Improved Projection GMM-LM Tests for Linear Restrictions*

Saraswata Chaudhuri†

Date: Dec 15, 2016

Abstract

Let the unknown true value of a $d_\theta \times 1$ parameter θ be θ^0. Let θ^0 satisfy $d_g \geq d_\theta$ moment restrictions. Consider a linear null hypothesis $H_0 : R\theta^0 = r_0$ where R is a fixed, full row-rank, $d_R \times d_\theta$ known matrix and r_0 is a $d_R \times 1$ known vector. The conventional projection test rejects H_0 at the level α if there does not exist any $d_\theta \times 1$ value θ_0 in a $(1-\alpha)$-level confidence set for θ^0 such that the restriction $R\theta_0 = r_0$ holds. The probability with which this test rejects H_0, when it is true, cannot exceed α asymptotically if the asymptotic coverage for the confidence set for θ^0 is no smaller than $1-\alpha$. However, this test can be conservative and computationally inconvenient. We propose an improved projection method based on a function of Neyman's C-alpha statistic to address both problems. The test is less conservative than the conventional projection test, but still allows to enforce a user-specified upper bound on its asymptotic rejection probability of the true H_0 irrespective of any identification failure of θ^0. Indeed, under conditions ensuring the local optimality of the classical plug-in-based Wald, likelihood ratio and score tests for $H_0 : R\theta^0 = r_0$, the improved projection test is asymptotically equivalent to them. This test also addresses the computational problem by requiring the projection from a smaller dimension $d_\theta - d_R$ instead of d_θ.

*I am extremely grateful to J-M Dufour, E. Renault, J. Robins, T. Richardson, V. Zinde-Walsh and E. Zivot for helpful discussions.

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

1 Introduction

Consider a parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ whose unknown true value θ^0 satisfies the following moment restrictions:

$$E[g(Z_t;\theta^0)] = 0 \tag{1}$$

where $\{Z_t\}_{t=1}^T$ are $\mathbb{R}^{d_z}$-valued random vectors, $g(.;\theta) : \mathbb{R}^{d_z} \times \Theta \mapsto \mathbb{R}^{d_g}$ is a known (up to θ) function, and $d_g \geq d_\theta$.

All expectations are taken with respect to the true joint distribution, call it F_T, of $\{Z_t\}_{t=1}^T$. Limits are taken by letting $T \to \infty$. Note that (1) allows for identification failures of θ^0 for a given T or asymptotically as in, e.g., Stock and Wright (2000), Kleibergen (2005), Antoine and Renault (2012), Andrews and Guggenberger (2014), etc.

Under this setup, to which we will subsequently impose certain standard restrictions on $\{F_T : T \geq 1\}$ but not identification of θ^0, the goal of this paper is to improve the conventional projection test for the null hypothesis

$$H_0 : R\theta^0 = r_0 \text{ against the alterative hypothesis } H : R\theta^0 \neq r_0 \tag{2}$$

where R is a fixed, full row-rank, $d_R \times d_\theta$ known matrix, and r_0 is a $d_R \times 1$ known vector.[1] The case for $d_R = d_\theta$ does not require a non-trivial projection, and is not considered. Instead we maintain that $d_R < d_\theta$ and $2 \leq d_\theta \leq d_g$.

The conventional projection test rejects H_0 at the level α if there does not exist any θ_0 satisfying $R\theta_0 = r_0$ in a $(1-\alpha)$-level confidence set for θ^0. Dufour (1997), Dufour and Jasiak (2001), Dufour et al. (2006), Dufour and Taamouti (2005, 2007), Dufour et al. (2013), etc. extensively document its usefulness under identification failure of θ^0. Given a confidence set for θ^0 with asymptotic coverage $1-\alpha$, the asymptotic size of the conventional projection test for H_0 in (2), based on this confidence set, cannot exceed α. Such confidence sets for θ^0, even under identification failure of θ^0, can be obtained with varying degree of computational ease by inverting, e.g., the S-test of Stock and Wright (2000), the K-test of Kleibergen (2005), the modifications of Moreira (2003)'s conditional likelihood ratio (CLR) test as in Kleibergen (2005), Andrews and Guggenberger (2014, 2015), etc.[2,3]

However, it is known that the conventional projection test can often be very conservative, and indeed needlessly so if there is no or mild (e.g., Hahn and Kuersteiner (2002), Caner (2010))) identification failure. Also, without a convenient way of imposing H_0, the test can in general be computationally inconvenient if d_θ is not very small.

We address these two problems – conservativeness and computational inconvenience – in the context of testing (2) by means of an improved projection (GMM-LM) test that is based on a function of Neyman (1959)'s C-alpha statistic and that uses the conventional fixed (χ^2) critical values.

The construction of the C-alpha statistic for testing H_0 in (2) is important in its own right beyond the paradigm

[1] The slightly awkward representation of H_0 is still in line with Section 9.1 of Newey and McFadden (1994). We use it to emphasize that we consider the true θ, and hence $R\theta$, as fixed but let the hypothesized value r_0 vary, possibly with sample size T, which is what determines if H_0 is true or false. Accordingly, the assumptions maintained in this paper focus on the fixed true θ^0. Contiguity arguments, when they appear, reflect local deviations of the null from the fixed truth. To avoid confusing the reader due to this, we do not use the term "local alternatives". We note that varying the r_0 in practice also directly leads to inversion of the concerned test.

[2] Also see Staiger and Stock (1997), Dufour (1997), Kleibergen (2002, 2007), Dufour and Taamouti (2005, 2007), Guggenberger and Smith (2005, 2008), Andrews et al. (2006), Otsu (2006), Mikusheva (2010), Beaulieu et al. (2013), among many others.

[3] On the other hand, for not to be over-sized, the conventional plug-in tests based on fixed critical values often crucially depend on consistent estimation of θ, which is not necessarily guaranteed under identification failure. Thus, the easy to conduct plug-in tests such as the Wald and score tests, and the computationally less easy plug-in quasi likelihood ratio test – see Newey and McFadden (1994) for precise definitions – can be badly over-sized. See Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Zivot et al. (1998), Stock and Wright (2000), Kleibergen (2002, 2004, 2005), Moreira (2003), Zivot et al. (2006), Guggenberger et al. (2012a), etc.

of projection tests or identification failure. In the special case of $R = [I_{d_R}, 0]$, the construction of the C-alpha statistic can be easily reconciled with the idea of the efficient score for the parameter $\beta := [I_{d_R}, 0]\theta$ treating the parameter $\gamma := [0, I_{d_\theta - d_R}]\theta$ as unknown, where the efficient score is the residual of the population regression of the score for β on the score for γ. To our knowledge, this reconciliation is not apparent from the literature on the C-alpha statistic for the case of a general R, as in (2), and in the context of moment conditions models, as in (1).

Our paper pays special attention to this reconciliation. We work with a generalization of Neyman's C-alpha statistic that has been studied extensively by Smith (1987) and Dagenais and Dufour (1991) (in the likelihood context), and then establish its numerical equivalence with the efficient score statistic in a re-parameterized model. As a consequence, the improved projection test can be recast in the re-parameterized model as the test in Chaudhuri and Zivot (2011) that was developed in Chaudhuri (2008) based on the original work of Robins (2004). (Also see Zivot and Chaudhuri (2009), Chaudhuri et al. (2010), Chaudhuri and Renault (2011), etc.) Subsequently, we allow for a very general structure for the identification failure of the elements of θ^0 following Antoine and Renault (2012) and Andrews and Guggenberger (2014), and study the asymptotic properties of the improved projection test.

The rest of the paper is organized as follows. In Section 2 we describe the improved projection test and the key idea behind it, we establish its numerical equivalence with the statistic considered in Chaudhuri and Zivot (2011) in a re-parameterized model, and provide an intuitive justification for its main asymptotic properties. We abstract from any identification failure in this section to fix ideas. Section 3 states the precise technical assumptions that are maintained in the paper. Section 3 also formally develops the asymptotic properties of the test and concludes by discussing the closely related recent literature. Proofs of all the technical results are collected in the Appendix.[4]

2 The Improved Projection test: Motivation and Definition

2.1 The key idea

The idea behind the reduction of – (a) conservativeness and (b) computational inconvenience – is best explained by maintaining the classical conditions from, e.g., Newey and McFadden (1994)'s Theorem 9.2 that, importantly, rule out any identification failure of θ^0. These conditions are referred to as the NM-9.2 conditions henceforth.

