c, c_r$. Furthermore, 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, since $s_1(\iota_{\max}) \geq s_1(c)$, we have

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

The first conclusion follows with $a_0 = \ln|V_\theta(c_r)|$, $a_1 = 2(s_1(\iota_{\max}) - s_1(c))(\delta_1 - \delta_2)$, and $a_{1T} = \ln|\hat{K}_{\theta,T}|$. We can therefore conclude that $mRMSC(c_r) < mRMSC(c)$ with probability approaching 1 as $T \to \infty$. □

Proof of Proposition 4.4. The first conclusion is a mere consequence of Theorems 4.2 and 4.3. Since $\mathbb{P}_T(\hat{c}_T = c_r) \to 1$ as $T \to \infty$, the second part follows directly from Lemma 1 of Pötscher (1991). □

Proof of Proposition 4.5. All the stochastic order of magnitude in this proof are under $\mathbb{P}_T$. 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

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

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,

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

In addition,

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

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

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

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

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

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 $\Sigma(\phi) = \mathbb{E}(\phi_{tT}(\theta_0)\phi_{tT}(\theta_0)')$. By a mean-value expansion, we have

$$vec\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{tT}(\hat{\theta}_T(\phi))\phi_{tT}(\hat{\theta}_T(\phi))'\right) = vec(\Sigma(\phi)) + vec\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{tT}(\theta_0)\phi_{tT}(\theta_0)' - \Sigma(\phi)\right)$$
$$+ \frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial \theta'}vec[\phi_{tT}(\theta)\phi_{tT}(\theta)']\Big|_{\theta=\ddot{\theta}}(\hat{\theta}_T(\phi) - \theta_0)$$
$$= vec(\Sigma(\phi)) + O_P(T^{-\frac{1}{2}+\delta_2})$$
$$= \Sigma(\phi) + O_P(T^{-\frac{1}{2}-\delta_1+2\delta_2}),$$

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 $\phi_{tT}(\theta_0)$. 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

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

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). □

APPENDIX C. Mixed Identification Strength of Arbitrary Number of Levels

This section establishes the consistency of mRMSC in a more general framework of moment condition models with mixed identification strength of arbitrary number of levels. We claim

that a moment condition model represented by the $\mathbb{R}^k$-valued estimating function $\phi(\cdot)$ has mixed identification strength if, for some $l \in \mathbb{N}$: $\phi \equiv (\phi_1', \ldots, \phi_l')' \in \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_l}$:

$$\mathbb{E}(\phi_i(Y_{tT}, \theta)) = \frac{\rho_i(\theta)}{T^{\delta_i}}, \quad i = 1, \ldots, l, \quad \text{and} \quad 0 \leq \delta_1 \leq \cdots \leq \delta_l < \frac{1}{2}. \tag{C.1}$$

The case studied in the main body of the paper corresponds to $l = 2$.

We assume that the set of available moment restrictions to be selected from are collected in the $\mathbb{R}^{k_{\max}}$-valued estimating function $\phi_{\max}$ satisfying the following assumption.

Assumption C.1. (i) $\phi_{\max}(\cdot)$ satisfies (C.1), that is, $\phi_{\max} \equiv (\phi_{\max,1}', \ldots, \phi_{\max,l}')' \in \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_l}$:

$$\mathbb{E}\left(\phi_{\max,i}(Y_{tT}, \theta)\right) = \frac{\rho_{\max,i}(\theta)}{T^{\delta_i}}, \quad i = 1, \ldots, l, \text{ and } 0 \leq \delta_1 \leq \cdots \leq \delta_l < \frac{1}{2},$$

where $\rho_{\max}(\cdot)$ is an $\mathbb{R}^{k_{\max}}$-valued function defined on the compact parameter set $\Theta \subset \mathbb{R}^p$.

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

(iii) $\sup_{\theta \in \Theta} \sqrt{T} \left\| \bar{\phi}_{\max,T}(\theta) - \mathbb{E}\left(\phi_{\max}(Y_{tT}, \theta)\right) \right\| = O_P(1)$ under $\mathbb{P}_T$, with $\bar{\phi}_{\max,T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{\max}(Y_{tT}, \theta)$.

