Subsampling tests of parameter hypotheses and overidentifying restrictions with possible failure of identification

Wolf, Michael

Abstract: We introduce a general testing procedure inmodels with possible identification failure that has exactasymptotic rejection probability under the null hypothesis. The procedure iswidely applicable and in this paper we apply it to tests of arbitrary linear parameter hypotheses as well as to tests of overidentification in time series models given by unconditional moment conditions. The main idea is to subsample classical tests, like for example theWald or the J test.More precisely, instead of using critical values based on asymptotic theory,we compute data-dependent critical values based on the subsampling technique. We show that under full identification the resulting tests are consistent against fixed alternatives and that they have exact asymptotic rejection probabilities under the null hypothesis independent of identification failure. Furthermore, the subsampling tests of parameter hypotheses are shown to have the same local power as the original tests under full identification. An algorithm is provided that automates the block size choice needed to implement the subsampling testing procedure. A Monte Carlo study shows that the tests have reasonable size properties and often outperform other robust tests in terms of power.

DOI: https://doi.org/10.1016/j.ijar.2012.12.003

Subsampling Tests of Parameter Hypotheses and Overidentifying Restrictions with Possible Failure of Identification

Michael Wolf *
Department of Economics
University of Zurich
Switzerland

Abstract

We introduce a general testing procedure in models with possible identification failure that has exact asymptotic rejection probability under the null hypothesis. The procedure is widely applicable and in this paper we apply it to tests of arbitrary linear parameter hypotheses as well as to tests of overidentification in time series models given by unconditional moment conditions. The main idea is to subsample classical tests, like for example the Wald or the J test. More precisely, instead of using critical values based on asymptotic theory, we compute data-dependent critical values based on the subsampling technique.

We show that under full identification the resulting tests are consistent against fixed alternatives and that they have exact asymptotic rejection probabilities under the null hypothesis independent of identification failure. Furthermore, the subsampling tests of parameter hypotheses are shown to have the same local power as the original tests under full identification.

An algorithm is provided that automates the block size choice needed to implement the subsampling testing procedure. A Monte Carlo study shows that the tests have reasonable size properties and often outperform other robust tests in terms of power.

JEL Classification: C12, C15, C52.

Keywords: Hypothesis Testing, Nonlinear Moment Conditions, Overidentifying Restrictions, Partial Identification, Subsampling, Weak Identification.

*I would like to thank Patrik Guggenberger for helpful comments. Research has been supported by the NCCR Finrisk project "New Methods in Theoretical and Empirical Asset Pricing".

1 Introduction

Since Phillips' (1989) seminal paper on the consequences of identification failure on the distribution of point estimators and test statistics a vast literature on partially or weakly identified models has developed.[1] A major finding of this literature is that in models with possible identification failure, point estimates can be severely biased and classical tests of parameter hypotheses or overidentifying restrictions can be extremely size distorted in finite samples.

In response to the unreliability of classical tests of parameter hypotheses in models with identification failure, such as Wald or likelihood ratio tests, several new tests for simple full vector parameter hypotheses have recently been introduced whose rejection probabilities under the null hypothesis are (asymptotically) unaffected by identification failure.[2]

However, to the best of our knowledge, no test of overidentifying restrictions, that is consistent under full identification and robust to identification failure, has been introduced in the literature. Furthermore, generalizations in the literature of the above mentioned tests of simple full vector parameter hypotheses to more general parameter hypotheses, either require additional assumptions or are too conservative. For example, Kleibergen's (2004, 2005) test can be used to test simple subvector hypotheses under the additional assumption that the parameters not under test are strongly identified. Dufour (1997) suggests a projection-based testing procedure for general parameter hypotheses that works without additional assumptions but leads to conservative tests. Furthermore, in general the projection idea is computationally cumbersome.[3]

In this paper, we address the need for robust tests of general linear hypotheses and overidentifying restrictions. More precisely, we introduce a general testing procedure in models with possible identification failure that has exact asymptotic rejection probability under the null hypothesis. We then apply the procedure to *tests of arbitrary linear parameter hypotheses* as well as for *tests of overidentifying restrictions*

[1] See among others, Nelson and Startz (1990), Choi and Phillips (1992), Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Forchini and Hillier (2003), and for recent reviews of the weak identification literature see Stock et al. (2002) and Zivot et al. (2006). A recent paper by Chao and Swanson (2006) brings together the many (Bekker, 1994) and weak instruments literature.

[2] Besides the early contribution of Anderson and Rubin (1949) see among others, Stock and Wright (2000), Kleibergen (2002, 2005), Caner (2010), Guggenberger (2003), Moreira (2003), Otsu (2006), Dufour and Taamouti (2005, 2007), and Guggenberger and Smith (2005). Andrews et al. (2006) investigate robust hypothesis testing with optimal power properties when the instrumental variables might be weak.

[3] One exception is the Anderson and Rubin (1949) statistic for scalar linear hypotheses where a closed-form solution is available, see Dufour and Taamouti (2005, 2007).

in time series models given by unconditional moment restrictions. The main idea of our procedure is to apply subsampling to classical tests. More specifically, instead of using critical values based on asymptotic theory, we compute data-dependent critical values based on the subsampling technique. The test statistic under consideration is evaluated on all (overlapping) blocks of the observed data sequence, where the common block size is small compared to the sample size. The critical value is then obtained as an empirical quantile of the resulting block (or subsample) test statistics.

We first introduce a general definition of identification failure that brings together Phillips' (1989) notion of partial identification with Stock and Wright's (2000) notion of weak identification. We then apply the subsampling method to the Wald and the J test of Hansen (1982), also see Newey (1985)) and show that under full identification the resulting tests are consistent against fixed alternatives and have exact asymptotic rejection probabilities under the null hypothesis independent of identification failure. Furthermore, we show that the subsampling version of the Wald test has the same local power as the Wald test under full identification. Our analysis is done in time series models given by nonlinear moment conditions. Throughout the paper, we use the linear single equation instrumental variables model as an illustrative example of the general time series model. Our parameter tests can be applied to general linear hypotheses without additional identification assumptions. In particular, unlike Kleibergen (2004, 2005), no additional identification assumptions are required for subvector testing. Also, in a linear single equation instrumental variables model we can test simultaneous hypotheses on the coefficients of the exogenous and endogenous variables. A further advantage of the subsampling approach considered here is its robustness to the model assumptions. For example, we show that the sizes of the subsampling tests are not affected (asymptotically) by instrument exclusion in the reduced form of a linear single equation instrumental variables model. This last advantage also holds true for the Anderson and Rubin (1949) statistic (in the case of a simple full vector hypothesis) but not for the tests by Kleibergen (2002, 2004) or Moreira (2003); see Dufour and Taamouti (2007).

We assess the finite sample performance of several parameter subvector tests in a Monte Carlo study using Dufour and Taamouti's (2007) linear design with two endogenous variables on the right side of the structural equation. We find that their projected Anderson and Rubin test is dominated in terms of power by Kleibergen's (2003) test across every single scenario. In all scenarios, where the parameter not under test is only weakly identified, our subsampled Wald test is the clear winner among the three statistics and the power gains can be dramatic in these cases. If this parameter is strongly identified, then Kleibergen's (2003) test typically has slightly better power properties than our test. In an additional Monte Carlo experiment we

assess the power loss of the subsampling procedure in a scenario where subsampling does not enjoy a comparative advantage, namely when testing a simple full vector hypothesis in a linear *i.i.d.* model. We find that even in this disadvantageous setup, subsampling still performs competitively but is often outperformed in terms of power by Moreira (2003). Lastly, we conduct an experiment to assess the size properties of tests of overidentifying restrictions. Again, our Monte Carlo results are consistent with our theory: While the classical J test oftentimes severely overrejects, we find that subsampling has generally very reliable size properties.

Besides all the advantages of the subsampling technique mentioned above, there are also general drawbacks. Firstly, compared to tests that are given in closed form, a relative disadvantage of the subsampling approach is its computational burden. However, this disadvantage is shared with other popular resampling methods, such as the bootstrap and the jackknife. Secondly, an application of the subsampling method requires the choice of a block size b, which can be considered a model parameter. To overcome that problem, we provide a data-dependent method to automate this choice. Thirdly, under full identification, weak regularity conditions, and one-sided alternatives, the error in rejection probability under the null for tests based on subsampling is typically of order $O_p(b^{-1/2})$ compared to the faster $O_p(n^{-1/2})$ of standard approaches, where b and n denote the block and sample size, respectively, see Politis et al. (1999, chapter 10.1). This slower convergence under full identification is the price that has to be paid for making the procedure robust to identification failure. Lastly, in our specific application, the finite sample power function of a subsampling version of a test is oftentimes below the one of the original test in strongly identified situations. However, compared to other tests that are robust to identification failure, our Monte Carlo study indicates that oftentimes there can be tremendous power gains of the subsampling approach.

In the Econometrics literature, subsampling has now been suggested in a variety of situations for the construction of confidence intervals or hypotheses tests where it is at least questionable whether the bootstrap would work. Some recent examples include, Romano and Wolf (2001) who use subsampling to construct confidence intervals for the autoregressive coefficient in an AR(1) model with a possible unit root. Andrews (2003) introduces a subsampling-like testing method for structural instability of short duration. Choi (2005) uses subsampling for tests of linear parameter constraints in a vector autoregression with potential unit roots and Gonzalo and Wolf (2005) suggest subsampling for the construction of confidence intervals for the threshold parameter in threshold autoregressive models with potentially discontinuous autoregressive function.

Related to our paper is Kleibergen (2003) who derives higher order expansions

of various statistics that are robust to weak instruments and suggests the bootstrap to further improve on the size properties of tests based on these statistics. He also provides insight as to why the bootstrap is not expected to improve on the size distortion of classical tests, like a Wald test. In an *i.i.d.* linear model with one endogenous right hand side variable, Moreira et al. (2004) go one step further by providing a *formal proof* of the validity of Edgeworth expansions for the score and conditional likelihood ratio statistics when instruments may be weak. These statistics are known to be robust to weak instruments. They show the validity of the bootstrap for the score test and the validity of the conditional bootstrap for various conditional tests. On the other hand, our paper shows that in general time series moment condition models subsampling fixes the size distortion of classical tests of general hypotheses that are not robust to weak identification, like a Wald test.

The remainder of the paper is organized as follows. In Section 2, the model is introduced, the testing problems are described, and a general definition of identification failure is provided. In order to be self contained, in Section 3 we first briefly review the basic theory of subsampling for time series data. We then derive the asymptotic distribution of some classical test statistics under the general asymptotic framework of identification failure to show that the tests are generally size distorted under identification failure. We then apply subsampling to those tests in Subsections 3.2 (overidentifying restrictions) and 3.3 (general linear parameter hypotheses) to cure the problem of size distortion. In Section 4 we provide a data-driven choice of the block size needed to implement the subsampling procedure. Section 5 describes the simulation results. All proofs are relegated to Appendix, Part (C) while Appendix, Part (A) and Part (B) contain discussion of our assumption on identification failure and contiguity, respectively.

The following notation and terminology is used in the paper. The symbols " $\rightarrow_d$ ", " $\rightarrow_p$ ", and " $\Rightarrow$ " denote convergence in distribution, convergence in probability, and weak convergence of empirical processes, respectively. For the latter, see Andrews (1994) for a definition. For "with probability 1" we write "w.p.1" and "a.s." stands for "almost surely". By $C^i(A, B)$ we denote the set of functions $f : A \rightarrow B$ that are i times continuously differentiable. If $B = \mathbb{R}$, the set of real numbers, we simply write $C^i(A)$ for $C^i(A, B)$. By id we denote the identity map and by $O(i)$ the group of orthogonal $i \times i$ matrices. By $e_j \in \mathbb{R}^p$ we denote the p-vector $(0, ..., 1, ..., 0)'$ with 1 appearing at position j. For a matrix M, $M > 0$ means that M is positive definite and $[M]_{i,j}$ denotes the element of M in row i and column j. By I_i we denote the i-dimensional identity matrix. Furthermore, $vec(M)$ stands for the column vectorization of the $k \times i$ matrix M, i.e. if $M = (m_1, ..., m_i)$ then $vec(M) = (m_1', ..., m_i')'$. By P_M we denote the orthogonal projection onto the range space of M. Finally,

$||M||$ equals the square root of the largest eigenvalue of $M'M$ and "$\otimes$" denotes the Kronecker product.

2 The Model, Tests, and Identification Failure

2.1 The Model

We consider models specified by a finite number of unconditional moment restrictions. Let $\{z_i : i = 1, ..., n\}$ be $\mathbb{R}^l$-valued data and, for each $n \in \mathbb{N}$, $g_n : G \times \Theta \to \mathbb{R}^k$, where $G \subset \mathbb{R}^l$ and $\Theta \subset \mathbb{R}^p$ denotes the parameter space. The model has a true parameter θ_0 for which the moment condition

$$Eg_n(z_i, \theta_0) = 0 \tag{2.1}$$

is satisfied for all $i = 1, ..., n$. For $g_n(z_i, \theta)$ we usually simply write $g_i(\theta)$. For example, moment conditions may result from conditional moment restrictions. Assume $E[h(Y_i, \theta_0)|F_i] = 0$, where $h : H \times \Theta \to \mathbb{R}^{k_1}$, $H \subset \mathbb{R}^{k_2}$, and F_i is the information set at time i. Let Z_i be a k_3-dimensional vector of instruments contained in F_i. If $g_i(\theta) := h(Y_i, \theta) \otimes Z_i$, then $Eg_i(\theta_0) = 0$ follows by taking iterated expectations. In (2.1), $k = k_1 k_3$ and $l = k_2 + k_3$. A second important example of model (2.1) is given by the following:

Example 2.1 (I.i.d. linear instrumental variable (IV) model): Consider the linear model with *i.i.d.* observations given by the structural equation

$$y = Y\beta_0 + X\gamma_0 + u \tag{2.2}$$

and the reduced form for Y

$$Y = Z\Pi + X\Phi + V, \tag{2.3}$$

where $y, u \in \mathbb{R}^n$, $Y, V \in \mathbb{R}^{n \times v_1}$, $X \in \mathbb{R}^{n \times v_2}$, $Z \in \mathbb{R}^{n \times j}$, $\Phi \in \mathbb{R}^{v_2 \times v_1}$, and $\Pi \in \mathbb{R}^{j \times v_1}$. Let $p := v_1 + v_2$, $k := j + v_2$, $\theta = (\beta', \gamma')'$, and $\theta_0 = (\beta_0', \gamma_0')'$. The matrix Y contains the endogenous and the matrix X contains the exogenous variables. The variables Z constitute a set of instruments for the endogenous variables Y. For the model to be identified it is necessary that $j \geq v_1$. Denote by Y_i, V_i, Z_i, ... $(i = 1, ..., n)$ the i^{th} row of the matrix Y, V, Z, ... written as a column vector and similarly for analogous expressions. Assume $E(Z_i', X_i')'u_i = 0$ and $E(Z_i', X_i')'V_i' = 0$. The first condition implies that $Eg_i(\theta_0) = 0$, where for each $i = 1, ..., n$

$$g_i(\theta) := (Z_i', X_i')'(y_i - Y_i'\beta - X_i'\gamma).$$

Note that in this example $g_i(\theta)$ depends on n if the reduced form coefficient matrix Π is modeled to depend on n, see Staiger and Stock (1997).

2.2 Hypothesis Tests

Interest focuses on two separate testing problems in a context that allows for identification failure: (i) testing hypotheses involving the unknown parameter vector θ_0 (ii) testing the overidentifying restrictions assumption $Eg_n(z_i, \theta_0) = 0$ for some $\theta_0 \in \Theta$ in (2.1), when the model is overidentified, that is when $k > p$. More precisely, the testing problems are

$$\text{(i) } H_0 \quad : \quad R\theta_0 = q \text{ versus } H_1 : R\theta_0 \neq q, \tag{2.4}$$

$$\text{(ii) } H_0 \quad : \quad \exists \theta \in \Theta, Eg_i(\theta) = 0 \text{ versus } H_1 : \forall \theta \in \Theta, Eg_i(\theta) \neq 0, \tag{2.5}$$

where in (2.4), $R \in \mathbb{R}^{r \times p}$ for a $1 \leq r \leq p$ is a matrix of maximal rank r and $q \in \mathbb{R}^r$ is an arbitrary vector.[4] For testing problem (ii) to make sense, one has to impose a stationarity assumption on the distribution of z_i, which we do below.