Under the NM-9.2 conditions, the efficient influence function for $R\theta^0$ is -$l(Z_t;\theta^0)$ (see Appendix A.1) where

$$l(Z_t;\theta) := R\left(G'(\theta)V^{-1}(\theta)G(\theta)\right)^{-1}G'(\theta)V^{-1}(\theta)g(Z_t;\theta),$$

$G(\theta) := \frac{\partial}{\partial\theta'}E\left[g(Z_t;\theta)\right]$, $V(\theta) := Var\left(g(Z_t;\theta)\right)$. Therefore, defining $\bar{g}_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} g(Z_t;\theta)$, the efficient GMM estimator of $R\theta^0$ has the asymptotically linear representation: $\sqrt{T}(\widehat{R\theta^0} - R\theta^0) = -\sqrt{T}l_T(\theta^0) + o_p(1)$ where

$$l_T(\theta) \quad := \quad \frac{1}{T}\sum_{t=1}^{T} l(Z_t;\theta) = R\left(G'(\theta)V^{-1}(\theta)G(\theta)\right)^{-1}G'(\theta)V^{-1}(\theta)\bar{g}_T(\theta).$$

[4] Given the numerical equivalence with the test statistic in Chaudhuri and Zivot (2011), for brevity we do not present any Monte Carlo results in this paper. Instead, we refer to Chaudhuri (2008), Zivot and Chaudhuri (2009), Chaudhuri et al. (2010), Chaudhuri and Zivot (2011) and Chaudhuri and Renault (2011) for extensive simulation evidence documenting the good finite-sample properties of the improved projection test. Thanks to the results established in Section 2, all such evidence apply directly to the test in our paper.

(a) The first part of the idea is that for local optimality, a test for H_0 can be based on an estimator of $l_T(\theta)$:

$$\widehat{l}_T(\theta) := R\left(\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)\right)^{-}\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\bar{g}_T(\theta)$$

where $\widehat{G}_T(\theta)\xrightarrow{P}G(\theta)$ and $\widehat{V}_T(\theta)\xrightarrow{P}V(\theta)$ uniformly in θ, at least in an open neighborhood of θ^0. $(.)^-$ stands for the g-inverse. When $d_R>1$, the test has to be based on a quadratic form of the standardized $\widehat{l}_T(\theta)$, i.e., on

$$LM_T(\theta) := T\times\widehat{l}'_T(\theta)\left(R\left(\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)\right)^{-}R'\right)^{-}\widehat{l}_T(\theta). \tag{3}$$

This is Smith (1987)'s LLM_T, Dagenais and Dufour (1991)'s PC or Newey and McFadden (1994)'s LM_{2n} statistic for testing linear restrictions. It falls under the class of Neyman (1959)'s C-alpha statistic.

In this paper we will always maintain that $V(\theta)$ is nonsingular, unlike in Andrews and Guggenberger (2015). However, in Section 3 (but not in Section 2) we will allow for the column-rank deficiency of $G(\theta)$ following Antoine and Renault (2012) and Andrews and Guggenberger (2014) to characterize the identification failure of θ. Our use of the g-inverse in the definitions of $\widehat{l}_T(\theta)$ and $LM_T(\theta)$ reflects our asymmetric treatment of $V(\theta)$ and $G(\theta)$.

Define $P(D):=D(D'D)^-D'$ as the projection matrix for any matrix D. If D is positive semidefinite then let $D^{1/2}$, upper triangular, be such that $D=D^{1/2'}D^{1/2}$. Then a familiar and equivalent representation of $LM_T(\theta)$ is

$$LM_T(\theta) := T\times\left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right)'P\left(\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)\left(\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)\right)^{-}R'\right)\left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right). \tag{4}$$

For illustration, let $d_\theta=2$ and $d_R=1$, and to characterize the null hypothesis H_0 consider two simple but common cases: (i) $R=[1,0]$ and (ii) $R=[1,1]$. Also, let $\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)$ be nonsingular (almost surely). Then it is straightforward to see that $LM_T(\theta)=t^2_{LM}(\theta)$ where, under cases (i) and (ii), $t_{LM}(\theta)$ is respectively:

$$(i)\ t_{LM}(\theta)=\frac{\widehat{\omega}_{22}(\theta)\xi_{T,1}(\theta)-\widehat{\omega}_{12}(\theta)\xi_{T,2}(\theta)}{\sqrt{\widehat{\omega}_{11}(\theta)\widehat{\omega}_{22}(\theta)-\widehat{\omega}^2_{12}(\theta)}\sqrt{\widehat{\omega}_{22}(\theta)}},\ (ii)\ t_{LM}(\theta)=\frac{(\widehat{\omega}_{22}(\theta)-\widehat{\omega}_{12}(\theta))\xi_{T,1}(\theta)+(\widehat{\omega}_{11}(\theta)-\widehat{\omega}_{12}(\theta))\xi_{T,2}(\theta)}{\sqrt{\widehat{\omega}_{11}(\theta)\widehat{\omega}_{22}(\theta)-\widehat{\omega}^2_{12}(\theta)}\sqrt{\widehat{\omega}_{11}(\theta)+\widehat{\omega}_{22}(\theta)-2\widehat{\omega}_{12}(\theta)}},$$

$\widehat{\omega}_{ij}(\theta)$ is the (i,j)-th element of $\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)$ and $\xi_{T,i}$ is the i-th row of $\widehat{G}'_T(\theta)\widehat{V}_T^{-1}(\theta)\sqrt{T}\bar{g}_T(\theta)$ for $i,j=1,2$.

(b) The second part of the idea is to re-parameterize the system in (1) in order to explicitly impose H_0 and thereby facilitate computation by allowing us to directly work with a reduced dimensional space from which we do the projection. Accordingly, consider a $(d_\theta-d_R)\times d_\theta$ matrix S such that the $d_\theta\times d_\theta$ matrix $A_S=[R',S']'$, indexed by S, is nonsingular. S exists since R is full row-rank (e.g., rows of S form a basis of the null space of R). Now, for this S, use the linear restrictions imposed by H_0 to rotate the original parameter vector θ and define:

$$(\beta',\gamma'_S)' := A_S\theta. \tag{5}$$

This rotation is different from that considered in Sargan (1983), Phillips (1989), and later in Choi and Phillips (1992), Zivot et al. (2006), Antoine and Renault (2009, 2012), Andrews and Cheng (2012, 2014), Cheng (2015), etc. The rotation in (5) isolates the directions in θ that are identified by the null hypothesis in (2) regardless of the identification that is due to the model (1), whereas in the aforementioned papers, a rotation is employed to isolate the directions identified by the model itself. Importantly, the rotation in (5) is enforceable in practice.

(1) and (5) imply that $\beta^0 := R\theta^0$ and $\gamma_S^0 := S\theta^0$ are the true values for β and γ_S. The parameter space for $(\beta', \gamma_S')'$ is $\mathcal{B} \times \Gamma_S$ where $\mathcal{B} := \{R\theta : \theta \in \Theta\} \subset \mathbb{R}^{d_R}$ and $\Gamma_S := \{S\theta : \theta \in \Theta\} \subset \mathbb{R}^{d_\theta - d_R}$. Obviously, by definition,

$$LM_T(\theta) \equiv LM_T\left(A_S^{-1}(\beta', \gamma_S')'\right).$$

(This is not the invariance to *reformulation* of H_0.) The same holds for all other quantities that are functions of θ.

2.2 An equivalence relation based on (5) through an alternative construction

The null hypothesis in terms of the new parameters $(\beta', \gamma_S')'$ is $H_0 : \beta^0 = r_0$. Thus β is the parameter of interest, and γ_S is an unknown key nuisance parameter. This directly fits into the framework of Chaudhuri and Zivot (2011) where the construction of a similar GMM-LM statistic adhered more closely to Neyman (1959)'s original C-alpha construction. By contrast, a similar adherence is not apparent from the construction of $LM_T(\theta)$ in (3). In this subsection we reconcile this difference by establishing a suitable equivalence relation between the two constructions.

Note that, given the choice of S in the re-parameterization (5), the scores for β and γ_s, by which we mean here the population version of the optimal rotations, in the efficient GMM sense, of $\bar{g}_T(\theta^0)$ along the directions of β^0 and γ_s^0, are $l_{\beta,S,T}(\theta^0) := \frac{1}{T}\sum_{t=1}^T l_{\beta,S}(Z_t;\theta^0)$ and $l_{\gamma_S,S,T}(\theta^0) := \frac{1}{T}\sum_{t=1}^T l_{\gamma_S,S}(Z_t;\theta^0)$ respectively, where

$$l_{\beta,S}(Z_t;\theta) := R_S^{1'} G'(\theta)V^{-1}(\theta)g(Z_t;\theta), \quad \text{and} \quad l_{\gamma_S,S}(Z_t;\theta) := S_S^{1'} G'(\theta)V^{-1}(\theta)g(Z_t;\theta),$$

and R_S^1 and S_S^1 are respectively $d_\theta \times d_R$ and $d_\theta \times (d_\theta - d_R)$ fixed, known matrices such that

$$A_S^{-1} = [R_S^1, S_S^1].$$

(Quantities dependent on the choice of S in (5) are indexed throughout by the subscript S. While the definition of $\beta := R\theta$ does not depend on S, the score for β in the re-parameterized model does depend on S through R_S^1.)