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

(v) $\frac{\partial \rho_{\max}}{\partial \theta'}(\theta_0)$ is full column rank and, for $i = 1, \ldots, l$, $\mathbb{E}\left(\frac{\partial \phi_{\max,i}(Y_{tT}, \theta_0)}{\partial \theta'}\right) = T^{-\delta_i} \frac{\partial \rho_{\max,i}}{\partial \theta'}(\theta_0) + o(T^{-\delta_i})$ and $\sqrt{T} \sup_{\theta \in \mathcal{N}_{\theta_0}} \left\| \frac{\bar{\phi}_{\max,T}(\theta)}{\partial \theta'} - \mathbb{E}\left(\frac{\partial \phi_{\max}(Y_{tT}, \theta)}{\partial \theta'}\right) \right\| = O_P(1)$ under $\mathbb{P}_T$, with $\bar{\phi}_{\max,i,T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{\max,i}(Y_{tT}, \theta)$.

(vi) $\delta_l < \frac{1}{4} + \frac{\delta_1}{2}$.

(vii) For all k, $1 \leq k \leq k_i (i = 1, \ldots, l)$,

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

under $\mathbb{P}_T$, 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 kth component of $\bar{\phi}_{\max,i,T}(\theta)$.

(viii) $\Sigma(\phi_{\max}) = \lim_{T \to \infty} Var(\sqrt{T} \bar{\phi}_{\max,T}(\theta_0))$ is positive-definite, with variance taken under $\mathbb{P}_T$.

For $i = 1, \ldots, l$, define

$$J_{\max,i} = \left(\frac{\partial \rho_{\max,1}(\theta_0)'}{\partial \theta} \cdots \frac{\partial \rho_{\max,i}(\theta_0)'}{\partial \theta} \right)'.$$

Let $s_1, \ldots, s_l$ be such that, for all $i = 1, \ldots, l$, $Rank(J_{\max,i}) = s_1 + \cdots + s_i$, with $Rank(J_{\max,l}) = s_1 + \cdots + s_l = p$. Let $R_{\max} \equiv (R_1 \cdots R_l)$ be an orthogonal (p,p)-matrix such that R_i is a (p, s_i)-full-column-rank matrix and, for all $j = 2, \ldots, l$,

$$J_{\max,j-1} R_j = 0.$$

Let $J_{\max}$ be the block-diagonal $(k_{\max}, p)$-matrix with diagonal blocks $(\partial \rho_{\max,i}(\theta_0)/\partial \theta')R_i$, $i = 1, \ldots, l$, and let $\Lambda_T(\phi_{\max})$ be the block diagonal (p,p)-matrix with diagonal blocks $T^{\frac{1}{2}-\delta_i} I_{s_i}$, $i = 1, \ldots, l$.

Assumption C.1 is sufficient—if we add that $\sqrt{T}\bar{\phi}_{\max,T}(\theta_0) \xrightarrow{d} N(0, \Sigma(\phi_{\max}))$ under $\mathbb{P}_T$—to establish the asymptotic normality of the two-step GMM estimator, $\hat{\theta}_{\max}$. (See Antoine and Renault (2012, Thm. 4.3).) More specifically,

$$\Lambda_T(\phi_{\max}) R_{\max}^{-1}(\hat{\theta}_{\max} - \theta_0) \xrightarrow{d} N(0, (J_{\max}' \Sigma(\phi_{\max})^{-1} J_{\max})^{-1}).$$