Problem (i) with $r < p$ contains as a particular subcase simple subvector tests in which case the rows of R are a subset of the rows of I_p. Subvector testing in the context of weak identification has attracted a lot of attention in the recent literature, see, for example, Kleibergen (2004, 2005), Dufour and Taamouti (2005, 2007), Guggenberger and Smith (2005), and Zivot et al. (2006). Note also that we allow for null hypotheses in (i) that, in the case of the linear model (2.2), may involve both the unknown parameters of the exogenous and endogenous variables. Many test statistics in the literature are designed for the linear model where the included exogenous variables have been projected out in a first step, thereby ruling out a test of a hypothesis that involves both parameters of the exogenous and endogenous variables, see for example Kleibergen's (2002, 2004) test.

2.3 Identification Failure

As is now widely documented, classical tests of the hypotheses in (2.4) and (2.5), such as the Wald, likelihood ratio, and J test (Hansen, 1982) can suffer from severe size distortion in situations where the model is not identified or "close to being not identified". In model (2.1) identification failure means that besides θ_0 there are other $\theta \in \Theta$ that satisfy the moment condition. The abstract meaning of weak identification is that there are other $\theta \in \Theta$ that satisfy the moment condition in the limit $n \to \infty$. The classical identification condition, the so called "rank condition of identification",

[4] While in this paper we only deal with two-sided alternatives, our approach can also be applied to one-sided alternatives of the form $H_1 : R\theta_0 < q$ or $H_1 : R\theta_0 > q$, if there is only one restriction under test, that is $r = 1$. Furthermore, using more complicated assumptions in the theorems below, our approach could even be adapted to nonlinear parameter hypotheses.

states that the matrix $(\partial E g_i/\partial\theta)(\theta_0)$ has full column rank p. In the linear model, violation of the rank condition immediately implies that the model is not identified.

Much of the literature on weak identification has focused on the particular case where the parameter vector θ_0 has a decomposition $\theta_0 = (\theta_{01}, \theta_{02})$ into some weakly, θ_{01}, and some strongly identified components, θ_{02}. Namely, the definition of weak identification introduced in Stock and Wright (2000) for nonlinear models focuses on this case[5]. Define

$$\widehat{g}(\theta) := n^{-1} \sum_{i=1}^{n} g_i(\theta).$$

As discussed in Appendix, Part (A), Assumption C, applied to the linear model, implies that $(\partial E\widehat{g}/\partial\theta') = (0, M)$, where M is a matrix of maximal rank. On the other hand, Phillips (1989) and Choi and Phillips (1992) allow for a linear model with general failure of the rank condition in what they call "partial identification". In their model, $(\partial E\widehat{g}/\partial\theta')$ can be of non-maximal rank without being of the particular form $(0, M)$. We now introduce a general version of identification failure in nonlinear models that brings together this partially identified and Stock and Wright's (2000) weakly identified model. We show in the next section that the subsampling tests are robust against this general version of identification failure. A more detailed discussion of Assumption ID is relegated to Appendix, Part (A).

Assumption ID: There exist a coordinate change[6] $T \in O(p)$ such that $T(\overline{\Theta}) = \Theta$, where $\overline{\Theta}$ is a compact product set $\overline{\Theta} = \overline{\Theta}_1 \times \overline{\Theta}_2 \subset \mathbb{R}^{p_1+p_2} = \mathbb{R}^p$, and functions $\overline{m}_{1n}$, $\overline{m}_1 : \overline{\Theta} \to \mathbb{R}^k$, and $\overline{m}_2 : \overline{\Theta}_2 \to \mathbb{R}^k$ such that for

$$\begin{aligned} \overline{\theta} :&= (\overline{\theta}_1, \overline{\theta}_2) := T^{-1}(\theta) \text{ and } \overline{\theta}_0 := (\overline{\theta}_{01}, \overline{\theta}_{02}) := T^{-1}(\theta_0) \\ \overline{\widehat{g}}(\cdot) :&= \widehat{g}(T(\cdot)) : \overline{\Theta} \to \mathbb{R}^k \end{aligned}$$

(i) $\overline{m}_1 \in C^0(\overline{\Theta}, \mathbb{R}^k)$, $\overline{m}_2 \in C^0(\overline{\Theta}_2, \mathbb{R}^k) \cap C^1(\mathcal{N}, \mathbb{R}^k)$ for a neighborhood $\mathcal{N}$ of $\overline{\theta}_{02}$,

(ii) $E\overline{\widehat{g}}(\overline{\theta}) = n^{-1/2}\overline{m}_{1n}(\overline{\theta}) + \overline{m}_2(\overline{\theta}_2)$, $\overline{m}_{1n}(\overline{\theta}) \to \overline{m}_1(\overline{\theta})$ uniformly on $\overline{\Theta}$,

[5] **Assumption C**, Stock and Wright (2000, p. 1061)**:** Decompose $\theta = (\theta_1', \theta_2')'$, $\theta_0 = (\theta_{01}', \theta_{02}')'$ and $\Theta = \Theta_1 \times \Theta_2$. (i) $E\widehat{g}(\theta) = n^{-1/2} m_{1n}(\theta) + m_2(\theta_2)$, where m_{1n}, $m_1 \in C^0(\Theta, \mathbb{R}^k)$, and $m_2 \in C^0(\Theta_2, \mathbb{R}^k)$, such that $m_{1n}(\theta) \to m_1(\theta)$ uniformly on Θ, $m_1(\theta_0) = 0$ and $m_2(\theta_2) = 0$ if and only if $\theta_2 = \theta_{02}$. (ii) $m_2 \in C^1(\mathcal{N}, \mathbb{R}^k)$ for a neighborhood $\mathcal{N} \subset \Theta_2$ of θ_{02} and $(\partial m_2/\partial\theta_2')(\theta_{02})$ has full column rank.

In the linear model with no included exogenous variables, Assumption C boils down to a decomposition for Π into $\Pi_n = (n^{-1/2}\Pi_A, \Pi_B)$, where Π_A and Π_B are fixed matrices with p_1 and p_2 columns and Π_B has full column rank, see Stock and Wright (2000, Section 3).

[6] For notational convenience we denote by T the linear map $T : \mathbb{R}^p \to \mathbb{R}^p$ *and* the uniquely defined matrix in $\mathbb{R}^{p\times p}$ that defines this map. Assumption ID could be generalized to allow for possibly nonlinear coordinate changes T.

(iii) $\overline{m}_1(\overline{\theta}_0) = 0$, $\overline{m}_2(\overline{\theta}_2) = 0$ if and only if $\overline{\theta}_2 = \overline{\theta}_{02}$, and $\overline{M}_2(\overline{\theta}_{02})$ has full column rank, where $\overline{M}_2(\overline{\theta}_2) := (\partial \overline{m}_2 / \partial \overline{\theta}_2')(\overline{\theta}_2) \in \mathbb{R}^{k \times p_2}$.

Assumption ID contains as a subcase the case of a fully identified model ($T \equiv id$ and $p_1 = 0$) and the case of a totally unidentified model ($T \equiv id$, $p_1 = p$, and $\overline{m}_{1n} \equiv 0$). T is a change of the coordinate system such that in the new coordinate system the identified components of the parameter vector θ_0 are singled out. ID essentially boils down to Assumption C in Stock and Wright (2000) if we set $T \equiv id$. If $T \equiv id$ then[7], by ID (ii)–(iii), the first components θ_{01} of $\theta_0 = (\theta_{01}, \theta_{02})$ are only weakly identified. Clearly, no information on θ_{01} can be gained from the term m_2. Therefore, all the identifying information on θ_{01} from the condition $E\widehat{g}(\theta) = 0$ has to come from the term $n^{-1/2} m_{1n}(\theta)$. But this term vanishes with increasing sample size.

ID is more general than Assumption C in Stock and Wright (2000) because unlike C it comprises the partially identified model of Phillips (1989). It is more general than the latter because it allows for nonlinear moment conditions and weak identification. For every finite sample size n, the model may be fully identified through the term $n^{-1/2}\overline{m}_{1n}(\overline{\theta})$. But the information contained in $n^{-1/2}\overline{m}_{1n}(\overline{\theta})$ fades away with n going to infinity leading to a partially identified model asymptotically.

3 Subsampling Tests Under Weak Identification

The main reason for the size distortion of classical tests (Wald, likelihood ratio, J test) under identification failure is that parameter estimates of θ_0 have a non-normal asymptotic distribution when the classical identification assumption is close to being violated which implies that the tests are no longer asymptotically χ^2 under weak identification, see Theorems 3.2(ii) and 3.3 below. Subsampling can cure the problem of size distortion: instead of critical values based on asymptotic theory, data-dependent critical values based on the subsampling technique are employed. The test statistic under consideration is evaluated on all (overlapping) blocks of the observed data sequence, where the common block size is small compared to the sample size. The critical value is then obtained as an empirical quantile of the resulting subsample test statistics. In this section we describe in more detail how to use subsampling to construct tests that have exact (asymptotic) rejection probabilities under the null hypothesis, both under full identification and identification failure. For a general

[7] Whenever $T = id$, the new coordinates are the same as the original ones and therefore, throughout the paper, we leave out the bars in the notation in this case.

reference on subsampling see Politis et al. (1999). Our approach is to present a high level theorem and then verify/illustrate it in the particular settings we are interested in.

One observes a stretch of vector-valued data $z_1, \ldots, z_n$. Denote the unknown probability mechanism generating the data by P. It is assumed that P belongs to a certain class of mechanisms $\mathbf{P}$. The null hypothesis H_0 asserts $P \in \mathbf{P_0}$ and the alternative hypothesis H_1 asserts $P \in \mathbf{P_1}$, where $\mathbf{P_0}, \mathbf{P_1} \subset \mathbf{P}$, $\mathbf{P_0} \cap \mathbf{P_1} = \emptyset$, and $\mathbf{P_0} \cup \mathbf{P_1} = \mathbf{P}$. The goal is to construct a test with exact asymptotic rejection probability under the null hypothesis based on a given test statistic

$$D_n = D_n(z_1, \ldots, z_n).$$

Let $C_n(P)$ denote the sampling distribution of D_n under P, that is,

$$C_n(x, P) := Prob_P\{D_n(z_1, \ldots, z_n) \leq x\}.$$

It will be assumed that under the null hypothesis $C_n(P)$ converges in distribution to a continuous limit law $C(P)$. The $1 - \alpha$ quantile of this limit law is denoted by $c(1 - \alpha, P)$ and defined as

$$c(1 - \alpha, P) := \inf\{x : C(x, P) \geq 1 - \alpha\}.$$

To describe the subsampling test construction, denote by $Q_1, \ldots, Q_N$ the $N := n - b + 1$ blocks of size b of the observed data stretch $\{z_1, \ldots, z_n\}$; that is, $Q_a = \{z_a, \ldots, z_{a+b-1}\}$ for $a = 1, \ldots, N$. The model parameter b is called the block size. We will discuss its choice in Section 4.

Let $D_{b,a}$ be equal to the statistic D_b evaluated at the block Q_a. The sampling distribution of D_n is then approximated by[8]

$$\widehat{C}_{n,b}(x) := N^{-1} \sum_{a=1}^{N} 1\{D_{b,a} \leq x\}.$$

The critical value for the test is obtained as the $1 - \alpha$ quantile of $\widehat{C}_{n,b}$, that is

$$\widehat{c}_{n,b}(1 - \alpha) := \inf\{x : \widehat{C}_{n,b}(x) \geq 1 - \alpha\},$$

and the test arrives at the following decision:

$$\text{Reject } H_0 \text{ at nominal level } \alpha \text{ if and only if } D_n > \widehat{c}_{n,b}(1 - \alpha). \tag{3.6}$$

[8] In the special case of i.i.d. data, one could theoretically use all $\binom{n}{b}$ blocks of size b rather than only the N blocks used in the general time series context. Computationally however, it is generally not feasible to use all $\binom{n}{b}$ blocks.

If our only concern was to construct a test with correct null rejection probability, it could be achieved trivially: generate a uniform (0,1) variable and reject the null hypothesis if the outcome is smaller than α. But, obviously, we also want to achieve power when the model is identified. To formally establish power, we make the further assumption that the test statistic can be written as

$$D_n(z_1, \ldots, z_n) = n^{\beta} d_n(z_1, \ldots, z_n) \text{ for some } \beta > 0, \tag{3.7}$$

where

$$d_n(z_1, \ldots, z_n) \to_p d(P) \text{ satisfying } \begin{cases} d(P) = 0 \text{ if } P \in \mathbf{P_0} \\ d(P) > 0 \text{ if } P \in \mathbf{P_1} \end{cases}. \tag{3.8}$$

The following theorem gives the consistency of the procedure under the null, under a fixed alternative, and under a sequence of contiguous alternatives.[9]

Theorem 3.1 *Assume the sequence $\{z_i\}$ is strictly stationary and strongly mixing*[10] *and that the block size satisfies $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.*

(i) Assume that for $P \in \mathbf{P_0}$, $C_n(P)$ converges weakly to a continuous limit law $C(P)$ whose cumulative distribution function is $C(\cdot, P)$ and whose $1-\alpha$ quantile is $c(1-\alpha, P)$. Then, if $P \in \mathbf{P_0}$,

$$\widehat{c}_{n,b}(1-\alpha) \to_p c(1-\alpha, P)$$

and

$$Prob_P\{D_n > \widehat{c}_{n,b}(1-\alpha)\} \to \alpha \text{ as } n \to \infty.$$

(ii) If (3.7) and (3.8) hold and $P \in \mathbf{P_1}$, then

$$Prob_P\{D_n > \widehat{c}_{n,b}(1-\alpha)\} \to 1 \text{ as } n \to \infty.$$

(iii) Suppose P_n is a sequence of alternatives such that, for some $P \in \mathbf{P_0}$, $\{P_n^{[n]}\}$ is contiguous to $\{P^{[n]}\}$. Here, $P_n^{[n]}$ denotes the law of the finite segment $\{z_1, \ldots, z_n\}$ when the law of the infinite sequence $\{\ldots, z_{-1}, z_0, z_1, \ldots\}$ is given by P_n. The meaning of $\{P^{[n]}\}$ is analogous. Then,

$$\hat{c}_{n,b}(1-\alpha) \to c(1-\alpha, P) \text{ in } P_n^{[n]}\text{-probability.}$$

Hence, if D_n converges in distribution to D under P_n, then

$$Prob_{P_n^{[n]}}\{D_n > \hat{c}_{n,b}(1-\alpha)\} \to Prob\{D > c(1-\alpha, P)\}.$$

[9] In Appendix, Part (B) we provide some background information on *contiguity.*

[10] Alternatively, *strongly mixing* is sometimes called *α-mixing*, see Politis et al. (1999, p. 315) for a definition.

The interpretation of part (iii) of Theorem 3.1 is the following. Suppose, instead of using the subsampling construction, one could use the "oracle" test that rejects when $D_n > c_n(1-\alpha, P)$, where $c_n(1-\alpha, P)$ is the exact $1-\alpha$ quantile of the true sampling distribution $C_n(\cdot, P)$, where $P \in \mathbf{P_0}$. Of course, this test is not available in general because P is unknown and so is $c_n(1-\alpha, P)$. Then, the limiting power of the subsampling test against a sequence of contiguous alternatives $\{P_n\}$ to P with $P \in \mathbf{P_0}$ is the same as the limiting power of this fictitious "oracle" test against the same sequence of alternatives. Hence, to the order considered, there is no loss in efficiency in terms of power.