The residual from the population regression of $l_{\beta,S}(Z_t;\theta^0)$ on $l_{\gamma_S,S}(Z_t;\theta^0)$ is $l_{\beta.\gamma_S,S}(Z_t;\theta^0)$ where

$$l_{\beta.\gamma_S,S}(Z_t;\theta) := l_{\beta,S}(Z_t;\theta) - \left(R_S^{1'}\Omega(\theta)S_S^1\right)\left(S_S^{1'}\Omega(\theta)S_S^1\right)^{-1} l_{\gamma_S,S}(Z_t;\theta),$$

and $\Omega(\theta) := G'(\theta)V^{-1}(\theta)G(\theta)$. Under the NM-9.2 conditions, it follows that $\frac{1}{\sqrt{T}}\sum_{t=1}^T l_{\beta.\gamma_S}(Z_t;\theta^0) \xrightarrow{d} N(0, \Xi_S(\theta^0))$ where

$$\Xi_S(\theta) := \left(R_S^{1'}\Omega(\theta)R_S^1\right) - \left(R_S^{1'}\Omega(\theta)S_S^1\right)\left(S_S^{1'}\Omega(\theta)S_S^1\right)^{-1}\left(S_S^{1'}\Omega(\theta)R_S^1\right).$$

Let $\widehat{\Omega}_T(\theta) := \widehat{G}_T'(\theta)\widehat{V}_T^{-1}(\theta)\widehat{G}_T(\theta)$ be a consistent estimator of $\Omega(\theta)$ and accordingly, let

$$\begin{aligned}\widehat{\Xi}_{S,T}(\theta) &:= \left(R_S^{1'}\widehat{\Omega}_T(\theta)R_S^1\right) - \left(R_S^{1'}\widehat{\Omega}_T(\theta)S_S^1\right)\left(S_S^{1'}\widehat{\Omega}_T(\theta)S_S^1\right)^{-1}\left(S_S^{1'}\widehat{\Omega}_T(\theta)R_S^1\right) \\ &= R_S^{1'}\widehat{G}_T'(\theta)\widehat{V}_T^{-1/2'}(\theta)\left(I_{d_g} - P\left(\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)S_S^1\right)\right)\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)R_S^1\end{aligned}$$

be a consistent estimator for $\Xi_S(\theta)$. Then a sample version of $l_{\beta.\gamma_S,T}(\theta) := \frac{1}{T}\sum_{t=1}^{T} l_{\beta.\gamma_S}(Z_t;\theta)$ naturally is:

$$\widehat{l}_{\beta.\gamma_S,T}(\theta) = R_S^{1'}\widehat{G}_T'(\theta)\widehat{V}_T^{-1/2'}(\theta)\left(I_{d_g} - P\left(\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)S_S^1\right)\right)\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta),$$

and the GMM-LM statistic in Chaudhuri and Zivot (2011) is a quadratic form of the standardized $\widehat{l}_{\beta.\gamma_S,T}(\theta)$:

$$\begin{aligned} LM_{T,S}^{\text{alt}}(\theta) &:= T \times \widehat{l}'_{\beta.\gamma_S,T}(\theta)\ \widehat{\Xi}_{S,T}^{-1}(\theta)\ \widehat{l}_{\beta.\gamma_S,T}(\theta), \\ &= T \times \left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right)' P\left(\left(I_{d_g} - P\left(\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)S_S^1\right)\right)\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)R_S^1\right)\left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right) \end{aligned} \tag{6}$$

which they refer to as the efficient score statistic, since they refer to $l_{\beta.\gamma_S,T}(\theta)$ as the efficient score for β.

Lemma 2.1 *For a given $\theta \in$ interior(Θ) and $T \geq 1$, let $\widehat{\Omega}_T(\theta)$ be positive definite almost surely. Then $LM_{T,S}^{alt}(\theta)$ is numerically invariant almost surely for any choice of S for which $[R', S']'$ is nonsingular.*

This result builds on Dagenais and Dufour (1991). The other quantities defined so far are not invariant. Using Lemma 2.1 we then have an equivalence result between the two alternative constructions $LM_T(\theta)$ and $LM_{T,S}^{\text{alt}}(\theta)$.

Proposition 2.2 *For a given $\theta \in$ interior(Θ) and $T \geq 1$, let $\widehat{\Omega}_T(\theta)$ be positive definite almost surely. Then $LM_T(\theta) = LM_{T,S}^{alt}(\theta)$ almost surely for any choice of S for which $[R', S']'$ is nonsingular.*

The final equivalence relation in Proposition 2.2 mimics the same in the likelihood context for the LM (score) statistics constructed as a standardized quadratic form of either the efficient score function or the efficient influence function (the latter having a close and direct resemblance with the Wald statistic). In this sense it reaffirms the connection between the original C-alpha construction of Neyman (1959) and the subsequent ones in Smith (1987) and Dagenais and Dufour (1991). Although intuitively expected, to our knowledge, this reconciliation is new.

2.3 The projection tests

Our improved projection test for H_0 can be conducted in two steps as follows. For some $\epsilon, \alpha > 0$ and $\epsilon + \alpha < 1$:

$$\begin{aligned} &\text{Step 1: obtain a nominal } (1-\epsilon)\text{-level confidence set } CI_T(\gamma_S;\epsilon) \text{ for } \gamma_S^0; \\ &\text{Step 2: reject } H_0 \text{ if } CI_T(\gamma_S;\epsilon) \text{ is empty or if } \inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > \chi^2_{d_R}(1-\alpha) \end{aligned} \tag{7}$$

where $\chi^2_{d_R}(1-\alpha)$ is the $(1-\alpha)$-th quantile of a central χ^2 distribution with d_R degrees of freedom.

$CI_T(\gamma_S;\epsilon)$ can be obtained by inverting e.g. the S-test, the K-test, modifications of Moreira (2003)'s CLR test (see Kleibergen (2005), Andrews and Guggenberger (2014, 2015)) for γ_S, while treating $\beta = r_0$ as known. In practice, the operations required in steps 1 and 2 can be simultaneously conducted since, to fail to reject H_0, it is sufficient to find a single point γ_0 that would belong in $CI_T(\gamma_S;\epsilon)$ and also satisfy the condition from Step 2.

On the other hand, in this GMM-LM context the conventional projection test rejects H_0 at the level α if:

$$\inf_{\theta_0 \in \Theta: R\theta_0 = r_o} \widetilde{LM}_T(\theta_0) > \chi^2_{d_\theta}(1-\alpha) \quad \text{or equivalently,} \quad \inf_{\gamma_0 \in \Gamma_S} \widetilde{LM}_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > \chi^2_{d_\theta}(1-\alpha) \tag{8}$$

where (see the LM_{3n} statistic in Newey and McFadden (1994) and the K statistic in Kleibergen (2005)):

$$\widetilde{LM}_T(\theta) := T \times \left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right)' P\left(\widehat{V}_T^{-1/2}(\theta)\widehat{G}_T(\theta)\right)\left(\widehat{V}_T^{-1/2}(\theta)\bar{g}_T(\theta)\right). \tag{9}$$

2.4 Convenience in computation

The first version in (8) is the standard representation of the conventional projection score test. It is exactly the same as rejecting H_0 if there does not exist any θ_0 inside the confidence set $\{\theta \in \Theta : \widetilde{LM}_T(\theta) \leq \chi^2_{d_\theta}(1-\alpha)\}$ for θ such that $R\theta_0 = r_0$. We present the second version to exploit the re-parameterization in (5) and thereby impose H_0 in order to facilitate computation. In general, these tests have to be conducted by searching over a grid of, say, m points along each dimension – larger m gives better accuracy (ceteris paribus). The computational advantage of the second version is now evident by noting that while the order of magnitude of the numerical operations required by the first version is m^{d_θ}, it is only $m^{d_\theta - d_R}$ for the second one. Similar arguments imply that the computational advantage of the improved projection test in (7) over the first version in (8) is also of the same order of magnitude.