This asymptotic variance is the semiparametric efficiency bound for the estimation of $R_{\max}^{-1}\theta_0$ from (C.1). (See Dovonon et al. (2022).) As already mentioned in the case $l = 2$, this bound is the cornerstone of our moment selection procedure and it can be consistently estimated under the conditions in Assumption C.1. We can indeed claim, by relying on Lemma 4.1 of Antoine and Renault (2012), that

$$\hat{J}_{\max} \equiv \sqrt{T} \frac{\partial \bar{\phi}_{\max,T}(\hat{\theta}_{\max})}{\partial \theta'} R_{\max} \Lambda_T(\phi_{\max})^{-1} \xrightarrow{P} J_{\max}, \text{ under } \mathbb{P}_T,$$

and, hence,

$$\left(\hat{J}_{\max}' \hat{\Sigma}(\phi_{\max})^{-1} \hat{J}_{\max}\right)^{-1} \xrightarrow{P} \left(J_{\max}' \Sigma(\phi_{\max})^{-1} J_{\max}\right)^{-1}, \text{ under } \mathbb{P}_T.$$

As in Section 4.2, any candidate model can be represented by a specific selection vector $c \in \mathbb{R}^{k_{\max}}$ with entries 0's and 1's. The set of all possible selection vectors is denoted $\mathscr{C}$, and the candidate models are represented by $\phi_{\max}(\cdot, c)$, $c \in \mathscr{C}$.

Letting c be a selection vector, we write $c = (c_1', \ldots, c_l')' \in \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_l}$ and, for $j = 1, \ldots, l$, let $s_j(c)$ be defined as s_j but with $\rho_{\max}(\cdot, c)$ replacing $\rho_{\max}(\cdot)$. Similar to Definition 1, we introduce the following definition of a relevant moment restriction in the general context:

DEFINITION C.2. *A subset of moment restriction characterized by $c_r \in \mathscr{C}$ is said to be relevant if the following two properties hold:*

(i) $s_1(c_r)\delta_1 + \cdots + s_l(c_r)\delta_l = s_1(\iota_{\max})\delta_1 + \cdots + s_l(\iota_{\max})\delta_l$ *and* $V_\theta(\iota_{\max}) = V_\theta(c_r)$, *where* $\iota_{\max}$ *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) $s_1(c_r)\delta_1 + \cdots + s_l(c_r)\delta_l < s_1(c_{r,1})\delta_1 + \cdots + s_l(c_{r,1})\delta_l$,

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

Similar comments to those following Definition 1 stand here as well. Lemma C.1 ensures that $s_1(c_r)\delta_1 + \cdots + s_l(c_r)\delta_l = s_1(\iota_{\max})\delta_1 + \cdots + s_l(\iota_{\max})\delta_l$ is equivalent to $s_1(c_r) = s_1(\iota_{\max}), \ldots, s_l(c_r) = s_l(\iota_{\max})$. Therefore, the rotation matrix associated with $\phi_{\max}(\cdot, c_r)$ can be set to $R_{\max}$ and $V_\theta(\iota_{\max})$ and $V_\theta(c_r)$ in Definition C.2(i) are expressed in terms of the same rotation matrix. Similar arguments stand for Definition C.2(ii).

We base the estimation of c_r, the selection vector corresponding to the relevant set of moment conditions, on the *mRMSC* introduced by (24) with a penalization term $\kappa_T = \kappa(|c|, T)$:

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

where $\hat{I}_{\theta,T}(c)$ is given by (24) with $\phi(\cdot) = \phi_{\max}(\cdot, c)$. As in Section 4.2, the relevant model c_r is estimated by $\hat{c}_T$ defined by

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

To formulate our consistency theory, we let

$$\mathscr{C}_{\text{eff}} = \{c \in \mathscr{C} : s_1(c)\delta_1 + \cdots + s_l(c)\delta_l = s_1(\iota_{\max})\delta_1 + \cdots + s_l(\iota_{\max})\delta_l \text{ and } V_\theta(c) = V_\theta(\iota_{\max})\}$$

and

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

and make the following assumption, which is mere adaptation of Assumption 6 to the more general specification (C.1).