3.1 Classical Test Statistics

In the next subsections we introduce subsampling based testing procedures for the testing problems (2.4) and (2.5) that, unlike classical Wald, LR, and J tests, have exact (asymptotic) rejection probabilities under the null, independent of possible identification failure. The Anderson and Rubin (1949) statistic, recently reinvestigated by Dufour and Taamouti (2007), has an exact F-distribution in the linear model under normality for a simple hypothesis $H_0 : \theta_0 = q$ in (2.4) and therefore, under normality, leads to exact finite-sample sizes independent of identification failure. However, for tests of more general hypotheses, the (projected) Anderson and Rubin test is only conservative, even asymptotically. Other recent tests, for example Kleibergen's (2004, 2005) test, are not available for tests of general linear hypotheses. They can be generalized however to tests of simple subvector hypotheses with exact asymptotic null rejection probabilities but require the additional assumption that the parameters not under test are strongly identified. In contrast, our testing approach, based on subsampling classical statistics, like for example the Wald, LR, and J statistic, has exact (asymptotic) null rejection probabilities without further assumptions and is applicable to general linear hypotheses and overidentifying restrictions, respectively. In this subsection, we introduce the test statistics, focusing on the J and the Wald statistic[11].

As in Stock and Wright (2000), we focus on a GMM setup. Let

$$S_n(\theta) := n||A_n(e_n(\theta))^{1/2}\widehat{g}(\theta)||^2$$

[11] A similar analysis can be done for the LR test. We focus on the Wald statistic here because it does not involve the restricted estimator of θ_0 under the null hypothesis which simplifies the exposition. To test overidentifying restrictions, other test statistics besides the J test could be considered. See, for example, Imbens (1997), Kitamura and Stutzer (1997) or Imbens et al. (1998) who investigate several Lagrange multiplier and criterion function tests based on generalized empirical likelihood methods.

be the GMM criterion function that is pinned down by some data-dependent weighting matrix $A_n(e_n(\theta))^{1/2} \in \mathbb{R}^{k\times k}$ for a (possibly stochastic) function $e_n(\cdot) : \Theta \to \Theta$. More precisely, we allow for three different cases, namely one-step, two-step, and continuous updating (CU) GMM, see Hansen et al. (1996) for the latter. For one-step GMM $A_n(e_n(\theta))$ is typically chosen to be I_k or some other fixed positive definite nonstochastic matrix. Furthermore,

$$e_n(\theta) := \begin{cases} e_n \text{ for two-step GMM} \\ \theta \text{ for CU GMM,} \end{cases} \tag{3.9}$$

for some preliminary estimator e_n of θ_0. Therefore, for two-step GMM, $e_n(\cdot)$ does not depend on θ and for CU, $e_n(\cdot)$ is the nonstochastic identity map id.

Define the GMM estimator as a sequence of random variables $\widehat{\theta}_n$ satisfying

$$\widehat{\theta}_n \in \Theta \text{ and } S_n(\widehat{\theta}_n) \leq \arg\inf_{\theta\in\Theta} S_n(\theta) + o_p(1). \tag{3.10}$$

We often write $\widehat{\theta}$ for $\widehat{\theta}_n$. Let

$$\begin{aligned} \Psi_n(\theta) &:= n^{1/2}(\widehat{g}(\theta) - E\widehat{g}(\theta)), \\ \Omega(\theta,\theta^+) &:= \lim_{n\to\infty} E\Psi_n(\theta)\Psi_n(\theta^+)' \text{ and } \Omega(\theta) := \Omega(\theta,\theta) \in \mathbb{R}^{k\times k}. \end{aligned}$$

Also, from now on, a bar denotes expressions in new coordinates, see Assumption ID. For example, we write, $\overline{\Psi}_n(\cdot) := \Psi_n(T(\cdot))$, $\overline{\Psi}(\cdot) := \Psi(T(\cdot))$, $\overline{\Omega}(\cdot,\cdot) := \Omega(T(\cdot),T(\cdot))$, $\overline{A}(\cdot) := A(T(\cdot))$, and $\overline{A}_n(\cdot) := A_n(T(\cdot))$ for functions and $\overline{e_n(\theta)} := T^{-1}(e_n(\theta))$ for vectors, and similarly for other expressions. Note that by writing functions in new variables, for example, $\overline{\Psi}(\overline{\theta})$ instead of $\Psi(\theta)$, we do not change the value of the function, that means $\overline{\Psi}(\overline{\theta}) = \Psi(\theta)$; what we achieve by using the new coordinates is to single out identified from unidentified components in the parameter vector $\overline{\theta}_0$.

For testing problem (2.4) we now define the classical Wald statistic W_n based on the GMM estimator and for problem (2.5) we define the J statistic J_n (Hansen (1982)) as the GMM criterion function evaluated at the GMM estimator. More precisely,

$$W_n := n(R\widehat{\theta}_n - q)'[R\widehat{B}_n^{-1}\widehat{\Omega}_n\widehat{B}_n^{-1}R']^{-1}(R\widehat{\theta}_n - q), \tag{3.11}$$

$$J_n := S_n(\widehat{\theta}_n), \tag{3.12}$$

where

$$\begin{aligned} \widehat{G}_n &:= n^{-1}\textstyle\sum_{i=1}^n \dfrac{\partial g_i}{\partial\theta'}(\widehat{\theta}_n) \in \mathbb{R}^{k\times p}, \\ \widehat{B}_n &:= \widehat{G}_n' A_n(e_n(\theta))\widehat{G}_n \in \mathbb{R}^{p\times p}, \\ \widehat{\Omega}_n &:= \widehat{G}_n' A_n(e_n(\theta))K_n(\widehat{\theta})A_n(e_n(\theta))\widehat{G}_n \in \mathbb{R}^{p\times p}, \end{aligned} \tag{3.13}$$

and $K_n(\cdot)$ is a $\mathbb{R}^{k\times k}$-valued (stochastic) function on Θ and $K_n(\widehat{\theta}_n)$ an estimator of the long-run covariance matrix $\Omega(\theta_0)$. For example, in an *i.i.d.* model, a natural choice would be $K_n(\theta) := n^{-1}\sum_{i=1}^n g_i(\theta)g_i(\theta)' \in \mathbb{R}^{k\times k}$, whereas in a time series model one would typically use some version of a heteroskedasticity and autocorrelation consistent (HAC) estimator, see Andrews (1991).

From now on, we distinguish the following two polar-opposite cases of identification.

- **Full identification:** Assume ID with $p_1 = 0$ and $T = id$.
- **Identification failure:** Assume ID with $p_1 > 0$ and $\overline{m}_{1n} \equiv 0$.

In the next subsections, we show that the classical Wald test of parameter hypotheses and the J test of overidentifying restrictions are generally size distorted under Assumption ID. On the other hand, we establish that the subsampling versions of the Wald test and the J test are *consistent* under full identification and have (asymptotically) *exact rejection probabilities under the null hypothesis*, both under full identification and identification failure. Extrapolating from these polar-opposite cases of identification and non-identification, we interpret this as evidence that the tests based on subsampling continue to have the latter property in the intermediate case of weak identification (where $\overline{m}_{1n} \neq 0$) as defined in Assumption ID.

3.2 Testing Overidentifying Restrictions

In this subsection, we first derive the asymptotic distribution of the classical J statistic under Assumption ID and conclude that the J test is potentially size distorted under identification failure. We then use the asymptotic result to show that the subsampling version of the J test has exact (asymptotic) rejection probability under the null.

To derive the asymptotic distribution of the J statistic under Assumption ID we first need the one of the estimator $\widehat{\theta}_n$. We essentially make the same "high level" assumptions as Stock and Wright (2000, see Assumptions B and D).

Assumption PE (parameter estimates):[12] Assume ID. Suppose there exists a neighborhood $\overline{U}_2 \subset \overline{\Theta}_2$ of $\overline{\theta}_{02}$, such that for $\overline{\Theta}_{12} := \overline{\Theta}_1 \times \overline{U}_2$

[12] Weak convergence here is defined with respect to the sup-norm on function spaces and the Euclidean norm on $\mathbb{R}^k$. Also, note that Assumption PE could alternatively be stated in original coordinates.

(i) $\overline{\Psi}_n \Rightarrow \overline{\Psi}$, where $\overline{\Psi}$ is a Gaussian stochastic process on $\overline{\Theta}_{12}$ with mean zero, covariance function $E\overline{\Psi}(\overline{\theta})\overline{\Psi}(\overline{\theta}^+)' = \overline{\Omega}(\overline{\theta}, \overline{\theta}^+)$ for $\overline{\theta}, \overline{\theta}^+ \in \overline{\Theta}_{12}$, sample paths that are continuous w.p.1, and $\sup_{\overline{\theta}\in\overline{\Theta}} ||n^{-1/2}\overline{\Psi}_n(\overline{\theta})|| \to_p 0$;

(ii) $\sup_{\overline{\theta}\in\overline{\Theta}} ||\overline{A}_n(\overline{\theta}) - \overline{A}(\overline{\theta})|| \to_p 0$, $\overline{A}(\cdot) \in C^0(\overline{\Theta}, \mathbb{R}^{k\times k})$, $\overline{A}(\overline{\theta}) > 0$, and $\overline{A}_n(\overline{\theta}) > 0$ for all $\overline{\theta} \in \overline{\Theta}$ w.p.1;

(iii) $\overline{e_n(\cdot)} \Rightarrow \overline{e(\cdot)}$ jointly with the statement in (i).[13]

Assumption PE states that after a coordinate change the Assumptions B, D, and Assumptions made in Theorem 1 in Stock and Wright (2000) hold. Our assumption is slightly weaker because in PE(i) we do not require that convergence holds on the whole parameter space $\overline{\Theta}$ but only on $\overline{\Theta}_{12}$. For the J statistic we now have the following theorem.

Theorem 3.2 *Suppose Assumptions ID and PE hold. Let* $\overline{\widehat{\theta}} = (\overline{\widehat{\theta}}_1, \overline{\widehat{\theta}}_2) := T^{-1}(\widehat{\theta})$ *and assume that* $\overline{S}$ *in (6.21) in the Appendix satisfies the "unique minimum"*[14] *condition in (6.22). Then,*

(i) (*Asymptotic distribution of parameter estimates*)

$$(\overline{\widehat{\theta}}_1, n^{1/2}(\overline{\widehat{\theta}}_2 - \overline{\theta}_{02})) \to_d \overline{\theta}^* := (\overline{\theta}_1^*, \overline{\theta}_2^*),$$

where the nonstandard limit $\overline{\theta}^*$ *is defined in (6.23) and (6.24) and*

(ii) (*Asymptotic distribution of the J statistic*)

$$J_n \to_d J^* := \overline{S}(\overline{\theta}_1^*, \overline{\theta}_2^*, \overline{e(\theta_1^*, \theta_{02})}).$$

Part (i) shows that some components of the estimator in new coordinates, $\overline{\widehat{\theta}}_2$, are root-$n$ consistent for $\overline{\theta}_2$ yet are not asymptotically normally distributed due to the inconsistent estimation of the remaining components $\overline{\theta}_1$ by $\overline{\widehat{\theta}}_1$. Under full identification ($T \equiv id$ and $p_1 = 0$) and assuming that $e_n \to_p \theta_0$ for the two-step GMM case, equation (6.24) shows that $n^{1/2}(\widehat{\theta} - \theta_0) \to_d \theta^*$ which is distributed

[13] By definition

$$\overline{e_n(\theta)} = \begin{cases} \overline{e_n} \text{ for two-step GMM} \\ \overline{\theta} \text{ for CU GMM} \end{cases}$$

and therefore, for two-step GMM PE(iv) means $\overline{e}_n \to_d \overline{e}$ for some random variable $\overline{e}$ while for CU PE(iv) boils down to the trivially satisfied condition $\overline{e}_n(\cdot) = \overline{e}(\cdot) := id(\cdot)$.

[14] The "unique minimum" condition is used in the proof when we apply Lemma 3.2.1 in van der Vaart and Wellner (1996), as in Stock and Wright's (2000) proof of Theorem 1(ii).

as $N(0,(M_2'AM_2)^{-1}(M_2'A\Omega(\theta_0)AM_2)(M_2'AM_2)^{-1})$, where $M_2 := M_2(\theta_0)$ and $A := A(e(\theta_0))$.

Choi and Phillips (1992) and Stock and Wright (2000) (Theorem 1 (ii)) derive the limit distribution of the parameter estimates in the linear model under partial identification and in the nonlinear model under ID with $T \equiv id$, respectively.

Part (ii) corresponds to Corollary 4 (i) in Stock and Wright (2000), where the asymptotic distribution of the J statistic is derived under their Assumption C. Part (ii) shows that in general the J statistic has a nonstandard asymptotic distribution while under full identification and $A = \Omega(\theta_0)^{-1}$, we obtain the well known result that $J_n \to_d \chi^2(k-p)$. Therefore, generally, under identification failure, the J test does not have correct rejection probability under the null if inference is based on χ^2 critical values. As we show now, subsampling overcomes that problem. To formally establish power, we have to make the following assumption under the alternative H_1.

Assumption MM (misspecified model):

(i) the parameter space Θ is compact;

(ii) $Eg_i(\cdot) \in C^0(\Theta, \mathbb{R}^k)$ and $\sup_{\theta\in\Theta} ||\widehat{g}(\theta) - Eg_i(\theta)|| \to_p 0$;

(iii) there exists a nonstochastic function $A(\cdot) \in C^0(\Theta, \mathbb{R}^{k\times k})$ such that $\sup_{\theta\in\Theta} ||A_n(\theta) - A(\theta)|| \to_p 0$ and $A(\theta) > 0$ for $\theta \in \Theta$ w.p.1;

(iv) for $e_n(\theta)$ defined in (3.9) we have $e_n(\theta) \to_p e(\theta)$, where $e(\theta)$ is nonstochastic;[15]

(v) $\widetilde{\theta} := \arg\min_{\theta\in\Theta} ||A(e(\theta))^{1/2} Eg_i(\theta)||$ exists and is unique.

Given the previous theorem, the next statement is a corollary of Theorem 3.1. The test is $H_0 : \exists\theta \in \Theta, Eg_i(\theta) = 0$ versus $H_1 : \forall\theta \in \Theta, Eg_i(\theta) \neq 0$.

Corollary 3.1 *Suppose the sequence $\{z_i\}$ is both strictly stationary and strongly mixing. Assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Let $D_n = J_n$ of (3.12) and define the subsampling test by (3.6).*

(i) Under H_0 assume PE and that J^* in Theorem 3.2 is continuously distributed. Then the rejection probability converges to α as $n \to \infty$ both under full identification and identification failure.

[15] In other words, for 2-step GMM we assume that the preliminary estimator e_n converges in probability to an element $e \in \Theta$.

(ii) Under H_1 and Assumption MM the rejection probability converges to 1 as $n \to \infty$.

The corollary shows that the subsampling test of overidentifying restrictions is consistent against model misspecification and has asymptotically exact rejection probabilities under the null hypothesis both under full identification and identification failure. The test therefore improves on the classical J test or the tests of overidentifying restrictions suggested in Imbens et al. (1998) that are all size distorted under identification failure.

3.3 Testing Parameter Hypotheses

In this subsection, we derive the asymptotic distribution of the Wald statistic under Assumption ID and conclude that the Wald test is size distorted under identification failure. We then use this asymptotic result to show that the subsampling version of the Wald test has exact (asymptotic) rejection probability under the null.