2.5 Reduction in conservativeness

Note from (3) and (9) that $\widetilde{LM}_T(\widetilde{\theta}_T) = LM_T(\widetilde{\theta}_T)$ where $\widetilde{\theta}_T$ is the restricted-by-H_0 GMM estimator of θ_T (see Appendix A.3). Under the NM-9.2 conditions (still maintained here), $LM_T(\widetilde{\theta}_T)$ converges in distribution to a central $\chi^2_{d_R}$ distribution if H_0 is true, and to a non-central $\chi^2_{d_R}$ distribution under local deviations from the truth (see (10) and note that $\sqrt{T}(\widetilde{\theta}_T - \theta^0)$ is still $O_p(1)$ under local deviations of H_0 from the truth). Hence, the needless part of the conservativeness of the conventional projection test, that we try to address, is due to its use of a critical value from the $\chi^2_{d_\theta}$ distribution, while the test statistic is itself $\inf_{\theta_0\in\Theta:R\theta_0=r_o}\widetilde{LM}_T(\theta_0) \leq \widetilde{LM}_T(\widetilde{\theta}_T) = LM_T(\widetilde{\theta}_T)$.

On the other hand, for any $(r_0', \gamma_0')'$ in a $\sqrt{T}$-neighborhood of $(\beta^{0\prime}, \gamma_S^{0\prime})'$ with probability approaching one – more precisely, for $r_0 := \beta^0 + \mu_\beta/\sqrt{T}$ and $\gamma_0 := \gamma_S^0 + \mu_{\gamma_S}/\sqrt{T}$ for some constant μ_β, and $\mu_{\gamma_S} = O_p(1)$ – we have:

$$\theta_0 := A_S^{-1}(r_0', \gamma_0')' = R_S^1(\beta^0 + \mu_\beta/\sqrt{T}) + S_S^1(\gamma_S^0 + \mu_{\gamma_S}/\sqrt{T}) = \theta^0 + (R_S^1\mu_\beta + S_S^1\mu_{\gamma_S})/\sqrt{T}.$$

The NM-9.2 conditions give: $\widehat{G}_T(\theta_0) \xrightarrow{P} G(\theta^0)$, $\widehat{V}_T(\theta_0) \xrightarrow{P} V(\theta^0)$ and, crucially,

$$\sqrt{T}\widehat{l}_T(\theta_0) \xrightarrow{P} \sqrt{T}l_T(\theta^0) + \mu_\beta$$

by using the orthogonality $RS_S^1 = 0$ following from $A_S A_S^{-1} = I_{d_\theta}$. Hence

$$\begin{aligned} \sqrt{T}l_T(\theta^0) \quad &\xrightarrow{d} \quad N\left(0, R\left(G'(\theta^0)V^{-1}(\theta^0)G(\theta^0)\right)^{-1}R'\right) \text{ and, therefore,} \\ LM_T(\theta_0) \quad &\xrightarrow{d} \quad \chi^2_{d_R} \text{ with non-centrality parameter } \mu_\beta'\left(R\left(G'(\theta^0)V^{-1}(\theta^0)G(\theta^0)\right)^{-1}R'\right)^{-1}\mu_\beta. \end{aligned} \tag{10}$$

Crucially, the $\sqrt{T}$-deviation of γ_0 from γ_S^0 does not matter for the asymptotic distribution of $LM_T(\theta_0)$.

Under the same conditions and those that ensure the existence of a consistent estimator for γ_S^0, it can be shown that $CI_T(\gamma_S;\epsilon)$ defined in (7) belongs in a $\sqrt{T}$-neighborhood of γ_S^0 with probability approaching one since the test statistic for the test inverting which the confidence set is obtained diverges to $+\infty$ uniformly in γ_S when evaluated at (r_0 and) γ_S outside the $\sqrt{T}$-neighborhood of γ_S^0. (To fix ideas, ignore empty $CI_T(\gamma_S;\epsilon)$ for now.) Thus, $\gamma^{\dagger}_{S,T} := \arg\inf_{\gamma_0\in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0',\gamma_0')'\right) = \gamma_S^0 + \mu_{\gamma_S,T}/\sqrt{T}$ for some $\mu_{\gamma_S,T} = O_p(1)$. Hence by (10), the improved projection test is asymptotically equivalent to the locally optimal infeasible test, based on the infeasible efficient influence function and the unknown true γ_S^0, that rejects H_0 at the level α if for $\theta_0^{\text{infeas}} := A_S^{-1}(r_0',\gamma_S^{0'})'$,

$$LM_T^{\text{infeas}}(\theta_0^{\text{infeas}}) := T \times l_T'(\theta_0^{\text{infeas}}) \left(R\left(G'(\theta_0^{\text{infeas}})V^{-1}(\theta_0^{\text{infeas}})G(\theta_0^{\text{infeas}})\right)^{-1} R'\right)^{-1} l_T(\theta_0^{\text{infeas}}) > \chi^2_{d_R}(1-\alpha). \quad (11)$$

Our paper allows for identification failure of θ^0, and the optimality discussion above is only for the special case of no identification failure. The next section presents a more general treatment of the improved projection test.

3 The Improved Projection Test: Asymptotic Properties

The discussion and the clean expressions from the last section, although intuitive, may not be valid under possible identification failures of θ^0 as characterized by, e.g., Stock and Wright (2000), Kleibergen (2005), Antoine and Renault (2012), Andrews and Guggenberger (2014), etc. There is already a vast literature discussing the general problems in such cases. Thus, it is important to discuss the asymptotic properties of the improved projection test allowing for possible identification failures of θ^0, a problem encountered in numerous empirical applications.

The choice of $\widehat{G}_T(\theta)$ in the definition of $LM_T(\theta)$ in (3) is important, and to account for a possible identification failure of θ^0 it is imperative that the choice is based on Kleibergen (2005):

$$\begin{aligned}\widehat{G}_T(\theta) &= \left[\widehat{G}_{1,T}(\theta),\ldots,\widehat{G}_{d_\theta,T}(\theta)\right],\\ \text{where } \widehat{G}_{T,j}(\theta) &= \frac{\partial}{\partial\theta_j}\bar{g}_T(\theta) - \widehat{V}_{j,g,T}(\theta)\widehat{V}_T^{-1}(\theta)\bar{g}_T(\theta),\end{aligned}$$

$\widehat{V}_{j,g,T}(\theta)$ and $\widehat{V}_T(\theta)$ are respectively $d_\theta \times d_g$ and $d_g \times d_g$ matrices, and θ_j is the j-th element of θ for $j = 1,\ldots,d_\theta$. In particular, for $j = 1,\ldots,d_\theta$, we would require $\widehat{V}_{j,g,T}(\theta)$ and $\widehat{V}_T(\theta)$ to be estimators of respectively,

$$V_{j,g}(\theta) := \lim_{T\to\infty} T \times E\left[\left(\frac{\partial}{\partial\theta_j}\bar{g}_T(\theta) - E\left[\frac{\partial}{\partial\theta_j}\bar{g}_T(\theta)\right]\right)\bar{g}_T(\theta)'\right] \text{ and } V(\theta) := \lim_{T\to\infty} T \times E\left[\left(\bar{g}_T(\theta) - E[\bar{g}_T(\theta)]\right)\bar{g}_T(\theta)'\right],$$

provided they exist. Also applicable are the variety of choices of $\widehat{G}_T(\theta)$ considered in Guggenberger and Smith (2005, 2008) that only deviate from $\widehat{G}_T(\theta)$ defined above by an order of magnitude of $o_p(1/\sqrt{T})$.

We maintain high-level but standard assumptions on the joint distribution F_T of the data $\{Z_t\}_{t=1}^T$. Allowing for a drifting data generating process (DGP) in what follows is important, and to emphasize it we index by T the key parameters defined in terms of F_T; see, e.g., Stock and Wright (2000), Andrews and Guggenberger (2014).

We repeat here that irrespective of the drifting DGP $\{F_T\}$, but consistent with the GMM literature, we take the true value θ^0 satisfying the moment restrictions in (1) as fixed. The null H_0 in (2) is true if the hypothesized value r_0 is equal to $R\theta^0$, it is false otherwise. Apart from characterizing the false H_0 by locally deviating (to be made precise later) r_0 from $R\theta^0$, no other maintained assumptions involve r_0. For convenience, we maintain that:

Assumption O:

$\theta^0 \in \text{int}(\Theta)$ where Θ is compact in $\mathbb{R}^{d_\theta}$ and $\text{int}(\Theta) := \text{interior}(\Theta)$.