Assumption C.2. (i) c_r satisfies Definition C.2 and $\mathscr{C}_{\min} = \{c_r\}$; (ii) $\forall c \in \mathscr{C}$, $\rho_{\max}(\theta, c) = 0 \Leftrightarrow \theta = \theta_0$, and $\text{Rank}(\partial \rho_{\max}(\theta_0, c)/\partial \theta') = p$; (iii) $\hat{\Sigma}(c)$ converges in probability, under $\mathbb{P}_T$, to $\Sigma(c) \equiv \lim_{T\to\infty} Var(\sqrt{T}\bar{\phi}_{\max,T}(\theta_0, c))$, positive-definite, with variance taken under $\mathbb{P}_T$; (iv) $\hat{V}_\theta(c) = V_\theta(c) + O_P(\tau_{T,c}^{-1})$ under $\mathbb{P}_T$, where $\tau_{T,c} \to \infty$ as $T \to \infty$; and (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 $T \to \infty$.

Note that Assumption C.2 is the same as Assumption 6. We only replace in the latter the definition of the relevant model by the more general concept above, and $\phi_{\max}$ and $\rho_{\max}$ are, respectively, replaced by their more general version in Assumption C.1. Similar to Theorem 4.2, we can now claim the following result.

THEOREM C.1. *If Assumptions C.1 and C.2 hold, then $\hat{c}_T$ converges in probability to c_r as $T \to \infty$.*

This result shows that the optimal model with respect to mRMSC converges to the relevant model c_r as the sample size grows when selection is made among candidate models satisfying point identification and first-order local identification properties. This is an extension of Theorem 4.2 to the case where $l \geq 2$ and, in a similar way, is at the core of the consistency of mRMSC over the whole set of candidate models. Indeed, as shown by Theorem 4.3, we can also show in this context that, with probability approaching 1, c_r outperforms candidate models that fail point identification or first-order local identification properties. We can also claim that the selection procedure yields a model that is efficient since we can establish an analog of Proposition 4.4 in the current configuration. We do not propose a formal exposition of the analogs of Theorem 4.3 and Proposition 4.4 to save space.

Proofs

LEMMA C.1. *Let* $c = (c_1', \dots c_l')' \in \{0,1\}^{k_1} \times \cdots \times \{0,1\}^{k_l}$ *be a candidate model. For* $i = 1, \dots, l$, *define*

$$J_{\max,i} = \left(\frac{\partial \rho_{\max,1}(\theta_0)'}{\partial \theta} \cdots \frac{\partial \rho_{\max,i}(\theta_0)'}{\partial \theta} \right)' \quad \textit{and}$$

$$J_{\max,i}(c) = \left(\frac{\partial \rho_{\max,1}(\theta_0, c_1)'}{\partial \theta} \cdots \frac{\partial \rho_{\max,i}(\theta_0, c_l)'}{\partial \theta} \right)'.$$

Let $s_1, \dots, s_l$ *and* $s_1(c), \dots, s_l(c)$ *be such that*

$$Rank(J_{\max,i}) = s_1 + \cdots + s_i, \quad \textit{and} \quad Rank(J_{\max,i}(c)) = s_1(c) + \cdots + s_i(c).$$

Finally, let R_j $(j = 1, \dots, l)$ *be a collection of full-column-rank* (p, s_j)*-matrices such that, for each* $i = 1, \dots, l$, *the columns of* $(R_1 \cdots R_i)$ *span those of* $J_{\max,i}'$ *and* $J_{\max,i} R_j = 0$ *for all* $j = i+1, \dots, l$.

If $Rank(J_{\max,l}) = Rank(J_{\max,l}(c)) = p$ *and* $s_1\delta_1 + \cdots + s_l\delta_l = s_1(c)\delta_1 + \cdots + s_l(c)\delta_l$, *then*

$$s_1 = s_1(c), \dots, s_l = s_l(c).$$

Furthermore, we also have that, for each $i = 1, \dots, l$, *the columns of* $(R_1 \cdots R_i)$ *span those of* $J_{\max,i}(c)'$ *and* $J_{\max,i}(c) R_j = 0$, *for all* $j = i+1, \dots, l$.

