Score tests in GMM: Why Use Implied Probabilities?*

Saraswata Chaudhuri† and Eric Renault‡

Abstract

While simple to implement and thus attractive in practice, the GMM score test of Newey and West (1987) often displays upward size distortion under common scenarios involving skewed moment vectors or models with weak identification. Inference based on the Generalized Empirical Likelihood (GEL) is seen as a general solution to this problem. Kleibergen (2005) and, more generally, Guggenberger and Smith (2005) devised an elegant theory for the GEL score tests. However, strictly speaking, the GEL score tests do not nest the Newey-West score test. Our paper provides a unified framework for score tests in GMM that nests all of the above as special cases and helps us to understand the precise mechanism by which the standard first order asymptotic theory on size and power well approximates the finite sample behavior of some score tests (namely, a subset of the GEL score tests) but not others. Special attention is paid to models with weak identification. We also argue that the apparent computational burden of GEL can be overcome in practice by recognizing the fundamental common role played by the GEL implied probabilities under all special cases of our framework. In particular, we show that all the GEL implied probabilities are asymptotically equivalent at a higher order – both under the null and under appropriate sequences of alternatives – and thus are exchangeable across computationally burdensome (e.g. Empirical Likelihood) and easy (e.g. Euclidean Empirical Likelihood) GEL score tests without affecting the first order asymptotics. Extensive simulation evidence is provided to corroborate our theoretical results. The simulation results also support a simple and yet important insight on the power of the tests: the use of implied probabilities to efficiently estimate the components of the score statistic, namely, the Jacobian and the asymptotic variance of the moment vector, can significantly improve the power of the score test in finite samples.

JEL Classification: C12; C13; C30

Keywords: GEL; GMM; Implied probabilities; Score test

*We thank the editor and two anonymous referees for their helpful comments. We also gratefully acknowledge helpful comments from conference and seminar participants. This version partially replaces an earlier version of the paper titled "Finite-sample improvements of score tests by the use of implied probabilities from Generalized Empirical Likelihood".

†Department of Economics, McGill University, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

‡Corresponding author. Department of Economics, University of Warwick, Coventry, United Kingdom. Email: Eric.Renault@warwick.ac.uk.

1 Introduction

The focus of our interest in this paper is the test of a null hypothesis

$$H_0 : \theta = \theta_0 \tag{1.1}$$

about a vector $\theta \in \Theta \subset \mathbb{R}^p$ of p unknown parameters. The unknown true value θ^0 of θ is uniquely defined by $H > p$ moments conditions:

$$E\left[\psi\left(W_i, \theta\right)\right] = 0 \Longleftrightarrow \theta = \theta^0. \tag{1.2}$$

The true unknown value θ^0, and as a consequence also the hypothetical one θ_0, are assumed to belong to the interior of the compact set Θ of parameters. Almost certainly for the distribution of W_i, the function $\psi(w, \theta)$ is assumed to be continuously differentiable in θ on the interior of Θ.

For the sake of expositional simplicity, the observations $W_i, i = 1, \ldots, n$, will be treated throughout as an i.i.d. sequence[1] such that a uniform law of large numbers and central limit theorem will apply to the sequence $\psi_i(\theta) \equiv \psi(W_i, \theta)$ with an asymptotic covariance matrix[2]:

$$V(\theta) = Var\left[\sqrt{n}\bar{\psi}_n(\theta)\right] = Var\left[\psi_1(\theta)\right], \quad \text{where} \quad \bar{\psi}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\theta).$$

When faced with the testing problem (1.1) under the moment restrictions (1.2), one could take two popular and computationally attractive strategies that do not involve estimation of θ.

(i) Hotelling approach: Note that, under the null hypothesis (1.1), $n\bar{\psi}_n'(\theta_0)'V^{-1}(\theta^0)\bar{\psi}_n(\theta_0) \xrightarrow{d} \chi^2_H$ where χ^2_H denotes central χ^2 distribution with H degrees of freedom; and this provides a very simple and consistent testing strategy. It is indeed nothing but the classical Hotelling test of the null hypothesis:

$$H_0 : E(X_i) = 0$$

with $X_i = \psi\left(W_i, \theta_0\right)$. This approach has regained popularity, in particular since the seminal work of Stock and Wright (2000), due to its robustness to weak identification.

(ii) Score approach: However, it is worth keeping in mind that, even though the null hypothesis (1.1) can obviously be tested in many ways without any care for estimation of the p unknown parameters $\theta \in \Theta$, the philosophy of GMM (see e.g. Newey and West (1987), Bera et al. (2010)) is to provide powerful tests when sequence of local deviations are defined in terms of the parameters of interest:

$$\theta_0 = \theta^0 + \frac{\nu}{\sqrt{n}} \quad \text{for some } \nu \neq 0 \in \mathbb{R}^p.$$

(The rate $\sqrt{n}$ here corresponds to strong identification but, in general, needs to correspond to the underlying identification strength.) In other words, we are not interested in the general local deviations:

$$E\left[\psi\left(W_i, \theta\right)\right] = \frac{\delta}{\sqrt{n}} \quad \text{for some } \delta \neq 0 \in \mathbb{R}^H$$

but in specific ones:

$$E\left[\psi\left(W_i, \theta\right)\right] \sim E\left[\frac{\partial\psi\left(W_i, \theta^0\right)}{\partial\theta'}\right]\frac{\nu}{\sqrt{n}}.$$

Obviously, the two approaches are not equivalent when there is overidentification ($H > p$). It has

[1] In order to accommodate for weak identification, we will consider throughout some drifting data generating processes such that the probability distribution of $(W_i)_{1\leq i\leq n}$ is a product of n identical distributions that may shift with n. However, the only relevant impact of this shift is on the moment conditions $E\left[\psi_i(\theta)\right]$ and does not impair the law of large numbers for higher order moments.

[2] Most of our results can be extended to the case of estimating functions $\psi_i(\theta), i = 1, \ldots, n$ resulting from a preliminary smoothing, as put forward by Kitamura and Stutzer (1997), to get efficient inference in the presence of serial dependence.

been known since Newey and West (1987) that a test of the null (1.1) should not be directly based on a norm of the sample mean $\bar{\psi}_n(\theta_0)$ but should instead maximize local power by setting the focus on the moment conditions that are most informative about θ. For this purpose, Newey and West (1987) proposed a score test based on the derivative of the criterion function of the efficient two-step GMM.

Unfortunately, this means that the econometrician may be faced with a trade off between robustness and efficiency. On the one hand, the Hotelling approach, as popularized in econometrics by Anderson and Rubin (1949) for linear instrumental variables and by Stock and Wright (2000) for nonlinear GMM, has the advantage to be robust to weak identification in terms of size control, at the expense of local power loss (at least in the case of strong identification; see e.g. Kleibergen (2002)). On the other hand, the locally most powerful Newey and West (1987) score test has been shown to provide a very poor size control in the case of weak identification (see e.g. Wang and Zivot (1998)).

An important message of the recent strand of literature on Generalized Empirical Likelihood (GEL) is that it may alleviate this tension between robustness and efficiency; see e.g. Kleibergen (2005) and Guggenberger and Smith (2005). (GEL is popularized in econometrics by Smith (1997) and Newey and Smith (2004).) Thanks to this important development, it is now well known that the GEL finite-sample improvements are in general even more warranted in the case of weak identification.

A possible interpretation of these well-documented improvements is that the GEL optimization provides implied probabilities that may lead us to revise our empirical views about the data generating process and thereby ensure better performance in finite samples.

We provide in this paper a general framework to take advantage of these implied probabilities for improving the finite-sample performance of the score test of Newey and West (1987). The key intuition that distinguishes our paper from the literature cited above is that, while power of tests is improved by setting the focus on efficient directions (as in the Newey and West score test), it is improved even further by estimating these directions efficiently. Moreover, by erasing perverse correlation between the moment conditions and the estimated efficient directions, this efficient estimation provides an almost perfect hedge against size distortion effects due to weak identification. We describe and emphasize the central role played by the GEL implied probabilities for such improvements of the score test.

The theoretical contribution of the paper is as follows. We derive a comprehensive theory of the asymptotic behavior of implied probabilities. Since our focus of interest is not only size but also power of tests, it takes a theory of these probabilities not only when computed at the true value of θ as hypothesized by the null (1.1), but also under convenient sequences of alternatives. These alternatives may be only local, in case of strong identification of all structural parameters θ, but also more global in case of weak identification of any component of the vector θ of parameters.

In this respect, our paper can be seen as following up on the important contribution of Guggenberger and Smith (2005). Like in their paper, our results are valid for any GEL criterion inspired by the Cressie-Read family of statistical discrepancies. However, for the sake of notational simplicity, we sometimes exclude the case of exponential tilting. For inference with weak instruments, exponential tilting has been the focus of interest of Caner (2010). An innovation with respect to Guggenberger and Smith (2005) is that, extending the results of Ramalho and Smith (2006), we establish a uniform asymptotic equivalence of the implied probabilities, so that we do not need to deal with the genuine GEL score vectors but we can instead freely work with different implied probabilities for estimating the Jacobian matrix and the variance matrix respectively.

For the purpose of user-friendly practical applications, our contribution is to illustrate the excellent performance of 3SEEL (Three Step Euclidean Empirical Likelihood) with shrinkage proposed by Antoine et al. (2007) (henceforth, ABR-07). Thanks to the closed form formulas for implied probabilities stemming from quadratic optimization, EEL does display significant computational advantages with respect to competitors like EL or GS (GS stands for EL that is partially modified by Guggenberger and Smith (2005)). Joint use of shrinkage and 3SEEL allows us to resort to user-friendly non-negative implied probabilities that are well suited to improve the estimators of both Jacobian and variance in the GMM score function. While this shrunk 3SEEL is shown to be first order asymptotically equivalent with its main competitors, EL and GS, it is not only much more user friendly but also displays better (size-corrected) finite-sample power, according to our Monte Carlo experiments.

The paper is organized as follows. Section 2 provides an overview of the implied probabilities with focus on EL and EEL. Section 3 puts on the table several alternatives to the Newey-West score test that take advantage of the implied probabilities based on the hypothesis (1.1). Section 4 discusses these score tests, and presents a Monte Carlo comparison of their finite-sample performance. Section 5 concludes. The appendix, referred to as Appendix A, of the paper develops the underlying asymptotic theory of the implied probabilities that we apply in Section 4 to describe the properties of score tests.

Supplemental materials are presented online in Appendices B, C, D and E. Appendix B proves the results from Section 2. Appendix C proves the results from Appendix A. Appendix D lists the standard assumptions maintained to describe these properties of score tests, and then proves the result from Section 4 by applying the results from Appendix A. Appendix E reports two tables containing empirical size and twelve figures containing size-corrected power plots for the simulation experiment. (Due to space limitation, we had to report these large number of tables and figures in this online Appendix E outside of the main text. We apologize for the inconvenience to the reader; however, we do make a strong effort in clearly describing their key take away message in the main text itself.)

2 Implied probabilities: Overview

2.1 Cressie-Read Lagrange multipliers

The information theoretic approaches to inference in moment condition models have become popular in econometrics since the seminal paper by Imbens et al. (1998). The idea in the context of general moment conditions (1.2) is first, for any given value of the vector θ of parameters, to look for implied probabilities $\hat{\pi}_n^{(\gamma)}(\theta) = (\hat{\pi}_{i,n}^{(\gamma)}(\theta))_{1\leq i\leq n}$ as solutions of:

$$\min_{\pi\in\mathbb{R}^n} \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^{n} \left[(n\pi_i)^{1+\gamma} - 1\right] \quad \text{subject to} \quad \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i \psi_i(\theta) = 0. \tag{2.1}$$

The objective function (2.1) is defined for any real γ, including the two limit cases $\gamma \to 0$ and $\gamma \to (-1)$. The family of these functions, indexed by γ, is generally referred to as the Cressie-Read family of power divergence statistics (see Imbens et al. (1998) and the references therein). It is well known (see for instance Schennach (2007) for a review with the same notations) that, up to a scaling factor, the vector $\hat{\lambda}_n(\theta)$ of H Lagrange multipliers associated to the constraints $\sum_{i=1}^{n} \pi_i \psi_i(\theta) = 0$ in the optimization program above is characterized as the solution of the dual problem:

$$\max_{\lambda\in\Lambda_n^{(\gamma)}(\theta)} \sum_{i=1}^{n} \varrho^{(\gamma)}\left[\lambda'\psi_i(\theta)\right] \tag{2.2}$$

with the following notations, for any $\gamma \in \mathbb{R}$:

$$\begin{aligned} Q^{(\gamma)} &= \{x \in \mathbb{R}; 1+\gamma x > 0\} && (2.3)\\ \Lambda_n^{(\gamma)}(\theta) &= \left\{\lambda \in \mathbb{R}^H; \lambda'\psi_i(\theta) \in Q^{(\gamma)}, \forall i = 1, ..., n\right\} && (2.4)\\ \frac{d\varrho^{(\gamma)}(x)}{dx} &= -\left[1+\gamma x\right]^{1/\gamma}, \forall x \in Q^{(\gamma)}. \end{aligned}$$

Note that, for all $\gamma \in \mathbb{R}$, $Q^{(\gamma)}$ is an open interval of the real line, containing $x = 0$, and:

$$\frac{d^2\varrho^{(\gamma)}(x)}{dx^2} = -\left[1+\gamma x\right]^{(1/\gamma-1)} < 0, \forall x \in Q^{(\gamma)}.$$

Thus, $\varrho^{(\gamma)}$ is a strictly concave function defined on $Q^{(\gamma)}$ with:

$$\frac{d\varrho^{(\gamma)}(0)}{dx} = \frac{d^2\varrho^{(\gamma)}(0)}{dx^2} = -1,$$

while the function $\varrho^{(\gamma)}$ itself is defined by (2.3) up to an additive constant, but, whenever it is convenient, will be normalized by:

$$\varrho^{(\gamma)}(0) = 0.$$

Define

$$\tau^{(\gamma)}(x) = \frac{d\varrho^{(\gamma)}}{dx}(x) = -[1+\gamma x]^{1/\gamma}$$

(including the limit case $\varrho^{(\gamma)}(x) = 1 - \exp(x)$ for $\gamma \to 0$). Then, we easily check that:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta) = c_n(\theta)\tau^{(\gamma)}\left(\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right) = -c_n(\theta)\left[1+\gamma\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right]^{1/\gamma} \tag{2.5}$$

where the scaling constant $c_n(\theta)$ is determined by the normalization condition:

$$\sum_{i=1}^{n}\hat{\pi}_{i,n}^{(\gamma)}(\theta) = 1.$$

Our first technical result below in Proposition 1 is tightly related to the setup of Guggenberger and Smith (2005), but a comprehensive proof is provided in Appendix B for the sake of self-containedness. Our maintained assumption is the following:

Assumption 1: $\Theta_n, n = 1, 2, ...,$ is a sequence of subsets of Θ, containing the true unknown value θ^0, and such that:

(i) $\sup_{\theta\in\Theta_n}\left\|E\left(\bar{\psi}_n(\theta)\right)\right\| = O\left(\frac{1}{\sqrt{n}}\right)$.

(ii) $\max_{1\le i\le n}\sup_{\theta\in\Theta_n}\|\psi_i(\theta)\| = o_P(\sqrt{n})$.

(iii) $\sup_{\theta\in\Theta_n}\|\psi_i(\theta)\| = O_P(1)$ for each $i = 1, \ldots, n$.

(iv) $\sup_{\theta\in\Theta_n}\left\|\bar{\psi}_n(\theta) - E\left(\bar{\psi}_n(\theta)\right)\right\| = O_P\left(\frac{1}{\sqrt{n}}\right)$.

The interpretation of Assumption 1 obviously depends on the choice of the sequence of sets $\Theta_n, n = 1, 2, \ldots$

We first note that Assumption 1 is hardly restrictive if we choose $\Theta_n = \{\theta^0\}$ for all $n = 1, 2, \ldots$. For this choice, Assumptions 1(i) and 1(iii) are fulfilled by definition. Moreover, Assumptions 1(ii) and 1(iv) are then a direct consequence of the fact that $\psi(W_i, \theta^0)$ is an i.i.d. sequence with zero mean and finite variance. Assumption 1(ii) can be proved in this context by using the Borel-Cantelli Lemma (see Owen (1990), Lemma 3 revisited by ABR-07). Assumption 1(iv) is obviously implied by Lindeberg-Levy Central Limit Theorem.

However, the choice $\Theta_n = \{\theta^0\}$ is obviously not very useful when the focus of our interest is mainly power issues. For instance, when working with sequences Θ_n of $\sqrt{n}$-local alternatives in the case of strong identification, Assumption 1(i) will hold by the mean value formula. Besides the sequences of local alternatives naturally suggested by asymptotic power studies, we will also have to consider larger sets Θ_n when θ (or some components of θ) will be only weakly identified.

We can then prove:

Proposition 1: If Assumption 1 holds, then, for all $\theta \in \Theta_n$ and $\gamma \in \mathbb{R}$, there exists a sequence of vectorial functions $\hat{\lambda}_n^{(\gamma)}(\theta) \in \Lambda_n^{(\gamma)}(\theta)$ where $\Lambda_n^{(\gamma)}(\theta)$ is defined in (2.4), such that:

$$\lim_{n\to\infty}\Pr\left[\bigcap_{\theta\in\Theta_n}\left\{\sum_{i=1}^{n}\varrho^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right] \ge \sum_{i=1}^{n}\varrho^{(\gamma)}\left[\lambda'\psi_i(\theta)\right], \forall\lambda \in \Lambda_n^{(\gamma)}(\theta)\right\}\right] = 1$$

with:

$$\sup_{\theta\in\Theta_n}\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\| = O_P(1/\sqrt{n}). \ \blacksquare$$

Then we prove the following result, already derived by Newey and Smith (2004), that gives the vector of Lagrange multipliers as a simple re-scaling of the moment conditions.

Proposition 2: If Assumption 1 holds, then for all $\theta \in \Theta_n$:

$$\begin{aligned} \hat{\lambda}_n^{(\gamma)}(\theta) &= -\left(\tilde{\Omega}_n^{(\gamma)}(\theta)\right)^{-1} \bar{\psi}_n(\theta) \\ \text{where } \tilde{\Omega}_n^{(\gamma)}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} k^{(\gamma)}\left(\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right)\psi_i(\theta)\psi_i'(\theta) \\ \text{and } k^{(\gamma)}(x) &= -\frac{1}{x}\left[1+\tau^{(\gamma)}(x)\right]. \blacksquare \end{aligned} \tag{2.6}$$

The interpretation of Proposition 2 is even more transparent in two particular cases:

Case 1 (EEL): The function $k^{(\gamma)}(.)$ is constant.

With $k^{(\gamma)}(x) = -\frac{1}{x}\left[1+\tau^{(\gamma)}(x)\right] = -\frac{1-(1+\gamma x)^{1/\gamma}}{x}$, we see that it happens if and only if $\gamma = 1$, that is in the case of EEL. The criterion function (2.1) is then quadratic with respect to the probabilities π_i. In this case, $\tilde{\Omega}_n^{(1)}(\theta) = \hat{\Omega}_n(\theta)$, where $\hat{\Omega}_n(\theta)$ is the estimator of the covariance matrix $V(\theta)$ based on the empirical distribution that gives equal weights $\hat{\pi}_{i,n} = 1/n$ to each observation:

$$\hat{\Omega}_n(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}\psi_i(\theta)\psi_i'(\theta).$$

We have:

$$\hat{\lambda}_n^{(1)}(\theta) = -\hat{\Omega}_n(\theta)^{-1}\bar{\psi}_n(\theta).$$

Case 2 (EL): The function $k^{(\gamma)}(.)$ is proportional to the function $\tau^{(\gamma)}(.)$.

With $k^{(\gamma)}(x) = -\frac{1}{x}\left[1+\tau^{(\gamma)}(x)\right] = -\frac{1-(1+\gamma x)^{1/\gamma}}{x}$, we see that it happens if and only if $\gamma = -1$, that is in the case of EL. The criterion function (2.1) is then proportional to $\sum_{i=1}^{n} \ln \pi_i$. Then:

$$\tilde{\Omega}_n^{(-1)}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[1-\hat{\lambda}_n^{(-1)}(\theta)'\psi_i(\theta)\right]^{-1}\psi_i(\theta)\psi_i'(\theta)$$

is proportional to:

$$\hat{\Omega}_n^{(-1)}(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(-1)}(\theta)\,\psi_i(\theta)\psi_i'(\theta)$$

that is an alternative estimator of the covariance matrix $V(\theta)$ of the moment vector $\psi(W_i,\theta)$, taking advantage of the implied probabilities. We deduce that:

$$\hat{\lambda}_n^{(-1)}(\theta) \propto -[\hat{\Omega}_n^{(-1)}]^{-1}(\theta)\bar{\psi}_n(\theta). \tag{2.7}$$

Hence, we conclude that in both cases of EL and EEL (and seemingly only in these two cases), the vector of Lagrange multipliers is proportional to $\hat{V}_n^{-1}(\theta)\bar{\psi}_n(\theta)$ for some estimator $\hat{V}_n(\theta)$ of $V(\theta)$.

We will set a special focus on these two cases, in particular because they allow us to extend the seminal argument of Back and Brown (1993) about taking advantage of the informational content of moment restrictions to estimate additional moments.

2.2 Implied Probabilities from EL and EEL

Assume that we want to estimate (again, for a given value of θ) the expectation of a known vector function $g(W_i,\theta)$ of data and parameters, while taking advantage of the information that $\psi(W_i,\theta)$ has

a zero expectation. The idea put forward by Back and Brown (1993) is then to consider an extended set of moments conditions, that is not only:

$$E[\psi(W_i, \theta)] = 0 \tag{2.8}$$

but also:

$$E[\mu - g(W_i, \theta)] = 0 \tag{2.9}$$

for an extended set $(\theta', \mu')'$ of parameters of interest.

Revisiting the minimization problem (2.1) with not only the constraints produced by (2.8), namely:

$$\sum_{i=1}^{n} \pi_i \psi(W_i, \theta) = 0 \tag{2.10}$$

but also the constraints produced by (2.9), that is:

$$\sum_{i=1}^{n} \pi_i [\mu - g(W_i, \theta)] = 0, \tag{2.11}$$

we will in general end up with different solutions $\pi_i, i = 1, 2, \ldots, n$. However, for given data $W_i, i = 1, 2, \ldots, n$, there is one specific value of μ such that the additional constraints (2.11) do not change the solution described above. It is the case where the additional constraints are not binding because the given value of μ is nothing but:

$$\hat{\mu}_n^{(\gamma)}(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\, g(W_i, \theta).$$

In this case, the Lagrange multipliers associated to constraints (2.11) will be all nil. This allows us to derive the following result.

Proposition 3: If Assumption 1 holds, then, for all $\theta \in \Theta_n$, we have for $\gamma = \pm 1$:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\, g(W_i, \theta) = \bar{g}_n(\theta) - \hat{Cov}_{n,\theta}^{(\gamma)} [g(W_i, \theta), \psi(W_i, \theta)] [\hat{V}_{n,\theta}^{(\gamma)}(\theta)]^{-1} \bar{\psi}_n(\theta) \tag{2.12}$$

with:

$$\begin{aligned}
\hat{Cov}_{n,\theta}^{(1)} [g(W_i, \theta), \psi(W_i, \theta)] &= \frac{1}{n} \sum_{i=1}^{n} g(W_i, \theta) \left[\psi_i(\theta) - \bar{\psi}_n(\theta)\right]', \\
\hat{Cov}_{n,\theta}^{(-1)} [g(W_i, \theta), \psi(W_i, \theta)] &= \sum_{i=1}^{n} \hat{\pi}_i^{(-1)}(\theta)\, g(W_i, \theta) [\psi_i(\theta)]', \\
\hat{V}_n^{(1)}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \psi_i(\theta) \left[\psi_i(\theta) - \bar{\psi}_n(\theta)\right]', \\
\hat{V}_{n,\theta}^{(-1)}(\theta) &= \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(-1)}(\theta)\, \psi_i(\theta) [\psi_i(\theta)]'. \ \blacksquare
\end{aligned}$$

Formula (2.12), albeit algebraically true for any θ, should be statistically interpreted for $\theta = \theta^0$, the true unknown value such that $E[\psi(W_i, \theta^0)] = 0$. Then, under standard regularity conditions, $\hat{V}_n^{(1)}(\theta^0)$ and $\hat{V}_{n,\theta}^{(-1)}(\theta^0)$ will be both consistent estimators of the variance matrix $V(\theta^0)$ of the vector $\psi(W_i, \theta^0)$. Similarly, $\hat{Cov}_{n,\theta}^{(1)} \left[g(W_i, \theta^0), \psi(W_i, \theta^0)\right]$ and $\hat{Cov}_{n,\theta}^{(-1)} \left[g(W_i, \theta^0), \psi(W_i, \theta^0)\right]$ are consistent estimators of the covariance matrix $Cov \left[g(W_i, \theta^0), \psi(W_i, \theta^0)\right]$. In other words, (2.12) can be read at the true value

as:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}\left(\theta^0\right) g(W_i,\theta^0) = \frac{1}{n}\sum_{i=1}^{n}\left[g(W_i,\theta^0) - \hat{b}_n^{(\gamma)}(\theta^0)'\psi(W_i,\theta^0)\right], \text{ for } \gamma = \pm 1$$

where $\hat{b}_n^{(\gamma)}(\theta^0)$ is a consistent estimator of the regression coefficient of $g(W_i,\theta^0)$ on $\psi(W_i,\theta^0)$. In yet other words, the two formulas encapsulated in (2.12) point out that when estimating the expectation $E[g(W_i,\theta^0)]$ (for known θ^0), the naive estimator, that is the sample average of observed $g(W_i,\theta^0)$, can be optimally improved (to get an asymptotic minimum variance among consistent estimators) by control variables $\psi(W_i,\theta^0)$, the expectation of which is known to be zero. While for $\gamma = 1$ formula (2.12) had been extensively discussed in ABR-07 for an efficient use of the informational content of moment equations in the context of EEL, the same formula for $\gamma = -1$ is the extension of ABR-07 to genuine EL. However, as can be seen from (2.14) below, the EL case does not actually deliver closed form formulas for $\hat{\pi}_{i,n}^{(-1)}(\theta)$ unlike in the EEL case because it must be kept in mind that the corresponding second moments in Proposition 3 and, hence, $\hat{b}_n^{(-1)}(\theta)$ themselves involve $\hat{\pi}_{i,n}^{(-1)}(\theta)$.

Up to this additional difficulty, formulas of Proposition 3 can, both for $\gamma = 1$ and for $\gamma = -1$, be also interpreted in terms of tilting the empirical distribution $\hat{\pi}_{i,n} = 1/n$ as follows.