Notation: We suppress the triangular array $\{Z_{t,T} : t = 1, \ldots, T; T \geq 1\}$ notation, and instead denote $Z_{t,T}$ by Z_t. For any matrix D, define $\|D\| := \sqrt{\text{trace}(D'D)}$ as the matrix norm. For any $a \times b$ matrix $D = [D_1, \ldots, D_b]$ define $D_{(j:k)} := [D_j, \ldots, D_l]$ for $1 \leq j \leq k \leq b$. Allow for $D_{(1:0)}$ and $D_{(b+1:b)}$ to be empty matrices with no column. For an $(ab) \times 1$ vector $D = [d_1, \ldots, d_{ab}]'$, define $devec_b(D) := [(d_1, \ldots, d_b)', (d_{b+1}, \ldots, d_{2b})', \ldots, (d_{(a-1)b+1}, \ldots, d_{ab})']$. To impose generic bounds on quantities uniformly with respect to T or sometimes with respect to θ, we use the quantities $\underline{c}$ and $\bar{c}$ where $\underline{c}$ and $\bar{c}$ are small and large (respectively), fixed, positive, real numbers.

3.1 Rejection of the true null hypothesis by the improved projection test

Assumption M:

M1. $g(z;\theta)$ is differentiable in θ at θ^0 for each $z \in \mathbb{R}^{d_z}$. For $\bar{G}_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \frac{\partial}{\partial \theta'} g(Z_t;\theta)$:

$$\begin{bmatrix} \sqrt{T}\bar{g}_T(\theta^0) \\ \sqrt{T}vec(\bar{G}_T(\theta^0) - E_T[\bar{G}_T(\theta^0)]) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \psi \\ \psi_G \end{bmatrix} \sim N(0, \Sigma)$$

where

$$\lim_{T\to\infty} Var_T \begin{pmatrix} \sqrt{T}\bar{g}_T(\theta^0) \\ \sqrt{T}vec(\bar{G}_T(\theta^0)) \end{pmatrix} \equiv \lim_{T\to\infty} \begin{bmatrix} V_T & V_{gG,T} \\ V_{Gg,T} & V_{GG,T} \end{bmatrix} = \Sigma := \begin{bmatrix} V & V_{gG} \\ V_{Gg} & V_{GG} \end{bmatrix}.$$

M2. $\|(E_T[\bar{G}_T(\theta^0)], V_T, V'_{Gg,T})\| \leq \bar{c}$, $\|\widehat{V}_T - V_T\| + \|\widehat{V}_{Gg,T} - V_{Gg,T}\| = o_p(1)$, and V are positive-definite.

To characterize the possible identification failure, let $G_T := E_T[\bar{G}_T(\theta^0)]$ and following Andrews and Guggenberger (2014)'s generalization of Kleibergen (2005)'s setup, consider the singular value decomposition of $V_T^{-1/2}G_T$:

$$V_T^{-1/2}G_T = C_T\bar{\Delta}_T B_T' \tag{12}$$

where C_T and B_T are non-random $d_g \times d_g$ and $d_\theta \times d_\theta$ orthogonal matrices whose columns are respectively the eigen-vectors of the matrices $V_T^{-1/2}G_T G_T' V^{-1/2'}$ and $G_T' V^{-1} G_T$. The $d_g \times d_\theta$ non-random matrix $\bar{\Delta}_T := [\Delta_T, 0]'$ where $\Delta_T := \text{diag}\left(\delta_{T,1}, \ldots, \delta_{T,d_\theta}\right)$ is the $d_\theta \times d_\theta$ diagonal matrix with its diagonal elements (always from the top) as $\delta_{T,1} \geq \delta_{T,2} \geq \ldots \geq \delta_{T,d_\theta}$ (≥ 0, without loss of generality) as the singular values of $V_T^{-1/2}G_T$.

Assumption M: (continued)

M3. For the singular value decomposition in (12), there exists a $p \in \{0, 1, \ldots, d_\theta\}$ such that:

(a) $\delta_{T,j} \to \delta_j$, a constant, and $\sqrt{T}\delta_{T,j} \to \infty$ for $j = 1, \ldots, p$ as $T \to \infty$ (this assumption is void if $p = 0$);

(b) $\sqrt{T}\delta_{T,j} \to l_j$, a constant, for $j = p+1, \ldots, d_R$ as $T \to \infty$ (this assumption is void if $p = d_\theta$);

(c) $C_T \to C$ and $B_T \to B$ as $T \to \infty$ where B is a nonsingular matrix;

(d) The $d_g \times d_\theta$ matrix $G^* := [C_{(1:p)}, C_{(p+1:d_\theta)}L + V^{-1/2}(\theta^0)devec_{d_g}(\psi_G - V_{Gg}(\theta^0)V^{-1}(\theta^0)\psi)B_{(p+1:d_\theta)}]$ has full column-rank d_θ almost surely, where $L := \text{diag}\left(l_{p+1}, \ldots, l_{d_\theta}\right)$ is a $(d_\theta - p) \times (d_\theta - p)$ diagonal matrix with $l_{p+1}, \ldots, l_{d_\theta}$ as its diagonal elements if $p < d_\theta$, and L is empty if $p = d_\theta$.

Remarks: p is the number of directions in θ that are better than weakly identified in the sense of Kleibergen (2005). The remaining $d_\theta - p$ directions in θ are at best weakly identified and necessitate the particular choice of $\widehat{G}_T(\theta)$. Assumption M3 and the representation involved in it are entirely based on the original work of Andrews and Guggenberger (2014) who systematically and rigorously develop the final assumption M3(d) in their Lemma 8.3(d) from primitive sufficient conditions; see, e.g., their equations (3.9), (3.10), Lemma 15.1, Corollary 15.2.[5] The other assumptions, O, M1 and M2, are standard; see, e.g., Kleibergen (2005), Guggenberger and Smith (2005).

Lemma 3.1 *Let assumptions O and M1-M3 hold. Then for $LM_T(\theta^0)$ defined in* (3), $LM_T(\theta^0) \xrightarrow{d} \chi^2_{d_R}$.

Proposition 3.2 *Let null hypothesis H_0 in* (2) *be true, i.e., $r_0 = R\theta^0$ for θ^0 defined in* (1). *Let the joint distribution $\{F_T : T \geq 1\}$ of $\{Z_t\}_{t=1}^T$ be constrained by the assumptions O and M1-M3. Let $\epsilon, \alpha > 0$ and $\epsilon + \alpha < 1$. Let $CI_T(\gamma_S; \epsilon)$ be a confidence set for γ_S defined in* (5) *with asymptotic coverage* $(1 - \epsilon)$ *for $\gamma_S^0 := S\theta^0$. Then the probability with which the improved projection test in* (7) *rejects H_0 cannot exceed* $(\epsilon + \alpha)$ *asymptotically.*

Remarks:

1. This is the most general and practically useful result on the improved projection test in this paper. It follows by Bonferroni's inequality applied to Lemma 3.1 and the asymptotic coverage of $CI_T(\gamma_S; \epsilon)$.[6] Importantly, the upper bound $(\epsilon + \alpha)$ for the rejection probability of the true H_0 is entirely under the control of the user.

2. An example of the first-step confidence set $CI_T(\gamma_S; \epsilon)$ that possesses the desired suitable property is:

$$CI_T^{SW}(\gamma_S; r_0, \epsilon) := \left\{\gamma_0 \in \Gamma_S : T \times Q_T(A_S^{-1}(r_0', \gamma_0)') \leq \chi^2_{d_g}(1 - \epsilon)\right\} \tag{13}$$

where the superscript SW stands for Stock and Wright (2000) who proposed such confidence sets based on the S-test (a non-linear generalization of the Anderson-Rubin test), r_0 is the hypothesized value of β that is implied by H_0 in (2) under the re-parameterization in (5), and

$$Q_T(\theta) := \bar{g}_T'(\theta)\widehat{V}_T^{-1}(\theta)\bar{g}_T(\theta) \tag{14}$$

is the standard continuously updated (CU-) GMM criterion function. Theorem 2 of Stock and Wright (2000) establishes that the asymptotic coverage of $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ for $\gamma_S^0 := S\theta^0$ is $(1 - \epsilon)$ when H_0 in (2) is true and when: (a) $\sqrt{T}\bar{g}_T(\theta^0) \xrightarrow{d} \psi$ and (b) $\widehat{V}_T(\theta^0) \xrightarrow{P} V(\theta^0)$. Since (a) and (b) are already assumed under M1 and M2, the asymptotic coverage for $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ holds under weaker conditions than what we maintain here.

Following Chaudhuri and Zivot (2011), we recommend the use of $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ in practice because of its: (i) validity under weak and general conditions, (ii) computational simplicity, and (iii) effectiveness in eliminating certain spurious declines in power of the GMM-LM test from the second step of our test. The ϵ in the upper bound in

[5] The convergence $\delta_{T,j} \to \delta_j$ for $j = 1, \ldots, p$ in M3(a), and $B_T \to B$ instead of $B_{T,(p+1:d_\theta)} \to B_{(p+1:d_\theta)}$ in M3(c) as $T \to \infty$ are slightly stronger than in Andrews and Guggenberger (2014). Strictly speaking, they are not necessary for the results in this section, but helps to avoid certain peripheral complications arising from the fact that $d_R < d_\theta$ (the main complications are addressed head on).