Proof of Lemma C.1. Since for each j, $J_{\max,j}$ is $J_{\max,j}(c)$ plus extra rows,

$$Rank(J_{\max,j}(c)) = s_1(c) + \cdots + s_j(c) \le s_1 + \cdots + s_j = Rank(J_{\max,j}).$$

We also have

$$\begin{aligned} 0 &= (s_1(c) - s_1)\delta_1 + (s_2(c) - s_2)\delta_2 + \cdots + (s_l(c) - s_l)\delta_l \\ &= \sum_{j=1}^{l-1} \left[\sum_{i=1}^{j} (s_i(c) - s_i) \right] (\delta_j - \delta_{j+1}) + \delta_l \sum_{i=1}^{l} (s_i(c) - s_i) \\ &= \sum_{j=1}^{l-1} \left[\sum_{i=1}^{j} (s_i(c) - s_i) \right] (\delta_j - \delta_{j+1}), \end{aligned}$$

where the last equality follows from the fact that $Rank(J_{\max,l}(c)) = s_1(c) + \cdots + s_l(c) = p = s_1 + \cdots + s_l = Rank(J_{\max,l})$. Since $\delta_j \le \delta_{j+1}$, each of the terms in this summation is nonnegative and it results that each of them is nil. We can then claim that $\sum_{i=1}^{j} (s_i(c) - s_i) = 0$ for each j or equivalently $s_i(c) = s_i$ for each $i = 1, \dots, l-1$. Moreover, since all the s_i's and $s_i(c)$'s add to p, this also shows that $s_l(c) = s_l$. This shows the first statement. The second one follows trivially. □

Proof of Theorem C.1. Analogous to previous notation, let

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

where ϕ in this definition includes only the components of $\phi_{\max}$ selected by c. Under Assumptions C.1 and C.2(ii), $\|\hat{\theta}_T(c)-\theta_0\| = O_P(T^{-\frac{1}{2}+\delta_l})$ under $\mathbb{P}_T$. Thanks to Lemma 4.1 of Antoine and Renault (2012), we can claim that $\sqrt{T}\frac{\partial \bar{\phi}_T}{\partial \theta'}(\hat{\theta}_T(c))R_{\max}(c)\Lambda_T(c)^{-1}$ converges in probability to $J(c) \equiv J_{\max}(c)$, and as a result, $\hat{V}_\theta(c)$ converges in probability to $(J(c)'\Sigma(c)^{-1}J(c))^{-1}$. Note that

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

Hence,

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

Thus,

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

and

$$mRMSC(c) = 2(s_1(c)\delta_1+\cdots+s_l(c)\delta_l) + \frac{\ln\left|\hat{V}_\theta(c)\right|}{\ln T} + \kappa(|c|,T). \tag{C.2}$$

Let

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

Thanks to Assumption C.2(i), $s_1(c)\delta_1+\cdots+s_l(c)\delta_l \le s_1(c_r)\delta_1+\cdots+s_l(c_r)\delta_l$; otherwise, $s_1(c)\delta_1+\cdots+s_l(c)\delta_l > s_1(\iota_{\max})\delta_1+\cdots+s_l(\iota_{\max})\delta_l$, which is impossible. This would indeed require that $s_j(c) > s_j(\iota_{\max})$ for some j.

We shall distinguish the following two cases: (1) $s_1(c)\delta_1+\cdots+s_l(c)\delta_l < s_1(c_r)\delta_1+\cdots+s_l(c_r)\delta_l$ and (2) $s_1(c)\delta_1+\cdots+s_l(c)\delta_l = s_1(c_r)\delta_1+\cdots+s_l(c_r)\delta_l$.

Case (1): $s_1(c)\delta_1+\cdots+s_l(c)\delta_l < s_1(c_r)\delta_1+\cdots+s_l(c_r)\delta_l$. Since $\hat{V}_\theta(c) \xrightarrow{P} V_\theta(c)$, and $\hat{V}_\theta(c_r) \xrightarrow{P} V_\theta(c_r)$ (both under $\mathbb{P}_T$ with finite limits) and $\kappa(|c|,T) \to 0$ as $T \to \infty$ for all c, we can claim that $\Delta_T(c,c_r) \xrightarrow{P} 2[(s_1(c)-s_1(c_r))\delta_1+\cdots+(s_l(c)-s_l(c_r))\delta_l] < 0$, meaning that c_r will be chosen over c as T gets large with probability approaching 1. The rest of the proof is similar to the case $l=2$.