Corollary 1: In the context of Proposition 3:

$$\hat{\pi}_{i,n}^{(1)}(\theta) = \frac{1}{n} - \bar{\psi}_n(\theta)'[\hat{V}_n^{(1)}(\theta)]^{-1}\frac{\psi_i(\theta) - \bar{\psi}_n(\theta)}{n} \tag{2.13}$$

and:

$$\hat{\pi}_{i,n}^{(-1)}(\theta) = \frac{1}{n} - \bar{\psi}_n(\theta)'[\hat{V}_{n,\theta}^{(-1)}(\theta)]^{-1}\hat{\pi}_{i,n}^{(-1)}(\theta)\,\psi_i(\theta). \blacksquare \tag{2.14}$$

Formulas (2.13) and (2.14) have quite similar interpretations. Assume for simplicity that $H = 1$ and $\bar{\psi}_n(\theta) > 0$. Then, a specific observation W_i will be downplayed (resp. magnified) by the tilting of probability weight from $\hat{\pi}_{i,n} = 1/n$ to implied probability $\hat{\pi}_{i,n}^{(\gamma)}(\theta), \gamma = \pm 1$, when the value of $\psi_i(\theta)$ is larger (resp. smaller) than $\bar{\psi}_n(\theta)$ (in case $\gamma = 1$) or zero (in case $\gamma = -1$) such that the weighted average using the tilted probabilities, in both cases, is identically zero.

The rationale of both tiltings is actually the same since in both cases we want to use the information that the population expectation of $\psi_i(\theta)$ is supposed to be zero. However, there is also an important difference between (2.13) and (2.14). While in (2.13), $\hat{\pi}_{i,n}^{(1)}(\theta)$ is explicitly given as a function of the observations $\psi_i(\theta)$, it is not the case for $\hat{\pi}_{i,n}^{(-1)}(\theta)$ in (2.14) since, as we just noted above Corollary 1, the implied probabilities $\hat{\pi}_{i,n}^{(-1)}(\theta)$ are also hidden in the definition of $\hat{V}_{n,\theta}^{(-1)}(\theta)$.

While on computational grounds, a closed-form formula to compute implied probabilities like (2.13) is of course very welcome, ((2.14) does not give such a formula) it comes with a cost. Non-negativity of the EEL implied probabilities is not warranted whereas it is warranted for the EL ones (since they maximize $\sum_{i=1}^{n}\ln\pi_i$). However, ABR-07 have shown that this flaw of the EEL implied probabilites can be fixed by a simple shrinkage, defining instead probabilities shrunk towards $(1/n)$ as:

$$\hat{\pi}_{i,n}^{(1),Sh}(\theta) = \frac{1}{1+\varepsilon_n(\theta)}\hat{\pi}_{i,n}^{(1)}(\theta) + \frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\hat{\pi}_{i,n} \tag{2.15}$$

where:

$$\varepsilon_n(\theta) = -n \times \min\left[\min_{1\le i\le n}\hat{\pi}_{i,n}^{(1)}(\theta), 0\right].$$

Note that, this shrinkage is performed in a minimum way so that the information content of the implied probabilities is not wasted asymptotically (see Propositions A.3 and A.4 in Appendix A). Therefore, one can also contemplate the use of shrunk probabilities $\hat{\pi}_{i,n}^{(1),Sh}(\theta)$ instead of the EEL ones, $\hat{\pi}_{i,n}^{(1)}(\theta)$, for the purpose of score testing.

3 Score vector using the implied probabilities

3.1 General framework

A score test for the hypothesis (1.1) on the full vector must be built from the first order conditions of a minimization program dedicated to the estimation of the unknown vector θ of parameters. Following the logic of Section 2.1, θ would be estimated through an information theoretic approach applied to the implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta)$ as follows:

$$\min_{\theta\in\Theta} \frac{1}{\gamma^*(\gamma^*+1)} \sum_{i=1}^{n} \left[\left\{ n\hat{\pi}_{i,n}^{(\gamma)}(\theta) \right\}^{1+\gamma^*} - 1 \right] \tag{3.1}$$

for some real number γ^*. Then, the first order conditions for computing an estimator of θ would be:

$$\sum_{i=1}^{n} \frac{\left[\hat{\pi}_{i,n}^{(\gamma)}(\theta) \right]^{\gamma^*} - 1}{\gamma^*} . \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} = 0. \tag{3.2}$$

Note that, except in the limit case $\gamma^* \to 0$, (3.2) can be rewritten in the more concise way:

$$\sum_{i=1}^{n} \left[\hat{\pi}_{i,n}^{(\gamma)}(\theta) \right]^{\gamma^*} \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} = 0 \tag{3.3}$$

since we have obviously:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta) = 1, \forall\ \theta \Rightarrow \sum_{i=1}^{n} \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} = 0. \tag{3.4}$$

Different, albeit first-order asymptotically equivalent, estimators of θ will be obtained, depending on the choice of a couple (γ, γ^*) of real numbers that define the two Cressie-Read distances at stake.

(i) The most common strategy is arguably to take the parameter γ^* of the Cressie-Read distance (3.1) equal to the parameter γ that has been previously at play to define the Lagrange multipliers vectors $\hat{\lambda}_n^{(\gamma)}(\theta)$ and the associated implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta)$.

(ii) An alternative strategy has been put forward by Schennach (2007). She actually recommends to choose $\gamma^* = -1$ (EL) to take advantage of the higher order efficiency properties of EL but uses the implied probabilities associated to exponential tilting (ET), that is $\gamma = 0$ (the KLIC minimization as put forward by Kitamura and Stutzer (1997)). Schennach (2007) shows that keeping $\gamma = 0$ for computing implied probabilities does not impair the higher order efficiency brought by $\gamma^* = -1$, while ensuring some robustness when moment conditions are misspecified. In a slightly different context with (local) conditional moment restrictions, Gagliardini et al. (2011) also remain true to ET for computing implied probabilities while they rather take $\gamma^* = 1$ (EEL) to estimate θ.

(iii) We will put forward in this paper a third strategy: we take advantage of the simplification brought by the condition $\gamma = \gamma^*$ to simplify the formula $\left[\hat{\pi}_{i,n}^{(\gamma)}(\theta) \right]^{\gamma^*}$, but we also keep the freedom of a different value of γ to compute $\frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta}$.

In other words, using (2.5) and disregarding the limit case $\gamma = 0$ for the sake of notational simplicity (see Section 3.3 for a special focus on this case), we can simplify the first order conditions (3.3) as follows:

$$\sum_{i=1}^{n} \left[1 + \gamma^* \hat{\lambda}_n^{(\gamma^*)}(\theta)' \psi_i(\theta) \right] . \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} = 0 \tag{3.5}$$

or, using (3.4),

$$\hat{\lambda}_n^{(\gamma^*)}(\theta)' \sum_{i=1}^{n} \psi_i(\theta) \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} = 0.$$

However, it is worth noting that:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\psi_i(\theta) = 0, \forall \theta \quad \Rightarrow \quad \sum_{i=1}^{n} \frac{\partial \hat{\pi}_{i,n}^{(\gamma)}(\theta)}{\partial \theta} \psi_i'(\theta) = -\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta) \frac{\partial \psi_i'(\theta)}{\partial \theta}.$$

Hence, we can eventually rewrite (3.5) as follows:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta) \frac{\partial \psi_i'(\theta)}{\partial \theta} . \hat{\lambda}_n^{(\gamma^*)}(\theta) = 0. \tag{3.6}$$

When remembering that for $\gamma = \pm 1$ we have:

$$\hat{\lambda}_n^{(\gamma)}(\theta) \propto [\hat{\Omega}_n^{(\gamma)}]^{-1}(\theta)\bar{\psi}_n(\theta),$$

playing with $\gamma, \gamma^* = \pm 1$ in (3.6) gives us several possible score vectors. As discussed below, formula (3.6) will allow us to bridge the gap between a score test implied by the information-theoretic approach (3.1) and the Newey-West score test. For the sake of improving the finite-sample performance, we will actually consider even more generally score vectors of the following form:

$$l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right) = \left\{ \sum_{i=1}^{n} \pi_{i,n}^G(\theta) G_i'(\theta) \right\} \left[\sum_{i=1}^{n} \pi_{i,n}^V(\theta) V_{i,n}(\theta) \right]^{-1} \sqrt{n}\bar{\psi}_n(\theta) \tag{3.7}$$

where:

$$G_i(\theta) = \frac{\partial \psi_i(\theta)}{\partial \theta'}, V_{i,n}(\theta) = \psi_i(\theta) \left(\psi_i(\theta) - \bar{\psi}_n(\theta)\right)'$$

and $\pi_{i,n}^G(\theta)$ and $\pi_{i,n}^V(\theta)$ may be different, but such that:

$$\pi_{i,n}^G(\theta), \pi_{i,n}^V(\theta) \in \left\{ \hat{\pi}_{i,n}^{(\gamma)}(\theta); \gamma \in \{\mathbb{R} \setminus \{0\}\} \right\} \cup \{\hat{\pi}_{i,n}\}. \tag{3.8}$$

Note that, the score vector (3.7) is consistent with the general framework (3.6) since, for all γ:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\psi_i(\theta)\psi_i'(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta) V_{i,n}(\theta).$$

It is only in the case of the unconstrained (naive) empirical probabilities $\hat{\pi}_{i,n}$ that we slightly modify the score vector (3.6) by considering moments of order two in mean deviation form $\hat{V}_n^{(1)}(\theta)$ instead of the naive estimator $\hat{\Omega}_n^{(1)}(\theta) = \hat{\Omega}_n(\theta)$.

3.2 Various score vectors

Out of the nine possible score vectors that can be deduced from (3.7) and the possible choices of the implied probabilities given by (3.8) with $\gamma = \pm 1$, we will pick five of them for reasons explained below and because they have already been considered in the literature. Note that, we use the acronyms 2SGMM (Two Step GMM), 3SEEL (Three Step EEL) that are typically used for estimation procedures based on a first-step estimator. However, as far as score testing of hypotheses such as (1.1) is concerned where θ is completely specified by the null hypothesis, there is no such thing as a multi-step procedure since we just plug in the hypothesized value θ_0 given by (1.1).

(i) S_n^{2SGMM} the score vector of the standard efficient 2SGMM of Hansen (1982):

$$S_n^{2SGMM}(\theta) = l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right); \pi_{i,n}^G(\theta) = \pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}.$$

The solution of the equation $S_n^{2SGMM}(\theta) = 0$ is known to reach the semi-parametric efficiency bound in terms of first order asymptotics, and is exactly the 2SGMM estimator only when the variance matrix $V_{i,n}(\theta)$ is independent of θ, up to a scaling factor (like for two stage least squares). Moreover, the score vector S_n^{2SGMM} is typically the one put forward by Newey and West (1987) for score testing. Note however that we apply 2SGMM with an efficient weighting matrix in which second order moments have been computed in mean deviation form. We follow in this respect the lesson of Hall (2000) and Chaudhuri and Renault (2015), with the hope of better finite-sample performance.

(ii) S_n^{EL} the score vector of EL:

$$S_n^{EL}(\theta) = l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right); \pi_{i,n}^G(\theta) = \pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}^{(-1)}(\theta).$$

The EL estimator $\hat{\theta}_n^{EL}$ is the solution of the equation $S_n^{EL}(\theta) = 0$. It is known (see Newey and Smith (2004)) to be higher order efficient (up to bias correction). As far as score testing is concerned, the score vector S_n^{EL} has been promoted by Guggenberger and Smith (2005).

(iii) S_n^{EEL} the score vector of EEL:

$$S_n^{EEL}(\theta) = l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right); \pi_{i,n}^G(\theta) = \hat{\pi}_{i,n}^{(1)}(\theta); \pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}.$$

Newey and Smith (2004) and ABR-07 have shown that the solution $\hat{\theta}_n^{CU}$ of the equation $S_n^{EEL}(\theta) = 0$ is the Continuously Updated (CU-) GMM Estimator of Hansen et al. (1996). By contrast with 2SGMM, it is numerically immaterial for CU-GMM to use an efficient weighting matrix in which the second order moments have been computed in mean deviation form or not. However, as shown by Newey and Smith (2004), the fact that $\hat{\theta}_n^{CU}$ is based on a naive estimation of this weighting matrix ($\pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}$) is in general responsible for some higher order bias that $\hat{\theta}_n^{EL}$ (or $\hat{\theta}_n^{3SEEL}$ defined below) does not display. However, Proposition 3 (applied with $g(W_i, \theta)$ given by any coefficient of the matrix $\psi_i(\theta^0)\psi_i'(\theta^0)$) shows that it is asymptotically harmless to use the naive empirical probabilities $\hat{\pi}_{i,n}$ for $\pi_{i,n}^V(\theta)$ (instead of $\hat{\pi}_{i,n}^{(\gamma)}(\theta), \gamma = \pm 1$) insofar as at the true value θ^0, the coefficients of $\psi_i(\theta^0)\psi_i'(\theta^0)$ are not correlated with the moments vector $\psi_i(\theta^0)$. As also noted by Newey and Smith (2004), it will be the case in particular when the latter are produced by cross-products with instruments $h(z_i)$ (functions of z_i's) for conditional moment restrictions:

$$E[u_i(\theta)\,|z_i] = 0$$

with an error term $u_i(\theta^0)$ that has a zero conditional-skewness given z_i. As far as score testing with weak instruments is concerned, the score vector S_n^{EEL} has been promoted by Kleibergen (2005). Note however that the latter paper gets this score vector not from a GEL approach but from CU-GMM.

(iv) S_n^{GS}, a score vector considered by Guggenberger and Smith (2005), has a hybrid form similar to EEL, but using instead the EL implied probabilities $\hat{\pi}_{i,n}^{(-1)}(\theta)$:

$$S_n^{GS}(\theta) = l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right); \pi_{i,n}^G(\theta) = \hat{\pi}_{i,n}^{(-1)}(\theta); \pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}.$$

Note that an estimator defined as the solution of the equation $S_n^{GS}(\theta) = 0$ would have, for the same reasons, exactly the same bias properties as CU-GMM described above: an additional higher order bias term by contrast with $\hat{\theta}_n^{EL}$ (or $\hat{\theta}_n^{3SEEL}$), no additional bias in the "symmetric" case (i.e., when the coefficients of $\psi_i(\theta^0)\psi_i'(\theta^0)$ are not correlated with the moments vector $\psi_i(\theta^0)$).

(v) S_n^{3SEEL} the score vector of the 3SEEL:

$$S_n^{3SEEL}(\theta) = l_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right); \pi_{i,n}^G(\theta) = \pi_{i,n}^V(\theta) = \hat{\pi}_{i,n}^{(1)}(\theta).$$

ABR-07 have introduced the 3SEEL estimator $\hat{\theta}_n^{3SEEL}$ and shown that it is as efficient as $\hat{\theta}_n^{EL}$ in terms of second order asymptotics, and albeit much more user-friendly on computational grounds. Even though they suggest to plug in the 2SGMM estimator in the defining equations of $\hat{\theta}_n^{3SEEL}$ (only $\bar{\psi}_n(\theta)$ would depend on the unknown θ) to get even simpler computations (hence the terminology three-step), we can consider a higher-order asymptotically equivalent estimator $\hat{\theta}_n^{3SEEL}$ as the solution of the equations $S_n^{3SEEL}(\theta) = 0$. It follows the logic of the EL estimator $\hat{\theta}_n^{EL}$ defined above while replacing the EL probabilities $\hat{\pi}_{i,n}^{(-1)}(\theta)$ by the more user friendly ones $\hat{\pi}_{i,n}^{(1)}(\theta)$.

Note that, for the purpose of testing the null hypothesis H_0, the five scores vectors listed above, namely $S_n^{2SGMM}(\theta)$, $S_n^{EL}(\theta)$, $S_n^{EEL}(\theta)$, $S_n^{GS}(\theta)$ and $S_n^{3SEEL}(\theta)$ will be all used after plugging in the hypothesized value $\theta = \theta_0$. They all differ on their way to sometimes replace the naive probabilities $\hat{\pi}_{i,n}$ by either the EL implied probabilities $\hat{\pi}_{i,n}^{(-1)}(\theta)$ or the EEL implied probabilities $\hat{\pi}_{i,n}^{(1)}(\theta)$. This can be done either to improve the estimation of the Jacobian matrix of the moment vector (probabilities: $\pi_{i,n}^G(\theta)$) or to improve the estimation of the variance matrix of the moment vector (probabilities: $\pi_{i,n}^V(\theta)$), or both. In the simpler context of a non-random Jacobian matrix of the moment vector (minimum distance estimation), Chaudhuri and Renault (2015) extensively document the impact of the choice of probabilities $\pi_{i,n}^V(\theta)$, especially to improve the finite-sample performance of score tests in the presence of skewness. This is the reason why we will set the priority in this paper on the choice of probabilities $\pi_{i,n}^G(\theta)$, while still considering the additional impact of choice of probabilities $\pi_{i,n}^V(\theta)$. But we will never adopt an approach similar to Guay and Pelgrin (2015) where the focus would be set on the choice of probabilities $\pi_{i,n}^V(\theta)$ while sticking to the naive ones for $\pi_{i,n}^G(\theta)$. Note that, besides higher order improvement issues, it is quite natural to set in priority the focus on probabilities $\pi_{i,n}^G(\theta)$ in order to devise inference procedures that are well-behaved even in the presence of weak identification. Kleibergen (2002, 2005)'s and Moreira (2003)'s key contribution had been to note that in order to get a score test robust to weak identification, the estimator of the Jacobian matrix of the moment vector had to be made asymptotically independent of the sample mean of the moment vector. Our Proposition 3 actually confirms that implied probabilities delivered by EEL will exactly perform the job of orthogonalization put forward in the aforementioned papers. However, Proposition 3 also shows the newer result that implied probabilities delivered by EL will also do exactly the same job, albeit even more efficiently (through the efficient estimation of the regression coefficients). Of course, this efficiency gain in the case of EL comes at the cost of the absence of a closed form formula and more computational burden.

This is the reason why we study extensively in this paper the five possible strategies listed above as (i) to (v) and summarized as follows:

Score	π^G	π^V
2SGMM	naive	naive
EL	EL	EL
EEL	EEL	naive
GS	EL	naive
3SEEL	EEL	EEL

Table 1: Score vectors considered in the paper

3.3 Score vectors based on Kullback Leibler Information Criterion (KLIC)

Even though we omit the KLIC case, corresponding to the limit case of $\gamma \to 0$, from Section 3.2 and the subsequent technical results in Section 4.2 and the asymptotic theory of implied probabilities

developed in Appendix A, we know from Proposition 2 that it corresponds to the following functions:

$$\begin{aligned}\varrho^{(0)}(x) &= 1-\exp(x)\\ \tau^{(0)}(x) &= \frac{d\varrho^{(0)}(x)}{dx} = -\exp(x)\\ k^{(0)}(x) &= -\frac{1+\tau^{(0)}(x)}{x} = -\frac{1-\exp(x)}{x}\end{aligned}$$

$$\text{so that:}\quad \hat{\pi}_{i,n}^{(0)}(\theta) \propto \tau^{(0)}\left(\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)\right) \quad\Longrightarrow\quad \hat{\pi}_{i,n}^{(0)}(\theta) = \frac{\exp\left(\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)\right)}{\sum_{j=1}^n \exp\left(\hat{\lambda}_n^{(0)}(\theta)'\psi_j(\theta)\right)} \tag{3.9}$$

with the Lagrange multipliers:

$$\begin{aligned}\hat{\lambda}_n^{(0)}(\theta) &= -\left(\tilde{\Omega}_n^{(0)}(\theta)\right)^{-1}\bar{\psi}_n(\theta)\\ \text{where}\quad \tilde{\Omega}_n^{(0)}(\theta) &= -\frac{1}{n}\sum_{i=1}^n \frac{1-\exp\left(\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)\right)}{\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)}.\psi_i(\theta)\psi_i'(\theta).\end{aligned}$$

Formula (3.9) leads to characterize the KLIC implied probabilities as "Exponential Tilting" (ET). It is then natural to consider additional score vectors by allowing γ and/or γ^* to be zero in the formula:

$$l_n(\theta,\gamma,\gamma^*) = \sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma)}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta}.\sqrt{n}\hat{\lambda}_n^{(\gamma^*)}(\theta).$$

Note that, we do not use the notation $l_n\left(\theta,\pi^G(\theta),\pi^V(\theta)\right)$ anymore because the weights that define $\tilde{\Omega}_n^{(0)}(\theta)$ cannot be interpreted as implied probabilities. Nevertheless, we are led to consider the three following score vectors:

(i) S_n^{KLIC} the score vector corresponding to the genuine KLIC minimization ($\gamma=\gamma^*=0$):

$$S_n^{KLIC}(\theta) = l_n(\theta,0,0) = \left\{\sum_{i=1}^n \hat{\pi}_{i,n}^{(0)}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta}\right\}\left[\frac{1}{n}\sum_{i=1}^n \frac{1-\exp\left(\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)\right)}{\hat{\lambda}_n^{(0)}(\theta)'\psi_i(\theta)}.\psi_i(\theta)\psi_i'(\theta)\right]^{-1}\sqrt{n}\bar{\psi}_n(\theta).$$

This score vector has been studied by Caner (2010). One may, however, argue that it is a shame to weight the matrices $\psi_i(\theta)\psi_i'(\theta)$ by coefficients that cannot be interpreted as implied probabilities.

This suggests to use the ET probabilities not only to estimate the Jacobian matrix but also to build an estimator of the variance, similarly to the estimators of variance that we had considered for EL or 3SEEL. This is the main intuition behind the score vector S_n^{ET} defined below.

(ii) S_n^{ET} the score vector where the ET probabilities are used at the two levels:

$$S_n^{ET}(\theta) = \left\{\sum_{i=1}^n \hat{\pi}_{i,n}^{(0)}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta}\right\}\left[\sum_{i=1}^n \hat{\pi}_{i,n}^{(0)}(\theta)\psi_i(\theta)\psi_i'(\theta)\right]^{-1}\sqrt{n}\bar{\psi}_n(\theta).$$

(iii) S_n^{Kl-ET} the score vector where the ET probabilities are used only to estimate the Jacobian matrix while the covariance matrix is estimated with the naive empirical probabilities:

$$S_n^{Kl-ET}(\theta) = \left\{\sum_{i=1}^n \hat{\pi}_{i,n}^{(0)}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta}\right\}\left[\frac{1}{n}\sum_{i=1}^n \psi_i(\theta)\psi_i'(\theta)\right]^{-1}\sqrt{n}\bar{\psi}_n(\theta).$$

One may note the striking analogy between the score vectors S_n^{Kl-ET} and S_n^{EEL} (considered in Section 3.2) with only $\hat{\pi}_{i,n}^{(1)}(\theta)$ replaced by $\hat{\pi}_{i,n}^{(0)}(\theta)$. Since the latter corresponds to CU-GMM and had been

put forward by Kleibergen (2005), Caner (2010) dubbed Kleibergen-ET the score vector S_n^{Kl-ET}.

4 Testing for $H_0 : \theta = \theta_0$

We revisit in this section the main result of Guggenberger and Smith (2005), namely the fact that all the GEL test statistics have the standard chi-square asymptotic null distributions independent of the strength or weakness of identification. We point out that this achievement of size control even in case of weak identification is basically an extension of Kleibergen (2005)'s original idea in the context of CU-GMM. What really matters to be robust to weak identification is to use a score vector in which the Jacobian matrix has been estimated with implied probabilities $\pi_{i,n}^G(\theta)$, irrespective of the choice of these probabilities. By drawing from the underlying asymptotic theory of the implied probabilities developed in Appendix A, we show that all the GEL score statistics, including those involving the shrunk Euclidean probabilities, deliver the correct asymptotic null distribution, even in case of weak identification. By contrast, the use of implied probabilities $\pi_{i,n}^V(\theta)$ is not really needed for the estimation of the variance matrix, at least as far as the first order asymptotics are concerned.

This is the reason why score testing based on four strategies (ii)-(v) of obtaining the score vector that we have proposed in Section 3.2 will all deliver this size control, while, on the other hand, strategy (i) will not, which will preclude its asymptotic equivalence with the rest under weak identification.

Our equivalence result is also relevant for inference under strong (resp. weak) identification; it will imply that all the score test statistics in (ii)-(v) are also equivalent under sequences of local (resp. global) alternatives. The said asymptotic equivalence actually holds uniformly in a suitable neighborhood around the true θ^0. While this uniformity is not required to establish the asymptotic equivalence of the tests against suitable "given" local alternatives, it has implications if one considers confidence intervals obtained by inverting the tests following the strategies (ii)-(v).