[6] While the formulation of the problem and the subsequence argument in Andrews and Guggenberger (2014) could be employed to state this result as the upper bound on the asymptotic size (limit of the exact size) of the improved projection test, since this is only an upper bound and the test is not asymptotically similar at this level of generality, we take a less rigourous approach for brevity.

Proposition 3.2 is, in practice, primarily due to the fact that $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ can be empty with nonzero probability (that increases with ϵ). This feature is actually useful for (iii) and also for (ii), and hence is accommodated in the definition of the improved projection test in (7). Thus, our recommendation is in spite of the concern raised in Davidson and MacKinnon (2014) and Muller and Norets (2016) (page 2184) that $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ does not properly reflect the parameter uncertainty, since this concern is at least partly addressed by the second step of our test.

Other choices of $CI_T(\gamma_S;\epsilon)$ include those proposed by Kleibergen (2005) based on the GMM-LM principle and on Moreira (2003)'s conditional likelihood ratio principle. See Andrews and Guggenberger (2014) for precise conditions under which they possess the property suitable for Proposition 3.2. When this happens, these confidence sets, by definition, always contain the CU-GMM estimator of γ_S restricted by H_0 in (2), i.e.,

$$\widehat{\gamma}_{S,T}(r_0) := \arg\min_{\gamma\in\Gamma_S} Q_T(A_S^{-1}(r_0',\gamma')') \equiv \arg\min_{\gamma\in\Gamma_S} Q_T(R_S^1 r_0 + S_S^1\gamma). \tag{15}$$

(If non-empty, then $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ also always contains $\widehat{\gamma}_{S,T}(r_0)$.) Sometimes the upper bound in Proposition 3.2 can be sharpened to α by the use of such confidence sets.[7] However, based on our experience with simulations, albeit under limited scenarios, we still prefer the use of $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ in practice for reasons (i)-(iii) stated above.[8]

3.2 Rejection of the false null hypothesis by the improved projection test

Without identification failure, Section 2.5 demonstrated that when the null deviates ($\sqrt{T}-$) locally from the truth, the improved projection test is asymptotically equivalent to the infeasible test that rejects H_0 if $LM_T^{\text{infeas}}(\theta_0^{\text{infeas}}) > \chi^2_{d_R}(1-\alpha)$ provided that $CI_T(\gamma_S;\epsilon)$ belongs in a $\sqrt{T}$-neighborhood of γ_S^0 with probability approaching one.

The purpose of this section is to allow for identification failure and still obtain analogous results. To use contiguity arguments reflecting local deviations, we rule out weak or worse identification of θ^0. In terms of assumption M3, it means $p = d_\theta$. With this condition imposed, the characterization of identification failure in M3 does not provide much additional generality over the characterization of identification failure in Antoine and Renault (2012), since taking $p = d_\theta$ rules out the cases that Andrews and Guggenberger (2014) refer to as "joint weak identification" in their equations (2.1)(ii) and (2.5)(iv). Furthermore, the local nature of M3 does not allow us to determine the distance from γ_S^0 of an arbitrary sequence of points inside $\{CI_T(\gamma_S;\epsilon) : T \geq 1\}$, which is essential for establishing the desired asymptotic equivalence result. For these reasons, the characterization in Antoine and Renault (2012), that directly models $E_T[\bar{g}_T(\theta)]$ for $\theta \in \Theta$ globally to characterize identification failure (better than weak), seems appropriate for our purpose. Accordingly, for some $\rho : \Theta \mapsto \mathbb{R}^{d_g}$ and a sequence of diagonal matrices $\{\Lambda_T : T \geq 1\}$ (to be made precise below), let

$$E_T\left[\bar{g}_T(\theta)\right] = \frac{\Lambda_T}{\sqrt{T}}\rho(\theta). \tag{16}$$

[7] For example, consider the canonical case $R = [I_{d_R}, 0]$, $S = [0, I_{d_\theta - d_R}]$ and let $\gamma_S := S\theta$ be strongly identified. Under a special case of our setup and in this context as in Chaudhuri and Zivot (2011), Theorem 2 of Kleibergen (2005) and Theorem 6 in Guggenberger and Smith (2005) show that $LM_T(r_0, \widehat{\gamma}_{S,T}(r_0)) \xrightarrow{d} \chi^2_{d_R}$ under H_0 (and when M4 holds). Then, exactly following the steps in the current proof of our Proposition 3.2 but with γ_S^0 replaced by $\widehat{\gamma}_{S,T}(r_0)$, it can be shown that the upper bound in Proposition 3.2 is α.

[8] While it is clear that $CI_T(\gamma_S;\epsilon)$ based on Kleibergen (2005)'s GMM-LM principle cannot be helpful for (iii) in general, it should be noted that $CI_T(\gamma_S;\epsilon)$ based on Moreira (2003)'s conditional likelihood ratio principle may not also be helpful for (iii). Simulation evidence and discussion on this can be found in Section 7.2.1 of Andrews (2016b), and empirical evidence can be found in Figure 2 of Chaudhuri and Rose (2009) (Supplemental Appendix). Both such $CI_T(\gamma_S;\epsilon)$'s can also be less appealing in terms of (i) and (ii).

Notation: For any $1 \times c$ vector $a = (a_1, \ldots, a_c)$, let $\text{diag}(a)$ denote the $c \times c$ diagonal matrix with a_i as its i-th diagonal element for $i = 1, \ldots, c$. Let 1_c denote the $1 \times c$ vector with all elements equal to 1.

Assumption N:

N1. $\rho(\theta) = 0$ if and only if $\theta = \theta^0$.

N2. $\psi_T(\theta) := \sqrt{T}\left(\bar{g}_T(\theta) - E_T\left[\bar{g}_T(\theta)\right]\right) \Rightarrow \psi(\theta)$ where $\psi(\theta)$ is a Gaussian process on Θ with mean zero and covariance function $E[\psi(\theta_1)\psi(\theta_2)'] = V(\theta_1, \theta_2)$, and $V(\theta^0, \theta^0) = V$ (as in M1).

N3. $\{\Lambda_T : T \geq 1\}$ is a deterministic sequence of $d_g \times d_g$ diagonal matrix with positive diagonal elements. I^* is a $d_g \times d_g$ matrix whose rows are a suitable permutation of the rows of I_{d_g} giving $I^* \Lambda_T I^{*'} = \text{diag}(\lambda_{T,1} 1_{k_1}, \ldots, \lambda_{T,l} 1_{k_l})$ where $k_j > 0$ for $j = 1, \ldots, l$ and $\sum_{j=1}^{l} k_j = d_g$, and, furthermore, such that $\lambda_{T,j} = o(\lambda_{T,j+1})$ for $j = 1, \ldots, l-1$, $\lim_T \lambda_{T,1} = \infty$ but $\lim_T \lambda_{T,l}/\sqrt{T} < \infty$.[9]

N4. The $d_g \times d_\theta$ matrix $\rho_\theta(\theta) := \frac{\partial}{\partial \theta'} \rho(\theta)$ exists and is continuous in $\theta \in \text{int}(\Theta)$. $\rho_\theta(\theta^0)$ is full column-rank d_θ.

N5. $g(z; \theta)$ is differentiable in $\theta \in \text{int}(\Theta)$ for each $z \in \mathbb{R}^{d_z}$.

N6. $\frac{\partial}{\partial \theta'} \psi_T(\theta^0) = \sqrt{T}\left[\bar{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}} \rho_\theta(\theta^0)\right] = O_p(1)$.

Remarks: Assumptions N1-N6 restate assumptions 1, 2, 3 and 5 of Antoine and Renault (2012) excluding their assumptions 3(iv) and 5(i), whose counterparts will be introduced later, and keeping their assumption 3(iii) local.

A crucial assumption in Antoine and Renault (2012) is the orthogonality condition in their assumption 6(i). They provide an extensive discussion of it, along with sufficient conditions (assumption 6*(i)), and relate it to the orthogonality condition in Andrews (1994). Our presentation deviates here slightly in that we will try to note the tradeoff between the smoothness of the moment vector with respect to θ, and the slowest rate at which the expected moment vector, after a suitable rotation, moves away from zero when θ moves away from θ^0. (Due to this suitable rotation (made precise below), this rate is not necessarily $\lambda_{T1}/\sqrt{T}$ (from N3).) It is important to recognize the aforementioned tradeoff; and our presentation tries to make it clear why, for example, in the case of linear instrumental variables regression (a common use of GMM) that involves the most smoothness, the slowest allowed rate of deviation could be anything faster than $T^{-1/2}$, while under the standard conditions of nonlinear GMM (as in Antoine and Renault (2009, 2012)) this rate should typically be faster than $T^{-1/4}$.