To derive the asymptotic distribution of the Wald statistic we need the following additional assumption besides Assumption PE. If they exist, denote by $(\partial\overline{g_i}/\partial\overline{\theta}_1')(\overline{\theta}) \in \mathbb{R}^{k\times p_1}$ and $(\partial\overline{g_i}/\partial\overline{\theta}_2')(\overline{\theta}) \in \mathbb{R}^{k\times p_2}$ the partial derivatives of $\overline{g_i}$ with respect to the first p_1 and last p_2 components of $\overline{\theta}$, respectively, where we use the notation of Assumption ID. Define

$$N := diag(n_{jj}) \in \mathbb{R}^{p\times p}, \tag{3.14}$$

where $n_{jj} = n^{1/2}$ if $j \leq p_1$ and $n_{jj} = 1$ otherwise for $j = 1, ..., p$.

Assumption WS (Wald statistic)**:** Assume ID and suppose there exists a neighborhood $\overline{U}_2 \subset \overline{\Theta}_2$ of $\overline{\theta}_{02}$, such that for $\overline{\Theta}_{12} := \overline{\Theta}_1 \times \overline{U}_2$

(i)
$$[(n^{-1/2}\textstyle\sum_{i=1}^n vec\dfrac{\partial\overline{g_i}}{\partial\overline{\theta}_1'}(\overline{\theta}))', \overline{\Psi}_n(\overline{\theta})']' \Rightarrow \overline{\Phi}(\overline{\theta})$$

holds jointly with PE(iii), where $\overline{\Phi}$ is a $k(p_1+1)$-dimensional Gaussian stochastic process on $\overline{\Theta}_{12}$ with sample paths that are continuous w.p.1, a certain (possibly nonzero) mean function, and covariance function $\overline{\Delta}(\overline{\theta}, \overline{\theta}^+) := E\overline{\Phi}(\overline{\theta})\overline{\Phi}(\overline{\theta}^+)'$ for $\overline{\theta}, \overline{\theta}^+ \in \overline{\Theta}_{12}$;

(ii)
$$\sup_{\overline{\theta}=(\overline{\theta}_1,\overline{\theta}_2)\in\overline{\Theta}_{12}} ||n^{-1}\textstyle\sum_{i=1}^n \dfrac{\partial\overline{g_i}}{\partial\overline{\theta}_2'}(\overline{\theta}) - \overline{M}_2(\overline{\theta}_2)|| \to_p 0$$

and $\overline{M}_2(\overline{\theta}_2)$ has maximal column rank for all $\overline{\theta} \in \overline{\Theta}_{12}$;

(iii) by (i), (ii), and Theorem 3.2 $(n^{-1}\sum_{i=1}^{n}(\partial\overline{g_i}/\partial\overline{\theta}')(\overline{\widehat{\theta}}))N \in \mathbb{R}^{k\times p}$ converges in distribution to a random variable with realizations in $\mathbb{R}^{k\times p}$. Assume the realizations have full column rank a.s.;

(iv) there exists a nonstochastic function $\Lambda : T(\overline{\Theta}_{12}) \to \mathbb{R}^{k\times k}$ such that

$$\sup_{\theta\in T(\overline{\Theta}_{12})} ||K_n(\theta) - \Lambda(\theta)|| \to_p 0 \tag{3.15}$$

and $\Lambda(\theta)$ has full rank for all $\theta \in T(\overline{\Theta}_{12})$.

We now discuss Assumption WS. WS(i) generalizes PE(i) by including a portion of the first derivative matrix in the functional central limit theorem (FCLT). Joint CLTs of g_i and (portions of) its derivative matrix have also been assumed by Kleibergen (2005, Assumption 1) and Guggenberger and Smith (2005, Assumption M_θ(vii)). However, instead of a FCLT, these papers only require a joint CLT at θ_0. We require a FCLT because instead of evaluating our test statistic at a fixed hypothesized parameter vector, our test statistic is evaluated at an estimated parameter vector. As shown in Theorem 3.2, this estimator is in general not consistent. Note that we do not have to subtract off the mean in the FCLT from the derivative component; under weak technical conditions that allow the interchange of differentiation and integration, ID(ii) implies that $n^{-1/2}\sum_{i=1}^{n} E(\partial\overline{g_i}/\partial\overline{\theta}_1')(\overline{\theta}) \to \overline{M}_1(\overline{\theta})$, where $\overline{M}_1(\overline{\theta}) \in \mathbb{R}^{k\times p_1}$ denotes the derivative of $\overline{m}_1(\overline{\theta})$ with respect to the first p_1 coordinates. Then the mean function of $\overline{\Phi}(\overline{\theta})$ equals $[(vec\overline{M}_1(\overline{\theta}))', 0']'$.

Assumptions WS(ii) and (iv) state uniform law of large numbers. In WS(ii), the series converges to $\overline{M}_2(\overline{\theta}_2)$ which assumes that one can interchange the order of integration and differentiation. We make this assumption to economize on notation, but everything that follows would go through if convergence was instead to a different full rank non-stochastic function, $\overline{G}_2(\overline{\theta}_2)$ say, instead of $\overline{M}_2(\overline{\theta}_2)$. On the other hand, note that in (iv) we do *not* require that $\Lambda(\theta_0)$ is the long-run covariance matrix $\Omega(\theta_0)$ of $g_i(\theta_0)$. Our theory goes through in the general time series context, even if a simple sample average $K_n(\theta) = n^{-1}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'$ is used in a time series context as long as $K_n(\theta)$ converges uniformly to a full rank nonstochastic matrix.

Example 2.1 (cont.): In the linear model, the upper-left kp_1-dimensional square submatrix of $\overline{\Delta}(\cdot,\cdot)$ and $\overline{M}_2(\cdot)$ from Assumptions WS(i) and (ii) do not depend on the argument $\overline{\theta}$. This implies an easy sufficient condition for WS(iii) as stated in the next Lemma. Furthermore, WS(i)–(ii) hold automatically.

Lemma 3.1 *In the linear model of Example 2.1. assume i.i.d. data,* $E(Z_i', X_i')'(u_i, V_i') = 0$, *and* $E||(Z_i', X_i')'(Z_i', X_i', u_i, V_i')||^2 < \infty$ *and set* $K_n(\theta) := n^{-1}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'$.

Then, under Assumption ID, it follows that WS(i)–(ii) hold. If, in addition, the upper-left kp_1-dimensional square submatrix of $\overline{\Delta}$ is positive definite, then WS(iii) holds. Finally, WS(iv) holds if $\lim_{n\to\infty} Eg_i(\theta)g_i(\theta)'$ is positive definite for all $\theta \in \Theta$.

Besides mild additional assumptions, the lemma states the main assumptions that are needed for the subsampling approach to work, when applied to the Wald test and parameter hypotheses in the linear model; see Corollary 3.2 below. We can now formulate the following theorem that derives the asymptotic distribution of the Wald statistic under ID.

Theorem 3.3 (*Asymptotic distribution of the Wald statistic*) *Assume the assumptions of Theorem 3.2 and Assumption WS hold. Then, under the null hypothesis $R\theta_0 = q$, we have*

$$W_n \to_d W^*,$$

where the limit W^* is defined in (6.25) in the Appendix.

Theorem 3.3 generalizes an analogous result about the Wald statistic in Staiger and Stock (1997, Theorem 1, (c)) from the linear model with only weakly identified parameters to the GMM setup under ID. Phillips (1989) and Choi and Phillips (1992) derive the asymptotic distribution of the Wald statistic that tests hypotheses on the coefficients of either the exogenous or endogenous regressors in the linear model under partial identification. For example, they show that in the totally unidentified case, the Wald statistic converges to a random variable that can be written as a continuous function of random variables that are distributed as noncentral Wishart and multivariate t (Phillips (1989, Theorem 2.8.)).

Theorem 3.3 shows that the Wald statistic has a nonstandard asymptotic distribution under identification failure. On the other hand, under full identification and assuming that $\Lambda(\theta_0) = \Omega(\theta_0)$, the proof of the theorem contains the well known result that the Wald statistic is asymptotically distributed as $\chi^2(r)$. A test based on the Wald statistic using critical χ^2-values is likely to be size distorted when identification fails. On the other hand, as we will show now, the subsampling test has rejection probabilities under the null that are asymptotically exact even under identification failure. What is crucial (and sufficient under very mild additional assumptions) for the subsampling approach to have exact (asymptotic) rejection probabilities under the null, is that the test statistics we apply subsampling to, converge to an asymptotic distribution independent of the particular assumption in ID; see part (i) of Corollaries 3.2 and 3.1.

Given the previous theorem the following statement is a corollary of Theorem 3.1. The test is $H_0 : R\theta_0 = q$ versus the two-sided alternative $H_1 : R\theta_0 \neq q$.

Corollary 3.2 *Assume PE, WS, and that W^* in Theorem 3.3 is continuously distributed. Suppose the sequence $\{z_i\}$ is both strictly stationary and strongly mixing. Assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Let $D_n = W_n$ of (3.11) and define the subsampling test by (3.6). Then*

(i) Under H_0 the rejection probability converges to α as $n \to \infty$ both under full identification and identification failure.

(ii) Under H_1 the rejection probability converges to 1 as $n \to \infty$ under full identification.

(iii) Consider a sequence of contiguous alternatives under full identification. Then the limiting rejection probability of the subsampling test (3.6) is equal to that of the Wald test.

The corollary shows that the subsampling test of parameter hypotheses is consistent against fixed alternatives under full identification and has asymptotically exact rejection probabilities under the null hypothesis both under full identification and identification failure. Furthermore, it has the same limiting power against contiguous alternatives under full identification as the original Wald test. As a special case for this last statement consider again Example 2.1. Assume a parametric distribution for z_i indexed by θ, $\{P_\theta : \theta \in \Theta\}$, that is differentiable in quadratic mean around a particular parameter θ_0 which satisfies $R\theta_0 = q$ (see Appendix, Part (B)). Denote by $\chi^2_{1-\alpha}(r)$ the $1-\alpha$ quantile of a χ^2 distribution with r degrees of freedom and let W be a random variable that follows a noncentral $\chi^2(r, \delta)$ distribution for some noncentrality parameter δ. Furthermore, assume the data is generated according to a Pitman drift $\theta_n = \theta_0 + h/\sqrt{n}$ for some $h \in \mathbb{R}^p$. Assuming various regularity conditions given in Newey and West (1987, Theorem 2), $\Lambda(\theta_0) = \Omega(\theta_0)$, and $e_n \to_p \theta_0$ in the two-step GMM case, the corresponding limiting power for both the classical Wald and the subsampling test is given by $P\{W > \chi^2_{1-\alpha}(r)\}$, where $\delta := h'R'[R\{M_2(\theta_0)\Omega(\theta_0)^{-1}M_2(\theta_0)'\}^{-1}R']^{-1}Rh$.

3.4 Limitation to Pointwise Asymptotics

It is important to point out that our results only cover *pointwise asymptotics*. We show, in general, that for any *fixed* $P \in \mathbf{P_0}$, the rejection probability of a subsampling test is no larger than the nominal significance level asymptotically.

We do no provide any results covering *uniform asymptotics*. This would require to show that the *supremum* of the rejection probabilities over all $P \in \mathbf{P_0}$ is no larger than the nominal significance level asymptotically.

The latter problem is more challenging and the reader is referred to Andrews and Guggenberger (2010a, 2010b) for corresponding results.

4 Choice of the Block Size

An application of the subsampling method requires a choice of the block size b. Unfortunately, the asymptotic requirements $b/n \to \infty$ and $b \to \infty$ as $n \to \infty$ offer little practical guidance. We propose to select b by a *calibration method*, an idea dating back to Loh (1987).

It is our goal to construct a test with nominal size α. However, this can only be achieved exactly as the sample size tends to infinity. The actual size in finite sample, denoted by λ, typically differs from α. The crux of the calibration method is to adjust the block size b in a manner such that the actual size λ will hopefully be close to the nominal size α. To this end consider the *calibration function* $h(b) = \lambda$. This function maps the block size onto the actual size of the test, considering the underlying probability mechanism and the sample size fixed. If $h(\cdot)$ were known, one could construct an "optimal" test by finding $\tilde{b}$ that minimizes $|h(b) - \alpha|$ and use $\tilde{b}$ as the block size; note that $|h(b) - \alpha| = 0$ may not always have a solution.

In principle, we could simulate $h(\cdot)$ if in return P were known by generating data of size n according to P and constructing subsampling hypothesis tests for H_0 for a number of different block sizes b. This process is then repeated many times and for a given b one estimates $h(b)$ as the fraction of tests that reject the null. The method we propose is identical except that P is replaced by an estimate $\hat{P}_n$ that is consistent for P, at least under the null. The choice of $\hat{P}_n$ should be made on a case-by-case analysis; further details are given below.

Algorithm 4.1 (Choice of the Block Size)

1. *Fix a set B of reasonable block sizes b, where $b_{low} := \min B$ and $b_{up} := \max B$.*
2. *From the original data, $z_1, \ldots, z_n$, generate L pseudo sequences $z^*_{l,1}, \ldots, z^*_{l,n}$, $l = 1, \ldots, L$ according to $\hat{P}_n$. For each sequence, $l = 1, \ldots, L$, and for each $b \in B$, construct a subsampling hypothesis test for H_0, $\phi_{l,b}$ say, in the way described in the beginning of Section 3. Note that the specific form of H_0 is allowed to depend upon $\hat{P}_n$ here. In particular, $\phi_{l,b} = 1$ if H_0 is rejected and $\phi_{l,b} = 0$ otherwise.*
3. *Define $\hat{h}(b) := L^{-1} \sum_{l=1}^{L} \phi_{l,b}$.*
4. *Calculate $\tilde{b} := \arg\min_{b \in B} |\hat{h}(b) - \alpha|$.*

We recommend to use $L \geq 1,000$ in practice. In step 2 of the algorithm it is noted that H_0 may depend upon $\hat{P}_n$. See subsection 4.1 for an example.

Remark 4.1 *Strictly speaking, Theorem 3.1 and, as consequence, Corollaries 3.1 and 3.2, assume an a priori determined sequence of block sizes b as $n \to \infty$. In practice, however, the choice of b will typically be data-dependent, such as given by Algorithm 4.1. As discussed in Politis et al. (1999, Section 3.6), such a data-dependent choice of block size does not affect the asymptotic validity of subsampling inference with strong mixing data as long as $b_{low} \to \infty$ and $b_{up}/n^{1/2} \to 0$ as $n \to \infty$.*

We now give some further details of the block size choice for the two main applications in the paper, namely, parameter testing and tests of overidentifying restrictions.

4.1 Choice of the Block Size for Testing Parameter Hypotheses

For simplicity, our proposal for $\hat{P}_n$ is to resample from the observed data $\{z_1, \ldots, z_n\}$ via the stationary bootstrap of Politis and Romano (1994). In the special case of *i.i.d.* data one should use the *i.i.d.* bootstrap of Efron (1979) instead.[16] The corresponding null hypothesis for use in Algorithm 4.1 then is $H_0 : R\theta_0 = R\hat{\theta}_n$. Since we resample from the observed data, the parameter θ corresponding to $\hat{P}_n$, denoted by $\theta(\hat{P}_n)$, is given by $\hat{\theta}_n$. But even if the null hypothesis is true, $R\hat{\theta}_n \neq q$ in general. This explains why one should use $R\hat{\theta}_n$ instead of q as the hypothesized value in step 2 of the algorithm.

Another possibility would be to generate pseudo data from a distribution $\hat{P}_{n,0}$ that satisfies the constraints of the null hypothesis, namely, $R\theta(\hat{P}_{n,0}) = q$, where $\theta(\hat{P}_{n,0})$ denotes the parameter vector θ that corresponds to the probability mechanism $\hat{P}_{n,0}$. In that case the null hypothesis for use in Algorithm 4.1 would be $H_0 : R\theta_0 = q$. However, this approach is more cumbersome and in some simulations that we tried in the context of Example 2.1, it did not work any better than resampling from the observed data as described above.

4.2 Choice of the Block Size for Testing Overidentifying Restrictions

Here the null hypothesis is not expressed in terms of the parameter vector θ_0. Therefore, we have to go through the effort of resampling from a distribution $\hat{P}_n$ that satisfies the constraints of H_0. The reason is that the simpler solution of resampling from the observed data in conjunction with adjusting the parameter vector for Algorithm 4.1 to $\hat{\theta}_n$ is not available.