4.1 Score test statistics

As discussed in Section 3, the score statistic for testing $H_0 : \theta = \theta_0$ is designed as a quadratic form of $l_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right)$ in (3.7) with respect to the inverse of an (consistent under the null) estimator of its asymptotic variance. From the formula (3.7), it is natural to estimate the asymptotic variance as:

$$\begin{aligned} I_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right) &= \left\{\sum_{i=1}^n \pi_{i,n}^G(\theta_0) G_i'(\theta_0)\right\} \left[\sum_{i=1}^n \pi_{i,n}^V(\theta_0) V_{i,n}(\theta_0)\right]^{-1} \hat{V}ar_n\left[\sqrt{n}\bar{\psi}_n(\theta_0)\right] \\ &\quad \times \left[\sum_{i=1}^n \pi_{i,n}^V(\theta_0) V_{i,n}(\theta_0)\right]^{-1} \left\{\sum_{i=1}^n \pi_{i,n}^G(\theta_0) G_i(\theta_0)\right\} \end{aligned}$$

where $\hat{V}ar_n\left[\sqrt{n}\bar{\psi}_n(\theta_0)\right]$ stands for an (consistent under the null) estimator of the asymptotic variance of $\sqrt{n}\bar{\psi}_n(\theta_0)$. While one is free to use different sets of weights to reweigh the estimators of the Jacobian and/or the variance matrix appearing in $l_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right)$ and $I_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right)$ respectively, we choose, for the sake of notational simplicity and for the purpose of our simulations as well, to use the same set of weights, so that we take:

$$I_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right) = \left\{\sum_{i=1}^n \pi_{i,n}^G(\theta_0) G_i'(\theta_0)\right\} \left[\sum_{i=1}^n \pi_{i,n}^V(\theta_0) V_{i,n}(\theta_0)\right]^{-1} \left\{\sum_{i=1}^n \pi_{i,n}^G(\theta_0) G_i(\theta_0)\right\}$$

and the associated score test statistic:

$$LM_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right) = l_n'\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right) \left[I_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right)\right]^{-1} l_n\left(\theta_0, \pi^G(\theta_0), \pi^V(\theta_0)\right).$$

4.2 Kleibergen-type statistics

We want to show now that all the score tests statistics defined above are first-order asymptotically equivalent to the K-statistic of Kleibergen (2005), insofar as probabilities are chosen as follows:

$$\begin{aligned} \pi_{i,n}^{G}(\theta_0) &\in \left\{\hat{\pi}_{i,n}^{(\gamma)}(\theta_0); \gamma \in \{\mathbb{R} \setminus \{0\}\}\right\} \cup \{\hat{\pi}_{i,n}^{(1),Sh}(\theta_0)\} \\ \pi_{i,n}^{V}(\theta_0) &\in \left\{\hat{\pi}_{i,n}^{(\gamma)}(\theta_0); \gamma \in \{\mathbb{R} \setminus \{0\}\}\right\} \cup \{\hat{\pi}_{i,n}^{(1),Sh}(\theta_0)\} \cup \{\hat{\pi}_{i,n}\}. \end{aligned} \tag{4.1}$$

Note that, the K-statistic of Kleibergen (2005) itself, even though it was introduced directly from a CU-GMM optimization, is actually nested in our framework with the choice:

$$\pi_{i,n}^{G}(\theta_0) = \hat{\pi}_{i,n}^{(1)}(\theta_0) \text{ and } \pi_{i,n}^{V}(\theta_0) = \hat{\pi}_{i,n}.$$

In this case, we will use the notation:

$$LM_n\left(\theta_0, \hat{\pi}_{.,n}^{(1)}(\theta_0), \hat{\pi}_{.,n}\right) = K_n(\theta^0).$$

We will actually be able to show also the asymptotic equivalence with a simplified version of the K-statistic of Kleibergen (2005) in which the weights are taken as the implied probabilities $\hat{\pi}_{i,n}^{(1)}(\theta_0)$ only for the components of the score vector corresponding to weak identification, while we can safely keep the naive empirical probabilities $\hat{\pi}_{i,n}$ for the other components. To see that, we will consider a setup of weak identification similar to the one proposed by Stock and Wright (2000):

Assumption ID: (Weak identification)
The parameter space Θ is a Cartesian product: $\Theta = \Theta_w \times \Theta_s$ where $\Theta_w \subset \mathbb{R}^{p_w}$ and $\Theta_s \subset \mathbb{R}^{p_s}$ and, accordingly, we have a partition of the general and true unknown values of $\theta = (\theta_w', \theta_s')'$ as: $\theta^0 = (\theta_w^{0\prime}, \theta_s^{0\prime})'$ where $\theta_w^0 \in Int(\Theta_w)$ and $\theta_s^0 \in Int(\Theta_s)$, and a decomposition of the expectation of the moment vector as:

$$E\left[\bar{\psi}_n(\theta)\right] = m(\theta_s) + \frac{1}{\sqrt{n}} m_n^w(\theta)$$

where $m(\theta_s)$ is such that $m(\theta_s) = 0 \Longleftrightarrow \theta_s = \theta_s^0$, and the sequence $m_n^w(\theta)$ converges uniformly towards a continuous function $m^w(\theta)$: $\sup_{\theta \in \Theta} \|m_n^w(\theta) - m^w(\theta)\| = o(1)$ satisfying $m^w(\theta^0) = 0$. ■

The above partition of the parameters leads to a corresponding partition of the Jacobian matrix:

$$G_i(\theta) = \frac{\partial \psi_i(\theta)}{\partial \theta'} = \left[\frac{\partial \psi_i(\theta)}{\partial \theta_w'}, \frac{\partial \psi_i(\theta)}{\partial \theta_s'}\right].$$

We can then formalize as follows the general idea of taking implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta_0)$ only for the components of the score vector corresponding to weak identification, while safely keeping the naive probabilities $\hat{\pi}_{i,n}$ for the other components. Of course, the case without weak identification is the particular case where the dimension p_w is zero. Accordingly, let us define another score vector as:

$$l_n\left(\theta, \tilde{\pi}^G(\theta), \pi^V(\theta)\right) = \left\{ \begin{array}{c} \sum_{i=1}^n \pi_{i,n}^G(\theta) \frac{\partial \psi_i'(\theta)}{\partial \theta_w} \\ \sum_{i=1}^n \hat{\pi}_{i,n} \frac{\partial \psi_i'(\theta)}{\partial \theta_s} \end{array} \right\} \left[\sum_{i=1}^n \pi_{i,n}^V(\theta) V_{i,n}(\theta)\right]^{-1} \sqrt{n}\bar{\psi}_n(\theta)$$

and, with obvious notations, define a corresponding score test statistic as:

$$LM_n\left(\theta_0, \tilde{\pi}^G(\theta_0), \pi^V(\theta_0)\right) = l_n'\left(\theta_0, \tilde{\pi}^G(\theta_0), \pi^V(\theta_0)\right)\left[I_n\left(\theta_0, \tilde{\pi}^G(\theta_0), \pi^V(\theta_0)\right)\right]^{-1} l_n\left(\theta_0, \tilde{\pi}^G(\theta_0), \pi^V(\theta_0)\right).$$

In this expression of the score test statistic, we use $\tilde{\pi}^G(\theta_0)$ instead of the $\pi^G(\theta_0)$ introduced earlier in (4.1). This is done to emphasize that here the columns corresponding to θ_w (the weakly identified

elements of θ) in the Jacobian matrix are weighted using the implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta_0)$, while the remaining columns are weighted by the naive empirical probabilities $\hat{\pi}_{i,n}$.

Then our simplified K-statistic will be:

$$\tilde{K}_n(\theta_0) = LM_n\left(\theta_0, \tilde{\pi}^G(\theta_0), \hat{\pi}_{.,n}\right) \quad \text{with} \quad \pi^G(\theta_0) = \hat{\pi}_{.,n}^{(1)}(\theta_0).$$

We want to show the general asymptotic equivalence between any score test statistic conformable to (4.1) with both $K_n(\theta_0)$ and $\tilde{K}_n(\theta_0)$ under two types of possible circumstances:

(i) The true unknown value θ^0 is the hypothesized value θ_0.

(ii) The true unknown value θ^0 differs from the hypothesized value θ_0 such that the p_w weakly identified components may differ arbitrarily (inside Θ_w) while the p_s strongly identified components may differ along a sequence of $\sqrt{n}$-local alternatives.

The bottom line is the definition of a set Θ_n of possible hypothesized values θ_0 such that the moments of interest are local to zero in the sense of Assumption 1(i). For this purpose we define Θ_n as follows. Let $r > 0$ be such that the open ball $B(\theta_s^0, r)$ of center θ_s^0 and of radius r is included in Θ_s. Then we define:

$$\Theta_n = \Theta_w \times B\left(\theta_s^0, \frac{r}{\sqrt{n}}\right). \tag{4.2}$$

The key idea is that both local alternatives and weak identification are settled for the moments to stay "local", which will allow us to apply the equivalence results based on the implied probabilities that are developed in Appendix A. Of course, further standard assumptions following Kliebergen (2005) and Andrews and Guggenberger (2017) are required: namely, the smoothness of the sample and population moment vectors (Assumptions SSM and SPM), the joint convergence in distribution of the sample moment vector and its derivative (Assumption CLT), the existence of uniformly (in Θ_n) consistent estimators for suitable elements of the asymptotic variance of this limiting distribution (Assumption O), and an almost-sure rank condition for the limit of the suitably scaled derivative of the sample moment vector (Assumption R). To avoid clutter in the main text, these assumptions along with the intermediate results leading to our final result, i.e., Theorem 1, are collected in Appendix D.

Theorem 1: Under Assumptions 1, 2, ID, SSM, SPM1-2, CLT, O and R (stated in Appendix D), we have for any choice of implied probabilities $\pi^G(\theta)$ and $\pi^V(\theta)$ conformable to (4.1), that:

$$\begin{aligned} \sup_{\theta\in\Theta_n} \left| LM_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right) - K_n(\theta) \right| &= o_P(1) \\ \sup_{\theta\in\Theta_n} \left| LM_n\left(\theta, \pi^G(\theta), \pi^V(\theta)\right) - \tilde{K}_n(\theta) \right| &= o_P(1). \ \blacksquare \end{aligned}$$

Recall that Kleibergen (2005)'s K test rejects the null hypothesis $H_0 : \theta = \theta_0$ at the level α if

$$K_n(\theta) > \chi_p^2(1-\alpha)$$

where $\chi_p^2(1-\alpha)$ is the $(1-\alpha)$-th quantile of a central χ^2 distribution with p degrees of freedom. In general, the asymptotic properties of this test for $\theta_0 = \theta^0$, i.e., when H_0 is true, and $\theta_0 = (\theta_w', (\theta_s^0 + \frac{d}{\sqrt{n}})')' \in \Theta_n$, i.e., under "suitable local" alternatives, are of interest.

Among other things, Theorem 1 implies that the K-test's first-order asymptotic size and power against local alternatives are all inherited by the score tests based on the four strategies (ii)-(v) put forward in our paper. Then, the question is: whether these four strategies that involve efficient estimation, as described in Sections 2 and 3, of the Jacobian and possibly the variance matrix are worthwhile? In a simulation exercise we now demonstrate that, in spite of the first-order asymptotic equivalence in Theorem 1, these four strategies provide demonstrable improvements not only for the finite-sample size but also, and even more so, for the finite-sample power of the score tests.

4.3 Monte Carlo experiments

4.3.1 Design

We use as the data generating process (DGP) for the observable variables $W_i = (y_i, X_i, Z_i')'$, for $i = 1, \ldots, n$, the following triangular system:

$$\begin{aligned} y_i &= X_i\theta^0 + u_i \\ X_i &= Z_i'\Pi + v_i \end{aligned} \tag{4.3}$$

whose inputs are i.i.d. draws of $(u_i, v_i, Z_i), i = 1 \ldots, n$.

In this system, there is a single right-hand-side (possibly) endogenous variable X_i ($X_i \in \mathbb{R}$) and no included exogenous variables, and H instrumental variables Z_i ($Z_i \in \mathbb{R}^H$).

By definition, the structural errors (u_i, v_i) are independent of the instruments Z_i. The endogeneity of X_i is governed by the correlation ϱ between u_i and v_i. We consider three levels of endogeneity: $\varrho = 0.9$ (highly endogenous), $\varrho = 0.5$ (moderately endogenous) and $\varrho = 0$ (not endogenous).

The moment vector is given by:

$$\psi_i(\theta) \equiv \psi\left(W_i, \theta\right) = Z_i\left(y_i - X_i\theta\right).$$

As noted before, it could be important to allow the implied probabilities to improve the estimation of the variance matrix. This implies considering moment conditions that may display some kind of multivariate asymmetry because for some $h, k, l = 1, \ldots, H$:

$$E\left[\psi_{i,h}(\theta^0)\psi_{i,k}(\theta^0)\psi_{i,l}(\theta^0)\right] \neq 0,$$

where $\psi_{i,h}(\theta)$ is the h-th element of $\psi_i(\theta)$ for $h = 1, \ldots, H$. By the law of iterated expectations:

$$E\left[\psi_{i,h}(\theta^0)\psi_{i,k}(\theta^0)\psi_{i,l}(\theta^0)\right] = E\left[Z_{i,h}.Z_{i,k}.Z_{i,l}\right]E[u_i^3].$$

Therefore, we take the following route to allow for a possible asymmetry in the distribution of the moment vector. First, we draw Z_i from a distribution with non zero mean. We actually take $Z_{i,h}, h = 1, \ldots, H$ mutually independent normal with unit mean and variance. Then, the wished multivariate asymmetry of the moment vector $\psi_i(\theta^0)$ is completely governed by the skewness of the error term u_i.

Note that, a common way to accommodate for an asymmetric moment vector is to endow u with the probability distribution of a demeaned even power of a normal (while the other error term v would be a standard normal). However, we do not want to do that since it would preclude endogeneity. This is the reason why we consider the two cases:

1st case: Symmetric: u_i and v_i are standard normal, with correlation ϱ.

2nd case: Asymmetric: Let $e_{1,i}, e_{2,i}, i = 1, \ldots, n$ stand for $(2n)$ independent draws in the exponential(1) distribution. For a given real number c, define the zero mean variables:

$$u_i = \frac{1}{\sqrt{2}}\left[e_{1,i} + e_{2,i} - 2\right] \quad \text{and} \quad v_i = \frac{1}{\sqrt{1+c^2}}\left[e_{1,i} + ce_{2,i} - (1+c)\right].$$

Then the correlation between u_i and v_i is:

$$\frac{1+c}{\sqrt{2}\sqrt{1+c^2}}.$$

We get the requested level ϱ of correlation by choosing c as a solution of: $2\varrho^2(1+c^2) = (1+c)^2$. We choose:

$$c = -\frac{1 - 2\varrho\sqrt{1-\varrho^2}}{1-2\varrho^2}.$$

It is worth noting that the level of skewness in u_i and v_i is not huge. For instance, the skewness

coefficient of u_i is equal to $\sqrt{2}$. This reasonable amount of skewness makes even more compelling the evidence of improvement that we will display in case of asymmetric moment conditions (see especially power comparisons below).

We consider the cases $H = 2$ and $H = 2$, and assume that the H instruments have the same degree of weakness by taking:

$$\Pi = \frac{a}{\sqrt{n}} 1_H$$

where 1_H is the H-dimensional vector with all components equal to 1. We consider two sample sizes: $n = 100$ (small sample) and $n = 1000$ (reasonable for a micro-econometric application).

For a given sample size, a governs the strength of identification through the concentration parameter:

$$\mu = \frac{1}{H} \Pi' \left(\sum_{i=1}^{n} Z_i Z_i' \right) \Pi.$$

We want to describe three different patterns of identification: strong, weak, and complete lack of identification. While each pattern will be studied through the draw of 10000 samples of size n each, we slightly modify the parameter a from one sample to the other in order to keep the same value of the concentration parameter for each draw in a given pattern:

$\mu = 10$ for strong identification, $\mu = 1$ for weak identification and $\mu = 0$ for lack of identification.

To summarize, it means that we are considering $72 = 2 \times 3 \times 2 \times 3 \times 2$ specifications of the DGP, for which we have 10000 replications:

- 2 values of the number H of instruments ($H = 2$ or 4),
- 3 levels of endogeneity ($\varrho = 0$ or $\varrho = 0.5$ or $\varrho = 0.9$),
- 2 patterns of symmetry: one symmetric distribution, one asymmetric distribution,
- 3 levels of instrument strength ($\mu = 10$ or $\mu = 1$ or $\mu = 0$),
- 2 possible sample sizes ($n = 100$ and $n = 1000$).

4.3.2 Size comparisons

Empirical sizes are first calculated using the 5% asymptotic critical values for the nine score tests of interest, namely:

- The five tests with score statistics corresponding to score vectors described in Section 3.2 (see Table 1) : 2SGMM, GS, EL, EEL and 3S which stands for 3SEEL.
- The modification of 3SEEL which uses shrunk probabilities as defined at the end of Section 2.2. The test statistic is denoted by 3S-Sh.
- The three score tests with statistics based on KLIC (see Section 3.3) namely KLIC, ET and Kl-ET.

The complete set of results on empirical size is reported in Tables 2 and 3 in Appendix E.

We do not impose conditional (on Z) homoskedasticity of the errors when computing the asymptotic variance matrix of the average moment vector. In this sense, the statistics belong to the genre what Guggenberger and Smith (2005, page 689-690) refer to as the K_{HET} statistic.

With a sample size $n = 1000$ and a symmetric moment vector, all empirical sizes of all the score tests except 2SGMM are controlled below 6%. This result is all the more compelling since, at the same time, size distortions of 2SGMM may go as far as 35.8% (case of complete lack of identification, strong endogeneity and 4 instruments) and 17.2% when, in the same circumstances, identification exists but only at a weak level. Generally speaking, when some endogeneity is at stake, the empirical size of 2SGMM is almost never below 6%, and these exceptions occur only when identification is strong. Note moreover that, none of the tests is overly conservative (empirical sizes are never below 4.4%). With

the same sample size $n = 1000$, the introduction of an asymmetric moment vector leads to slightly more oversized tests, but nothing dramatic except for the 2SGMM case (never beyond an empirical size of 6.4% for the rest).

Not surprisingly, the 5% asymptotic critical values are less reliable when the sample size is only $n = 100$. In this case, besides 2SGMM, the worst performance is achieved by the 3SEEL methods, with or without shrinkage. Empirical sizes are typically between 8% and 17.1%, which is not compellingly better than the 2SGMM. By contrast, the EEL approach, as promoted by Kleibergen (2005), does an excellent job: the empirical size is always below 6% in the symmetric case, below 8.5% in the asymmetric case. Interestingly enough, EEL dominates both the two kinds of empirical likelihood approaches (EL and GS) as well as the three kinds of ET approaches (KLIC, ET and Kl-ET). The small sample over-rejection seems especially significant for KLIC and ET (empirical size as large as 10% in the symmetric case, 11.1% in the asymmetric case). By contrast, GS, EL and Kl-ET all do a decent job (empirical size below 8% in the symmetric case, 9% in the asymmetric case). Note however that, except for the two 3SEEL methods (when $n = 100$), the performance of all methods based on the implied probabilities is still significantly much better than 2SGMM.

The somewhat disappointing performance of 3SEEL methods in small samples is not surprising. When the sample is too small, the probability of negative values of implied probabilities is far from negligible, and shrinking them just amounts to setting too many of them equal to the naive empirical probabilities. Over the 10000 replications of samples of size $n = 100$ (repeated for each specification), we actually met a significant proportion of them for which some EEL implied probabilities were negative: always between 5% and 6% of them in the symmetric case, and between 17% and 21% of them in the asymmetric case. By contrast, over the 10000 replications of samples of size $n = 1000$, negative EEL implied probabilities never show up in the symmetric case and they occur in less than 1% of the samples in the asymmetric case.

Note also that, it is only for small samples that increasing the number of instruments (from 2 to 4) leads to more over-rejection, in particular for the less reliable methods in small samples, namely 3S, 3S-Sh, KLIC, ET, GS and EL.

The bottom line is that only the 2SGMM tests have an over-rejection rate significantly increased by the weakness of instruments. The satisfactory performance in this respect for all eight alternative statistics is clearly explained by the fact that they all use implied probabilities for estimation of the Jacobian part. By contrast, it sounds relatively useless (or even detrimental in small samples) to use the implied probabilities for the estimation of the variance matrices. The excellent performance of the EEL/K statistic (and also Kl-ET) is especially convincing in this respect. The main message is that, as far as size control in the presence of weak identification is concerned, it is the use of the implied probabilities to estimate the Jacobian part (and, in particular, to ensure that this estimator is asymptotically independent of the moment conditions) that matters more than the use of the genuine GEL-based inference methods.

4.3.3 Power comparisons

Empirical size-adjusted power curves are calculated for the preceding eight 5% score tests that are alternative to 2SGMM. These eight tests were shown to have correct asymptotic size under scenarios that cover the cases considered in our Monte Carlo experiment. On the other hand, since 2SGMM suffers from severe size distortion, we have excluded it from the comparison of the size-adjusted power.

The complete set of results on empirical power is reported in Figures 1–12 in Appendix E.

With a sample size $n = 1000$ and a symmetric moment vector, the eight size-adjusted power curves are pretty much the same (see Figures 4,5,6 and 10,11,12). By contrast, when considering asymmetric moment vectors, the 3SEEL approach and even more the 3SEEL with shrinkage (3S-Sh) display a better power performance (see Figures 1,2,3 and 7,8,9), especially when identification is weak ($\mu = 1$) or absent ($\mu = 0$). In other words, the use of the EEL implied probabilities to estimate not only the Jacobian matrix (as it comes for the EEL/K case) but also to improve the estimation of the variance

matrix is important as far as power is concerned. For the same reason EL (resp. ET or KLIC) dominates GS (resp. Kl-ET).

This remark is all the more interesting since it is worth reminding that, for such large sample sizes, there is no cost in terms of size control to re-use the same implied probabilities for estimation of both Jacobian and variance matrices, while it is not what the GEL approach would lead to do, except in the EL case. By the way, the power performance would rather lead us to use everywhere the EEL probabilities, that turn out to also be the most user friendly, since these are the only ones for which we have closed form formulas (both for the implied probabilities and for the estimated expectations, as residuals of regressions on the moment conditions). To put it differently, the naive least squares approach dominates the computationally more involved alternative approaches, i.e., the EL and KLIC ones. While the only cost of using the EEL implied probabilities might have been the finding of negative probabilities, the shrinkage trick fixes this issue without deteriorating the power performance, quite the contrary indeed.

By contrast, with a small sample size ($n = 100$), we have shown that both 3S approaches are a bit dominated by standard GEL approaches, in terms of size control. Interestingly enough, as far as size-adjusted power is concerned, both 3S approaches keep the edge over all other 6 tests under study. They still strictly dominate when moment vectors have an asymmetric probability distribution and identification is weak ($\mu = 1$) or absent ($\mu = 0$).

To be complete, it must be acknowledged that the genuine EL approach may display some edge over all the three Euclidean approaches in one circumstance, that is when we observe (see Figure 2, both for $n = 100$ and $n = 1000$), for a high degree of endogeneity ($\varrho = 0.9$), some spurious decline in power far from the null hypothesis. This effect was already documented for EEL by Kleibergen(2005), and may still show up, not only for EEL but also, to a lesser extent, for 3SEEL methods and even for the genuine EL. However, EL may sometimes do the best job in this case.

4.3.4 Main Message of the Monte Carlo study

While it had already been well documented (see in particular Guggenberger and Smith (2005) and Caner (2010)) that GEL approaches were able to do a much better job than 2SGMM in case of weak identification (much better size control and sufficiently good power performances), we argue from our results that this good performance is more due to the use of the implied probabilities than to the use of the genuine GEL inference. We actually show that there is no compelling reason to prefer the computationally and numerically involved implied probabilities provided by empirical likelihood or by exponential tilting, while the naive Euclidean probabilities do an excellent job. As already put forward by Kleibergen (2005), they do (at least for sufficient sample sizes) the best possible job when used to estimate the Jacobian matrix in order to get a size of the test that is robust to weak identification. With our so-called three steps EEL methods (following the terminology of ABR-07), we show that the Euclidean methods are also well suited for improving the estimation of the variance matrix (in case of asymmetrically distributed moment conditions) and, in turn, the power of the test. When it turns to the estimation of the variance, the shrinkage procedure proposed by ABR-07 to ensure the non-negativity of the implied probabilities is even more relevant.

5 Conclusion

While information theoretic approaches have become a popular alternative to GMM, their main use in econometrics is rather as a black box providing satisfactory solutions to the poor finite sample performance of GMM, in particular in the case of weak instruments. While the mechanism of this black box is built from the implied probabilities, the information content of these probabilities has been relatively little documented. In the context of score testing, this paper has promoted the use of implied probabilities in two ways:

First, it extends Kleibergen's (2005) seminal contribution, showing that the use of an estimator of the Jacobian that is uncorrelated with moment conditions allows a dramatic improvement of the size control in the case of weak identification.

Second, it draws the lesson from Newey and Smith (2004) as well as ABR-07 to note that the implied probabilities allow efficient estimation of any moment matrix, allowing a more accurate estimation of the selection matrix of the estimated equations, for a more powerful score test. In particular, as explained by these authors, it is in cases of skewness of the moment vectors that an efficient use of the implied probabilities for estimation of the variance matrix is especially relevant.

The value added by the present paper is threefold:

First, we show that the implied probabilities provided by EL produce an orthogonalization of the estimator of the Jacobian matrix with respect to the moment conditions, in the same way as the first order conditions of the continuously updated GMM for Kleibergen (2005) (or, equivalently, the estimator based on the probabilities provided by EEL for ABR-07). It is important to recognize that this orthogonalization does not need to appeal to asymptotics even in the case of EL, as long as we rely on the EL implied probabilities instead of the empirical probabilities.

Second, we notice that, more generally, the implied probabilities obtained by minimization of any Cressie Read discrepancy are asymptotically equivalent in the sense that they all produce asymptotically equivalent (up to the order $o_p(n^{-1/2})$) estimators of any moment functions. In particular, this follows from our result that the implied probabilities from different Cressie Read discrepancies may only differ by the order of $o_p(n^{-3/2})$.

Third, we document that the use of the implied probabilities is a valuable hedge against the size distortion effects of weak identification when it is performed for the estimation of the Jacobian matrix, while efficient estimation of the variance matrix is not necessary for this purpose. By contrast, the latter is important for better finite-sample power of the score tests. The role of the implied probabilities as a hedge against weak identification is even more explicitly pinned down by showing that, in a setting of a vector of parameters a la Stock and Wright (2000), with only a subvector that is weakly identified, it is for the weakly identified components that the efficient estimation of the expected partial derivatives matters.

Moreover, an important technical insight of this paper is, as a generalization of the approach of the GEL theory of Guggenberger and Smith (2005), an asymptotic theory with results that are uniform not only under the null but also under a sequence of alternatives relevant for the power of tests. These alternatives are local in the case of strong identification but global in the case of weak identification.

As already mentioned, an innovation with respect to Guggenberger and Smith (2005) is that we do not need to deal with the genuine GEL score vectors but we can instead freely work with different implied probabilities for estimating the Jacobian matrix and the variance matrix respectively. We document in particular that the use of the implied probabilities based on EEL, with a possible shrinking a la ABR-07 for non-negativity, may be very efficient for score testing, especially for the power of tests based on skewed moments.

A byproduct of the degree of freedom regarding the choice of well suited implied probabilities is a computational advantage, since, for instance, the implied probabilities based on EEL amount to linear regressions. However, we only consider in this paper the score tests for null hypotheses that specify the value of the full vector of parameters. Testing on the value of a subvector (or more generally on a function of the parameters) would be more demanding since it may require some numerically involved constrained estimation. A companion paper, Chaudhuri and Renault (2018), proposes to use the tool of $C(\alpha)$ tests to circumvent the computational issues. Moreover, testing on functions of the parameters paves the way for important cases where, as studied by Antoine and Renault (2012), different identification strengths may be at stake in different directions in the parameter space.

References

Anderson, T. W. and Rubin, H. (1949). Estimation of the Paramters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics*, 20:46–63.

Andrews, D. W. K. and Guggenberger, P. (2017). Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models. *Econometric Theory*, 33:1046–1080.

Antoine, B., Bonnal, H., and Renault, E. (2007). On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics*, 138:461–487.

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

Back, K. and Brown, D. (1993). Implied Probabilities in GMM estimators. *Econometrica*, 61:971–976.

Bera, A., Montes-Rojas, G., and Sosa-Escudero, W. (2010). General Specification Testing with Locally Misspecified Models. *Econometric Theory*, 26: 1838–1845.

Caner, M. (2010). Exponential tilting with weak instruments: Estimation and testing. *Oxford Bulletin of Economics and Statistics*, 72:307–326.

Chaudhuri, S. and Renault, E. (2015). Shrinkage of Variance for Minimum Distance Based Tests. *Econometric Reviews*, 34:328–351.

Chaudhuri, S. and Renault, E. (2018). Tests for non-linear restrictions with heterogeneity in identification strengths. Mimeo.

Gagliardini, P., Gourieroux, C., and Renault, E. (2011). Efficient Derivative Pricing by the Extended Method of Moments. *Econometrica*, 79:1181–1232.

Guggenberger, P. and Smith, R. (2005). Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification. *Econometric Theory*, 21:667–709.

Hall, A. R. (2000). Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test. *Econometrica*, 68:1517–1527.

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

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics*, 14:262–280.

Hausman, J., Lewis, R., Menzel, K., and Newey, W. (2011). Properties of the CUE estimator and a modification with moments. *Journal of Econometrics*, 165:45–57.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information Theoretic Approaches to Inference in Moment Condition Models. *Econometrica*, 66:333–357.

Kitamura, Y. and Stutzer, M. (1997). An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica*, 65:861–874.

Kleibergen, F. (2002). Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression. *Econometrica*, 70:1781–1803.