To proceed, we will first need to characterize the local deviation of the null from the truth such that the directions of the deviation are "efficient" in the sense of Antoine and Renault (2012). It is useful for this characterization (and also in various proofs of our results including that of Lemma 3.1) to consider the following constructions that are adapted from the original work of Antoine and Renault (2012), Andrews and Cheng (2014), Cheng (2015), etc.

Let $\{W_T = [W_{T,1}, \ldots, W_{T,m_T}] : T \geq 1\}$ be a sequence of $r \times c$ (for some r, c) matrix of full row-rank $r (\leq c)$ where $W_{T,j}$ is $r \times c_{T,j}$ (and empty if $c_{T,j} = 0$) for $j = 1, \ldots, m_T$ and such that $\sum_{j=1}^{m_T} c_{T,j} = c$ for each $T \geq 1$.

UBT-Construction: An upper block-triangular (UBT) construction

We construct a sequence of $r \times r$ matrix $\{\Pi_T = [\Pi_{T,1}, \ldots, \Pi_{T,m_T}] : T \geq 1\}$ such that the $c \times r$ matrix $W_T' \Pi_T$ has an upper block-triangular structure for each $T \geq 1$. For any given T, the following steps give such a Π_T.

[9] $I^{*-1} = I^{*'}$. I^* is not unique unless $k_1 = \ldots = k_l = 1$ and thus $l = d_g$. The multiplicity of the elements can be made dependent on T and θ at the cost of significantly involved notation, but such generalizations may not be relevant in practice.

- Let $\text{rank}(W_{T,m_T}) = c^*_{T,m_T} \leq \min(r, c_{m_T})$. Define Π_{T,m_T} as the $r \times c^*_{T,m}$ matrix such that its columns form an orthogonal basis for the column space of W'_{T,m_T}. Stop if $m_T = 1$, otherwise proceed to the next step.
- Let $\text{rank}([W_{T,m_T-1}, W_{T,m_T}]) - \text{rank}(W_{T,m_T}) = c^*_{T,m_T-1} \leq \min(r, c_{m_T-1})$. Define Π_{T,m_T-1} as the $c \times c^*_{T,m_T-1}$ matrix such that the columns of $[\Pi_{T,m_T-1}, \Pi_{T,m_T}]$ form an orthogonal basis for the column space of $[W_{T,m_T-1}, W_{T,m_T}]'$. Stop if $m_T = 2$, otherwise proceed to the next step.
- Continue step-by-step, as above, for $j = m_T - 2, \ldots, 1$ and for each j, define $\Pi_{T,j}$ as the $r \times c^*_{T,j}$ matrix, where $c^*_{T,j} = \text{rank}([W_{T,j}, \ldots, W_{T,m_T}]) - \text{rank}([W_{T,j+1}, \ldots, W_{T,m_T}) \leq \min(r, c_{T,j})$, such that the columns of $[\Pi_{T,j}, \ldots, \Pi_{T,m_T}]$ form an orthogonal basis for the column space of $[W_{T,j}, \ldots, W_{T,m_T}]'$.

As a convention we take $\Pi_{T,j}$ as an empty matrix if $c^*_{T,j} = 0$. Π_T is an orthogonal matrix by construction and

(i) for some integer $q_T \in \{1, \ldots, \min(r, m_T)\}$, the q_T blocks $W'_{T,j_{k,T}} \Pi_{T,j_{k,T}}$ for $k = 1, \ldots, q_T$, and where $1 \leq j_{1,T} < \ldots < j_{q_T,T} \leq m_T$, each has full column-rank $c^*_{T,j_{k,T}} > 0$ satisfying $\sum_{k=1}^{q_T} c^*_{T,j_{k,T}} = r$;

(ii) $W'_{T,j} \Pi_{T,k} = 0$, a zero matrix of suitable (according to the above) dimension, for all $1 \leq k < j \leq m_T$.

LBT-Construction: A lower block-triangular (LBT) construction

We construct a sequence of $r \times r$ matrix $\{\Pi_T = [\Pi_{T,1}, \ldots, \Pi_{T,m_T}] : T \geq 1\}$ such that the $c \times r$ matrix $W'_T \Pi_T$ has a lower block-triangular structure for each $T \geq 1$. For any given T, the following steps give such a Π_T. (This is essentially the same as the UBT-Construction, but in reverse order. Hence to save new notation, we continue to use the same notation as in the UBT-Construction and hope that this is not confusing.)

- Let $\text{rank}(W_{T,1}) = c^*_{T,1} \leq \min(r, c_1)$. Define $\Pi_{T,1}$ as the $r \times c^*_{T,1}$ matrix such that its columns form an orthogonal basis for the column space of $W'_{T,1}$. Stop if $m_T = 1$, otherwise proceed to the next step.
- Let $\text{rank}([W_{T,1}, W_{T,2}]) - \text{rank}(W_{T,1}) = c^*_{T,2} \leq \min(r, c_2)$. Define $\Pi_{T,2}$ as the $c \times c^*_{T,2}$ matrix such that the columns of $[\Pi_{T,1}, \Pi_{T,2}]$ form an orthogonal basis for the column space of $[W_{T,1}, W_{T,2}]'$. Stop if $m_T = 2$, otherwise proceed to the next step.
- Continue step-by-step, as above, for $j = 3, \ldots, m_T$ and for each j, define $\Pi_{T,j}$ as the $r \times c^*_{T,j}$ matrix, where $c^*_{T,j} = \text{rank}([W_{T,1}, \ldots, W_{T,j}]) - \text{rank}([W_{T,1}, \ldots, W_{T,j-1}) \leq \min(r, c_{T,j})$, such that the columns of $[\Pi_{T,1}, \ldots, \Pi_{T,j}]$ form an orthogonal basis for the column space of $[W_{T,1}, \ldots, W_{T,j}]'$.

As a convention we take $\Pi_{T,j}$ as an empty matrix if $c^*_{T,j} = 0$. Π_T is an orthogonal matrix by construction and

(i) for some integer $q_T \in \{1, \ldots, \min(r, m_T)\}$, the q_T blocks $W'_{T,j_{k,T}} \Pi_{T,j_{k,T}}$ for $k = 1, \ldots, q_T$, and where $1 \leq j_{1,T} < \ldots < j_{q_T,T} \leq m_T$, each has full column-rank $c^*_{T,j_{k,T}} > 0$ satisfying $\sum_{k=1}^{q_T} c^*_{T,j_{k,T}} = r$;

(ii) $W'_{T,j} \Pi_{T,k} = 0$, a zero matrix of suitable (according to the above) dimension, for all $1 \leq j < k \leq m_T$.

The local deviation of the null from the truth:

We consider the local deviation (of r_0 from the truth $R\theta^0$) that is efficient in the sense of Antoine and Renault (2012). Roughly speaking, it captures the direction along which the local asymptotic power of the improved projection test increases at the fastest rate. To obtain this deviation we apply the UBT and LBT constructions, in that order, as follows. First, let $\rho_\theta := \rho_\theta(\theta^0)$ (i.e., $\partial\rho(\theta^0)/\partial\theta'$), and by using N3 write $I^* \Lambda_T \rho_\theta \equiv I^* \Lambda_T I^{*'} I^* \rho_\theta = \left[\lambda_{T,1} \rho'_{\theta,1}, \ldots, \lambda_{T,l} \rho'_{\theta,l}\right]'$ where $\rho_{\theta,j}(\theta)$ is a $k_j \times d_\theta$ matrix for $j = 1, \ldots, l$. Take $W_T = \left[\rho'_{\theta,1}, \ldots, \rho'_{\theta,l}\right] = (I^* \rho_\theta)'$ (not depending on T) in the UBT-Construction. Thus $r = d_\theta$, $c = d_g$ and $m_T = l$ in terms of the notation from the UBT-Construction. W_T is full row-rank r (i.e. d_θ) by N4.

$$\Pi_{\rho_\theta} = [\Pi_{\rho_\theta,1}, \ldots, \Pi_{\rho_\theta,l}] \text{ is the } d_\theta \times d_\theta \text{ matrix } \Pi_T \text{ from the UBT-Construction with } W_T = \left(I^* \rho_\theta(\theta^0)\right)'. \tag{17}$$