[16] Efron's bootstrap is a special case of the stationary bootstrap, namely when the (expected) block length of the stationary bootstrap is set equal to one.

Unfortunately, the particular form of imposing H_0 onto $\hat{P}_n$ has to depend on the situation at hand. The general idea is to transform the observed data "as little as possible" to satisfy the constraints of H_0 in the empirical distribution of the transformed data and then to resample from the transformed data. We give here a specific description for Example 2.1. The observed data are (y, Y, X, Z). The null hypothesis states that $E(Z_i', X_i')'u_i = 0$. Let $\hat{\theta}_n = (\hat{\beta}_n', \hat{\gamma}_n')'$ be the 2SLS estimator of θ_0 and let $\hat{u} := y - Y\hat{\beta}_n - X\hat{\gamma}_n$ be the vector of corresponding residuals. By construction, $\hat{u}'X = 0$ (in case there are any included exogenous variables to begin with). On the other hand, $\hat{u}'Z \neq 0$ in general. So the empirical distribution of the observed data does not satisfy the constraints of H_0.

Therefore, we transform $\hat{u}$ in the least possible way to make it orthogonal to Z by projecting it onto the null space of $[XZ]$. The thus transformed residuals, in return, imply a transformed y vector. So let $\tilde{u} := (I - P_{[XZ]})\hat{u}$ and let $\tilde{y} := Y\hat{\beta}_n + X\hat{\gamma}_n + \tilde{u}$. The transformed data set from which we resample using Efron's bootstrap then is $(\tilde{y}, Y, X, Z)$. Since $(\tilde{y} - Y\hat{\beta}_n - X\gamma_n)'[XZ] = 0$, the constraints of H_0 are satisfied by the empirical distribution of the transformed data set.[17]

5 Monte Carlo Experiments

To assess the finite sample performance of the subsampling tests introduced above, we conduct a set of Monte Carlo experiments.

(I) In the first experiment we look at a simple full vector parameter hypothesis. The data generating process (DGP) is given by model (2.2) and (2.3), where

$$\begin{aligned} v_1 &= 1,\ v_2 = 0 \text{ (that is one endogenous, no exogenous variable)} \\ \beta_0 &= 0 \text{ (structural parameter value)} \\ Z &\sim N(0, I_j \otimes I_n), \text{ where } j = 1 \text{ or } 3, \\ n &= 100 \text{ (sample size), and} \\ (u_i, V_i)' &\sim i.i.d.\ N(0, \Sigma), \text{where } \Sigma = \begin{pmatrix} 1 & .25 \\ .25 & 1 \end{pmatrix}. \end{aligned}$$

The j-vector Π equals $(\pi, ..., \pi)$, where π equals $0, .01, .05, .1, .5$, or 1. Interest focuses on testing the scalar null hypothesis

$$H_0 : \beta_0 = 0 \text{ versus } H_1 : \beta_0 \neq 0.$$

[17] If there are no included exogenous variables in the model, the modifications are the obvious ones. Let $\hat{u} := y - Y\hat{\beta}_n$, $\tilde{u} := (I - P_Z)\hat{u}$, and $\tilde{y} := Y\hat{\beta}_n + \tilde{u}$. The transformed data set then is $(\tilde{y}, Y, Z)$.

We also explore the impact of conditional heteroskedasticity on the performance of the test statistics by replacing u_i by $\widetilde{u}_i := ||Z_i||u_i$. In total we are looking at 24 different DGPs (different j and π values and homo/heteroskedasticity). We compare the size and power performance of the following four statistics:

- The subsampling method (3.6) is applied to the standard homoskedastic version of the Wald statistic W. This approach is denoted Sub. Empirical null rejection probabilities are obtained via the data-dependent choice of block size of Algorithm 4.1.[18]
- The K test by Kleibergen (2005), based on a heteroskedasticity robust estimator of the covariance matrix.
- The empirical likelihood based test LM_{EL} by Guggenberger and Smith (2005).
- The conditional likelihood ratio test by Moreira (2003), denoted by LR_M.

See Guggenberger and Smith (2005, Section 5.2) for a precise definition of the latter three tests. The case of a simple full vector hypothesis test is not an application where we expect subsampling to have a comparative advantage over other tests robust to weak identification recently introduced in the literature, on the contrary. However, while subsampling is applicable to tests of general linear hypotheses, these other tests are not. This experiment is used to investigate the premium price (in terms of power loss) for the robustness of the subsampling approach, in a scenario, where we expect the performance of the test to be at its worst relative to these other statistics. For that reason we include the LR_M test: this test is size distorted under conditional heteroskedasticity but is known to be uniformly most powerful unbiased for two sided alternatives in the case $j = 1$, normal reduced form errors with known covariance matrix, and nonstochastic exogenous variables, see Andrews et al. (2006).

(II) The second experiment looks at a simple subvector hypothesis test, which is a scenario where we recommend application of the subsampling approach. The DGP is given by model (2.2) and (2.3) considered in Example 2.1 above and the

[18] In Algorithm 4.1, we use $B := \{4, 6, 8, 10, 15, 20, 25, 30, 35\}$ as the set of input block sizes, $L = 250$ as the number of repetitions, and Efron's (1979) i.i.d. bootstrap to resample the data. Even though we have i.i.d. data, we only use the $n - b + 1$ blocks of consecutive data rather than all the possible $\binom{n}{b}$ subsamples to approximate the sampling distribution of the Wald statistic.

When calculating power curves it is too computer intensive to implement Algorithm 4.1. Therefore, empirical power was obtained by using the fixed block size $b \in B$ that resulted in the empirical size closest to the empirical size obtained by the data-dependent block choice method. Also, while we recommend to use $L \geq 1,000$ for a practical application, this choice was not feasible for the computational demands of a large-scale simulation study.

parameter specifications are similar to the setup in Dufour and Taamouti (2007), *viz.* in Example 2.1 we choose

$$\begin{aligned} v_1 &= 2,\ v_2 = 1 \text{ (that is two endogenous, one exogenous variable)} \qquad (5.16)\\ \beta_0 &= (0,0)', \gamma_0 = 0,\ \Phi = (.1,.5) \text{ (parameter values)}\\ X &= 1_n, \text{ an } n\text{-column of ones, } Z \text{ is a } (n\times 2)\text{-matrix of } i.i.d.\ N(1,1) \text{ variables,}\\ n &= 100 \text{ (sample size), and} \end{aligned}$$

$$(u_i, V_i')' \sim i.i.d.\ N(0,\Sigma), \text{ where } \Sigma := \begin{pmatrix} 1 & .8 & .8 \\ .8 & 1 & .3 \\ .8 & .3 & 1 \end{pmatrix}. \qquad (5.17)$$

Our simulation study varies over different Π matrices thereby investigating the effects of weak identification or identification failure. More specifically, for $\pi_1 = 0, .01, .05, .1, .5, 1$ and $\pi_2 = 0, .01, .05, .1, .5, 1$ we take all 71 possible combinations of Π matrices defined as[19]

$$\bar{\Pi} = \begin{pmatrix} \pi_1 & 2\pi_2 \\ 2\pi_1 & \pi_2 \end{pmatrix} \text{ or } \tilde{\Pi} = \begin{pmatrix} 2\pi_1 & \pi_2 \\ \pi_1 & 2\pi_2 \end{pmatrix}. \qquad (5.18)$$

Interest focuses on testing the scalar null hypothesis

$$H_0 : \beta_{01} = 0 \text{ versus the alternative hypothesis } H_1 : \beta_{01} \neq 0.$$

We compare the size and power performance of the following four test statistics:

- The classical Wald statistic based on the two stage least squares (2SLS) estimator $\widehat{\theta} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\gamma})'$ of $\theta_0 = (\beta_0', \gamma_0)' = 0$ using a homoskedastic covariance matrix estimator
$$W = n\widehat{\beta}_1^2[(Y,X)'P_{Z,X}(Y,X)]_{1,1}/\widehat{\sigma}^2, \qquad (5.19)$$
where $\widehat{\sigma}^2 := (n-3)^{-1}\sum_{i=1}^n (y_i - (Y_i', X_i)\widehat{\theta})^2$ denotes the sum of squared residuals divided by $n-3$.[20]

- The subsampling method (3.6) is applied to the Wald statistic W in (5.19). This approach is denoted *Sub*. Again, empirical null rejection probabilities are obtained via the data-dependent choice of block size of Algorithm 4.1, with the set of input block sizes given by $B = \{6, 8, 10, 15, 20, 25, 30, 35\}$ and with the number of repetitions given by $L = 250$. Empirical power was calculated as noted above in experiment (I).

[19] The case $\pi_1 = \pi_2 = 0$ leads to the same Π matrix in both designs.

[20] We also experimented with the classical likelihood ratio statistic and its subsampling counterpart but did not find an advantage over the Wald statistic approach.

- Kleibergen's (2004) subvector statistic, denoted K, defined in his equation (17)[21].
- A projected version of the Anderson and Rubin (1949) statistic, denoted AR_P, as suggested in Dufour and Taamouti (2007).

We investigate the subvector case rather than a more general linear hypothesis to have the K test available as a competitor. Recall that the K test can not be applied in the latter case. No instruments are excluded from the reduced form to satisfy a main assumption for the K test to work properly. We look at only two endogenous variables because the power properties of the AR_P test are likely to be better in this case than in a scenario where one has to "project out" more dimensions. Therefore, if anything, we believe that this Monte Carlo design works in favor of the competitors of *Sub*. Note that Moreira's (2003) test can not be applied in this scenario.

(III) In the third experiment we investigate the performance of tests of overidentifying restrictions. The DGP is as in experiment (II), (5.16)–(5.17), except that we add two additional excluded exogenous variables, that is, Z is now a $(n \times 4)$-matrix of *i.i.d.* $N(1,1)$ variables. Instead of $n = 100$, we work with $n = 200$. Again, the study varies over different Π matrices. More specifically, for π_1 and π_2 as in (II) we take all 36 possible combinations of Π matrices defined as

$$\Pi = \begin{pmatrix} \pi_1 & 2\pi_2 \\ 2\pi_1 & \pi_2 \\ .0001 & .0001 \\ .0001 & .0001 \end{pmatrix}. \tag{5.20}$$

The hypothesis under test is

$$H_0 : \exists \theta \in \Theta, Eg_i(\theta) = 0 \text{ versus } H_1 : \forall \theta \in \Theta, Eg_i(\theta) \neq 0.$$

We compare the size performance of the following two statistics:

- The classical J statistic $J = n\hat{g}(\hat{\theta})'[\hat{\sigma}^2(Z,X)'(Z,X)/n]^{-1}\hat{g}(\hat{\theta})$ based on the 2SLS estimator $\widehat{\theta} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\gamma})'$ of $\theta_0 = (\beta_0', \gamma_0)' = 0$ using a homoskedastic covariance matrix estimator, where $\widehat{\sigma}^2 := n^{-1}\sum_{i=1}^n (y_i - (Y_i', X_i)\widehat{\theta})^2$ denotes the sum of squared residuals divided by n.
- The subsampling method (3.6) applied to this J statistic. This approach is denoted *Sub*. Again, empirical null rejection probabilities are obtained via the

[21] Kleibergen's (2004) subvector statistic is defined in a linear model with no exogenous variables. In case there are exogenous variables, he suggests to project them out. Therefore, in our study, we project out the constant X when calculating the K statistic.

data-dependent choice of block size of Algorithm 4.1, with the set of input block sizes given by $B = \{10, 30, 50, 70, 90, 110\}$ and with the number of repetitions given by $L = 250$.

5.1 Size and Power Comparison

In all experiments sizes are calculated at the 5% nominal level. We use $R = 2,000$ repetitions for experiments (I) and (II) and $R = 1,000$ for (III).

(I) We first discuss the results for experiment (I) starting with size. The results for $j = 1$ and $j = 3$ are not qualitatively different and therefore, we only report the results for $j = 3$, see Table 1, where the empirical rejection probabilities (ERP) under the null hypothesis are summarized. The size results can be quickly summarized. As discussed above already, the version of LR_M employed here is not robust to conditional heteroskedasticity and consequently, the test overrejects severely under conditional heteroskedasticity. Theory says that the ERPs under the null of all other tests should not be affected by the strength of identification and indeed all the ERPs under the null come close to the nominal level across all scenarios we looked at, for example, the ERPs of Sub, K, and LM_{EL} across the scenarios in Table 1 fall into the intervals $[4.3\%, 7.3\%]$, $[4.5\%, 6.5\%]$, and $[4.3\%, 6.1\%]$, respectively.

We now discuss the power results of experiment (I). ERPs for the four test statistics are calculated for the true β_0 being an element of $\{-1, -0.9, -0.8, \ldots, 0.9, 1\}$ and the null hypothesis being $H_0 : \beta_0 = 0$. There is no qualitative difference in the power results for $j = 1$ and $j = 3$ and we therefore focus on the latter. Figures I(a)–(d) contain power curves for the cases $j = 3$, $\pi = .1$ and 1 under both homo-and heteroskedasticity. For $\pi = 0, .01$, and .05 we obtain essentially flat power curves at the empirical null rejection probability of each test. The case $\pi = .5$ is qualitatively similar to $\pi = 1$ with lower power for all tests. While for $\pi = .1$ the power curves are still relatively flat, especially under heteroskedasticity (see Figures I(a)–(b)), the tests have high power and are U-shaped for $\pi = 1$ (see Figures I(c)–(d)). In the case $\pi = 1$, the Sub test is dominated by the other tests for most β_0. The power loss is higher for positive β_0 and under heteroskedasticity. In fact, for negative $\beta_0 < -.5$, Sub has higher power than LM_{EL} and K under homoskedasticity. These results are rather encouraging for the subsampling approach because they indicate that even in a scenario where the competitors are known to perform best relative to Sub, their power advantage over Sub is not overwhelming.

(II) We now discuss the results for experiment (II) starting with size. There are no significant differences in the results for the two different designs of the Π matrix and therefore we only report results for the first design, where $\Pi = \bar{\Pi}$, see

Table 2. Theory predicts that the Wald statistic is size distorted if at least one of the parameters π_1 or π_2 is small, that the K statistic is size distorted if the parameter not under test is only weakly identified, that is π_2 is small, and that AR_P is generally conservative. On the other hand, the subsampling approach, Sub should lead to exact sizes (at least asymptotically) under all scenarios considered. The ERP, summarized in Tables 2, are consonant with this prediction. Across all experiments, ERPs for the Wald test fall into the interval [.6%,41.9%]. The test severely overrejects in cases where π_1 is (relatively) small, and typically underrejects when π_2 is small and π_1 is large. The K test leads to reliable size results except for cases where π_2 is small, where the test severely underrejects; for example, in all experiments with $\pi_2 = 0$ or $\pi_2 = .01$ the ERP is .4%. The AR_P test severely underrejects. Across all experiments, ERPs fall into the interval [.0%,1.5%]! Finally, the subsampling procedure seems to have the best overall size properties; there is no clear pattern of size-distortion, but still, the size results for Sub are not perfect either and there are various under- and overrejections for certain parameter combinations. For example, for $\pi_1 = .05$ and $\pi_2 = 0$ the ERP is about 2% for both designs of the Π matrix. This is also consonant with theory that states that (for one-sided alternatives) under the null the error in rejection probability of tests based on the subsampling approach is typically of order $O_p(b^{-1/2})$ compared to the faster $O_p(n^{-1/2})$ of standard approaches (and a qualitatively analogous statement holds for two-sided alternatives).

The potentially severe size distortion of the Wald test under weak identification rules out its application in situations where the strength of identification is under doubt. We still include the Wald test into the following power study as a benchmark that allows us to quantify how much power tests, that are robust to weak identification, lose with respect to classical procedures that are size distorted under weak identification.