Kleibergen, F. (2005). Testing Parameters In GMM Without Assuming That They Are Identified. *Econometrica*, 73:1103–1123.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, 71:1027–1048.

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

Newey, W. K. and West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, 28:777–787.

Owen, A. B. (1990). Empirical Likelihood Ratio Confidence Regions. *Annals of Statistics*, 18:90–120.

Ramalho, J. and Smith, R. (2006). Goodness of Fit Tests for Moment Condition Models. Mimeo.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35:634–672.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal*, 107:503–519.

Stock, J. H. and Wright, J. H. (2000). GMM with Weak Identification. *Econometrica*, 68:1055–1096.

Wang, J. and Zivot, E. (1998). Inference on a Structural Parameter in Instrumental Variables Regression with Weak Instruments. *Econometrica*, 66:1389–1404.

A Appendix A: Asymptotic theory of implied probabilities

A.1 General Motivation

While it is well known that for any $\gamma \neq 0$ (we omit the case of exponential tilting for the sake of notational simplicity), minimization of the Cressie-Read discrepancy delivers implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta), i = 1, \ldots, n$, defined by (2.5) (up to normalization), little has been done to compare these different probability distributions for different values of the real number γ. Hausman et al. (2011) stated (with our notations) that, if $\hat{\theta}_n^{EL}$ (resp. $\hat{\theta}_n^{CU}$) stands for the EL (resp. EEL) estimator of θ:

$$\hat{\pi}_{i,n}^{(-1)}(\hat{\theta}_n^{EL}) = \hat{\pi}_{i,n}^{(1)}(\hat{\theta}_n^{CU}) + O_P(1/n), \tag{A.1}$$

which is little informative for our purpose, since our goal is to improve upon the naive empirical probabilities $\hat{\pi}_{i,n} = (1/n)$ which also fulfill for any γ and for any $\sqrt{n}$-consistent estimator $\hat{\theta}_n$:

$$\hat{\pi}_{i,n}^{(\gamma)}(\hat{\theta}_n) = \hat{\pi}_{i,n} + O_P(1/n).$$

For the purpose of efficient use of the information content of the moment conditions, what really matters is that the asymptotic distribution of the estimator:

$$\bar{g}_n^{(\gamma)}(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\, g(W_i, \theta) \tag{A.2}$$

is an improvement upon the naive estimator $\bar{g}_n(\theta)$ (see Proposition 3 for the particular cases $\gamma = \pm 1$). Our main goal in this section is to show that all these estimators $\bar{g}_n^{(\gamma)}(\theta)$ for any $\gamma \neq 0$ are asymptotically equivalent, at least for θ in a convenient neighborhood of the true unknown value θ^0, and indeed asymptotically more accurate than $\bar{g}_n(\theta)$.

This result will be applied to the various score tests in the sense that all score test statistics involve some sample means whose asymptotic variance can be reduced by using instead the optimally weighted averages as (A.2). The key intuition is that the improvement of accuracy of estimators like (A.2) should result in a power gain for corresponding score tests, leading to score tests that are more powerful than the standard score test of Newey and West (1987). Moreover, thanks to the aforementioned asymptotic equivalence, Proposition 3 shows that in estimators like (A.2), any perverse correlation with moment conditions has been erased, which should give us a hedge against bad size distortions.

Since the ultimate goal is to compare asymptotic distributions, we want to get equivalence results warranting that for all θ in some neighborhood of θ^0 and for any pair (γ, γ^*) of non-zero real numbers:

$$\sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)}(\theta)\, g(W_i, \theta) - \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma^*)}(\theta)\, g(W_i, \theta) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

Since we are summing over n terms, such a degree of asymptotic equivalence requires intuitively that for $i = 1, 2, \ldots, n$:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta) - \hat{\pi}_{i,n}^{(\gamma^*)}(\theta) = o_P\left(\frac{1}{n\sqrt{n}}\right). \tag{A.3}$$

Note that, the equivalence result (A.3) is much more powerful than (A.1). In particular, it would not work with the naive (unconstrained) probabilities $\hat{\pi}_{i,n} = (1/n)$. We will prove it in the next subsection. Note, however, that we do not prove that:

$$\max_{1 \leq i \leq n} \left[\hat{\pi}_{i,n}^{(\gamma)}(\theta) - \hat{\pi}_{i,n}^{(\gamma^*)}(\theta)\right] = o_P\left(\frac{1}{n\sqrt{n}}\right).$$

Therefore, (A.3) does not directly provide the asymptotic equivalence result between the weighted averages that we are looking for. It will take additional assumptions and proofs in a further subsection.

A.2 Multipliers and Implied Probabilities

We derive in this subsection a formal theory of Lagrange multipliers and implied probabilities that will justify the informal analysis of Appendix A.1. In particular, while we maintain the regularity conditions announced in the Introduction, as well as the concept of unique true unknown value defined by the moment conditions, we want to allow for various identification patterns, involving both strong and weak identification. Typically, the definition of the latter involves drifting data generating processes and thus the introduction of double arrays of observations indexed by both n and $i = 1, \ldots, n$. We will keep the notations simple by not making explicit these double arrays, keeping the notation $(W_i)_{1 \leq i \leq n}$ instead of the correct notation $(W_{i,n})_{1 \leq i \leq n}$.

For the purpose of the results from here onward, it is useful to bestow Θ_n defined in Assumption 1 with two additional properties.

Assumption 2: Let the sequence $\Theta_n, n = 1, 2, \ldots$ defined in Assumption 1 also satisfy:

(i) $\sup_{\theta \in \Theta_n} \left\| \hat{\Omega}_n(\theta) - V(\theta) \right\| = o_p(1)$, $\sup_{\theta \in \Theta_n} \left\| \hat{V}_n^{(1)}(\theta) - V(\theta) \right\| = o_p(1)$,
$\sup_{\theta \in \Theta_n} \left\| \left[\hat{\Omega}_n(\theta)\right]^{-1} - V^{-1}(\theta) \right\| = o_p(1)$ and $\sup_{\theta \in \Theta_n} \left\| \left[\hat{V}_n^{(1)}(\theta)\right]^{-1} - V^{-1}(\theta) \right\| = o_p(1)$.

(ii) $0 < \inf_{\theta \in \Theta_n} \gamma_{\min}(\theta) < \sup_{\theta \in \Theta_n} \gamma_{\max}(\theta) < +\infty$ where $\gamma_{\min}(\theta)$ and $\gamma_{\max}(\theta)$ stand for the smallest and largest eigenvalues respectively of $V(\theta)$.

Assumption 2 (i) only maintains the validity of a uniform law of large numbers for the sample covariance matrix, with a population covariance matrix. The two convergence on the first line of (i) are equivalent, thanks to Assumption 1 (i) and (iv). The same is true for the convergence on the second line under the additional condition of Assumption 2 (ii), which maintains that the population covariance matrix is finite and positive definite.

We first prove the asymptotic equivalence of the sequences of Lagrange multipliers $\hat{\lambda}_n^{(\gamma)}(\theta)$ for $\theta \in \Theta_n$, defined in Proposition 1.

Proposition A.1: Under Assumptions 1 and 2, for any $\gamma \neq 0$:

$$\sup_{\theta \in \Theta_n} \left\| \hat{\lambda}_n^{(\gamma)}(\theta) - \hat{\lambda}_n^{(1)}(\theta) \right\| = o_P(1/\sqrt{n})$$

with $\hat{\lambda}_n^{(1)}(\theta) = -\hat{\Omega}_n(\theta)^{-1}\bar{\psi}_n(\theta)$ and $\hat{\Omega}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_i(\theta)\psi_i(\theta)'$. ■

Note that, while we only knew from Proposition 1 that for all γ:

$$\sup_{\theta\in\Theta_n} \left\| \hat{\lambda}_n^{(\gamma)}(\theta) \right\| = O_P(1/\sqrt{n}),$$

Proposition A.1 implies in addition a smaller order of magnitude for the difference between the Lagrange multipliers corresponding to the different values of γ:

$$\gamma, \gamma^* \in \{\mathbb{R} \setminus \{0\}\} \Rightarrow \sup_{\theta\in\Theta_n} \left\| \hat{\lambda}_n^{(\gamma)}(\theta) - \hat{\lambda}_n^{(\gamma^*)}(\theta) \right\| = o_P(1/\sqrt{n}).$$

In terms of the implied probabilities, Proposition A.1 implies directly:

Proposition A.2: Under Assumptions 1 and 2, we have for any $\gamma \neq 0$ and every $i = 1, \ldots, n$:

$$\sup_{\theta\in\Theta_n} \left| \hat{\pi}_{i,n}^{(\gamma)}(\theta) - \hat{\pi}_{i,n}^{(1)}(\theta) \right| = o_P\left(\frac{1}{n\sqrt{n}}\right) \quad \text{and} \quad \sup_{\theta\in\Theta_n} \left| \hat{\pi}_{i,n}^{(\gamma)}(\theta) - \frac{1}{n} \right| = O_P\left(\frac{1}{n\sqrt{n}}\right)$$

with $\hat{\pi}_{i,n}^{(1)}(\theta) = \frac{1}{n} - \bar{\psi}_n(\theta)'[\hat{V}_n^{(1)}(\theta)]^{-1}\frac{\psi_i(\theta)-\bar{\psi}_n(\theta)}{n}$, and where $\hat{V}_n^{(1)}(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_i(\theta)[\psi_i(\theta) - \bar{\psi}_n(\theta)]'$. ■

Note that, the result of Proposition A.2 is similar to Lemma 14 of Ramalho and Smith (2006). As mentioned in Section 2.1, one may also contemplate the use of the shrunk probabilities $\hat{\pi}_{i,n}^{(1),Sh}(\theta)$ to avoid the perverse possible negativity of the EEL probabilities $\hat{\pi}_{i,n}^{(1)}(\theta)$. The nice thing with the shrunk probabilities is that they also define the implied probabilities that are asymptotically equivalent to the EEL probabilities at the same order as in Proposition A.2.

Proposition A.3: Under Assumptions 1 and 2, for every $i = 1, \ldots, n$:

$$\sup_{\theta\in\Theta_n} \left| \hat{\pi}_{i,n}^{(1)}(\theta) - \hat{\pi}_{i,n}^{(1),Sh}(\theta) \right| = o_p\left(\frac{1}{n\sqrt{n}}\right). \blacksquare$$

A.3 Laws of Large Numbers and Central Limit Theorems

Propositions A.2 and A.3 imply that we can safely build consistent estimators of population expectations by using any family of implied probabilities. More precisely:

Proposition A.4: Let Assumptions 1 and 2 hold. Let $(Y_i)_{1\leq i\leq n}$ be a sequence of i.i.d. square integrable random vectors. Then for any $\gamma \neq 0$:

$$\sup_{\theta\in\Theta_n} \left\| \sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma)}(\theta) Y_i - \frac{1}{n}\sum_{i=1}^n Y_i \right\| = o_P(1) \quad \text{and} \quad \sup_{\theta\in\Theta_n} \left\| \sum_{i=1}^n \hat{\pi}_{i,n}^{(1),Sh}(\theta) Y_i - \frac{1}{n}\sum_{i=1}^n Y_i \right\| = o_P(1). \blacksquare$$

In particular, Proposition A.4 implies that all estimators of population expectations, based on the implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}(\theta)$ or $\hat{\pi}_{i,n}^{(1),Sh}(\theta)$, are uniformly consistent like the ones based on the naive empirical probabilities $\hat{\pi}_{i,n} = (1/n)$ for which the said uniformity holds trivially. The proof of Proposition A.4 is quite straightforward:

$$\left\| \sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma)}(\theta) Y_i - \frac{1}{n}\sum_{i=1}^n Y_i \right\| \leq \left[\max_{1\leq i\leq n} \|Y_i\| \right] \sum_{i=1}^n \left| \hat{\pi}_{i,n}^{(\gamma)}(\theta) - \frac{1}{n} \right|$$

with:

$$\sum_{i=1}^n \left| \hat{\pi}_{i,n}^{(\gamma)}(\theta) - \frac{1}{n} \right| \leq nO_P\left(\frac{1}{n\sqrt{n}}\right),$$

which gives the required result since, when $(Y_i)_{1\leq i\leq n}$ is a sequence of i.i.d. square integrable random vectors:

$$\max_{1\leq i\leq n}\|Y_i\| = o_P(\sqrt{n}).$$

To move from the laws of large numbers to the central limit theorems, it would take to be able to replace $o_P(1)$ by $o_P(1/\sqrt{n})$ in Proposition A.4. Of course, it is not true when comparing with the naive weights $\hat{\pi}_{i,n} = (1/n)$, since implied probabilities have precisely been used to provide a more accurate estimator (see Proposition 3). But it might become true when comparing two different sets of implied probabilities for the Cressie-Read parameters γ and γ^*:

$$\left\|\sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma)}(\theta)Y_i - \sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma^*)}(\theta)Y_i\right\| \leq \max_{1\leq i\leq n}\left|\hat{\pi}_{i,n}^{(\gamma)}(\theta) - \hat{\pi}_{i,n}^{(\gamma^*)}\right| \sum_{i=1}^n \|Y_i\|.$$

Since $\sum_{i=1}^n \|Y_i\| = O_P(n)$, we would get the required result if we knew that:

$$\max_{1\leq i\leq n}\left|\hat{\pi}_{i,n}^{(\gamma)}(\theta) - \hat{\pi}_{i,n}^{(\gamma^*)}\right| = o_P\left(\frac{1}{n\sqrt{n}}\right).$$

Unfortunately, Propositions A.2 and A.3 only give this order of magnitude for each $i = 1,\ldots,n$ but there is no such thing as a uniform upper bound for $i = 1,\ldots,n$. Hence, we must resort to direct proofs of the requested equivalences. We first show that, as already pointed out in ABR-07, shrinking the implied probabilities is immaterial in terms of first order asymptotics.

Proposition A.5: Let Assumptions 1 and 2 hold. Consider a sequence $(Y_i)_{1\leq i\leq n}$ of i.i.d. square integrable random vectors such that:

$$\begin{aligned}\sup_{\theta\in\Theta_n}\left\|\frac{1}{n}\sum_{i=1}^n Y_i(\psi_i(\theta) - \bar{\psi}_n(\theta))' - Cov\left[Y_i,\psi_i(\theta)\right]\right\| &= o_P(1),\\ \sup_{\theta\in\Theta_n}\|Cov\left[Y_i,\psi_i(\theta)\right]\| &= O(1).\end{aligned}$$

Then:

$$\sup_{\theta\in\Theta_n}\left\|\sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta)Y_i - \sum_{i=1}^n \hat{\pi}_{i,n}^{(1),Sh}(\theta)Y_i\right\| = o_P\left(\frac{1}{\sqrt{n}}\right). \blacksquare$$

Using different sets of implied probabilities is also immaterial, insofar as the probability distribution of Y_i does not display overly fat tails.

Proposition A.6: Let Assumptions 1 and 2 hold. Consider a sequence $(Y_i)_{1\leq i\leq n}$ of i.i.d. random vectors such that $E\left[\|Y_i\|^4\right] < \infty$. Furthermore, suppose that a suitable functional central limit theorem gives:

$$\sup_{\theta\in\Theta_n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\psi_i(\theta)Y_i - E[\psi_i(\theta)Y_i]\}\right\| = O_p(1). \tag{A.4}$$

Then for any $\gamma,\gamma^* \neq 0$:

$$\sup_{\theta\in\Theta_n}\left\|\sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma)}(\theta)Y_i - \sum_{i=1}^n \hat{\pi}_{i,n}^{(\gamma^*)}(\theta)Y_i\right\| = o_P\left(\frac{1}{\sqrt{n}}\right). \blacksquare$$

The condition in (A.4) is apparently strong. It is only required to show the uniformity in θ for the closeness between the two weighted averages in the result of Proposition A.6. On the other hand,

if focus lies in the said closeness for a given $\theta \in \Theta_n$, then one could accordingly weaken (A.4) which would then amount to a standard central limit theorem in the case $\Theta_n = \{\theta^0\}$.

Propositions A.5 and A.6 will allow us to show that all these estimators are asymptotically normal with the same asymptotic variances as follows:

Proposition A.7: Let Assumptions 1 and 2 hold. Consider a sequence $(Y_i)_{1\leq i\leq n}$ of i.i.d. random vectors with $E\left[\|Y_i\|^4\right] < \infty$ and such that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} Y_i - E[Y_i] \\ \psi_i(\theta) - E[\psi_i(\theta)] \end{bmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} Var(Y_i) & Cov[Y_i, \psi_i(\theta)] \\ Cov[\psi_i(\theta), Y_i] & Var(\psi_i(\theta)) \end{bmatrix}\right)$$

for each $\theta \in \Theta_n$. Then, for all $\theta \in \Theta_n$, the vector:

$$U_n = \sqrt{n}\begin{bmatrix} \sum_{i=1}^{n}\hat{\pi}_{i,n}^{(1)}(\theta)Y_i - \left(E[Y_i] - Cov[Y_i, \psi_i(\theta)][Var(\psi_i(\theta))]^{-1}E[\psi_i(\theta)]\right) \\ \bar{\psi}_n(\theta) - E[\psi_i(\theta)] \end{bmatrix}$$

is asymptotically normal with mean zero, a block diagonal variance matrix and the North-West Block of the variance matrix is given by:

$$Var(Y_i) - Cov[Y_i, \psi_i(\theta)]\left[Var\left(\psi_i(\theta)\right)\right]^{-1} Cov[\psi_i(\theta), Y_i]. \blacksquare$$

Note that, the statement above involves an abuse of notation since $\psi_i(\theta)$ depends indirectly on n (for θ picked in Θ_n). However the statement is clear if we rather say that the characteristic function of U_n minus the characteristic function of the normal distribution described above converges to zero. When describing asymptotic properties of the score test in Section 4, we make assumptions such that this convergence is uniform (valid, as for all the propositions above, for the supremum over $\theta \in \Theta_n$).

Note also that, Proposition A.7 confirms that the use of the implied probabilities provides a more accurate estimator of $E[Y]$; its asymptotic variance is smaller than $Var(Y)$, albeit at the cost of introducing an asymptotic bias that is due to the non-zero $E[\psi(\theta)]$ (more precisely, non-$o_p(1/\sqrt{n})$ $E[\bar{\psi}_n(\theta)]$). In our application where we use the implied probabilities for the score test, this bias only has first order asymptotic effect if some elements of θ are weakly identified, in which case, however, their use turns out to be even more important since not doing so leads to an over-sized score test (see Kleibergen (2005)).

Online supplemental material for: Score tests in GMM: Why Use Implied Probabilities?

Saraswata Chaudhuri[3] and Eric Renault[4]

Supplemental materials are presented here in Appendices B, C, D and E. Appendix B proves the results from Section 2. Appendix C proves the results from Appendix A. Appendix D lists the standard assumptions maintained to describe these properties of score tests, and then proves the result from Section 4 by applying the results from Appendix A. Appendix E reports two tables containing empirical size and twelve figures containing size-corrected power plots for the simulation experiment.

Table of content:

B Appendix B: Proofs of the results in Section 2

Proof of Proposition 1: The proof comes in four steps.

Step 1: We define a sequence of compact subsets $B_n, n = 1, 2, ..$ of $\mathbb{R}^H$ by:

$$B_n = \left\{\lambda \in \mathbb{R}^H; \|\lambda\| \leq r_n\right\}$$

with:

$$(r_n)^{-2} = \sqrt{n} \sup_{\theta\in\Theta_n} \max_{1\leq i\leq n} \|\psi_i(\theta)\| = o_P(n)$$

where the second inequality comes from Assumption 1(ii) and we assume throughout:

$$\sup_{\theta\in\Theta_n} \max_{1\leq i\leq n} \|\psi_i(\theta)\| \neq 0.$$

This latter assumption can be maintained without loss of generality since all the results become straightforward otherwise.

Then, we show that for all $\gamma \in \mathbb{R}$:

$$\lim_{n=\infty} \Pr\left[B_n \subset \bigcap_{\theta\in\Theta_n} \Lambda_n^{(\gamma)}(\theta)\right] = 1.$$

Proof for Step 1:

For all $\lambda \in B_n$:

$$\sup_{\theta\in\Theta_n} \max_{1\leq i\leq n} \left|\lambda'\psi_i(\theta)\right| \leq \|\lambda\| \sup_{\theta\in\Theta_n} \max_{1\leq i\leq n} \|\psi_i(\theta)\| \leq \frac{(r_n)^{-1}}{\sqrt{n}} = o_P(1) \tag{B.1}$$

[3]Department of Economics, McGill University, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

[4]Corresponding author. Department of Economics, University of Warwick, Coventry, United Kingdom. Email: Eric.Renault@warwick.ac.uk.

by definition of the sequence r_n. Therefore:

$$\forall \varepsilon > 0, \lim_{n=\infty} \Pr\left[\sup_{\lambda \in B_n} \sup_{\theta \in \Theta_n} \max_{1 \leq i \leq n} \left|\lambda' \psi_i(\theta)\right| < \varepsilon\right] = 1.$$

For a given $\gamma \in \mathbb{R}$, let us consider $\varepsilon = (2|\gamma|)^{-1}$ (ε arbitrary positive number when $\gamma = 0$). We deduce from above that:

$$\lim_{n=\infty} \Pr\left[\inf_{\lambda \in B_n} \inf_{\theta \in \Theta_n} \min_{1 \leq i \leq n} \left[1 + \gamma \lambda' \psi_i(\theta)\right] > \frac{1}{2}\right] = 1$$

which implies that:

$$\lim_{n=\infty} \Pr\left[B_n \subset \bigcap_{\theta \in \Theta_n} \Lambda_n^{(\gamma)}(\theta)\right] = 1.$$

Step 2: We recall that the sets B_n and $\Lambda_n^{(\gamma)}(\theta)$ are random since they depend on the state of nature through $\psi_i(\theta), i = 1, ..., n$.

We denote by A_n the set of states of nature such that:

$$B_n \subset \bigcap_{\theta \in \Theta_n} \Lambda_n^{(\gamma)}(\theta).$$

For all $\omega \in A_n, \gamma \in \mathbb{R}, \theta \in \Theta_n$, the function $h_{n,\theta}^{(\gamma)}(\lambda) = \sum_{i=1}^{n} \varrho^{(\gamma)}\left[\lambda' \psi_i(\theta)\right]$ is welll defined for all $\lambda \in B_n$ and can be maximized on this compact set to define:

$$\lambda_n^{*(\gamma)}(\theta) = \arg\max_{\lambda \in B_n} \sum_{i=1}^{n} \varrho^{(\gamma)}\left[\lambda' \psi_i(\theta)\right].$$

Note that the random variable $\lambda_n^{*(\gamma)}(\theta), \theta \in \Theta_n$, is defined only on the set A_n of states of nature but, by definition, this set has asymptotically a probability one. Therefore, it makes sense to study the stochastic convergence of the sequence $\lambda_n^{*(\gamma)}(\theta), n = 1, 2, ...$ Note also that this sequence is uniquely defined since the function $\varrho^{(\gamma)}$ is strictly concave. We first show:

$$\sup_{\theta \in \Theta_n} \left\|\lambda_n^{*(\gamma)}(\theta)\right\| = O_P\left(1/\sqrt{n}\right).$$

Proof for Step 2: By definition of $\lambda_n^{*(\gamma)}(\theta)$, we have

$$\sum_{i=1}^{n} \varrho^{(\gamma)}\left[\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right] \geq \sum_{i=1}^{n} \varrho^{(\gamma)}\left[0'\psi_i(\theta)\right] = 0 = \varrho^{(\gamma)}[0]. \tag{B.2}$$

We can write the Taylor expansion:

$$\varrho^{(\gamma)}\left[\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right] = \varrho^{(\gamma)}[0] + \frac{\partial \varrho^{(\gamma)}[0]}{\partial x}\left(\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right) + \frac{1}{2}\frac{\partial^2 \varrho^{(\gamma)}[u_{in}(\theta)]}{\partial x^2}\left(\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right)^2 \tag{B.3}$$

with:

$$0 \leq \sup_{\theta \in \Theta_n} \max_{1 \leq i \leq n} |u_{in}(\theta)| \leq \sup_{\theta \in \Theta_n} \max_{1 \leq i \leq n} \left|\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right| \leq \frac{(r_n)^{-1}}{\sqrt{n}} = o_P(1)$$

where the last inequality is deduced from (B.1). Therefore, we have by continuity:

$$\frac{\partial^2 \varrho^{(\gamma)}[0]}{\partial x^2} = -1 \Rightarrow \lim_{n=\infty} \Pr\left[\sup_{\theta \in \Theta_n} \max_{1 \leq i \leq n} \frac{\partial^2 \varrho^{(\gamma)}[u_{in}(\theta)]}{\partial x^2} \leq -\frac{1}{2}\right] = 1. \tag{B.4}$$

Considering together (B.2),(B.3),(B.4) and the fact that:

$$\varrho^{(\gamma)}[0] = 0, \frac{\partial \varrho^{(\gamma)}[0]}{\partial x} = -1,$$

we deduce:

$$\lim_{n=\infty} \Pr\left[\sup_{\theta\in\Theta_n}\left\{\frac{1}{4n}\sum_{i=1}^{n}\left(\lambda_n^{*(\gamma)\prime}(\theta)\psi_i(\theta)\right)^2 - \lambda_n^{*(\gamma)\prime}(\theta)\bar{\psi}_n(\theta)\right\} \leq 0\right] = 1$$

that is:

$$\lim_{n=\infty} \Pr\left[\sup_{\theta\in\Theta_n}\left\{\lambda_n^{*(\gamma)\prime}(\theta)\Omega_n(\theta)\lambda_n^{*(\gamma)} - 4\lambda_n^{*(\gamma)\prime}(\theta)\bar{\psi}_n(\theta)\right\} \leq 0\right] = 1.$$

In particular, if we introduce:

$$\inf_{\theta\in\Theta_n} \gamma_{\min}(\theta) = \eta > 0,$$

we have:

$$\lim_{n=\infty} \Pr\left[\sup_{\theta\in\Theta_n}\left\{\eta\left\|\lambda_n^{*(\gamma)}(\theta)\right\|^2 - 4\left\|\lambda_n^{*(\gamma)}(\theta)\right\|\left\|\bar{\psi}_n(\theta)\right\|\right\} \leq 0\right] = 1.$$

Thus:

$$\lim_{n=\infty} \Pr\left[\sup_{\theta\in\Theta_n}\left\{\left\|\lambda_n^{*(\gamma)}(\theta)\right\| - \frac{1}{4\eta}\left\|\bar{\psi}_n(\theta)\right\|\right\} \leq 0\right] = 1.$$

Note that, by virtue of Assumption 1 we have:

$$\sup_{\theta\in\Theta_n}\left\|\bar{\psi}_n(\theta)\right\| = O_P\left(1/\sqrt{n}\right).$$

Hence the announced result for step 2.