Case (2): $s_1(c)\delta_1+\cdots+s_l(c)\delta_l = s_1(c_r)\delta_1+\cdots+s_l(c_r)\delta_l$. Lemma C.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 $R_{\max}$. By definition, $V_\theta(c_r) = V_\theta(\iota_{\max})$ and, considering $V_\theta(c)$ as expressed in terms of $R_{\max}$ 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 semidefinite. We further consider these two cases.

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

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

where we use Assumption C.2(iv). By Assumption C.2(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 Magnus and Neudecker (2002, Thm. 22), $|V_\theta(c)| > |V_\theta(c_r)|$ and we have

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

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 $\hat{c} \xrightarrow{P} c_r$ under $\mathbb{P}_T$ as $T \to \infty$. □

SUPPLEMENTARY MATERIAL

Dovonon, P., F. Doko Tchatoka, and M. Aguessy (2022). Supplement to "Relevant moment selection under mixed identification strength," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466622000640.

REFERENCES

Andrews, D.W. & J.H. Stock (2007) Testing with many weak instruments. *Journal of Econometrics* 138(1), 24–46.

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

Andrews, D.W.K. & 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. & E. Renault (2009) Efficient GMM with nearly-weak instruments. *The Econometrics Journal* 12, S135–S171.

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

Antoine, B. & E. Renault (2017) On the relevance of weaker instruments. *Econometric Reviews* 36, 928–945.

Antoine, B. & E. Renault (2020) Testing identification strength. *Journal of Econometrics* 218, 271–293.

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

Breusch, T., H. Qian, P. Schmidt, & D. Wyhowski (1999) Redundancy of moment conditions. *Journal of Econometrics* 91, 89–111.

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

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

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

Cheng, X. & 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.

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

Dovonon, P. & Y.F. Atchadé (2020) Efficiency bounds for semiparametric models with singular score functions. *Econometric Reviews* 39, 612–648.

Dovonon, P., Y.F. Atchadé, and F. Doko Tchatoka (2022) Efficiency bounds for moment condition models with mixed identification strength. Technical report, Department of Economics, Concordia University.

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

Dovonon, P. and E. Renault (2020). GMM overidentification test with first-order underidentification. Technical report, Department of Economics, Concordia University.

Gagliardini, P., C. Gourieroux, & E. Renault (2011) Efficient derivative pricing by the extended method of moments. *Econometrica* 79(4), 1181–1232.

Hahn, J. & G. Kuersteiner (2002) Discontinuities of weak instrument limiting distributions. *Economics Letters* 75(3), 325–331.

Hall, A.R., A. Inoue, K. Jana, & 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., A. Inoue, & C. Shin (2008) Entropy-based moment selection in the presence of weak identification. *Econometric Reviews* 27(4–6), 398–427.

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

Han, S. & A. McCloskey (2019) Estimation and inference with a (nearly) singular Jacobian. *Quantitative Economics* 10(3), 1019–1068.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4), 1029–1054.

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

Kabaila, P. & H. Leeb (2006) On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* 101(474), 619–629.

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

Magnus, J.R. & H. Neudecker (2002) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley.

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

Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory* 7, 163–185.

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

Stock, J.H. & M. Yogo (2005) Testing for weak instruments in linear IV regression. In D.W.K. Andrews & J.H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80–108. Cambridge University Press.

Windmeijer, F., H. Farbmacher, N. Davies, & D.G. Smith (2019) On the use of the Lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* 114, 1339–1350.

Ziegler, K. (1997) Functional central limit theorems for triangular arrays of function-indexed processes under uniformly integrable entropy conditions. *Journal of Multivariate Analysis* 62(2), 233–272.