We now discuss the 71 power results of experiment (II). ERPs for the four test statistics are calculated for the true β_{01} being an element of $\{-1, -0.9, -0.8, \ldots, 0.9, 1\}$ and the null hypothesis being $H_0 : \beta_{01} = 0$. As for the size results there are no qualitative differences for the two designs of the Π matrix and we therefore focus our discussion on the first design $\bar{\Pi}$. A subset of the power results for the five parameter combinations ($\pi_1 = 0, .05, .5$ and $\pi_2 = 0$) and ($\pi_1 = .1, 1$ and $\pi_2 = 1$) is given in Figures II(a)–(e).

If the parameter under test is not or very weakly identified, $\pi_1 = 0$ or .01, then the power curves of Sub, K, and AR_P are essentially horizontal lines through the ERPs under the null; in particular, the value of the AR_P power curve is typically far below 5% in all these cases while the one of the power curve based on Sub is close to 5%. It is intuitive that these tests do not have any power for small π_1: if

the parameter under test is not or only very weakly identified, we can not expect to learn much about it from the data. The power curve of the Wald test has an entirely different shape. It is a convex function that takes on its minimum for $\beta_{01} < -.5$ and then grows as β_{01} increases, taking on values far bigger than 5% under the null and reaching ERP of up to about 65% as β_{01} reaches 1. The Wald test is therefore severely biased but seems to be symmetric about its argmin. A representative figure for these cases is Figure II(a) where $\pi_1 = \pi_2 = 0$.

If the strength of identification of the parameter under test is increased further and π_1 equals .05 or .1, these observations are still true with the following modifications. The power curve of the Wald test still takes on its minimum value at a negative β_{01} but this β_{01} is now closer to zero in absolute value. While still being flat for positive β_{01} values, the power curve of the Sub test has a peak at about at about $\beta_{01} = -.5$ for small π_2 values. See Figure II(b), where $\pi_1 = .05$ and $\pi_2 = 0$, for a representative case. For larger π_2 values the K and AR_P test also pick up power for negative β_{01} with the K test outperforming Sub and AR_P, the latter being the worst in terms of power. See Figure II(c), where $\pi_1 = .1$ and $\pi_2 = 1$, where these features are displayed.

Finally, we discuss the cases where $\pi_1 = .5$ or 1. The main power advantage of the Sub test appears in those many cases where the parameter under test is strongly identified relative to the parameter not under test, that is, all the cases where $\pi_1 \geq .5$ and $\pi_2 < .5$. In these scenarios, the power curves of the K and AR_P statistics are still relatively flat (with the former always uniformly outperforming the latter in terms of power) with power well below 20% in most cases while Sub takes on power of up to 60%! See Figure II(d), where $\pi_1 = .5$ and $\pi_2 = 0$. In these cases, the Wald test is the best procedure, with a U-shaped power curve centered at $\beta_{01} = 0$ and power reaching up to 80% if $|\beta_{01}| = 1$. Finally, if the parameter not under test is strongly identified, that is π_2 is large, $\pi_2 \geq .5$, the power of the K test improves dramatically and its power curve then almost coincides with the one of the Wald test. In these cases all power curves are U-shaped and centered at $\beta_{01} = 0$ with AR_P and Sub outperformed by K.

In summary, the AR_P test is dominated by the K test across every single scenario and based on this Monte Carlo study we can not recommend its use. In all scenarios where the parameter not under test is only weakly identified, $\pi_2 < .5$, the Sub test is the clear winner among the three statistics that are robust to weak instruments. The power gains over the K test can be dramatic in these cases, as shown in Figure II(d). If π_2 increases further, then the K test sometimes has slightly better power properties than the Sub test. If both π_1 and π_2 are large, the Wald test is very competitive; however, in cases of weak identification, the Wald test is biased and severely size

distorted.

(III) Finally, we discuss the size results for the tests of overidentifying restrictions of experiment (III), see Table 3. The classical J test experiences size distortion, especially but not exclusively, in some of the weakly identified scenarios. The ERP is bigger than 15%, 10%, and 8% in 5, 9, and 16 of the 36 scenarios, respectively. Subsampling almost uniformly improves on the size properties of the J test. In particular, its ERP is bigger than 15%, 10%, and 8% in 0, 1, and 2 of the 36 scenarios, respectively. As to be expected from our theoretical results, there is no pattern in the results that would indicate that the size properties of Sub depend on the degree of identification. As in our previous experiments we find that, while subsampling successfully improves the size problems of classical tests, it does not fully cure them in finite samples due to the slower rate of convergence to zero of the ERP. Still, the improvements are tremendous in many scenarios. For example for $\pi_1 = .05$, the ERPs of the J test over the different π_2-values are 16.2, 16.1, 17.0, 10.1, 7.4, and 4.7%. On the other hand, the corresponding numbers for the subsampling version are 8.2, 6.4, 4.7, 3.8, 5.1, and 4.6%! To the best of our knowledge, subsampling is the first approach to testing overidentifying restrictions that is robust to identification failure. In particular, the tests suggested in Imbens et al. (1998) are not robust to identification failure.

6 Conclusion

We introduce new subsampling based tests of parameter hypotheses and overidentifying restrictions that are robust to weak identification. The tests are applicable in a time series context given by unconditional moment restrictions. To the best of our knowledge, there are no other tests of overidentifying restrictions in the literature that are robust to weak identification and consistent under full identification. Furthermore, there are no other tests of general linear parameter hypotheses in the literature that are consistent under full identification and have exact (asymptotic) rejection probabilities under the null; for example, projection based tests are only conservative and our Monte Carlo study indicates that they typically have poor power properties under weak identification compared to the subsampling approach. In a linear single equation model, our approach can be used to simultaneously test hypotheses on the coefficients of the exogenous and endogenous variables. On the other hand, this can not be done with test procedures where the exogenous variables are projected out in a first step, see for example, Kleibergen (2002, 2004).

Roughly speaking, what is required for the subsampling approach to work, is that asymptotically, under the null hypothesis, the subsampling test statistic obeys

a continuous limit law. Given this weak assumption, it seems very likely that the subsampling tests would also be robust to the so called "many instrument problem", see Bekker (1994), Hahn and Inoue (2002), and Hansen et al. (2008). This question is currently under our investigation. The subsampling method is very general and could be applied to other testing problems in the context of weak and or many instruments, for example, to tests of exogeneity, see Staiger and Stock (1997, p.567).

Appendix

(A) Discussion and motivation of Assumption ID:

The linear model serves as a motivating example for Assumption ID.
Example 2.1 (cont.; based on Phillips (1989, p. 185–6)): In the above linear model simple calculations using $E(Z_i', X_i')'u_i = 0$ and $E(Z_i', X_i')'V_i' = 0$ yield

$$E\widehat{g}(\theta) = QF(\theta_0 - \theta),\ (\partial E\widehat{g}/\partial\theta') \equiv -QF \in \mathbb{R}^{(j+v_2)\times(v_1+v_2)},$$

where we set

$$F := \begin{pmatrix} \Pi & 0 \\ \Phi & I_{v_2} \end{pmatrix} \in \mathbb{R}^{(j+v_2)\times(v_1+v_2)},\ Q := E(Z_i', X_i')'(Z_i', X_i') = \begin{pmatrix} Q_{ZZ} & Q_{ZX} \\ Q_{XZ} & Q_{XX} \end{pmatrix},$$

and assume that the matrix Q has full rank. In the linear model the rank condition for identification is that Π has full column rank. Indeed, θ_0 is identified if and only if Π has full column rank. In the polar opposite case, $\Pi = 0$, β_0 is totally unidentified, while certain linear combinations of γ_0 may still be identified depending on the rank of the matrix Φ. More precisely, if $P = (P_1, P_2) \in O(v_2)$ and P_1 spans the null space of Φ', then $P_1'\gamma_0$ is identified while $P_2'\gamma_0$ is totally unidentified; for example, if $\Phi = 0$, then γ_0 is fully identified! Similarly, for general rank of Π, choose $S = (S_1, S_2) \in O(v_1)$ such that S_2 spans the null space of Π. Then $S_1'\beta_0$ is identified. Therefore, in the partially identified model the identifiable linear combinations of θ_0 can be retrieved *after* a rotation of the coordinate system.

Stock and Wright's (2000) Assumption C is a special case of the partially identified model in the sense that, according to C, in the *original* coordinates the components of θ_0 are either identified or (asymptotically) unidentified. Putting it differently, the matrix $(\partial E\widehat{g}/\partial\theta') \in \mathbb{R}^{k\times p}$ in Phillips (1989) can be of non-maximal rank without necessarily being of the particular form $(\partial E\widehat{g}/\partial\theta') = (0, M)$, where M is a matrix of maximal rank. On the other hand, such a decomposition into $(0, M)$ is implied by Assumption C in Stock and Wright (2000) as $n \to \infty$ (for $M = (\partial m_2/\partial\theta_2')$ under the weak technical assumption that $n^{-1/2}(\partial m_{1n}/\partial\theta')(\theta) \to 0$ uniformly), where in the limit the derivative of $E\widehat{g}(\theta)$ with respect to the weakly identified variables has to be constantly equal to 0. In the linear model $(\partial E\widehat{g}/\partial\theta') = -QF$ has a decomposition into $(0, M) \in \mathbb{R}^{k\times(p_1+p_2)}$ if and only if the first p_1 columns of F equal zero which holds if and only if the first p_1 columns of Π and Φ are zero. This is one particular case of, but does not account for the general case of identification failure. Therefore, Assumption ID is motivated by Phillips' (1989) treatment of the linear model that allows for more general forms of rank deficiency of Π and Θ. We finally discuss the

coordinate change T in the example of the linear model.

Example 2.1 (cont.): In the linear partially identified model choose a matrix $T = (T_1, T_2) \in O(p_1 + p_2)$ such that T_1 spans the null space of QF. Then

$$E\overline{\widehat{g}}(\overline{\theta}) = QFT(\overline{\theta}_0 - \overline{\theta}) = QFT_2(\overline{\theta}_{02} - \overline{\theta}_2),$$

$\overline{m}_2(\overline{\theta}_2) := QFT_2(\overline{\theta}_{02} - \overline{\theta}_2)$, and $\overline{m}_{1n} \equiv 0$. It follows that $\overline{M}_2(\overline{\theta}_2) \equiv -QFT_2$ has full column rank and that $\overline{m}_2(\overline{\theta}_2) = 0$ if and only if $\overline{\theta}_2 = \overline{\theta}_{02}$. The orthogonal map T transforms the coordinate system in $\mathbb{R}^{p_1+p_2}$ in such a way that in the decomposition $\overline{\theta}_0 := (\overline{\theta}_{01}, \overline{\theta}_{02})$ of the new variables, the first components $\overline{\theta}_{01}$ are unidentified while the remaining components $\overline{\theta}_{02}$ are identified.

(B) Some Words on Contiguity:

The notion of *contiguity* is a very useful tool to compute the limiting power of statistical tests against a certain class of "local" alternatives. Consider two sequences of probability measures $\{P_n\}$ and $\{Q_n\}$ defined on a common probability space. Then the sequence $\{Q_n\}$ is *contiguous* to the sequence $\{P_n\}$ if $P_n(E_n) \to 0$ implies $Q_n(E_n) \to 0$ for every sequence of (measurable) events $\{E_n\}$. Therefore, contiguity can be considered as an asymptotic version of one probability measure being absolutely continuous with respect to another one. Assume one knows the limiting distribution of a sequence of test statistic D_n under P_n but the behavior of D_n under Q_n is also required. Contiguity provides a means of performing the required calculation.

To verify contiguity in a particular setting, several high level conditions are available. One of them, and of particular interest to us, is the following. Consider a parametric family $\{P_\theta : \theta \in \Theta\}$ with corresponding densities $p_\theta(\cdot)$ with respect to a σ-finite measure. Assume θ_0 is an interior point of Θ and let $\theta_n = \theta_0 + h/\sqrt{n}$ for some $h \in \Theta$. Denote by P_θ^n the joint distribution of $\{z_1, \ldots, z_n\}$ when the z_i are *i.i.d.* from P_θ. Then, under general smoothness conditions, $\{P_{\theta_n}^n\}$ is contiguous to $\{P_{\theta_0}^n\}$. One such sufficient condition is that the parametric family $\{P_\theta : \theta \in \Theta\}$ be *differentiable in quadratic mean* in a neighborhood of θ_0. For example, this condition is satisfied by most exponential families, including the multivariate normal distribution.

For a detailed treatment of contiguity, differentiability in quadratic mean, and corresponding applications to compute the limiting power of tests against "local" alternatives, the reader is referred to Hájek et al. (1999, Chapter 7), van der Vaart (1998, Chapters 6, 7, 14, and 15), and Lehmann and Romano (2005, Chapter 12).

Contiguity, arguably, provides the most elegant tool to compute the limiting power against "local" alternatives, even though sometimes the calculations can be performed

by direct means. However, in doing so one has to account for the fact that the probability mechanism changes with n (using Lindeberg's central limit theorem, say).

(C) Proofs:

Proof of Theorem 3.1. See Theorem 3.5.1 of Politis et al. (1999). ■

Proof of Theorem 3.2. (i) We use the proof of Theorem 1 in Stock and Wright (2000) for the model in new coordinates. Set $\overline{S}_n(\overline{\theta}) := n||\overline{A}_n(\overline{e_n(\theta)})^{1/2}\overline{\widehat{g}}(\overline{\theta})||^2$ and note that the sequence $\overline{\widehat{\theta}}_n = T^{-1}(\widehat{\theta}_n)$ satisfies $\overline{\widehat{\theta}}_n \in \overline{\Theta}$ and $\overline{S}_n(\overline{\widehat{\theta}}_n) \leq \arg\inf_{\overline{\theta}\in\overline{\Theta}} \overline{S}_n(\overline{\theta}) + o_p(1)$. By assumption[22],

$$\overline{S}(\overline{\theta}_1, \overline{\theta}_2, \overline{e(\theta_1, \theta_{02})}) := ||\overline{A}(\overline{e(\theta_1, \theta_{02})})^{1/2}[\overline{\Psi}(\overline{\theta}_1, \overline{\theta}_{02}) + \overline{m}_1(\overline{\theta}_1, \overline{\theta}_{02}) + \overline{M}_2(\overline{\theta}_{02})\overline{\theta}_2]||^2 \tag{6.21}$$

satisfies the condition: There exists a random element $(\widetilde{\theta}_1, \widetilde{\theta}_2) \in \overline{\Theta}$ such that a.s.

$$\overline{S}(\widetilde{\theta}_1, \widetilde{\theta}_2, \overline{e(\widetilde{\overline{\theta}}_1, \theta_{02})}) < \inf_{(\overline{\theta}_1, \overline{\theta}_2)\in\overline{\Theta}\setminus\overline{G}} \overline{S}(\overline{\theta}_1, \overline{\theta}_2, \overline{e(\theta_1, \theta_{02})}) \tag{6.22}$$

for every open set $\overline{G}$ in $\overline{\Theta}$ that contains $(\widetilde{\theta}_1, \widetilde{\theta}_2)$. (This condition is needed when applying Lemma 3.2.1 in van der Vaart and Wellner (1996) in Stock and Wright's (2000) Theorem 1(ii).) Therefore, using ID and PE, the proof of Theorem 1 in Stock and Wright (2000) can be applied to the model in new coordinates yielding[23] $(\widetilde{\overline{\theta}}_1, n^{1/2}(\widetilde{\overline{\theta}}_2 - \overline{\theta}_{02})) \to_d (\overline{\theta}_1^*, \overline{\theta}_2^*)$, where

$$\overline{\theta}_1^* := \arg\min_{\overline{\theta}_1\in\overline{\Theta}_1} \overline{S}^*(\overline{\theta}_1, \overline{e(\theta_1, \theta_{02})}), \tag{6.23}$$

$$\overline{\theta}_2^* := -[\overline{M}_2(\overline{\theta}_{02})'\overline{A}(\overline{e(\theta_1^*, \theta_{02})})\overline{M}_2(\overline{\theta}_{02})]^{-1}\overline{M}_2(\overline{\theta}_{02})'\overline{A}(\overline{e(\theta_1^*, \theta_{02})})[\overline{\Psi}(\overline{\theta}_1^*, \overline{\theta}_{02}) + \overline{m}_1(\overline{\theta}_1^*, \overline{\theta}_{02})], \tag{6.24}$$

$$\overline{S}^*(\overline{\theta}_1, \overline{e(\theta_1, \theta_{02})}) := [\overline{\Psi}(\overline{\theta}_1, \overline{\theta}_{02}) + \overline{m}_1(\overline{\theta}_1, \overline{\theta}_{02})]'\overline{A}^*(\overline{\theta}_1, \overline{e(\theta_1, \theta_{02})})[\overline{\Psi}(\overline{\theta}_1, \overline{\theta}_{02}) + \overline{m}_1(\overline{\theta}_1, \overline{\theta}_{02})],$$

$$\overline{A}^*(\overline{\theta}_1, \overline{e(\theta_1, \theta_{02})}) :=$$
$$\overline{A}(\overline{e(\theta_1, \theta_{02})}) - \overline{A}(\overline{e(\theta_1, \theta_{02})}\overline{M}_2(\overline{\theta}_{02})[\overline{M}_2(\overline{\theta}_{02})'\overline{A}(\overline{e(\theta_1, \theta_{02})})\overline{M}_2(\overline{\theta}_{02})]^{-1}\overline{M}_2(\overline{\theta}_{02})'\overline{A}(\overline{e(\theta_1, \theta_{02})}).$$

(ii) The J statistic expressed in new coordinates reads $J_n = S_n(\widehat{\theta}_n) = \overline{S}_n(\overline{\widehat{\theta}}_n)$. Therefore, the statement follows from Corollary 4(i) and (j) in Stock and Wright (2000) applied to $\overline{S}_n(\overline{\widehat{\theta}}_n)$ ■

[22] For notational simplicity we write $\overline{\Psi}(\overline{\theta}_1, \overline{\theta}_{02})$ for $\overline{\Psi}((\overline{\theta}_1', \overline{\theta}_{02}')')$ and similarly in other expressions.