Step 3:

$$\lim_{n=\infty} \Pr\left[\bigcap_{\theta\in\Theta_n}\left\{\left\|\lambda_n^{*(\gamma)}(\theta)\right\| < r_n\right\}\right] = 1.$$

Proof for Step 3: Let $\varepsilon > 0$. We want to find $n_\varepsilon \in \mathbb{N}$ such that:

$$\forall n \geq n_\varepsilon, \Pr\left[\sup_{\theta\in\Theta_n}\left\|\lambda_n^{*(\gamma)}(\theta)\right\| < r_n\right] \geq 1-\varepsilon.$$

We know that we can find $n_\varepsilon \in \mathbb{N}$ and a finite number M_ε such that for all $n \geq n_\varepsilon$:

$$\Pr\left[\sup_{\theta\in\Theta_n}\sqrt{n}\left\|\lambda_n^{*(\gamma)}(\theta)\right\| \leq M_\varepsilon\right] \geq 1 - \frac{\varepsilon}{2}$$

and:

$$\Pr\left[M_\varepsilon < r_n\sqrt{n}\right] \geq 1 - \frac{\varepsilon}{2}.$$

The first inequality is implied by step 2 while the second inequality is implied by the fact (see step 1) that:

$$(r_n)^{-2} = o_P(n).$$

Then for $n \geq n_\varepsilon$:

$$\begin{aligned}\Pr\left[\sup_{\theta\in\Theta_n}\left\|\lambda_n^{*(\gamma)}(\theta)\right\| < r_n\right] &\geq \Pr\left[\left\{\sup_{\theta\in\Theta_n}\sqrt{n}\left\|\lambda_n^{*(\gamma)}(\theta)\right\| \leq M_\varepsilon\right\} \cap \left\{M_\varepsilon < r_n\sqrt{n}\right\}\right] \\ &\geq \Pr\left[\left\{\sup_{\theta\in\Theta_n}\sqrt{n}\left\|\lambda_n^{*(\gamma)}(\theta)\right\| \leq M_\varepsilon\right\}\right] + \Pr\left[\left\{M_\varepsilon < r_n\sqrt{n}\right\}\right] - 1 \geq 1-\varepsilon.\end{aligned}$$

Step 4: We conclude from steps 2 and 3 that for n sufficiently large and for all $\omega \in A_n, \gamma \in \mathbb{R}, \theta \in \Theta_n$, the vector $\lambda_n^{*(\gamma)}(\theta)$ that maximizes the function $h_{n,\theta}^{(\gamma)}(\lambda)$ on B_n is actually an interior point of B_n. Therefore, it must fulfill the first order condition:

$$\frac{\partial h_{n,\theta}^{(\gamma)}\left(\lambda_n^{*(\gamma)}(\theta)\right)}{\partial \lambda} = 0.$$

Since the function $h_{n,\theta}^{(\gamma)}(\lambda)$ is actually well defined and strictly concave on $\Lambda_n^{(\gamma)}(\theta)$, the first order condition is sufficient to ensure that $\lambda_n^{*(\gamma)}(\theta) = \hat{\lambda}_n^{(\gamma)}(\theta)$ maximizes this function on $\Lambda_n^{(\gamma)}(\theta)$. This completes the proof of Proposition 1. ■

Proof of Proposition 2: The first order conditions associated to (2.2) are:

$$\sum_{i=1}^{n} \tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)\prime}(\theta)\psi_i(\theta)\right]\psi_i(\theta) = 0.$$

We can rewrite them as:

$$\sum_{i=1}^{n} \left\{\tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)\prime}(\theta)\psi_i(\theta)\right] + 1\right\}\psi_i(\theta) - n\bar{\psi}_n(\theta) = 0$$

or:

$$\sum_{i=1}^{n} k^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)\prime}(\theta)\psi_i(\theta)\right]\psi_i(\theta)\left(\hat{\lambda}_n^{(\gamma)\prime}(\theta)\psi_i(\theta)\right) - n\bar{\psi}_n(\theta) = 0$$

or, equivalently:

$$n\tilde{\Omega}_n^{(\gamma)}(\theta)\hat{\lambda}_n^{(\gamma)}(\theta) - n\bar{\psi}_n(\theta) = 0,$$

which gives the announced formula (2.6) for $\hat{\lambda}_n^{(\gamma)}(\theta)$. ■

Proof of Proposition 3: From Proposition 2, the Lagrange multipliers associated with the constraints (2.11) correspond to the second set of coefficients of the matrix are:

$$\left(\tilde{\Omega}_n^{(\gamma)}(\theta)\right)^{-1}\bar{\Psi}_n(\theta)$$

where $\bar{\Psi}_n(\theta) = [\bar{\psi}_n'(\theta), \hat{\mu}_n^{(\gamma)\prime}(\theta) - \bar{g}_n'(\theta)]'$ stands for the sample counterpart of the augmented set of moment conditions.

Therefore, with the generic notation $\Sigma(\theta) = \tilde{\Omega}_n^{(\gamma)}(\theta)$, we have:

$$\hat{\Sigma}_n^{21}(\theta)\bar{\psi}_n(\theta) - \hat{\Sigma}_n^{22}(\theta)\left[\hat{\mu}_n^{(\gamma)}(\theta) - \bar{g}_n(\theta)\right] = 0$$

where $\Sigma^{21}(\theta)$ and $\Sigma^{22}(\theta)$ denote respectively the South-West and South-East blocks of the matrix $\Sigma^{-1}(\theta)$ according to the partition $[\psi(W_i,\theta)', g(W_i,\theta)']'$. Then:

$$\hat{\mu}_n^{(\gamma)}(\theta) = \bar{g}_n(\theta) + [\hat{\Sigma}_n^{22}(\theta)]^{-1}\hat{\Sigma}_n^{21}(\theta)\bar{\psi}_n(\theta) \tag{B.5}$$

that is:

$$\hat{\mu}_n^{(\gamma)}(\theta) = \bar{g}_n(\theta) - \hat{\Sigma}_{21n}(\theta)[\hat{\Sigma}_{11n}(\theta)]^{-1}\bar{\psi}_n(\theta) \tag{B.6}$$

where with obvious notations, (B.6) is deduced from (B.5) thanks to the identity:

$$\Sigma\Sigma^{-1} = Id_H \Rightarrow \Sigma^{21}\Sigma_{11} + \Sigma^{22}\Sigma_{21} = 0 \Leftrightarrow [\Sigma^{22}]^{-1}\Sigma^{21} = -\Sigma_{21}[\Sigma_{11}]^{-1}.$$

In the EL case ($\gamma = -1$), we know that:

$$\tilde{\Omega}_n^{(-1)}(\theta) = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(-1)}(\theta)\, \Psi_i(\theta)\Psi_i'(\theta) \quad \text{with} \quad \Psi_i(\theta) = [\psi_i'(\theta), g'(W_i,\theta)]'.$$

Thus, for $\gamma = -1$, (B.6) exactly coincides with the value for $\hat{\mu}_n^{(-1)}(\theta)$ stated by Proposition 3.

In the EEL case ($\gamma = 1$), we know that:

$$\tilde{\Omega}_n^{(1)}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \Psi_i(\theta)\Psi_i'(\theta).$$

Thus, for $\gamma = 1$,(B.6) does not exactly coincide with the value for $\hat{\mu}_n^{(1)}(\theta)$ stated by Proposition 3, since Proposition 3 considers sample counterparts of variances and covariances in mean deviation form.

However, this modification is immaterial for the following reason. Proposition 2 tells us that:

$$\hat{\lambda}_n^{(1)}(\theta) = \left(\tilde{\Omega}_n^{(1)}(\theta)\right)^{-1} \bar{\Psi}_n(\theta).$$

Let us define:

$$\hat{W}_n^{(1)}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \Psi_i(\theta)\left[\Psi_i(\theta) - \bar{\Psi}_n(\theta)\right]'.$$

Then, we get:

$$\bar{\Psi}_n(\theta) = \tilde{\Omega}_n^{(1)}(\theta)\hat{\lambda}_n^{(1)}(\theta) = \left[\hat{W}_n^{(1)}(\theta) + \bar{\Psi}_n(\theta)\bar{\Psi}_n(\theta)'\right]\hat{\lambda}_n^{(1)}(\theta).$$

Hence:

$$[\hat{W}_n^{(1)}(\theta)]^{-1}\bar{\Psi}_n(\theta) - [\hat{W}_n^{(1)}(\theta)]^{-1}\bar{\Psi}_n(\theta)\bar{\Psi}_n(\theta)'\hat{\lambda}_n^{(1)}(\theta) = \hat{\lambda}_n^{(1)}(\theta)$$

which can be rewritten:

$$\frac{\hat{\lambda}_n^{(1)}(\theta)}{1 - \bar{\Psi}_n(\theta)'\hat{\lambda}_n^{(1)}(\theta)} = [\hat{W}_n^{(1)}(\theta)]^{-1}\bar{\Psi}_n(\theta).$$

Therefore, the second set of coefficients of the vector $\hat{\lambda}_n^{(1)}(\theta)$ is nil if and only if the second set of coefficients of $[\hat{W}_n^{(1)}(\theta)]^{-1}\bar{\Psi}_n(\theta)$ is itself nil. In other words, when $\gamma = 1$, the formula (B.6) is also valid with $\Sigma(\theta) = \hat{W}_n^{(1)}(\theta)$, which gives the result in mean-deviation form as stated in Proposition 3. ■

C Appendix C: Proofs of the propositions in Appendix A.1-A.3

Proof of Proposition A.1: For expositional simplicity, we always consider a state of nature in the set A_n defined in step 2 of the proof of Proposition 1. This set is asymptotically of probability one and allows us to use that:

$$B_n \subset \bigcap_{\theta\in\Theta_n} \Lambda_n^{(\gamma)}(\theta).$$

In particular, for all $\gamma \in \{\mathbb{R}\backslash\{0\}\}$ and $\theta \in \Theta_n$, $\hat{\lambda}_n^{(\gamma)}(\theta)$ fulfills the first order conditions for maximization of the function $h_{n,\theta}^{(\gamma)}(\lambda) = \sum_{i=1}^{n} \varrho^{(\gamma)}\left[\lambda'\psi_i(\theta)\right]$, so that:

$$0 = \sum_{i=1}^{n} \varrho_1^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right]\psi_i(\theta)$$

where $\varrho_1^{(\gamma)}[.]$ stands for the first derivative of the function $\varrho^{(\gamma)}[.]$. Since $\varrho_1^{(\gamma)}[0] = -1$, a Taylor expansion around zero gives a set of H equations:

$$\begin{aligned} 0 &= -\bar{\psi}_n(\theta) + \frac{1}{n}\sum_{i=1}^{n} \varrho_2^{(\gamma)}\left[v_{i,n}^{(\gamma)}(\theta)\right]\psi_i(\theta)\psi_i(\theta)'\hat{\lambda}_n^{(\gamma)}(\theta) \\ &= -\bar{\psi}_n(\theta) - \frac{1}{n}\sum_{i=1}^{n} \psi_i(\theta)\psi_i(\theta)'\hat{\lambda}_n^{(\gamma)}(\theta) + R_n^{(\gamma)}(\theta) \\ R_n^{(\gamma)}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varrho_2^{(\gamma)}\left[v_{i,n}^{(\gamma)}(\theta)\right] - \varrho_2^{(\gamma)}[0]\right]\psi_i(\theta)\psi_i(\theta)'\hat{\lambda}_n^{(\gamma)}(\theta) \end{aligned}$$

where $v_{i,n}^{(\gamma)}(\theta)$ is for each equation (with a standard abuse of notation overlooking the fact that it may take different values for different components) a number between 0 and $\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)$. $\varrho_2^{(\gamma)}[.]$ stands for the second derivative of the function $\varrho^{(\gamma)}[.]$ and we have used the fact that $\varrho_2^{(\gamma)}[0] = -1$. From the second equation, we deduce that:

$$\begin{aligned} \hat{\lambda}_n^{(\gamma)}(\theta) &= -\left[\hat{\Omega}_n(\theta)\right]^{-1}\bar{\psi}_n(\theta) + \left[\hat{\Omega}_n(\theta)\right]^{-1} R_n^{(\gamma)}(\theta) \\ &= \hat{\lambda}_n^{(1)}(\theta) + \left[\hat{\Omega}_n(\theta)\right]^{-1} R_n^{(\gamma)}(\theta). \end{aligned}$$

Since:

$$\begin{aligned} \sup_{\theta\in\Theta_n}\left\|\left[\hat{\Omega}_n(\theta)\right]^{-1} - \Omega^{-1}(\theta)\right\| &= o_P(1) \\ \text{and } \inf_{\theta\in\Theta_n}\gamma_{\min}(\theta) &> 0, \end{aligned}$$

we will get the result of Proposition A.1 if we show that:

$$\sup_{\theta\in\Theta_n}\left\|R_n^{(\gamma)}(\theta)\right\| = o_P(1/\sqrt{n}).$$

However, by definition:

$$\begin{aligned} \left\|R_n^{(\gamma)}(\theta)\right\| &\leq \max_{1\leq i\leq n}\left|\varrho_2^{(\gamma)}\left[v_{i,n}^{(\gamma)}(\theta)\right] - \varrho_2^{(\gamma)}[0]\right|\left\|\left[\hat{\Omega}_n(\theta)\right]\right\|\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\| \\ &\leq \max_{1\leq i\leq n}\left|\varrho_3^{(\gamma)}\left[z_{i,n}^{(\gamma)}(\theta)\right]\right|\max_{1\leq i\leq n}\left|v_{i,n}^{(\gamma)}(\theta)\right|\left\|\left[\hat{\Omega}_n(\theta)\right]\right\|\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\| \end{aligned}$$

where $z_{i,n}^{(\gamma)}(\theta)$ is a number between 0 and $v_{i,n}^{(\gamma)}(\theta)$ while $\varrho_3^{(\gamma)}[.]$ stands for the third derivative of the function $\varrho^{(\gamma)}[.]$. Note that:

$$\begin{aligned} \sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\left\|z_{i,n}^{(\gamma)}(\theta)\right\| &\leq \sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\left\|v_{i,n}^{(\gamma)}(\theta)\right\| \leq \sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\left\|\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right\| \\ &\leq \sup_{\theta\in\Theta_n}\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\|\sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\|\psi_i(\theta)\| \\ &= O_P(1/\sqrt{n})o_P(\sqrt{n}) = o_P(1). \end{aligned} \tag{C.1}$$

Therefore, we can always consider a state of nature in a subset A_n^* of A_n that is also asymptotically of probability one and such that there exists a compact subset $K^{(\gamma)}$ of $Q^{(\gamma)}$ containing $z_{i,n}^{(\gamma)}(\theta)$ for all $\theta\in\Theta_n$ and $i = 1, ..., n$. Let us introduce:

$$b^{(\gamma)} = \sup_{x\in K^{(\gamma)}}\left|\varrho_3^{(\gamma)}[x]\right|.$$

Then:

$$\begin{aligned}
\sup_{\theta\in\Theta_n}\left\|R_n^{(\gamma)}(\theta)\right\| &\leq b^{(\gamma)}\sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\left|v_{i,n}^{(\gamma)}(\theta)\right|\left\|\left[\hat{\Omega}_n(\theta)\right]\right\|\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\| \\
&\leq b^{(\gamma)}\sup_{\theta\in\Theta_n}\left\|\left[\hat{\Omega}_n(\theta)\right]\right\|\sup_{\theta\in\Theta_n}\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\|^2\sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\|\psi_i(\theta)\| \\
&= O_P(1/n)o_P(\sqrt{n}) \\
&= o_P(1/\sqrt{n})
\end{aligned}$$

where we have used (C.1) again as well as the fact that, by assumption 2, $\left\|\left[\hat{\Omega}_n(\theta)\right]\right\|$ is upper bounded on Θ_n. ■

Proof of Proposition A.2: We know from (2.5) that:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta)=\frac{\tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right]}{\sum_{j=1}^n\tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right]}. \tag{C.2}$$

Each term in the sum in the denominator of (C.2) can be expanded as follows:

$$\tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right]=\tau^{(\gamma)}(0)+\tau_1^{(\gamma)}(0).\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right]+\left[\tau_1^{(\gamma)}(v_{j,n}(\theta))-\tau_1^{(\gamma)}(0)\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right]$$

with $v_{j,n}(\theta)$ in the interval between 0 and $\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)$ and:

$$\tau_1^{(\gamma)}(x)=\frac{d\tau^{(\gamma)}(x)}{dx}.$$

Hence:

$$\begin{aligned}
\tau^{(\gamma)}\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right] &= 1+\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)+\left[\tau_1^{(\gamma)}(v_{j,n}(\theta))-1\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right] \\
&= 1-\psi_j(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)+\psi_j(\theta)'u_n(\theta)+\left[\tau_1^{(\gamma)}(v_{j,n}(\theta))-1\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right]
\end{aligned}$$

where $u_n(\theta)=\hat{\lambda}_n^{(\gamma)}(\theta)-\hat{\lambda}_n^{(1)}(\theta)$ and, by Proposition A.1:

$$\sup_{\theta\in\Theta_n}\|u_n(\theta)\|=o_P\left(\frac{1}{\sqrt{n}}\right).$$

Then by dividing both numerator and denominator by $(1/n)$, we can rewrite (C.2) as:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta)=\frac{\frac{1}{n}\left[1-\psi_i(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)+R_{i,n}(\theta)\right]}{1-\bar{\psi}_n(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)+\frac{1}{n}\sum_{j=1}^n R_{j,n}(\theta)} \tag{C.3}$$

where:

$$\begin{aligned}
R_{i,n}(\theta) &= \psi_i(\theta)'u_n(\theta)+\left[\tau_1^{(\gamma)}(v_{i,n}(\theta))-1\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right] \qquad (C.4)\\
\frac{1}{n}\sum_{j=1}^n R_{j,n}(\theta) &= \bar{\psi}_n(\theta)'u_n(\theta)+\frac{1}{n}\sum_{j=1}^n\left[\tau_1^{(\gamma)}(v_{j,n}(\theta))-1\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right].
\end{aligned}$$

Note that by the definition of $v_{j,n}(\theta)$, Proposition 1 and Assumption 1:

$$\begin{aligned}\max_{1\leq j\leq n}\sup_{\theta\in\Theta_n}|v_{j,n}(\theta)| &\leq \max_{1\leq j\leq n}\sup_{\theta\in\Theta_n}\left|\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_j(\theta)\right| \\ &\leq O_P\left(\frac{1}{\sqrt{n}}\right)\max_{1\leq j\leq n}\sup_{\theta\in\Theta_n}\|\psi_j(\theta)\| = O_P\left(\frac{1}{\sqrt{n}}\right)o_P\left(\sqrt{n}\right) = o_P\left(1\right).\end{aligned}$$

Thus, there exists a closed ball B such that, for n sufficiently large, $v_{j,n}(\theta)$ belongs to B for all $j = 1, .., n$ and all $\theta \in \Theta_n$. Thus, for n sufficiently large:

$$\max_{1\leq j\leq n}\sup_{\theta\in\Theta_n}\left|\tau_1^{(\gamma)}(v_{j,n}(\theta)) - 1\right| \leq \sup_{x\in B}\left|\frac{d\tau_1^{(\gamma)}(x)}{dx}\right|\max_{1\leq j\leq n}\sup_{\theta\in\Theta_n}|v_{j,n}(\theta)| = o_P\left(1\right). \tag{C.5}$$

Hence, the first equation of (C.4) gives:

$$\sup_{\theta\in\Theta_n}|R_{i,n}(\theta)| \leq \sup_{\theta\in\Theta_n}\|\psi_i(\theta)\|\left\{\|u_n(\theta)\| + o_P\left(1\right)\left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\|\right\} = o_P\left(\frac{1}{\sqrt{n}}\right)$$

since, by Assumption 1:

$$\sup_{\theta\in\Theta_n}\|\psi_i(\theta)\| = O_P(1).$$

Moreover, using (C.5) again, the second equation of (C.4) gives:

$$\sup_{\theta\in\Theta_n}\left|\frac{1}{n}\sum_{j=1}^{n}R_{j,n}(\theta)\right| \leq \sup_{\theta\in\Theta_n}\left\|\bar{\psi}_n(\theta)\right\|\|u_n(\theta)\| + o_P(1)\sup_{\theta\in\Theta_n}\left|\hat{\lambda}_n^{(\gamma)}(\theta)'\bar{\psi}_n(\theta)\right| = o_P\left(\frac{1}{n}\right)$$

since, by Assumption 1:

$$\sup_{\theta\in\Theta_n}\left\|\bar{\psi}_n(\theta)\right\| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Thus, we have shown:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta) = \frac{\frac{1}{n}\left[1 - \psi_i(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)\right] + \alpha_{i,n}(\theta)}{1 - \bar{\psi}_n(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta) + \beta_n(\theta)} \tag{C.6}$$

with:

$$\begin{aligned}\alpha_{i,n}(\theta) &= \frac{1}{n}R_{i,n}(\theta) \\ \beta_n(\theta) &= \frac{1}{n}\sum_{j=1}^{n}R_{j,n}(\theta)\end{aligned}$$

where:

$$\begin{aligned}\sup_{\theta\in\Theta_n}|\alpha_{i,n}(\theta)| &= o_P\left(\frac{1}{n\sqrt{n}}\right) \\ \sup_{\theta\in\Theta_n}|\beta_n(\theta)| &= o_P\left(\frac{1}{n}\right).\end{aligned}$$

Now, consider the denominator of the expression (C.6) and, for brevity, write it as:

$$D_n(\theta) = 1 - \bar{\psi}_n(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta) + \beta_n(\theta).$$

Note that:

$$\sup_{\theta\in\Theta_n}|D_n(\theta)-1|\leq\sup_{\theta\in\Theta_n}\bar{\psi}_n(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)+o_P\left(\frac{1}{n}\right)=O_P\left(\frac{1}{n}\right) \tag{C.7}$$

since:

$$\sup_{\theta\in\Theta_n}\bar{\psi}_n(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)\leq\sup_{\theta\in\Theta_n}\left\|\bar{\psi}_n(\theta)\right\|^2\sup_{\theta\in\Theta_n}\left\|\hat{\Omega}_n^{-1}(\theta)\right\|=O_P\left(\frac{1}{n}\right)O_P\left(1\right)=O_P\left(\frac{1}{n}\right)$$

by Assumption 1 and noting in particular that $\left\|\hat{\Omega}_n^{-1}(\theta)\right\|$ is uniformly bounded in probability by virtue of Assumption 2. We deduce from (C.7) that:

$$\sup_{\theta\in\Theta_n}\frac{1}{|D_n(\theta)|}=O_P(1).$$

Hence:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta)=\frac{\frac{1}{n}\left[1-\psi_i(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)\right]}{D_n(\theta)}+o_P^U\left(\frac{1}{n\sqrt{n}}\right)$$

where we use the notation $o_P^U(.)$ to mean that the bound $o_P(.)$ is valid uniformly in $\theta\in\Theta_n$.

Therefore:

$$\begin{aligned}\hat{\pi}_{i,n}^{(\gamma)}(\theta)-\frac{1}{n} &= \frac{\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)}{D_n(\theta)}-\frac{\beta_n(\theta)}{nD_n(\theta)}+o_P^U\left(\frac{1}{n\sqrt{n}}\right)\\ &= \frac{\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)}{D_n(\theta)}+o_P^U\left(\frac{1}{n\sqrt{n}}\right)=O_P^U\left(\frac{1}{n\sqrt{n}}\right)\end{aligned}$$

where the last equality is obviously implied by assumptions 1 and 2.

Therefore, we deduce:

$$\begin{aligned}\hat{\pi}_{i,n}^{(\gamma)}(\theta)-\hat{\pi}_{i,n}^{(1)}(\theta) &= \left[\hat{\pi}_{i,n}^{(\gamma)}(\theta)-\frac{1}{n}\right]-\left[\hat{\pi}_{i,n}^{(1)}(\theta)-\frac{1}{n}\right]\\ &= \frac{\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)}{D_n(\theta)}+o_P^U\left(\frac{1}{n\sqrt{n}}\right)\\ &\quad-\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\left[\hat{V}_n^{(1)}(\theta)\right]^{-1}\bar{\psi}_n(\theta).\end{aligned}$$

Hence:

$$\begin{aligned}\hat{\pi}_{i,n}^{(\gamma)}(\theta)-\hat{\pi}_{i,n}^{(1)}(\theta) &= \frac{\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\left\{\hat{\Omega}_n^{-1}(\theta)-\left[\hat{V}_n^{(1)}(\theta)\right]^{-1}\right\}\bar{\psi}_n(\theta)}{D_n(\theta)}\\ &\quad+\left(-\frac{1}{n}\right)\left[\psi_i(\theta)-\bar{\psi}_n(\theta)\right]'\left[\hat{V}_n^{(1)}(\theta)\right]^{-1}\bar{\psi}_n(\theta).\frac{1-D_n(\theta)}{D_n(\theta)}+o_P^U\left(\frac{1}{n\sqrt{n}}\right).\end{aligned}$$

In this decomposition of the difference of implied probabilities, we see that the first term is $o_P^U\left(\frac{1}{n\sqrt{n}}\right)$ by assumptions 1 and 2, while, by taking also (C.7) into account, we get that the second term is $O_P^U\left(\frac{1}{n^2\sqrt{n}}\right)$. We have then proved that:

$$\hat{\pi}_{i,n}^{(\gamma)}(\theta)-\hat{\pi}_{i,n}^{(1)}(\theta)=o_P^U\left(\frac{1}{n\sqrt{n}}\right).\ \blacksquare$$

Proof of Proposition A.3: Recall that by definition (2.15):

$$\hat{\pi}_{i,n}^{(1),Sh}(\theta) = \frac{1}{1+\varepsilon_n(\theta)}\hat{\pi}_{i,n}^{(1)}(\theta) + \frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\hat{\pi}_{i,n}.$$

Then:

$$\hat{\pi}_{i,n}^{(1)}(\theta) - \hat{\pi}_{i,n}^{(1),Sh}(\theta) = \frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\left[\hat{\pi}_{i,n}^{(1)}(\theta) - \frac{1}{n}\right]$$

that is, by (2.13):

$$\hat{\pi}_{i,n}^{(1)}(\theta) - \hat{\pi}_{i,n}^{(1),Sh}(\theta) = -\frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\bar{\psi}_n(\theta)'[\hat{V}_n^{(1)}(\theta)]^{-1}\frac{\psi_i(\theta)-\bar{\psi}_n(\theta)}{n}.$$

Therefore:

$$\left|\hat{\pi}_{i,n}^{(1)}(\theta) - \hat{\pi}_{i,n}^{(1),Sh}(\theta)\right| \leq \frac{1}{n}\frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\left\|\bar{\psi}_n(\theta)\right\|\left\|\hat{V}_n^{(1)}(\theta)]^{-1}\right\|\left|\psi_i(\theta)-\bar{\psi}_n(\theta)\right|.$$

We deduce from Assumptions 1 and 2, continuity of the functions $\psi_i(\theta)$ and compactness of Θ that

$$\sup_{\theta\in\Theta_n}\left|\hat{\pi}_{i,n}^{(1)}(\theta) - \hat{\pi}_{i,n}^{(1),Sh}(\theta)\right| = \sup_{\theta\in\Theta_n}\left[\frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\right]O_P\left(\frac{1}{n\sqrt{n}}\right) \leq \sup_{\theta\in\Theta_n}\left[\varepsilon_n(\theta)\right]O_P\left(\frac{1}{n\sqrt{n}}\right) = o_P\left(\frac{1}{n\sqrt{n}}\right)$$

since:

$$\sup_{\theta\in\Theta_n}\left[\varepsilon_n(\theta)\right] = o_P(1). \tag{C.8}$$

To see that, note that:

$$\varepsilon_n(\theta) = \max\left\{\max_{1\leq i\leq n}\left(-n\hat{\pi}_{i,n}^{(1)}(\theta)\right), 0\right\}$$

with:

$$\max_{1\leq i\leq n}\left(-n\hat{\pi}_{i,n}^{(1)}(\theta)\right) = -1 + \max_{1\leq i\leq n}\left[\bar{\psi}_n(\theta)'[\hat{V}_n^{(1)}(\theta)]^{-1}\left(\psi_i(\theta)-\bar{\psi}_n(\theta)\right)\right]$$

so that:

$$\sup_{\theta\in\Theta_n}\max_{1\leq i\leq n}\left(-n\hat{\pi}_{i,n}^{(1)}(\theta)\right) = -1 + O_P\left(\frac{1}{\sqrt{n}}\right)$$

which obviously implies (C.8). ■

Proof of Proposition A.4: See the text below Proposition A.4 in Appendix A.3. ■

Proof of Proposition A.5: We have:

$$\sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta)Y_i - \sum_{i=1}^n \hat{\pi}_{i,n}^{(1),Sh}(\theta)Y_i = \frac{\varepsilon_n(\theta)}{1+\varepsilon_n(\theta)}\left[\frac{1}{n}\sum_{i=1}^n Y_i\left(\psi_i(\theta)-\bar{\psi}_n(\theta)\right)'\right][\hat{V}_n^{(1)}(\theta)]^{-1}\bar{\psi}_n(\theta).$$

Then, from Assumptions 1 and 2, (C.8), and the additional assumptions for Proposition A.5:

$$\sup_{\theta\in\Theta_n}\left\|\sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta)Y_i - \sum_{i=1}^n \hat{\pi}_{i,n}^{(1),Sh}(\theta)Y_i\right\| = o_P(1)O_P(1)O_P\left(\frac{1}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right)$$

which gives the announced result. ■

Proof of Proposition A.6: We have from the proof of Proposition A.2:

$$\begin{aligned}\hat{\pi}_{i,n}^{(\gamma)}(\theta) &= \frac{1}{nD_n(\theta)}\left[1-\psi_i(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)\right]+\frac{1}{nD_n(\theta)}R_{i,n}(\theta)\\ \Longrightarrow \sum_{i=1}^{n}\hat{\pi}_{i,n}^{(\gamma)}(\theta)Y_i &= A_n(\theta)+B_n(\theta)\end{aligned}$$

with:

$$\begin{aligned}A_n(\theta) &= \frac{1}{nD_n(\theta)}\sum_{i=1}^{n}R_{i,n}(\theta)Y_i\\ B_n(\theta) &= \frac{1}{nD_n(\theta)}\sum_{i=1}^{n}\left[1-\psi_i(\theta)'\hat{\Omega}_n^{-1}(\theta)\bar{\psi}_n(\theta)\right]Y_i.\end{aligned}$$

We first show that:

$$A_n(\theta)=o_P^U\left(\frac{1}{\sqrt{n}}\right).$$

By using (C.7), we only have to show that:

$$\sum_{i=1}^{n}R_{i,n}(\theta)Y_i=o_P^U\left(\sqrt{n}\right).$$

Recall from (C.4) that:

$$R_{i,n}(\theta)=\psi_i(\theta)'u_n(\theta)+\left[\tau_1^{(\gamma)}(v_{i,n}(\theta))-1\right].\left[\hat{\lambda}_n^{(\gamma)}(\theta)'\psi_i(\theta)\right]$$

so that:

$$\sum_{i=1}^{n}R_{i,n}(\theta)Y_i=A_{1,n}(\theta)+A_{2,n}(\theta)$$

with:

$$A_{1,n}(\theta)=u_n(\theta)'\sum_{i=1}^{n}\psi_i(\theta)Y_i=\sqrt{n}u_n(\theta)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\psi_i(\theta)Y_i-E\left[\psi_i(\theta)Y_i\right]\right\}+nu_n(\theta)'\frac{1}{n}\sum_{i=1}^{n}E\left[\psi_i(\theta)Y_i\right].$$

By the functional central limit theorem:

$$u_n(\theta)=o_P^U\left(\frac{1}{\sqrt{n}}\right)\Longrightarrow\sqrt{n}u_n(\theta)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\psi_i(\theta)Y_i-E\left[\psi_i(\theta)Y_i\right]\right\}=o_P^U(1)\leq o_P^U\left(\sqrt{n}\right).$$

On the other hand, applying the Cauchy-Schwartz and the triangle inequalities, and then using assumptions 1(i) and 2(ii) and the square integrability of Y_i, gives:

$$\left|nu_n(\theta)'\frac{1}{n}\sum_{i=1}^{n}E\left[\psi_i(\theta)Y_i\right]\right|\leq n\left\|u_n(\theta)\right\|\frac{1}{n}\sum_{i=1}^{n}\left\|E\left[\psi_i(\theta)Y_i\right]\right\|\leq n\left\|u_n(\theta)\right\|O_p^U\left(\frac{1}{\sqrt{n}}\right)O_p^U(1)=o_P^U\left(\sqrt{n}\right),$$

where, as with $o_p^U(.)$, here we use the notation $O_p^U(.)$ to mean that the bound $O_p(.)$ is valid uniformly in $\theta\in\Theta_n$.

Hence:

$$A_{1,n}(\theta)=o_P^U\left(\sqrt{n}\right).$$

On the other hand:

$$\begin{aligned}|A_{2,n}(\theta)| &= \left|\hat{\lambda}_n^{(\gamma)}(\theta)' \sum_{i=1}^n \psi_i(\theta) Y_i \left[\tau_1^{(\gamma)}(v_{i,n}(\theta)) - 1\right]\right| \\ &\leq \left\|\hat{\lambda}_n^{(\gamma)}(\theta)\right\| \max_{1\leq j\leq n} \sup_{\theta\in\Theta_n} \left|\tau_1^{(\gamma)}(v_{j,n}(\theta)) - 1\right| \left\|\sum_{i=1}^n \psi_i(\theta) Y_i\right\| = o_P^U\left(\frac{1}{\sqrt{n}}\right)\left\|\sum_{i=1}^n \psi_i(\theta) Y_i\right\| = o_P^U\left(\sqrt{n}\right),\end{aligned}$$

respectively by Proposition 1, (C.5) and since the argument above about $A_{1,n}(\theta)$ has shown that:

$$\left\|\sum_{i=1}^n \psi_i(\theta) Y_i\right\| = O_P^U(n).$$

We complete the proof by showing that:

$$B_n(\theta) = \sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta) Y_i + o_P^U\left(\frac{1}{\sqrt{n}}\right).$$

We have:

$$\begin{aligned}B_n(\theta) &= \frac{1}{nD_n(\theta)} \sum_{i=1}^n \left[1 - \psi_i(\theta)' \hat{\Omega}_n^{-1}(\theta) \bar{\psi}_n(\theta)\right] Y_i \\ &= \frac{1}{nD_n(\theta)} \sum_{i=1}^n \left[1 - \psi_i(\theta)' \left[\hat{V}_n^{-1}(\theta)\right]^{-1} \bar{\psi}_n(\theta)\right] Y_i \\ &\quad + \frac{1}{nD_n(\theta)} \sum_{i=1}^n \psi_i(\theta)' \left\{\left[\hat{V}_n^{-1}(\theta)\right]^{-1} - \hat{\Omega}_n^{-1}(\theta)\right\} \bar{\psi}_n(\theta) Y_i.\end{aligned}$$

We first note that:

$$\begin{aligned}\left|\sum_{i=1}^n \psi_i(\theta)' \left\{\left[\hat{V}_n^{-1}(\theta)\right]^{-1} - \hat{\Omega}_n^{-1}(\theta)\right\} \bar{\psi}_n(\theta) Y_i\right| &\leq \left\|\left[\hat{V}_n^{-1}(\theta)\right]^{-1} - \hat{\Omega}_n^{-1}(\theta)\right\| \left\|\sum_{i=1}^n \psi_i(\theta) Y_i\right\| \left\|\bar{\psi}_n(\theta)\right\| \\ &= o_P^U(1) O_P^U(n) O_P^U\left(\frac{1}{\sqrt{n}}\right) = o_P^U\left(\sqrt{n}\right)\end{aligned}$$

so that:

$$\begin{aligned}B_n(\theta) &= \frac{1}{nD_n(\theta)} \sum_{i=1}^n \left[1 - \psi_i(\theta)' \left[\hat{V}_n^{-1}(\theta)\right]^{-1} \bar{\psi}_n(\theta)\right] Y_i + o_P^U\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{nD_n(\theta)} \sum_{i=1}^n \left[1 - \left[\psi_i(\theta) - \bar{\psi}_n(\theta)\right]' \left[\hat{V}_n^{-1}(\theta)\right]^{-1} \bar{\psi}_n(\theta)\right] Y_i \\ &\quad - \frac{1}{D_n(\theta)} \bar{\psi}_n \theta)' \left[\hat{V}_n^{-1}(\theta)\right]^{-1} \bar{\psi}_n(\theta) \frac{1}{n} \sum_{i=1}^n Y_i + o_P^U\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{D_n(\theta)} \sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta) Y_i + O_P^U\left(\frac{1}{\sqrt{n}}\right) O_P^U\left(\frac{1}{\sqrt{n}}\right) O_P^U(1) + o_P^U\left(\frac{1}{\sqrt{n}}\right) \\ &= \sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta) Y_i + \sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta) Y_i \left[\frac{1 - D_n(\theta)}{D_n(\theta)}\right] + o_P^U\left(\frac{1}{\sqrt{n}}\right) \\ &= \sum_{i=1}^n \hat{\pi}_{i,n}^{(1)}(\theta) Y_i + o_P^U\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

where the last equality is a direct consequence of (C.7). ■

Proof of Proposition A.7: Follows directly from the central limit theorem and the definition of $\hat{\pi}_{i,n}^{(1)}(\theta)$ (see Proposition 3). ■

D Appendix D: Assumptions and the proof of Theorem 1

This Appendix lists all the standard assumptions and the intermediate results that lead to Theorem 1 in Section 4. To put all the assumptions together, we state here, again, Assumptions 1 and 2 that were not stated in Section 4 but stated earlier in Section 2.1 and Appendix A.2 respectively.

Assumption 1: $\Theta_n, n = 1, 2, ...,$ is a sequence of subsets of Θ, containing the true unknown value θ^0, and such that:

(i) $\sup_{\theta\in\Theta_n} \left\|E\left(\bar{\psi}_n(\theta)\right)\right\| = O\left(\frac{1}{\sqrt{n}}\right)$.

(ii) $\max_{1\leq i\leq n} \sup_{\theta\in\Theta_n} \|\psi_i(\theta)\| = o_P\left(\sqrt{n}\right)$.

(iii) $\sup_{\theta\in\Theta_n} \|\psi_i(\theta)\| = O_P(1)$ for each $i = 1, \ldots, n$.

(iv) $\sup_{\theta\in\Theta_n} \left\|\bar{\psi}_n(\theta) - E\left(\bar{\psi}_n(\theta)\right)\right\| = O_P\left(\frac{1}{\sqrt{n}}\right)$.

Assumption 2: Let the sequence $\Theta_n, n = 1, 2, \ldots$ defined in Assumption 1 also satisfy:

(i) $\sup_{\theta\in\Theta_n} \left\|\hat{\Omega}_n(\theta) - V(\theta)\right\| = o_p(1)$, $\sup_{\theta\in\Theta_n} \left\|\hat{V}_n^{(1)}(\theta) - V(\theta)\right\| = o_p(1)$,
$\sup_{\theta\in\Theta_n} \left\|\left[\hat{\Omega}_n(\theta)\right]^{-1} - V^{-1}(\theta)\right\| = o_p(1)$ and $\sup_{\theta\in\Theta_n} \left\|\left[\hat{V}_n^{(1)}(\theta)\right]^{-1} - V^{-1}(\theta)\right\| = o_p(1)$.

(ii) $0 < \inf_{\theta\in\Theta_n} \gamma_{\min}(\theta) < \sup_{\theta\in\Theta_n} \gamma_{\max}(\theta) < +\infty$ where $\gamma_{\min}(\theta)$ and $\gamma_{\max}(\theta)$ stand for the smallest and largest eigenvalues respectively of $V(\theta)$.

Now, returning to the discussion in Section 4, we note that a smoothness assumption on the population moment (SPM1) $m(\theta_s)$, coupled with Assumption ID, reinforces the key condition, Assumption 1(i), in Lemma 1.

Assumption SPM1: $M_s(\theta_s) = \frac{\partial}{\partial\theta_s'} m(\theta_s)$ is continuous in $\theta_s \in B\left(\theta_s^0, \frac{r}{\sqrt{n}}\right)$ and $M_s(\theta_s^0)$ is finite.

Lemma 1: Under Assumptions ID and SPM1, and for Θ_n defined by (4.2), we have:

$$\sup_{\theta\in\Theta_n} \left\|E\left(\bar{\psi}_n(\theta)\right)\right\| = O\left(\frac{1}{\sqrt{n}}\right). \blacksquare$$

Proof of Lemma 1: Using the uniform convergence of $m_n^w(\theta)$ to $m^w(\theta)$ where the latter is continuous in θ with $m^w(\theta^0) = 0$, it follows by the compactness of Θ that for any $\theta \in \Theta_n$:

$$\sqrt{n}E[\bar{\psi}_n(\theta)] = m_n^w(\theta) + \sqrt{n}m(\theta_s) = (m^w(\theta) + o(1)) + \sqrt{n}m(\theta_s).$$

Hence, by the definition and continuous differentiability of $m(\theta_s)$, in conjunction with the definition of Θ_n, we have:

$$\begin{aligned}\sqrt{n} \sup_{\theta\in\Theta_n} \|E[\bar{\psi}_n(\theta)]\| &\leq (\|m^w(\theta)\| + o(1)) + \sqrt{n}\|m(\theta_s^0)\| + \sqrt{n} \sup_{\theta\in\Theta_n} \|M_s(\theta_s)\| \frac{\|r\|}{\sqrt{n}} \\ &= O(1) + 0 + \|O(1)\|\|r\| = O(1)\end{aligned}$$

using the triangle inequality, assumptions ID and SPM1, and the Cauchy-Schwartz inequality. ■

To get the required equivalence result of the concerned score tests, we maintain, in addition to assumptions ID and SPM1, the following set of assumptions CLT, SSM, SPM2, O and R that are similar to Kleibergen (2005), although slightly strengthened in two directions: (i) Since we are interested

in equivalence results uniformly in Θ_n, the assumptions here are made accordingly under CLT and O. (ii) Assumption R augments the rank condition in Kleibergen (2005) in a way whose significance has been recently emphasized by Andrews and Guggenberger (2017). Equipped with these, we will then reinforce the result of Proposition A.7 in an uniform sense. Accordingly, we proceed as follows.

Let $\Longrightarrow$ denote the weak convergence of random functions on Θ_n with respect to the sup norm.

Assumption CLT:(Functional CLT) Consider the stochastic process $\left\{\Psi_n(\theta); \theta \in \Theta_w \times \left\{\theta_s^0\right\}\right\}$ with:

$$\Psi_n(\theta) = \sqrt{n}\left[\begin{array}{c} \bar{\psi}_n(\theta) - E\left[\bar{\psi}_n(\theta)\right] \\ vec\left(\frac{\partial \bar{\psi}_n(\theta)}{\partial \theta_w'} - E\left[\frac{\partial \bar{\psi}_n(\theta)}{\partial \theta_w'}\right]\right) \end{array}\right].$$

Then, $\Psi_n(\theta) \Longrightarrow \Psi(\theta)$ where $\Psi(\theta)$ is a Gaussian process with zero mean and covariance function $\mathcal{C}(\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \Theta_w \times \left\{\theta_s^0\right\}$. Let

$$\mathcal{C}(\theta) \equiv \mathcal{C}(\theta, \theta) = \left[\begin{array}{cc} V(\theta) & C(\theta)' \\ C(\theta) & D(\theta) \end{array}\right]. \blacksquare$$

Note that, Assumption CLT implicitly assumes that the function $\psi(W, \theta_w, \theta_s^0)$ is differentiable with respect to θ_w on some open set containing the compact set Θ_w. We will actually assume even more.

Assumption SSM: (Smoothness of sample moment) Almost surely in W, the function $\psi(W, \theta_w, \theta_s^0)$ is differentiable with respect to θ on $\Theta_w^* \times B(\theta_s^0, r)$ for some $r > 0$, where Θ_w^* is an open set containing Θ_w and $B(\theta_s^0, r)$ is as defined in (4.2). $\blacksquare$

Furthermore, to characterize the expected Jacobian $E\left[\frac{\partial \bar{\psi}_n(\theta)}{\partial \theta'}\right]$, which is the key to our local asymptotic analysis, we maintain, in addition to assumption SPM1, that:

Assumption SPM2: $M_s^w(\theta) = \frac{\partial}{\partial \theta_s'} m^w(\theta)$ and $M_w^w(\theta) = \frac{\partial}{\partial \theta_w'} m^w(\theta)$ exist and are continuous in $\theta \in \Theta_n$. Also, $\frac{\partial}{\partial \theta'} E[\psi_i(\theta)] = E\left[\frac{\partial}{\partial \theta'} \psi_i(\theta)\right]$ for $\theta \in \Theta_n$.

The other (O) assumptions below are in accordance with Propositions A.5 and A.6. In particular, the assumptions (O(i)-(ii)) related to $\frac{\partial \psi_i(\theta)}{\partial \theta_w'}$, i.e., the Jacobian for the weakly identified components, resemble that from Proposition A.6. For the other terms of interest, namely, $\frac{\partial \psi_i(\theta)}{\partial \theta_s'}$ and the third moment of $\psi_i(\theta)$, the assumptions (O(iii)-(v)) are weaker when contrasted with Assumption CLT.

Assumption O:

(i) $\sup_{\theta \in \Theta_n} \left\| \frac{1}{n} \sum_{i=1}^n \frac{\partial \psi_i(\theta)}{\partial \theta_w'} \left(\psi_i(\theta) - \bar{\psi}_n(\theta)\right)' - Cov\left(\frac{\partial \psi_i(\theta)}{\partial \theta_w'}, \psi_i(\theta)\right) \right\| = o_p(1)$.

ii) $\sup_{\theta \in \Theta_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\partial \psi_i(\theta)}{\partial \theta_w'} \psi_i'(\theta) - E\left[\frac{\partial \psi_i(\theta)}{\partial \theta_w'} \psi_i(\theta)'\right]\right) \right\| = O_p(1)$.

(iii) $\sup_{\theta \in \Theta_n} \left\| \frac{\partial}{\partial \theta_s'} \bar{\psi}_n(\theta) - E\left[\frac{\partial}{\partial \theta_s'} \bar{\psi}_n(\theta)\right] \right\| = o_p(1)$.

(iv) $\sup_{\theta \in \Theta_n} \left\| \frac{1}{n} \sum_{i=1}^n \left\{\frac{\partial}{\partial \theta_s'} \psi_i(\theta)\right\} \left(\psi_i(\theta) - \bar{\psi}_n(\theta)\right)' \right\| = O_p(1)$.

(v) $\sup_{\theta \in \Theta_n} \left\| \frac{1}{n} \sum_{i=1}^n \psi_i(\theta) \psi_i'(\theta) \otimes \left(\psi_i(\theta) - \bar{\psi}_n(\theta)\right) \right\| = O_p(1)$.

Assumptions SPM2 and O(iii) imply that $\sup_{\theta \in \Theta_n} \left\| \frac{\partial}{\partial \theta_s'} \bar{\psi}_n(\theta) - \left[M_s(\theta_s) + \frac{1}{\sqrt{n}} M_s^w(\theta)\right] \right\| = o_p(1)$.

To simplify the expressions for the relevant quantities obtained from assumptions CLT, SPM1-2 and O, use the notation $A_{c:d}(\theta)$ to denote the block of rows from c to d of any $a \times b$ matrix $A(\theta)$ with $a \leq c \leq d \leq b$, and define:

(i) $\Psi_{1:H}(\theta_w, \theta_s^0)$ is the $H \times 1$ random vector consisting of the first H rows of $\Psi(\theta_w, \theta_s^0)$.

(ii) $\xi(\theta) = (\xi_1(\theta), \ldots, \xi_{p_w}(\theta))$ is a $H \times p_w$ random matrix with:

$$\xi_j(\theta) = \Psi_{jH+1:(j+1)H}(\theta) - C_{(j-1)H+1:jH}(\theta) \left[V(\theta)\right]^{-1} \Psi_{1:H}(\theta) \text{ for } j = 1, \ldots p_w.$$

(iii) $G_w(\theta) = [G_{w,1}(\theta), \ldots, G_{w,p_w}(\theta)]$ is a $H \times p_w$ non-random matrix with:

$$G_{w,j}(\theta) = M^w_{w,j}(\theta) - C_{(j-1)H+1:jH}(\theta) \left[V(\theta)\right]^{-1} m^w(\theta) \text{ for } j = 1, \ldots p_w$$

where $M^w_{w,j}(\theta)$ denotes the j-th column of $M^w_w(\theta)$ for $j = 1, \ldots p_w$.

Lemma 2: Under Assumptions 1, 2, ID, SSM, SPM1-2, CLT and O, we have for any choice of implied probabilities $\pi^G(\theta)$ and $\pi^V(\theta)$ conformable to (4.1):

(i) Asymptotic independence: For $\theta \in \Theta_n$,

$$\sqrt{n}\left[\bar{\psi}_n(\theta) - E[\bar{\psi}_n(\theta)],\ \sum_{i=1}^n \pi^G_{i,n}(\theta)\frac{\partial \psi_i(\theta)}{\partial \theta'_w}\right] \Longrightarrow \left[\Psi_{1:H}(\theta_w, \theta^0_s), G_w(\theta_w, \theta^0_s) + \xi(\theta_w, \theta^0_s)\right],$$

where $\left[\Psi_{1:H}(\theta)', [vec\left(\xi(\theta)\right)]'\right]$ is a mean zero Gaussian process with covariance function $\mathcal{D}(\theta_1, \theta_2)$ for any $\theta_1, \theta_2 \in \Theta_w \times \left\{\theta^0_s\right\}$; with:

$$\mathcal{D}(\theta) \equiv \mathcal{D}(\theta, \theta) = \left[\begin{array}{cc} V(\theta) & 0 \\ 0 & D(\theta) - C(\theta)[V(\theta)]^{-1}C(\theta)' \end{array}\right].$$

(ii) Asymptotic equivalence of averages using the naive and the implied probabilities:

$$\sup_{\theta \in \Theta_n} \left\| \sum_{i=1}^n \hat{\pi}_{i,n} \frac{\partial \psi_i(\theta)}{\partial \theta'_s} - M_s(\theta_s) \right\| = \sup_{\theta \in \Theta_n} \left\| \sum_{i=1}^n \pi^G_{i,n}(\theta) \frac{\partial \psi_i(\theta)}{\partial \theta'_s} - M_s(\theta_s) \right\| = o_p(1),$$
$$\sup_{\theta \in \Theta_n} \left\| \sum_{i=1}^n \pi^V_{i,n}(\theta) V_{i,n}(\theta) - V(\theta) \right\| = o_p(1). \ \blacksquare$$

Proof of Lemma 2: (i) We maintain all the assumptions from Propositions A.5 and A.6 (and also Propositions A.3 and A.4, by default). Hence, for any two choices of $\pi^G_{i,n}(\theta)$ in (4.1), call them $\pi^G_{i,n,1}(\theta)$ and $\pi^G_{i,n,2}(\theta)$ respectively, it directly follows that:

$$\sqrt{n}\left\| \sum_{i=1}^n \pi^G_{i,n,1}(\theta) \frac{\partial \psi_i(\theta)}{\partial \theta'_w} - \sum_{i=1}^n \pi^G_{i,n,2}(\theta) \frac{\partial \psi_i(\theta)}{\partial \theta'_w} \right\| = o^U_p(1).$$

Now, noting that assumption CLT is actually a stronger version of the joint convergence assumption maintained in Proposition A.7, it follows from assumptions CLT and O that for any choice of $\pi^G_{i,n}(\theta)$, one would obtain:

$$\sqrt{n}\left[\bar{\psi}_n(\theta) - E[\bar{\psi}_n(\theta)],\ \sum_{i=1}^n \pi^G_{i,n}(\theta)\frac{\partial \psi_i(\theta)}{\partial \theta'_w} - E\left[\sum_{i=1}^n \pi^G_{i,n}(\theta)\frac{\partial \psi_i(\theta)}{\partial \theta'_w}\right]\right] \Longrightarrow \left[\Psi_{1:H}(\theta_w, \theta^0_s), \xi(\theta_w, \theta^0_s)\right].$$

Hence, the final result follows by considering the EEL implied probability for the choice of $\pi^G_{i,n}(\theta)$ since, then Lemma 1, assumptions CLT, SPM2 and O imply that for $j = 1, \ldots, p_w$:

$$\begin{aligned} \sqrt{n}E\left[\sum_{i=1}^n \pi^G_{i,n}(\theta)\frac{\partial \psi_i(\theta)}{\partial \theta_j}\right] &= \sqrt{n}\frac{\partial}{\partial \theta_j}E\left[\bar{\psi}_n(\theta)\right] - Cov\left(\frac{\partial \psi_i(\theta)}{\partial \theta_j}, \psi_i(\theta)\right)\left[V(\theta)\right]^{-1}\sqrt{n}E\left[\bar{\psi}_n(\theta)\right] + o^U(1) \\ &= G_{w,j}(\theta_w, \theta^0_s) + o^U(1). \end{aligned}$$

(ii) Directly follows as above, but without the use of Assumption CLT. ■

Now, combining the results of Lemma 2 (i) and (ii), it follows under the same assumptions that:

$$\begin{bmatrix} \sqrt{n}\sum_{i=1}^{n}\pi_{i,n}^{G}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta_w} \\ \sum_{i=1}^{n}\hat{\pi}_{i,n}\frac{\partial\psi_i'(\theta)}{\partial\theta_s} \end{bmatrix} \Longrightarrow [G^*(\theta_w,\theta_s^0)]' \quad \text{and} \quad \begin{bmatrix} \sqrt{n}\sum_{i=1}^{n}\pi_{i,n}^{G}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta_w} \\ \sum_{i=1}^{n}\pi_{i,n}^{G}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta_s} \end{bmatrix} \Longrightarrow [G^*(\theta_w,\theta_s^0)]'$$

for $(\theta_w',\theta_s')\in\Theta_n$, and where:

$$G^*(\theta)\equiv G^*(\theta_w,\theta_s)=\left[G_w(\theta)+\xi(\theta),M_s(\theta_s)\right].$$

Furthermore, $G^*(\theta_w,\theta_s^0)$ is asymptotically independent of $\sqrt{n}\bar{\psi}_n(\theta_w,\theta_s)$ for $(\theta_w',\theta_s')\in\Theta_n$.