[23] Note that Stock and Wright's (2000) proof can be adapted to our slightly different definition of the GMM estimator as an approximate (up to order $o_p(1)$) minimizer; all that changes is that on the right hand side of their equation (A.1) we have an $o_p(1)$ term rather than 0.

Proof of Corollary 3.1. By Theorem 3.3 and assumption, the test statistic $D_n = J_n$ has a continuous limit distribution J^* both under full identification and identification failure. So the proof of (i) follows from part (i) of Theorem 3.1.

To prove (ii), let $\beta = 1$ and $d_n = J_n/n$ in (3.7). By Newey and McFadden (1994, Theorem 2.1) and Assumption MM it follows that $\widehat{\theta}_n$ is consistent for $\widetilde{\theta}$. Therefore by Assumption MM d_n converges in probability to

$$d(P) := ||A(e(\widetilde{\theta}))^{1/2} E g_i(\widetilde{\theta})||.$$

Clearly, $d(P) > 0$ under H_1. On the other hand, under H_0, $\widetilde{\theta}$, as the unique minimizer of $||A(e(\theta))^{1/2} E g_i(\theta)||$, has to satisfy $E g_i(\widetilde{\theta}) = 0$ and therefore $d(P) = 0$. Part (ii) of Theorem 3.1 therefore proves the result. ■

Proof of Lemma 3.1 Omitted. ■

Proof of Theorem 3.3. In new coordinates, Assumption ID implies that (in the terminology of Stock and Wright (2000)) the first p_1 components of the parameter vector $\overline{\theta}_0$ are weakly identified, the remaining p_2 components are strongly identified and are root n-consistently estimated as shown in Theorem 3.2. Therefore, when deriving the asymptotic distribution of the Wald statistic, we have to renormalize certain expressions that have different convergence rates than others. To do that we use two matrices $N \in \mathbb{R}^{p \times p}$ defined in (3.14) and $M \in \mathbb{R}^{r \times r}$. To define M, let $L \subset \mathbb{R}^r$ be the linear subspace spanned by the first p_1 columns of the matrix RT. Set $p_{11} := \dim L$, for which $p_{11} \leq \min(r, p_1)$, and assume w.l.o.g. that the first p_{11} columns of $RT \in \mathbb{R}^{r \times p}$ form a basis B_1 of L. Because RT has maximal rank r, there are $r - p_{11}$ columns among the last $p - p_{11}$ columns of RT that together with B_1 form a basis for $\mathbb{R}^r$. W.l.o.g. assume the last $r - p_{11}$ columns of RT can be taken and call them B_2. Let $(B_2^\perp, B_1^\perp)$ be another basis of $\mathbb{R}^r$, $B_2^\perp \in \mathbb{R}^{r \times p_{11}}$, $B_1^\perp \in \mathbb{R}^{r \times (r - p_{11})}$ such that the columns of $B_i^\perp$ are orthogonal to the columns in B_i, $i = 1, 2$. Define $M := (n^{-1/2} B_2^\perp, B_1^\perp) \in \mathbb{R}^{r \times r}$. Note that in the case of full identification $M = I_r$.

Using $n^{-1} \sum_{i=1}^n (\partial \overline{g_i} / \partial \overline{\theta}')(\overline{\widehat{\theta}}_n) = \widehat{G}_n T$, the Wald statistic (3.11) under the null hypothesis reads in renormalized form (and with some factors expressed in new coordinates)

$$W_n = \xi_n^{*\prime} [R_n^* B_n^{*-1} \Omega_n^* B_n^{*-1} R_n^{*\prime}]^{-1} \xi_n^*,$$

where, for $\widehat{G}_n^*$ defined in (3.13), we set

$$\begin{aligned}
R_n^* &:= M' RTN \in \mathbb{R}^{r \times p}, \\
\xi_n^* &:= R_n^* N^{-1} n^{1/2} (\overline{\widehat{\theta}}_n - \overline{\theta}_0) \in \mathbb{R}^r, \\
B_n^* &:= N \widehat{G}_n^{*\prime} A_n(e_n(\theta)) \widehat{G}_n^* N \in \mathbb{R}^{p \times p}, \\
\Omega_n^* &:= N \widehat{G}_n^{*\prime} A_n(e_n(\theta)) K_n(\widehat{\theta}) A_n(e_n(\theta)) \widehat{G}_n^* N \in \mathbb{R}^{p \times p}.
\end{aligned}$$

By construction of M it follows that (the nonstochastic matrix) $NT'R'M \in \mathbb{R}^{p\times r}$ converges to a matrix that has maximal rank r; namely, the first p_{11} and last $r-p_{11}$ rows of $NT'R'M$ span $\mathbb{R}^{r'}$ and do not depend on n because the normalizations $n^{1/2}$ from N and $n^{-1/2}$ from M cancel out. In the other rows of $NT'R'M$ the normalizations have either cancelled out each other as well or the components of $NT'R'M$ are in $O(n^{-1/2})$, that is, they converge to zero. Therefore, $R_n^* \in \mathbb{R}^{r\times p}$ converges to a matrix that has maximal rank r. Regarding the second factor in ξ_n^*, Theorem 3.2 implies that $N^{-1}n^{1/2}(\overline{\widehat{\theta}}_n - \overline{\theta}_0) \to_d (\overline{\theta}_1^{*'} - \overline{\theta}_0', \overline{\theta}_2^{*'})'$.

Now, let $\theta_{12} := (\overline{\theta}_1^{*'}, \overline{\theta}_{02}')'$ for $\overline{\theta}_1^*$ defined in (6.23). By Assumption WS(i)–(ii) and Theorem 3.2(i) we have $e_i'B_n^*e_j \to_d \overline{\Phi}_i'(\theta_{12})A(e(\theta_{12}))\overline{\Phi}_j(\theta_{12})$ if both $i,j \le p_1$, where by $\overline{\Phi}_j \in \mathbb{R}^k$ we denote the subvector of $\overline{\Phi}$ from WS(i) containing the components $k(j-1)+1$ to kj, $e_i'B_n^*e_j \to_d \overline{\Phi}_i'(\theta_{12})A(e(\theta_{12}))\overline{M}_{2(j-p_1)}(\theta_{12})$ if $i \le p_1$ and $j > p_1$, where by $\overline{M}_{2j}(\cdot) \in \mathbb{R}^k$ we denote the j^{th} column of the matrix $\overline{M}_2(\cdot)$ in WS(ii), $e_i'B_n^*e_j \to_d \overline{M}_{2(i-p_1)}'(\theta_{12})A(e(\theta_{12}))\overline{\Phi}_j(\theta_{12})$ if $i > p_1$ and $j \le p_1$, and $e_i'B_n^*e_j \to_d \overline{M}_{2(i-p_1)}'(\theta_{12})A(e(\theta_{12}))\ \overline{M}_{2(j-p_1)}(\theta_{12})$ if both $i,j > p_1$. Similar statements hold for $e_i'\Omega_n^*e_j$; for example, by WS(iv) we have $e_i'\Omega_n^*e_j \to_d \overline{\Phi}_i'(\theta_{12})A(e(\theta_{12}))\Lambda(\theta_{12})A(e(\theta_{12}))\overline{\Phi}_j(\theta_{12})$ if both $i,j \le p_1$. Note that all the above limits hold jointly.

Denote by ξ^*, R^*, B^*, and Ω^* the just described limits of ξ_n^*, R_n^*, B_n^*, and Ω_n^*. Then by the continuous mapping theorem the asymptotic distribution of the Wald statistic is given by the distribution of W^*

$$W_n \to_d W^* := \xi^{*\prime}[R^*B^{*-1}\Omega^*B^{*-1}R^{*\prime}]^{-1}\xi^*, \tag{6.25}$$

if we can show that $R_n^*B_n^{*-1}\Omega_n^*B_n^{*-1}R_n^{*\prime}$ converges to a random variable (with realizations in $\mathbb{R}^{r\times r}$) whose realizations are invertible a.s.. We have shown already that R^* has full rank and it thus remains to show that B_n^* and Ω_n^* converge to random variables that have full rank a.s.. But this holds by Assumptions WS(iii)–(iv) and PE(ii) that establish that $\Lambda(\theta_{12})$, $A(e(\theta_{12}))$, and the limit of $\widehat{G}_n^*N \in \mathbb{R}^{k\times p}$ have full rank a.s.. ■

Proof of Corollary 3.2. By Theorem 3.3 and assumption, the test statistic $D_n = W_n$ has a continuous limit distribution W^* both under full identification and identification failure. So the proof of (i) follows from part (i) of Theorem 3.1.

To prove (ii), assume full identification. In (3.7) let $\beta = 1$ and $d_n = W_n/n$. Then d_n converges in probability to

$$d(P) := (R\theta_0 - q)'[RB^{*-1}\Omega^*B^{*-1}R']^{-1}(R\theta_0 - q),$$

where $B^* := plim(\widehat{B}_n)$ and $\Omega^* := plim(\widehat{\Omega}_n)$, defined in (6.25), are the probability limits under strong identification. Obviously, $d(P) = 0$ if and only if the null hypothesis is true. Now apply part (ii) of Theorem 3.1.

To prove (iii), assume full identification and consider a sequence of alternatives P_n that are contiguous to $P \in \mathbf{P_0}$. Denote by θ_n the parameter corresponding to P_n and by θ_0 the parameter corresponding to P. We have $R\theta_0 = q$. A test based on the Wald statistic W_n has exact asymptotic rejection probability under P when the critical value is the $1-\alpha$ quantile of the $\chi^2(r)$ distribution, denoted by $\chi^2_{1-\alpha}(r)$. So in the notation of part (iii) of Theorem 3.1, $C(P) = \chi^2(r)$. Now let W be a random variable whose distribution is the limiting distribution of W_n under P_n. Then the asymptotic power of the Wald test is given by $Prob\{D > \chi^2_{1-\alpha}(r)\}$. An application of part (iii) of Theorem 3.1, with $D_n = W_n$, $D = W$, and $c(1-\alpha, P) = \chi^2_{1-\alpha}(r)$, now implies that the limiting power of the subsampling test against the sequence P_n is also given by $Prob\{D > \chi^2_{1-\alpha}(r)\}$. ■

References

Anderson, T.W. and H. Rubin (1949): "Estimators of the parameters of a single equation in a complete set of stochastic equations", *The Annals of Mathematical Statistics* 21, 570–582.

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

——— (1994): "Empirical process methods in econometrics", in *Handbook of Econometrics*, Volume 4, ed. R.F. Engle and D. McFadden. North Holland: Amsterdam, 2247–2294.

——— (2003): "End-of-sample instability tests", *Econometrica* 71(6), 1661–1694.

Andrews, D.W.K. and P. Guggenberger (2010a): "Asymptotic size and a problem with subsampling and with the m out of n bootstrap", *Econometric Theory* 26(2), 426–468.

——— (2010b): "Hybrid and size-corrected subsample methods", *Econometrica* 77(3), 721–762.

Andrews, D.W.K., M. Moreira, and J.H. Stock (2006): "Optimal invariant similar tests for instrumental variables regression", *Econometrica* 74(3), 715–752.

Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators", *Econometrica* 62(3), 657–681.

Caner, M. (2010): "Exponential tilting with weak instruments: estimation and testing", *Oxford Bulletin of Economics and Statistics* 72(3), 307–325.

Chao, J.C. and N.R. Swanson (2006): "Asymptotic normality of single-equation estimators for the case with a large number of weak instruments", *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, ed. D. Corbae, S. N. Durlauf, and B. E. Hansen. Cambridge University Press: Cambridge, 82–124.

Choi, I. (2005): "Subsampling vector autoregressive tests of linear constraints", *Journal of Econometrics* 124(1), 55–89.

Choi, I. and P.C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations", *Journal of Econometrics* 51, 113–150.

Dufour, J. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models", *Econometrica* 65(6), 1365–1387.

——— (2003): "Identification, weak instruments and statistical inference in econometrics. Presidential Address to the Canadian Economics Association", *Canadian Journal of Economics* 36(4), 767–808.

Dufour, J. and M. Taamouti (2005): "Projection-based statistical inference in linear structural models with possibly weak instruments", *Econometrica* 73(4), 1351–1365.

——— (2007): "Further results on projection-based inference in IV regressions with weak, collinear or missing instruments", *Journal of Econometrics* 139(1), 133–153.

Efron, B. (1979): "Bootstrap methods: another look at the jackknife", *Annals of Statistics* 7, 1–26.

Forchini, G. and G. Hillier (2003): "Conditional inference for possibly unidentified structural equations", *Econometric Theory* 19(5), 707–743.

Gonzalo, J. and Wolf, M. (2005): "Subsampling inference in threshold autoregressive models", *Journal of Econometrics*, 81, 201–224.

Guggenberger, P. (2003): "Econometric essays on generalized empirical likelihood, long–memory time series, and volatility", Ph.D. thesis, Yale University.

Guggenberger, P. and R.J. Smith (2005): "Generalized empirical likelihood estimators and tests under partial, weak and strong identification", *Econometric Theory* 21, 667–709.

Hahn, J. and A. Inoue (2002): "A monte carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators", *Econometric Reviews* 21, 309–336.

Hájek, J., Z. Šidăk, and P. Sen. (1999): "Theory of Rank Tests", Academic Press: New York.

Hansen, C., J. Hausman, and W.K. Newey (2008): "Many instruments, weak instruments, and microeconometric practice", *Journal of Business & Economic Statistics* 26(4), 398–422.

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

Hansen, L.P., J. Heaton, and A. Yaron (1996): "Finite–sample properties of some alternative GMM estimators", *Journal of Business & Economic Statistics* 14(3), 262–280.

Imbens, G.(1997): "One–step estimators for over–identified generalized method of moments models", *Review of Economic Studies* 64, 359–383.