Utilizing this result and the notation introduced above, the standard rank condition for the Jacobian is modified in Assumption R below following Andrews and Guggenberger (2017) who showed that this is particularly important when considering the asymptotic size of the score test.

Assumption R: (Rank condition) $G^*(\theta)$ is full column rank almost surely for $\theta\in\Theta_w\times\{\theta_s^0\}$. ■

Under Assumptions 1, 2, ID, SSM, SPM1-2, CLT, O and R, we will now present the proof of the main result, Theorem 1, from Section 4.

Proof of Theorem 1: First note that, Lemma 1 and CLT give:

$$\sup_{\theta\in\Theta_n}\sqrt{n}\|\bar{\psi}_n(\theta)\|=O_p(1), \tag{D.1}$$

fixing the order of magnitude of the average moment vector occurring in the score vectors in all three types of statistics: $LM_n\left(\theta,\pi^G(\theta),\pi^V(\theta)\right)$, $K_n(\theta)$ and $\tilde{K}_n(\theta)$.

Now, consider the diagonal scaling matrix T_n with $\sqrt{n}$ in the first p_w diagonal elements and 1 in the rest. T_n is nonsingular for any given n. Pre-multiply by T_n all the occurrences of

$$\left\{\sum_{i=1}^{n}\pi_{i,n}^{G}(\theta)G_i'(\theta)\right\},\quad \left\{\sum_{i=1}^{n}\pi_{i,n}^{(1)}(\theta)G_i'(\theta)\right\} \quad \text{and} \quad \left\{\begin{array}{c} \sum_{i=1}^{n}\pi_{i,n}^{G}(\theta)\frac{\partial\psi_i'(\theta)}{\partial\theta_w} \\ \sum_{i=1}^{n}\hat{\pi}_{i,n}\frac{\partial\psi_i'(\theta)}{\partial\theta_s}\end{array}\right\}$$

in $LM_n\left(\theta,\pi^G(\theta),\pi^V(\theta)\right)$, $K_n(\theta)$ and $\tilde{K}_n(\theta)$ respectively. Note that in all three types of statistics, such occurrences happen only four times: once in each of the extreme terms (score vector) and twice in the middle term (estimator of the inverse of the variance of the score vector). Thus, for all n, this pre-multiplication is harmless as T_n and T_n^{-1} always cancel out. Hence, taking (D.1) into consideration, the result of Theorem 1 now directly follows from the intermediate results in Lemma 2. ■

E Appendix E: Tables and Figures for Section 4

The following tables and figures are referred to in the Monte Carlo experiment in Section 4.3. In particular, the tables serve as references for the study of empirical size in Section 4.3.2., while the figures for that of empirical power in Section 4.3.3.

			$n = 100$ & Asymmetric moment vector						$n = 100$ & Symmetric moment vector					
H	ϱ	μ	2SGMM	GS	EL	EEL	3S	3S-Sh	2SGMM	GS	EL	EEL	3S	3S-Sh
2	0.9	0	15.7	7.3	7.1	6.5	10	10.2	14.2	6.2	6.5	5.9	9.5	9.7
2	0.9	1	8	6.6	6.4	6.3	9.5	9.6	9.1	5.3	5.4	4.9	8.4	8.2
2	0.9	10	5.3	7	6.2	6.5	9.9	9.5	5.2	5.2	5.3	4.9	8.3	8
2	0.5	0	9.8	7.2	7.2	6.5	10.6	10.4	8.6	6.1	6.6	5.8	9.4	9.2
2	0.5	1	6.1	6.8	7.1	6.2	10.1	10	6.6	5.7	6	5.3	8.8	8.6
2	0.5	10	6.7	7.2	6.5	6.8	9.9	9.6	5.1	5.1	5.4	4.8	8.2	8
2	0	0	6.1	6.6	6.4	6.1	9.8	9.6	5.5	5.9	6.2	5.4	9.3	9.1
2	0	1	7.1	7.1	7	6.6	10.2	9.9	5.9	5.8	6.1	5.5	9.3	9.1
2	0	10	6.9	6.5	5.9	6	8.9	8.7	5.3	5.3	5.6	5	8.4	8.1
4	0.9	0	39.6	8.3	8.7	6.7	14.9	17.1	38.2	6.5	7.5	5.5	13.5	15.5
4	0.9	1	14	8.5	8.1	7.7	14.2	13.6	18.8	5.9	6.6	4.9	12.4	12
4	0.9	10	6.1	8.9	8	7.7	14.8	12.6	7.3	5.6	6.6	4.5	11.8	10.5
4	0.5	0	16.8	8.4	8.7	6.8	15.2	13.8	15.3	6.7	8	5.4	14	12.9
4	0.5	1	6.7	8.1	8.1	6.9	14.3	12.4	9.7	6	6.8	4.8	12.2	11.3
4	0.5	10	8.2	9.5	8.9	8.2	15.6	13.3	6.4	5.8	6.7	4.9	12	10.8
4	0	0	7.1	8.9	8.8	7	15.8	13.8	6.2	6.8	7.9	5.9	13.8	12.4
4	0	1	8.6	8.3	8.6	6.8	15.1	13	6.3	6.2	7.1	5.2	12.9	11.5
4	0	10	9.4	8.6	7.6	7.4	14	11.7	5.8	5.6	6.7	4.6	12.2	10.6
			$n = 1000$ & Asymmetric moment vector						$n = 1000$ & Symmetric moment vector					
H	ϱ	μ	2SGMM	GS	EL	EEL	3S	3S-Sh	2SGMM	GS	EL	EEL	3S	3S-Sh
2	0.9	0	13.2	5.2	5.2	5.2	5.6	5.7	13	4.9	4.9	4.8	5.2	5.2
2	0.9	1	8.4	5.2	5.1	5.1	5.6	5.6	8	4.5	4.4	4.4	4.8	4.8
2	0.9	10	4.3	5.1	4.8	5.1	5.1	5.1	4.9	5.1	5	5	5.3	5.3
2	0.5	0	7.5	4.9	4.7	4.7	5.2	5.2	7.8	4.7	4.6	4.7	5.1	5.1
2	0.5	1	5.7	5.3	5.2	5.3	5.8	5.8	6.3	5.2	5.2	5.1	5.5	5.5
2	0.5	10	5.2	5.6	5.1	5.5	5.7	5.6	5.2	5.1	5	5	5.3	5.3
2	0	0	5.4	5.6	5.3	5.4	5.9	5.9	5	5	4.9	4.9	5.3	5.3
2	0	1	5.2	5.2	5.2	5.2	5.7	5.7	5.3	5.2	5.1	5.1	5.6	5.6
2	0	10	5.2	5.2	5	5.1	5.4	5.4	5.3	5.3	5.3	5.2	5.6	5.6
4	0.9	0	35.5	5.5	5.2	5.2	5.9	6	35.8	5	4.9	4.8	5.7	5.7
4	0.9	1	15.6	5.4	5	5.3	5.7	5.8	17.2	5.3	5.2	5.1	5.6	5.6
4	0.9	10	4.9	5.4	4.7	5.3	5.6	5.6	6.2	5	4.9	4.9	5.4	5.4
4	0.5	0	14.3	5.3	5.1	5	5.8	5.8	13.9	5	4.8	4.8	5.4	5.4
4	0.5	1	7.3	5.4	5	5.2	5.8	5.8	8.9	5.1	4.9	5	5.5	5.5
4	0.5	10	5.1	5.5	5.3	5.4	6	6	5	4.5	4.5	4.4	5	5
4	0	0	5.5	5.8	5.5	5.5	6.4	6.3	4.7	4.9	4.8	4.7	5.5	5.5
4	0	1	5	5.1	5	4.8	5.7	5.7	5.2	5.2	5.1	5	5.7	5.7
4	0	10	5.3	5.1	4.7	5	5.4	5.3	5.1	5.1	5	4.9	5.5	5.5

Table 2: Empirical size of 5% score tests based on 10000 Monte Carlo trials. H: number of instruments (moments), ϱ: level of endogeneity, and μ: strength of instruments (moments).

			Asymmetric moment vector			Symmetric moment vector		
			$n = 100$			$n = 100$		
H	ϱ	μ	Kl-ET	KLIC	ET	Kl-ET	KLIC	ET
2	0.9	0	6.9	7.2	8.2	6.1	6.7	7.8
2	0.9	1	6.5	6.7	7.6	5.1	5.6	6.6
2	0.9	10	6.7	6.8	7.7	5.1	5.5	6.5
2	0.5	0	7	7.5	8.5	5.9	6.6	7.8
2	0.5	1	6.4	7	8.1	5.4	6	7.2
2	0.5	10	7	7	8	5	5.6	6.4
2	0	0	6.3	6.8	7.9	5.6	6.4	7.6
2	0	1	6.7	7.1	8.4	5.6	6.3	7.4
2	0	10	6.2	6.2	7.2	5.1	5.8	6.9
4	0.9	0	7.3	8.2	10.3	5.9	7.3	9.9
4	0.9	1	8.1	8	10.2	5.2	6.3	9
4	0.9	10	8.2	8.2	10	5	6	8.4
4	0.5	0	7.4	8.1	10.9	5.8	7.4	10
4	0.5	1	7.2	7.8	9.9	5.2	6.3	8.7
4	0.5	10	8.8	8.7	10.9	5.3	6.3	8.7
4	0	0	7.6	8.6	11.1	6.4	7.6	9.6
4	0	1	7.5	8.2	10.6	5.6	6.7	8.9
4	0	10	8.1	8	9.3	5.1	6.3	8.5
			Asymmetric moment vector			Symmetric moment vector		
			$n = 1000$			$n = 1000$		
H	ϱ	μ	Kl-ET	KLIC	ET	Kl-ET	KLIC	ET
2	0.9	0	5.2	5.3	5.5	4.9	4.9	5
2	0.9	1	5.1	5.3	5.4	4.4	4.5	4.6
2	0.9	10	5.1	5	4.9	5.1	5.1	5.2
2	0.5	0	4.8	4.8	4.9	4.7	4.8	4.8
2	0.5	1	5.3	5.4	5.5	5.2	5.3	5.4
2	0.5	10	5.5	5.4	5.4	5.1	5.1	5.1
2	0	0	5.5	5.5	5.5	4.9	5	5.1
2	0	1	5.2	5.2	5.5	5.1	5.2	5.3
2	0	10	5.1	5.1	5.2	5.3	5.4	5.4
4	0.9	0	5.3	5.4	5.5	4.9	5.1	5.3
4	0.9	1	5.4	5.2	5.3	5.2	5.3	5.4
4	0.9	10	5.3	5.1	5.2	5	5.1	5.2
4	0.5	0	5.1	5.2	5.3	4.9	5	5.1
4	0.5	1	5.3	5.2	5.3	5.1	5.1	5.2
4	0.5	10	5.4	5.4	5.6	4.5	4.6	4.7
4	0	0	5.6	5.6	5.8	4.8	5	5.2
4	0	1	5	5	5.3	5.1	5.3	5.4
4	0	10	5	4.9	5.1	5	5.1	5.3

Table 3: Empirical size of 5% score tests based on 10000 Monte Carlo trials. H: number of instruments (moments), ϱ: level of endogeneity, and μ: strength of instruments (moments). Jacobian is re-weighted by the ET implied probabilities for all three tests. K-ET: empirical variance matrix. ET: variance matrix re-weighted by $k_i/(\sum_j k_j)$. our-ET: variance matrix re-weighted by the ET implied probabilities. All re-weighting impose the null hypothesis.

Figure 1: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 0$: strength of instruments.

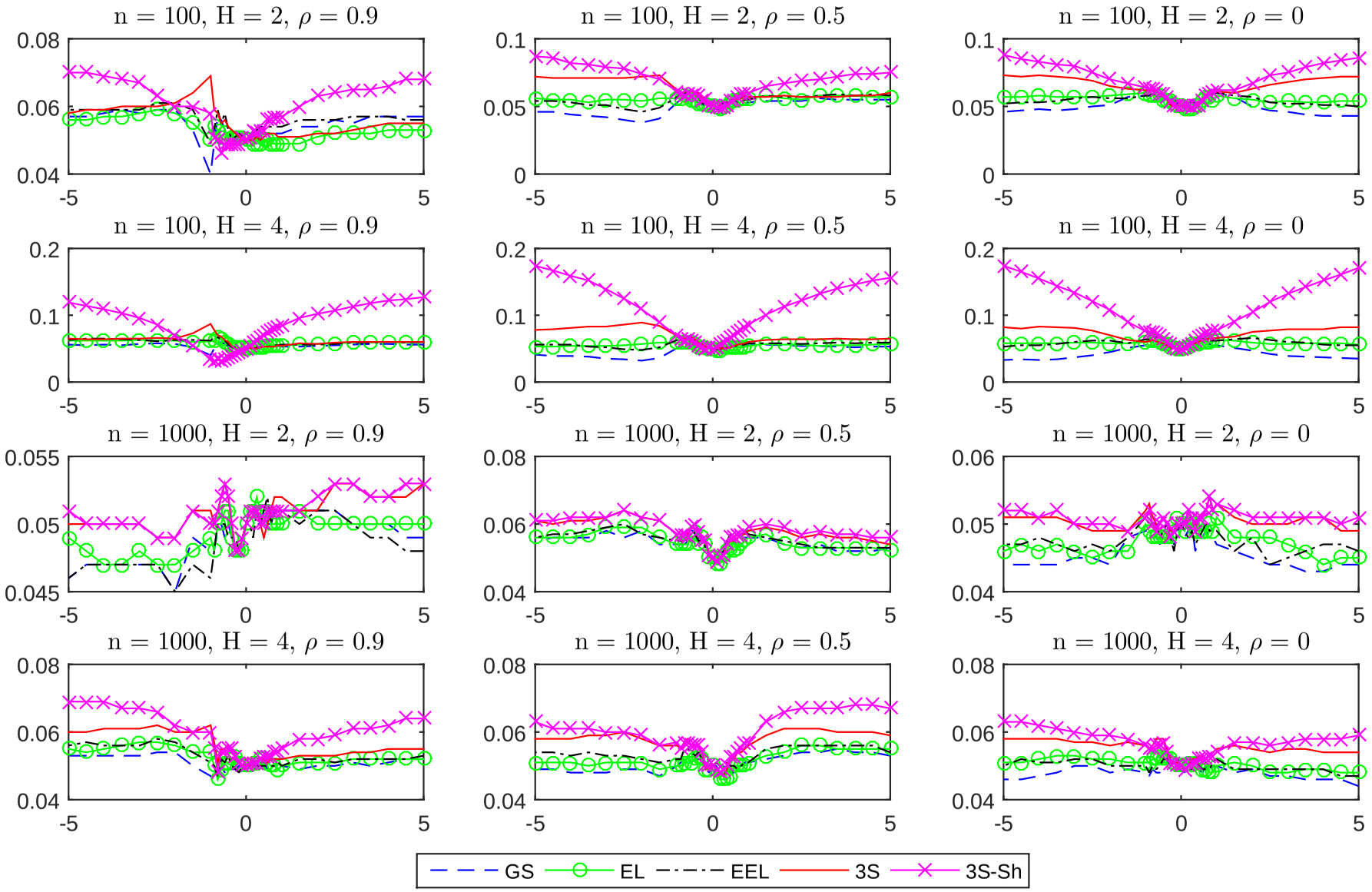

Figure 2: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 1$: strength of instruments.

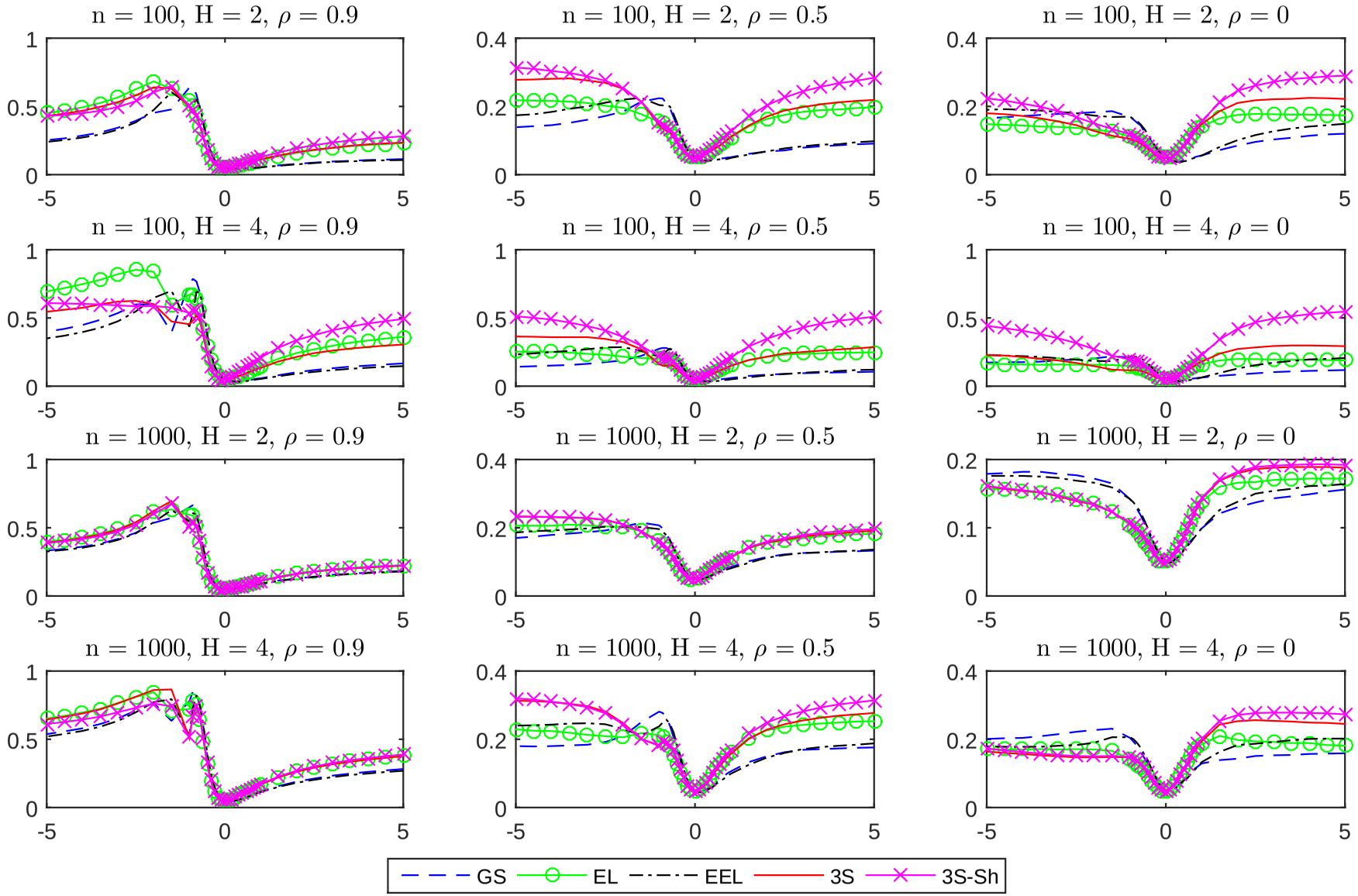

Figure 3: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 10$: strength of instruments.

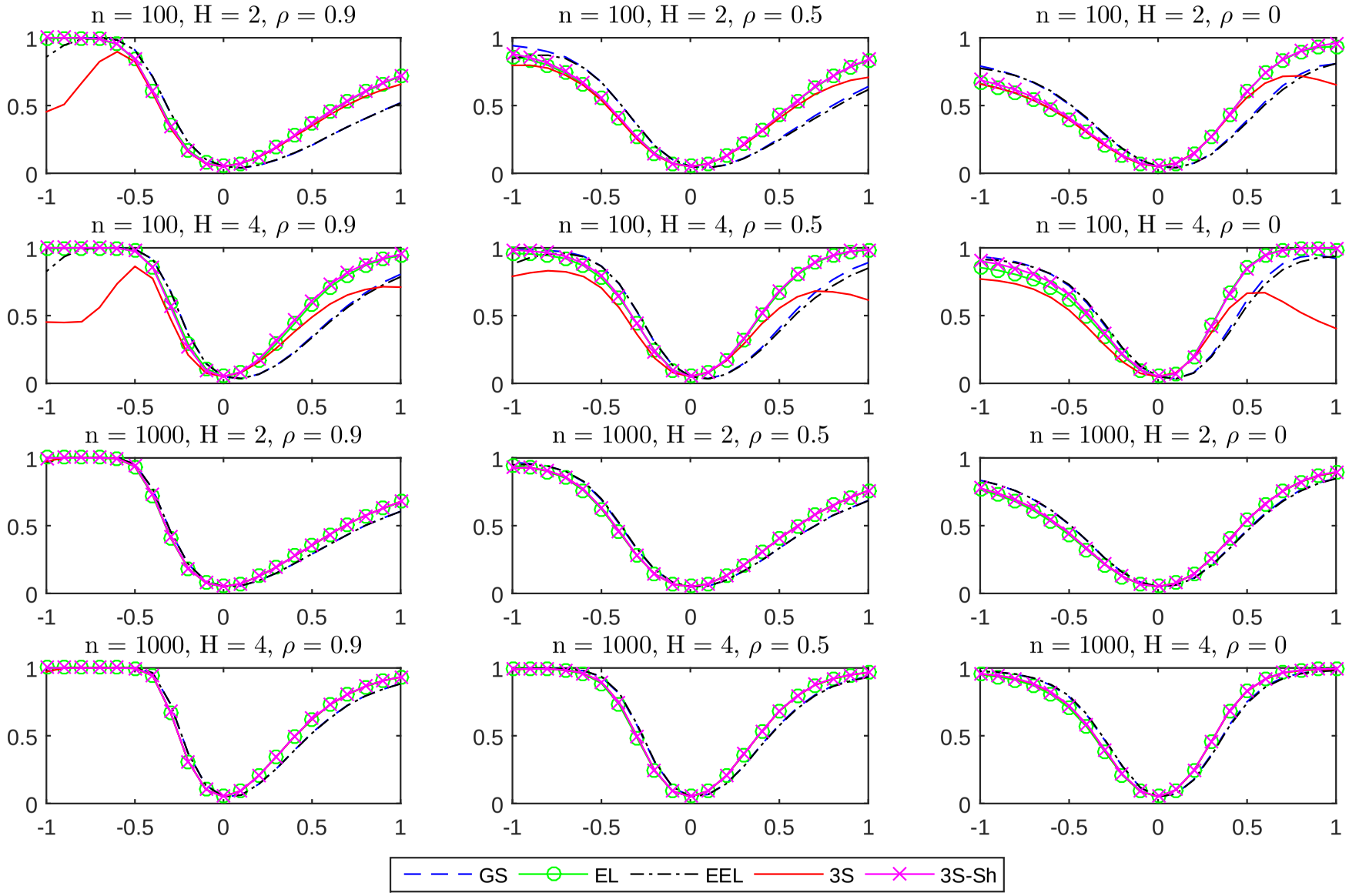

Figure 4: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 0$: strength of instruments.

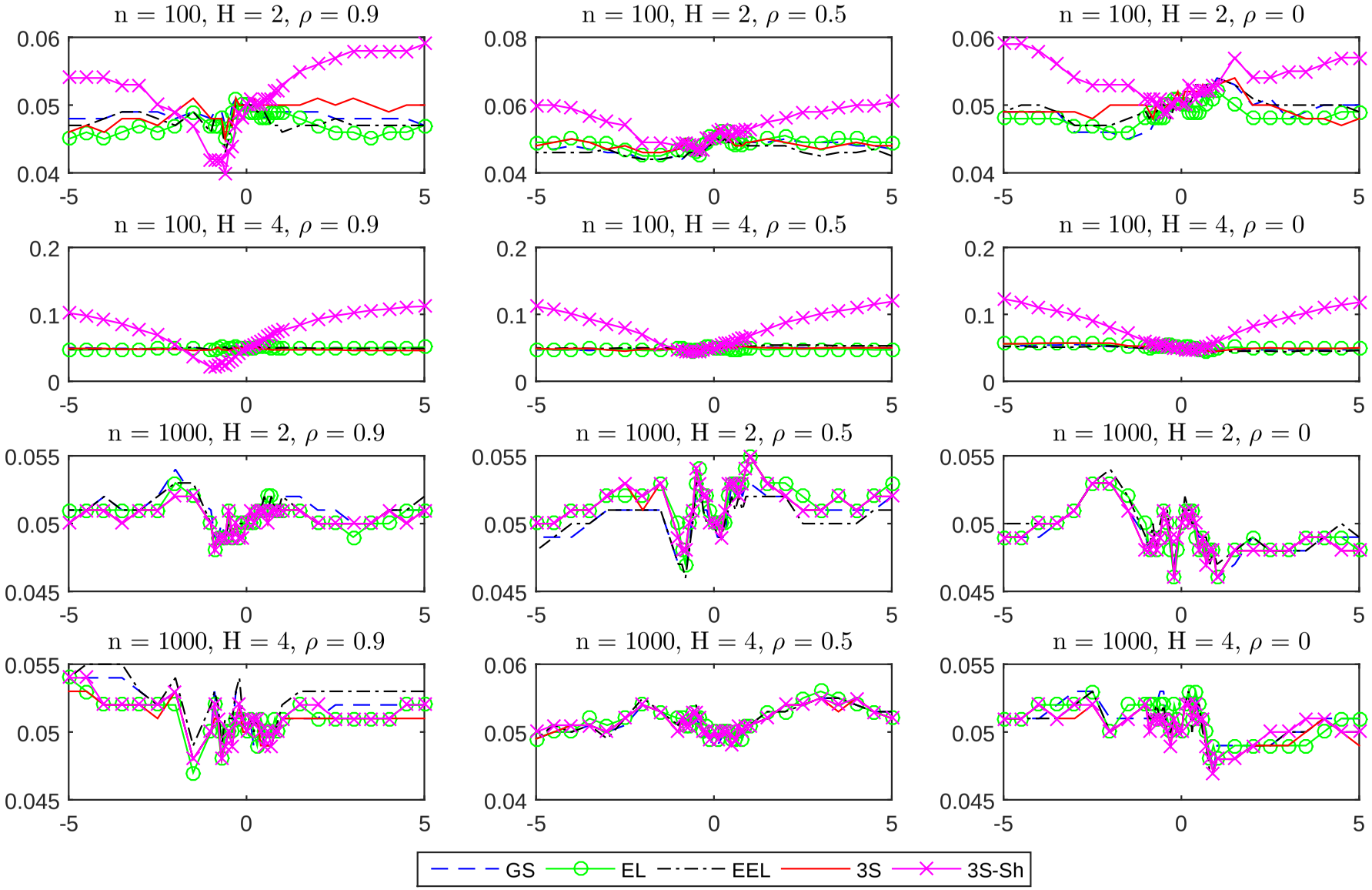

Figure 5: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 1$: strength of instruments.