Imbens, G., R.H. Spady, and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models", *Econometrica* 66(2), 333–357.

Kitamura, Y. and M. Stutzer (1997): "An information–theoretic alternative to generalized method of moments estimation", *Econometrica* 65(4), 861–874.

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression", *Econometrica* 70(5), 1781–1805.

——— (2003): "Expansions of GMM statistics that indicate their properties under weak and/or many instruments and the bootstrap", *Working Paper*, Department of Economics, Brown University.

——— (2004): "Testing subsets of structural parameters in the instrumental variables regression model", *Review of Economics and Statistics* 86, 418–423.

——— (2005): "Testing parameters in GMM without assuming that they are identified", *Econometrica* 73(4), 1103–1123.

Lehmann, E.L. and J.P. Romano (2005): "Testing Statistical Hypotheses", third edition. Springer: New York.

Loh, W.–Y. (1987): "Calibrating confidence coefficients", *Journal of the American Statistical Association* 82, 155–162.

Moreira, M.J. (2003): "A conditional likelihood ratio test for structural models", *Econometrica* 71(4), 1027–1048.

Moreira, M.J., J.R. Porter, and G. Suarez (2004): "Bootstrap and higher-order expansion validity when instruments may be weak", *Working Paper*, Department of Economics, Harvard University.

Nelson, C.R. and R. Startz (1990): "Some further results on the exact small sample properties of the instrumental variable estimator", *Econometrica* 58(4), 967–976.

Newey, W.K. (1985): "Generalized method of moments specification testing", *Journal of Econometrics* 29(3): 229–256.

Newey, W.K. and D. McFadden (1994): "Large sample estimation and hypothesis testing", in *Handbook of Econometrics*, Volume 4, ed. R.F. Engle and D. McFadden. North Holland: Amsterdam, 2111–2245.

Newey, W.K. and K.D. West (1987): "Hypothesis testing with efficient method of moments estimation", *International Economic Review* 28(3), 777–787.

Otsu, T. (2006): "Generalized empirical likelihood inference for nonlinear and time series models under weak identification", *Econometric Theory* 22, 513–527.

Phillips, P.C.B. (1989): "Partially identified econometric models", *Econometric Theory* 5, 181–240.

Politis, D.N. and J.P. Romano (1994): "The stationary bootstrap", *Journal of the American Statistical Association* 89, 1303–1313.

Politis, D.N., J.P. Romano, and M. Wolf (1999): "Subsampling", Springer: New York.

Romano, J.P. and M. Wolf (2001): "Subsampling intervals in autoregressive models with linear time trend", *Econometrica* 69, 1283–1314.

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

Zivot, E., R. Startz, and C.R. Nelson (2006): "Improved inference in weakly identified instrumental variables regression", *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, ed. D. Corbae, S. N. Durlauf, and B. E. Hansen. Cambridge University Press: Cambridge, 125–166.

Stock, J.H. and J.H. Wright (2000): "GMM with weak identification", *Econometrica* 68(5), 1055–1096.

Stock, J.H., J.H. Wright, and M. Yogo (2002): "A survey of weak instruments and weak identification in generalized method of moments", *Journal of Business & Economic Statistics* 20(4), 518–529.

van der Vaart, A.W. (1998): "Asymptotic Statistics", Cambridge University Press: Cambridge.

van der Vaart, A.W. and J.A. Wellner (1996): "Weak convergence and empirical processes", Springer: New York.

Table 1: Empirical null rejection probabilities in Monte Carlo experiment (I) for various tests with nominal size $\alpha = 5\%$. The number of repetitions is 2,000 per scenario.

π	Sub	Sub^*	K	LM_{EL}	LR_M
$j = 3$, Homoskedastic					
0	5.3	5.0	5.3	5.2	5.7
.01	5.2	4.3	5.6	5.4	6.2
.05	4.9	5.2	5.5	5.1	5.9
.1	6.0	5.1	5.0	4.9	5.8
.5	5.6	5.5	4.7	4.5	5.9
1	4.3	5.2	4.9	4.7	5.9
$j = 3$, Heteroskedastic					
0	5.4	4.9	4.8	4.9	20.0
.01	5.8	5.0	5.4	5.3	21.4
.05	5.9	4.4	5.3	5.1	19.8
.1	7.3	6.1	4.9	4.9	18.7
.5	7.2	5.3	6.5	6.1	17.9
1	4.5	5.1	4.5	4.3	14.5

Table 2: Empirical null rejection probabilities in Monte Carlo experiment (II) for various tests with nominal size $\alpha = 5\%$. The design of the Π matrix is $\overline{\Pi}$ of (5.18). The number of repetitions is 2,000 per scenario.

π_1	π_2	W	Sub	Sub^*	K	AR_P
0	0	41.9	3.2	5.4	0.4	0.0
0	.01	39.2	3.1	4.5	0.4	0.1
0	.05	21.2	2.5	4.4	2.1	0.2
0	.1	15.1	3.4	4.1	4.7	1.4
0	.5	12.9	3.7	4.3	5.0	1.3
0	1	10.0	4.0	4.6	5.5	1.5
.01	0	38.2	3.2	4.8	0.4	0.0
.01	.01	30.0	3.1	3.5	0.4	0.1
.01	.05	21.2	2.5	3.6	2.1	0.2
.01	.1	12.5	3.0	3.7	4.7	0.5
.01	.5	13.3	3.7	4.4	5.0	1.3
.01	1	12.7	3.8	5.1	5.3	1.5
.05	0	11.3	1.9	2.3	0.4	0.0
.05	.01	7.6	2.0	2.3	0.4	0.0
.05	.05	3.0	2.4	2.7	1.9	0.3
.05	.1	4.2	2.9	4.3	4.4	1.2
.05	.5	10.0	3.2	5.2	5.0	1.3
.05	1	11.0	2.9	5.3	5.7	1.4
.1	0	4.5	2.2	3.1	0.4	0.0
.1	.01	3.4	3.1	3.6	0.4	0.1
.1	.05	0.6	5.1	4.8	1.7	0.2
.1	.1	1.3	5.2	5.9	4.7	1.0
.1	.5	7.9	4.1	4.8	4.9	1.3
.1	1	8.5	3.7	5.5	5.3	1.5
.5	0	0.9	4.5	4.3	0.4	0.0
.5	.01	1.1	4.9	5.1	0.4	0.1
.5	.05	1.1	7.1	5.5	2.1	0.4
.5	.1	2.8	7.3	4.3	4.7	1.0
.5	.5	3.5	6.4	4.0	5.0	1.3
.5	1	4.7	6.2	5.7	5.3	1.5
1	0	0.9	4.1	5.7	0.4	0.0
1	.01	1.0	4.9	5.6	0.4	0.1
1	.05	2.1	7.3	5.4	1.7	0.3
1	.1	5.0	7.3	5.7	4.7	1.0
1	.5	4.4	6.5	6.6	5.0	1.3
1	1	4.6	5.4	5.0	5.3	1.5

Table 3: Empirical null rejection probabilities in Monte Carlo experiment (III) for various tests with nominal size $\alpha = 5\%$. The design of the Π matrix is $\overline{\Pi}$ of (5.20). The number of repetitions is 2,000 per scenario.

π_1	π_2	J	$Boot$	Sub	Sub^*
0	0	1.9	3.0	4.0	5.4
0	.01	4.4	7.1	6.8	4.5
0	.05	16.5	11.5	9.6	8.3
0	.1	8.2	4.7	4.4	5.1
0	.5	3.2	4.3	4.5	4.7
0	1	3.1	4.0	4.9	4.7
.01	0	6.2	8.5	7.2	5.8
.01	.01	13.2	13.7	12.1	12.5
.01	.05	16.8	11.9	8.5	8.0
.01	.1	8.6	5.5	4.4	5.2
.01	.5	3.0	4.0	4.3	4.7
.01	1	3.0	4.2	5.2	4.9
.05	0	17.3	12.9	9.8	9.4
.05	.01	18.3	11.5	8.9	8.0
.05	.05	17.6	9.5	6.7	4.7
.05	.1	10.0	5.8	4.5	4.8
.05	.5	6.1	6.6	6.2	4.2
.05	1	6.9	8.1	6.7	5.3
.1	0	8.1	5.4	5.0	5.5
.1	.01	8.0	4.7	4.4	4.6
.1	.05	10.9	7.0	5.7	5.7
.1	.1	8.9	5.2	4.2	4.6
.1	.5	8.5	8.5	7.1	4.7
.1	1	10.8	10.9	9.4	8.5
.5	0	3.6	4.2	4.4	5.4
.5	.01	2.6	3.7	4.2	5.1
.5	.05	6.6	7.3	6.5	4.4
.5	.1	8.8	8.5	6.5	4.9
.5	.5	5.8	5.5	4.9	4.6
.5	1	5.5	5.1	5.4	4.6
1	0	3.2	4.1	4.6	4.3
1	.01	2.7	3.7	4.2	5.2
1	.05	7.0	8.2	7.5	5.3
1	.1	9.3	9.2	8.1	7.5
1	.5	5.9	5.8	4.8	5.7
1	1	5.2	5.3	5.6	5.3

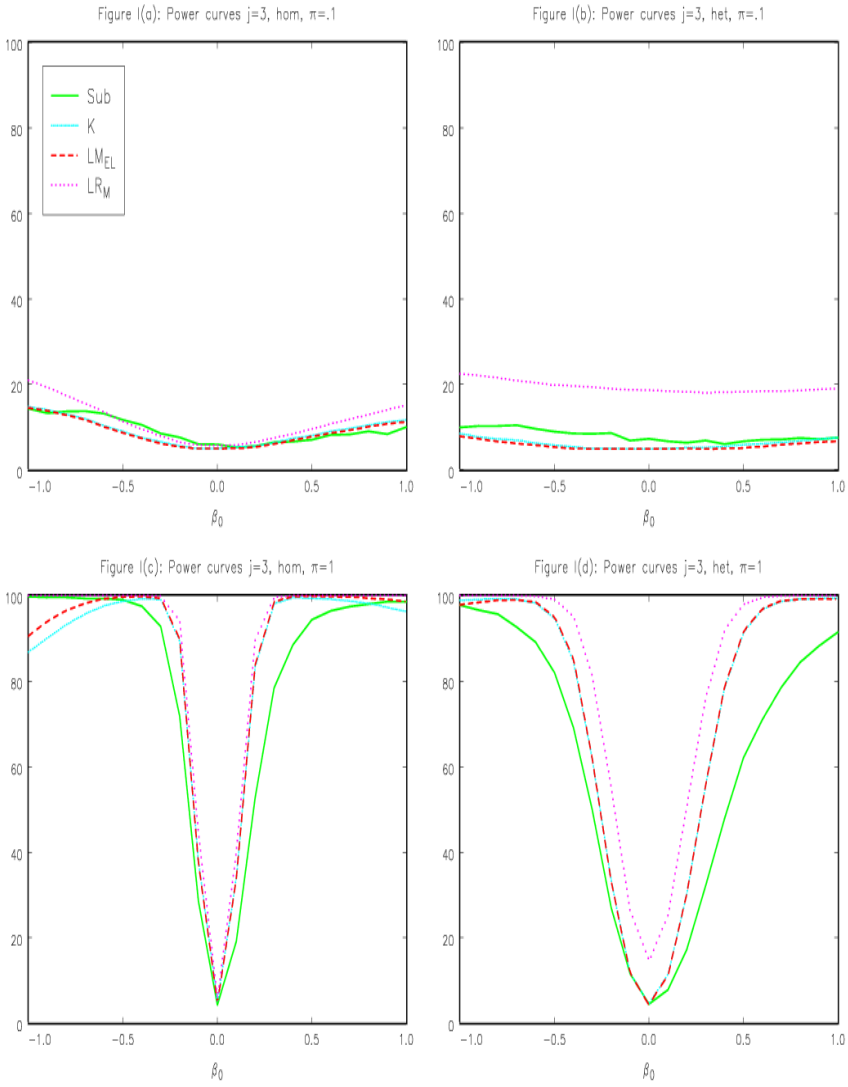

Figure I: Empirical power results for experiment (I).

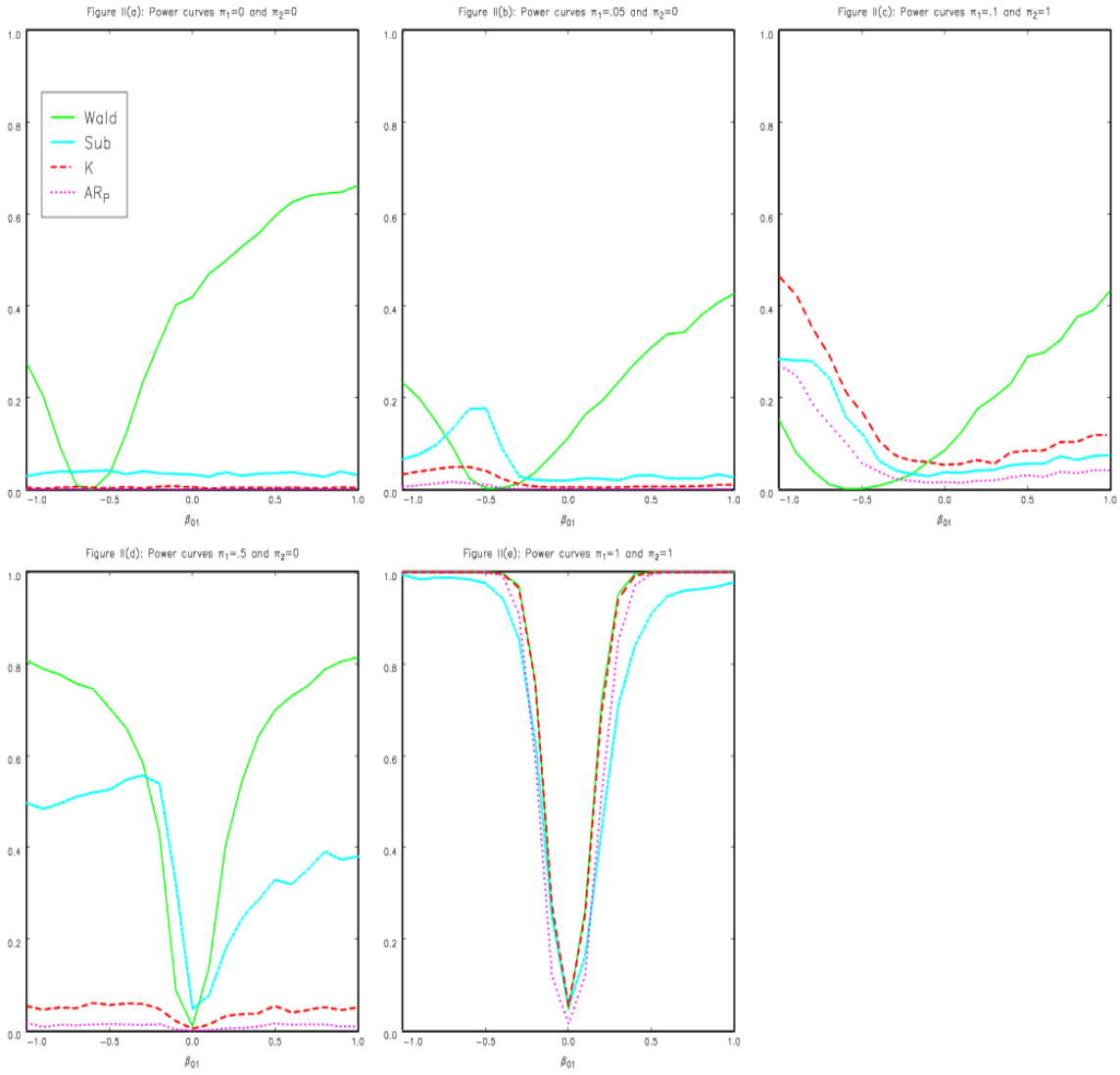

Figure II: Empirical power results for experiment (II).