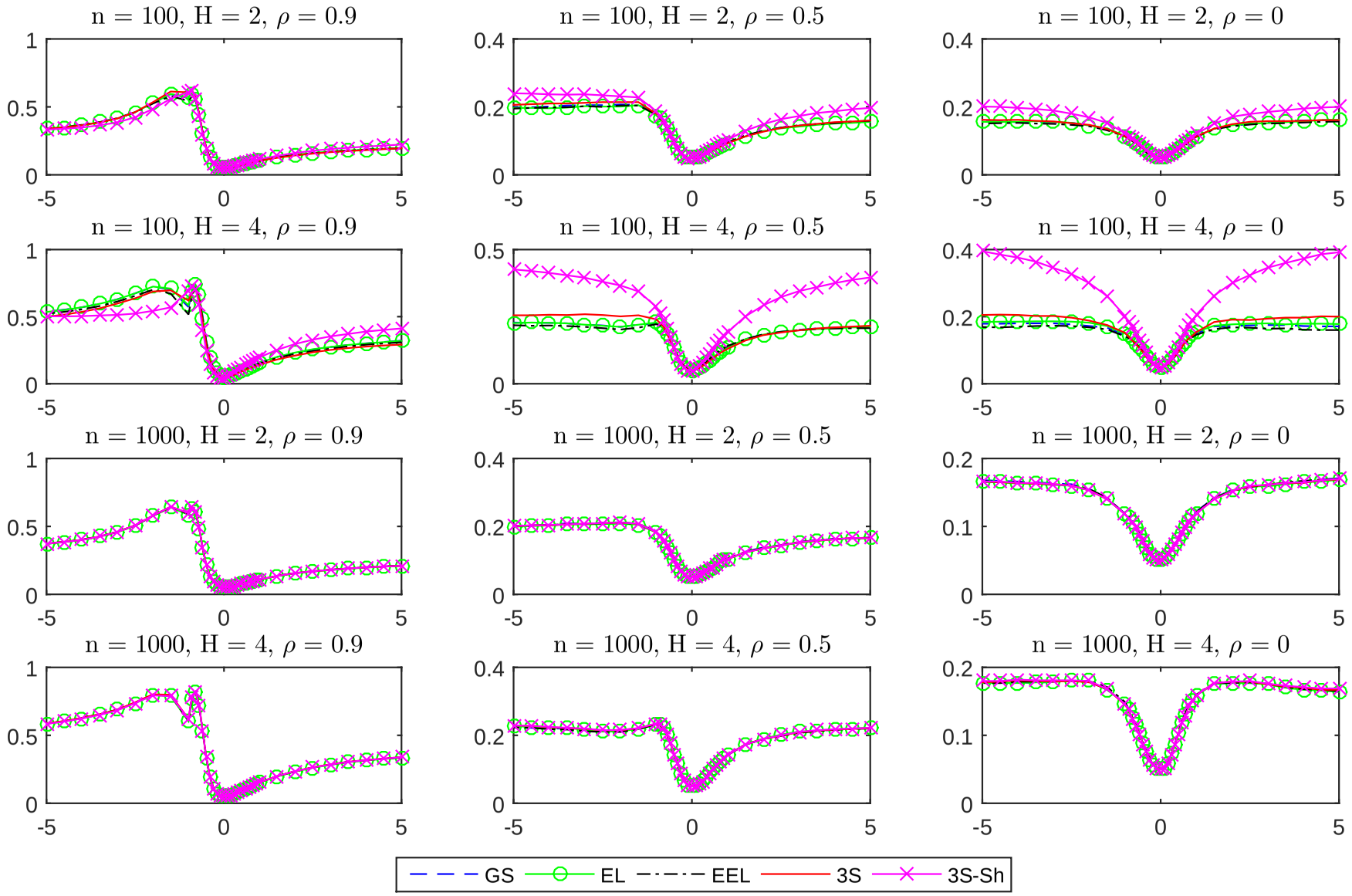

Figure 6: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 10$: strength of instruments.

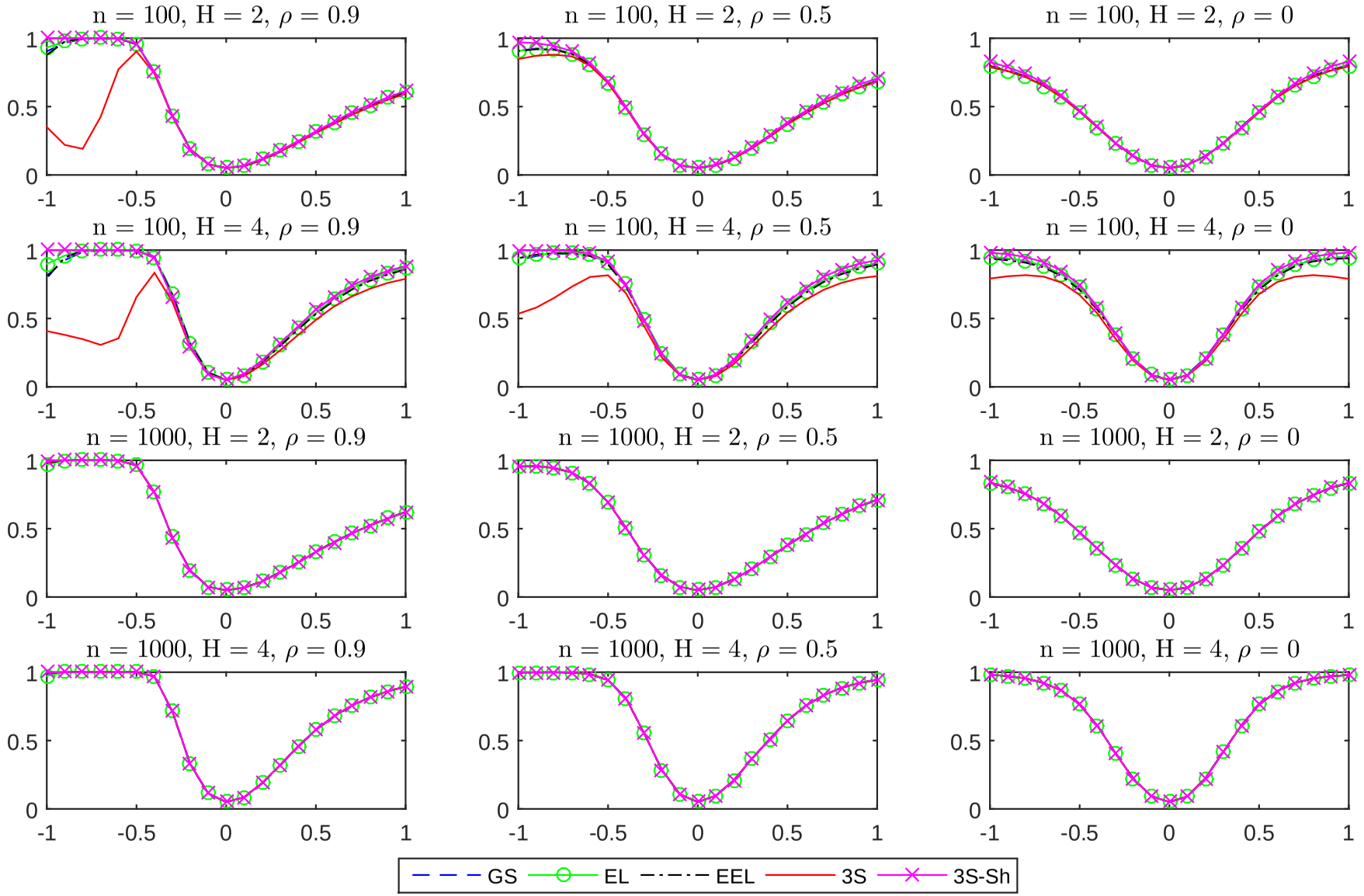

Figure 7: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 0$: strength of instruments.

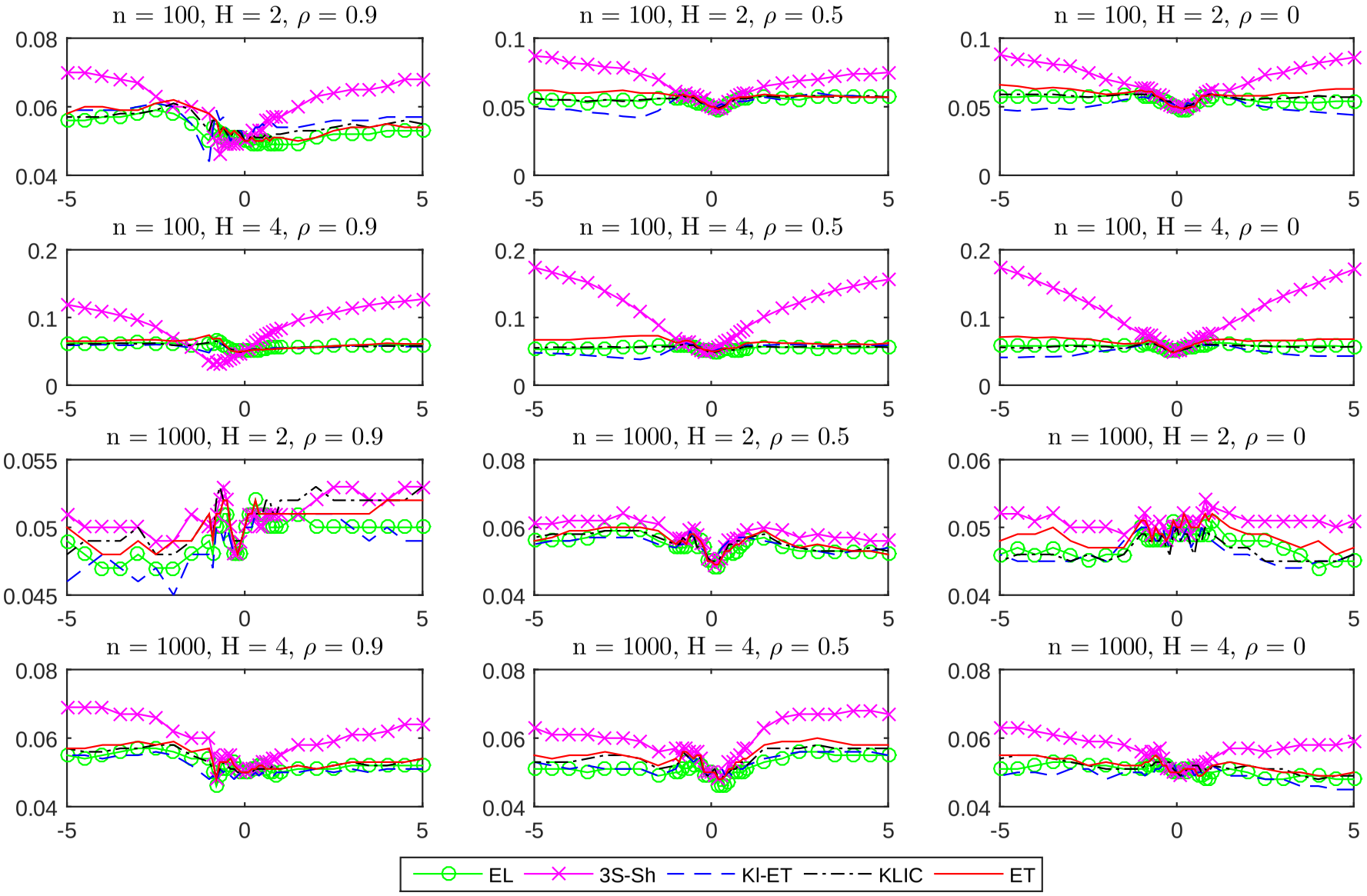

Figure 8: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 1$: strength of instruments.

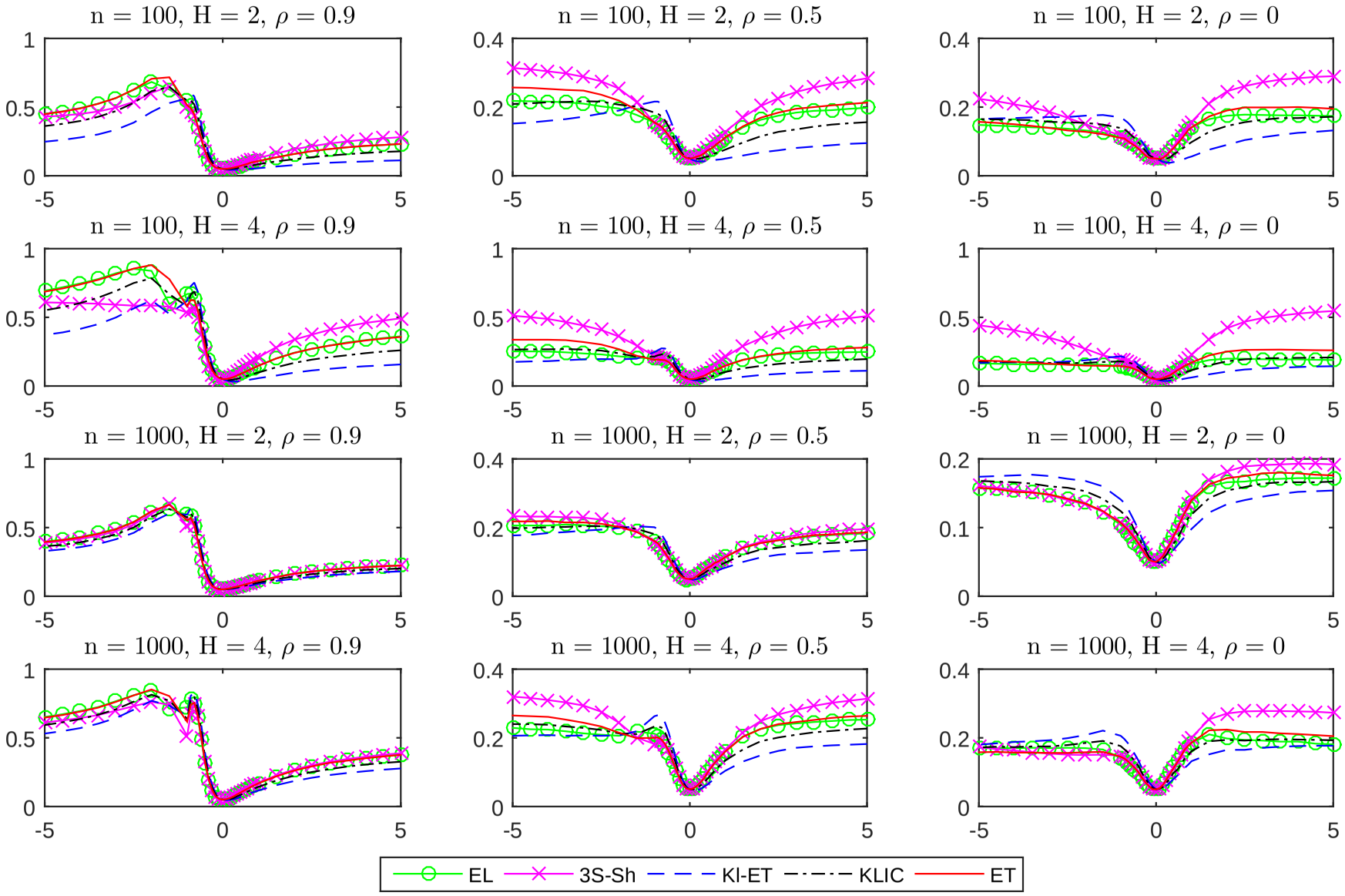

Figure 9: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Asymmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 10$: strength of instruments.

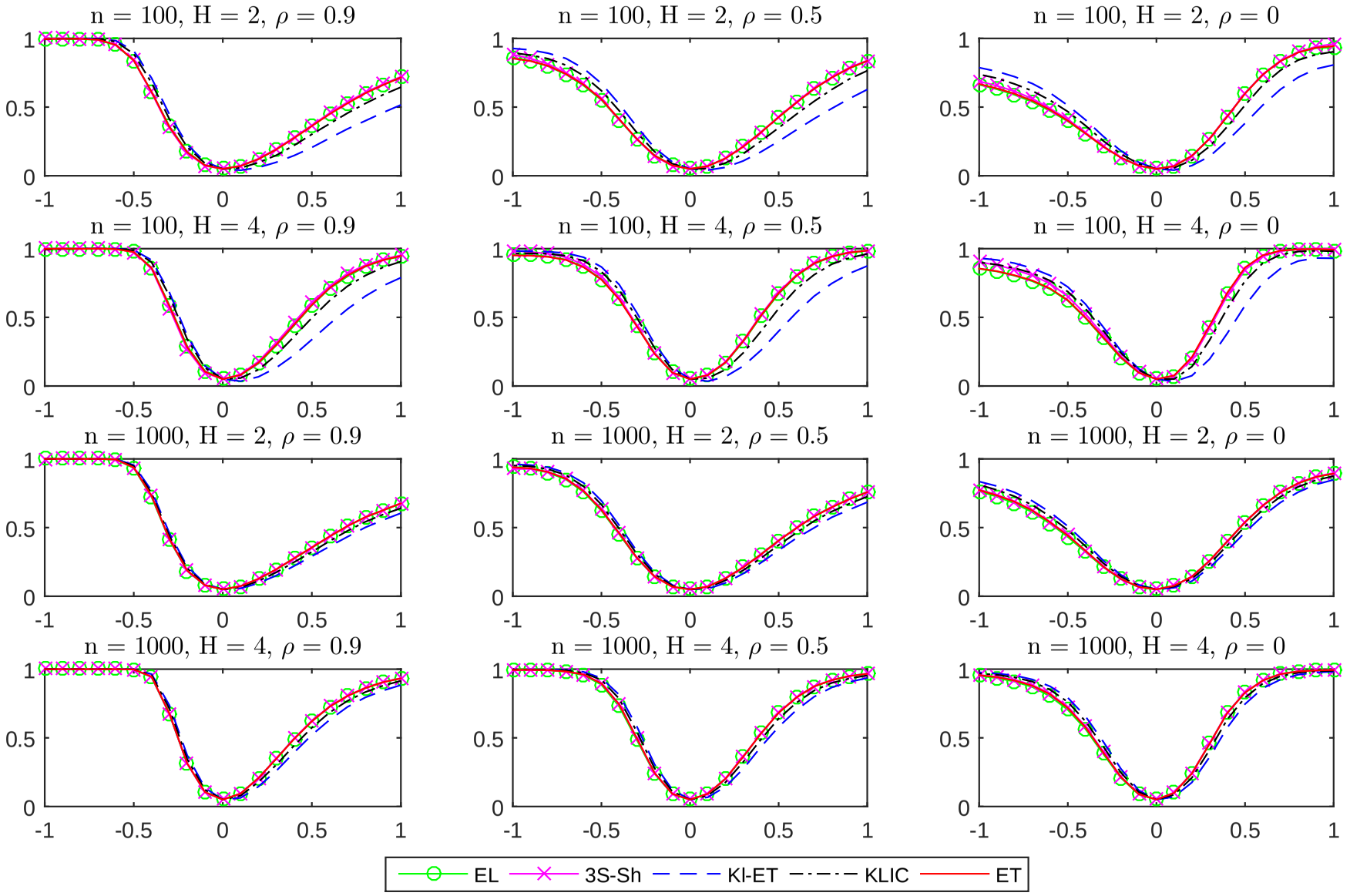

Figure 10: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 0$: strength of instruments.

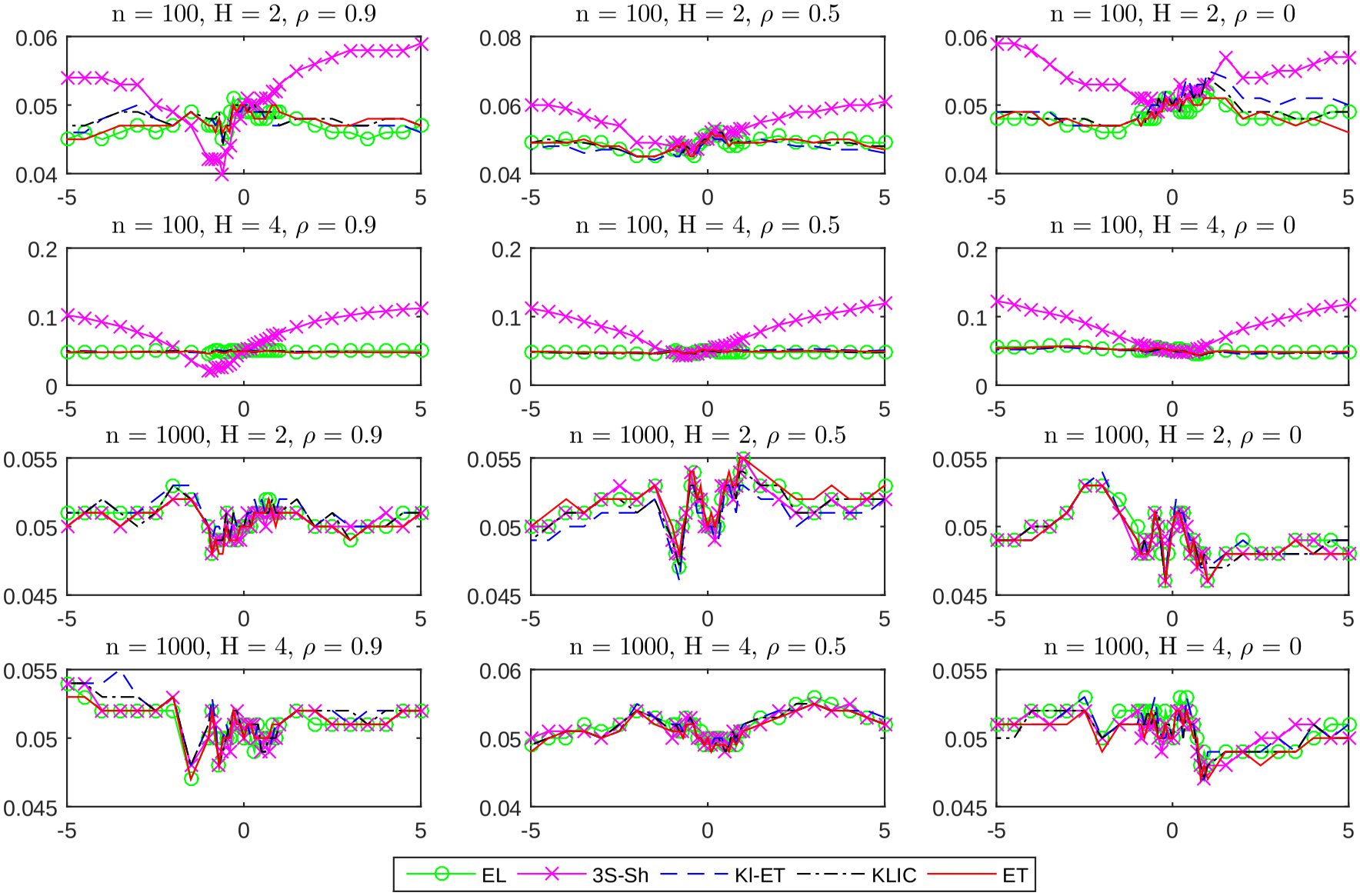

Figure 11: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 1$: strength of instruments.

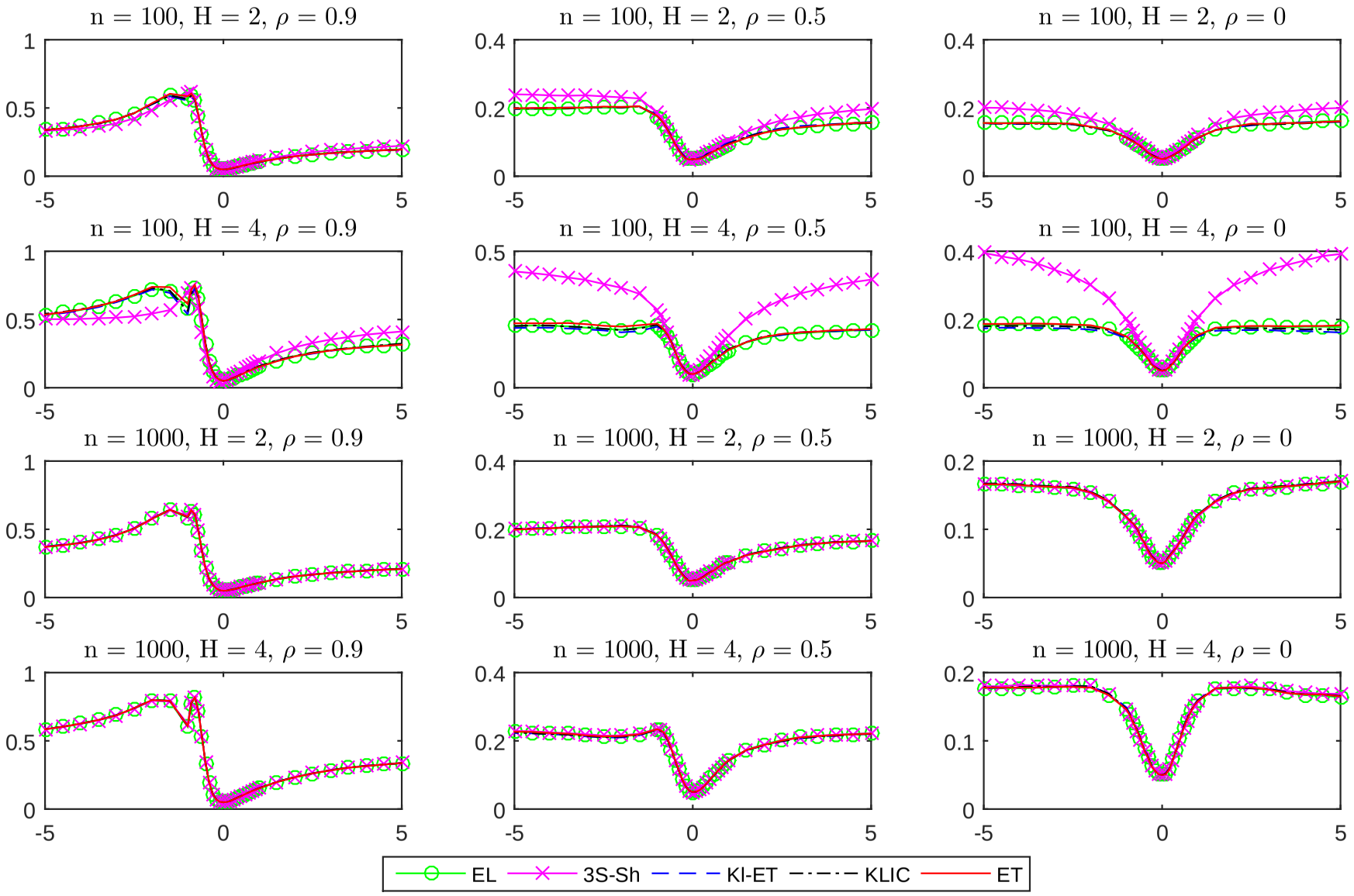

Figure 12: Size-adjusted power of 5% score tests based on 10000 Monte Carlo trials. Horizontal axis: true - hypothesized θ. Symmetric moment vector, H: number of instruments, ϱ: level of endogeneity, and $\mu = 10$: strength of instruments.