Let $c^*_{\rho_\theta,j} \equiv c^*_{\rho_\theta,T,j} = c^*_{T,j} \geq 0$ denote the number of columns of $\Pi_{\rho_\theta,j}$ for $j = 1, \ldots, l$, and $q_{\rho_\theta} \equiv q_{\rho_\theta,T} = q_T$ from (i) in this UBT construction. Let $(j_1, \ldots, j_{q_{\rho_\theta}})$ denote the indices such that the block $\rho_{\theta,j_i}\Pi_{\rho_\theta,j_i}$ of dimension $k_{j_i} \times c^*_{\rho_\theta,j_i}$ is full column-rank $c^*_{\rho_\theta,j_i} > 0$ for $i = 1, \ldots, q_{\rho_\theta}$ and $\sum_{i=1}^{q_{\rho_\theta}} c^*_{\rho_\theta,j_i} = d_\theta$. Thus, the corresponding block of $I^*\Lambda_T I^{*'} I^* \rho_\theta(\theta^0)\Pi_{\rho_\theta}$ is $\lambda_{T,j_i}\rho_{\theta,j_i}\Pi_{\rho_\theta,j_i}$ and, correspondingly, the columns from $(d_\theta - \sum_{i'=i}^{q_{\rho_\theta}} c^*_{\rho_\theta,j_{i'}})$ to $(d_\theta - \sum_{i'=i}^{q_{\rho_\theta}} c^*_{\rho_\theta,j_{i'}} + c^*_{\rho_\theta,j_i})$ for $i = 1, \ldots, q_{\rho_\theta}$ of $I^*\Lambda_T I^{*'} I^* \rho_\theta(\theta^0)\Pi_{\rho_\theta}$ are represented by the $d_g \times c^*_{\rho_\theta,j_i}$ matrix: $[\lambda_{T,1}(\rho_{\theta,1}\Pi_{\rho_\theta,1})', 0']'$ if $j_i = 1$, and $[\lambda_{T,1}(\rho_{\theta,1}\Pi_{\rho_\theta,j_i})', \ldots, \lambda_{T,j_i}(\rho_{\theta,j_i}\Pi_{\rho_\theta,j_i})', 0']'$ otherwise. In both cases: $j_i = 1$ and $j_1 > 1$, the 0's inside the big matrices denote sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding big matrix is d_g.

Now, conforming to this above structure, define a $d_\theta \times d_\theta$ matrix D_{T,ρ_θ} as

$$D_{T,\rho_\theta} := \sqrt{T}\text{diag}\left(\lambda^{-1}_{T,j_1} 1_{c^*_{\rho_\theta,j_1}}, \ldots, \lambda^{-1}_{T,j_{q_{\rho_\theta}}} 1_{c^*_{\rho_\theta,j_{q_{\rho_\theta}}}}\right) \tag{18}$$

such that

$$\frac{1}{\sqrt{T}} I^*\Lambda_T I^{*'} I^* \rho_\theta(\theta^0)\Pi_{\rho_\theta} D_{T,\rho_\theta} \to G^\dagger \text{ as } T \to \infty \tag{19}$$

where $G^\dagger$ is a $d_g \times d_\theta$ finite matrix of full column-rank, and its columns from $(d_\theta - \sum_{i'=i}^{q_{\rho_\theta}} c^*_{\rho_\theta,j_{i'}})$ to $(d_\theta - \sum_{i'=i}^{q_{\rho_\theta}} c^*_{\rho_\theta,j_{i'}} + c^*_{\rho_\theta,j_i})$ for $i = 1, \ldots, q_{\rho_\theta}$ are represented by the $d_g \times c^*_{\rho_\theta,j_i}$ matrix: $[(\rho_{\theta,1}\Pi_{\rho_\theta,1})', 0']'$ if $j_i = 1$, and $[0', (\rho_{\theta,j_i}\Pi_{\rho_\theta,j_i})', 0']'$ otherwise (the use of 0's to denote 0 sub-matrices follow the same convention as above (18)). Define the $d_g \times d_\theta$ finite matrix of full column-rank G^* as:

$$G^* := I^{*'} G^\dagger. \tag{20}$$

Now take $W_T = R\Pi_{\rho_\theta} = [W_{T,1} = R\Pi_{\rho_\theta,j_1}, \ldots, W_{T,q_{\rho_\theta}} = R\Pi_{\rho_\theta,j_{\rho_\theta}}]$ (not depending on T) in the LBT-Construction. (Note that $(j_1, \ldots, j_{q_{\rho_\theta}})$ are the indices defined immediately below (17).) Thus $r = d_R$, $c = d_\theta$ and $m_T = q_{\rho_\theta}$. W_T is full row-rank by the definition of R, Π_{ρ_θ} and Lemma 3.8 (in Appendix B).

$$\Pi_R = [\Pi_{R,1}, \ldots, \Pi_{R,q_{\rho_\theta}}] \text{ is the } d_R \times d_R \text{ matrix } \Pi_T \text{ from the LBT-Construction with } W_T = R\Pi_{\rho_\theta}. \tag{21}$$

Let $c^*_{R,j} \equiv c^*_{R,T,j} = c^*_{T,j} \geq 0$ denote the number of columns of $\Pi_{R,j}$ for $j = 1, \ldots, q_{\rho_\theta}$, and $q_R \equiv q_{R,T} = q_T$ from (i) in this LBT construction. Let $(j_{n_1}, \ldots, j_{n_{q_R}})$ denote the sub-indices of the indices $(j_1, \ldots, j_{q_{\rho_\theta}})$ (defined immediately below (17)) such that the block $\Pi'_{\rho_\theta,j_{n_i}} R'\Pi_{R,n_i}$ of dimension $c^*_{\rho_\theta,j_{n_i}} \times c^*_{R,n_i}$ is full column-rank $c^*_{R,n_i} > 0$ for $i = 1, \ldots, q_R$ and $\sum_{i=1}^{q_R} c^*_{R,n_i} = d_R$. Thus, the corresponding block of $D_{T,\rho_\theta}\Pi'_{\rho_\theta}R'\Pi_R$ is $\frac{\sqrt{T}}{\lambda_{T,j_{n_i}}}\Pi'_{\rho_\theta,j_{n_i}} R'\Pi_{R,n_i}$ and, correspondingly, the columns from $(d_R - \sum_{i'=i}^{q_R} c^*_{R,n_{i'}})$ to $(d_R - \sum_{i'=i}^{q_R} c^*_{R,n_{i'}} + c^*_{R,n_i})$ for $i = 1, \ldots, q_R$ of $D_{T,\rho_\theta}\Pi'_{\rho_\theta}R'\Pi_R$ are represented by the $d_\theta \times c^*_{R,n_i}$ matrix: $\left[0', \frac{\sqrt{T}}{\lambda_{T,j_{q_{\rho_\theta}}}}\left(\Pi'_{\rho_\theta,j_{q_{\rho_\theta}}} R'\Pi_{R,q_{\rho_\theta}}\right)'\right]'$ if

$n_i = q_{\rho\theta}$ and $\left[0', \frac{\sqrt{T}}{\lambda_{T,j_{n_i}}}\left(\Pi'_{\rho\theta,j_{n_i}} R'\Pi_{R,n_i}\right)', \ldots, \frac{\sqrt{T}}{\lambda_{T,j_{q_{\rho\theta}}}}\left(\Pi'_{\rho\theta,j_{q_{\rho\theta}}} R'\Pi_{R,n_i}\right)'\right]'$ otherwise (as above, 0 represents the sub-matrix of zeros with number of rows that make the number of rows of the corresponding matrix equal to d_θ).

Now, conforming to this above structure, define a $d_R \times d_R$ matrix $D_{T,R}$ as

$$D_{T,R} := T^{-1/2}\text{diag}\left(\lambda_{T,j_{n_1}} 1_{c^*_{R,n_1}}, \ldots, \lambda_{T,j_{n_{q_R}}} 1_{c^*_{R,n_{q_R}}}\right) \tag{22}$$

such that

$$D_{T,\rho\theta}\Pi'_{\rho\theta} R'\Pi_R D_{T,R} \to R^{*'} \tag{23}$$

where $R^{*'}$ is a $d_\theta \times d_R$ finite matrix of full column-rank, and its columns from $(d_R - \sum_{i'=i}^{q_R} c^*_{R,n_{i'}})$ to $(d_R - \sum_{i'=i}^{q_R} c^*_{R,n_{i'}} + c^*_{R,n_i})$ for $i = 1, \ldots, q_R$ are represented by the $d_\theta \times c^*_{R,n_i}$ matrix: $\left[0', \left(\Pi'_{\rho\theta,j_{q_{\rho\theta}}} R'\Pi_{R,q_{\rho\theta}}\right)'\right]'$ if $n_i = q_{\rho\theta}$ and $\left[0', \left(\Pi'_{\rho\theta,j_{n_i}} R'\Pi_{R,n_i}\right)', 0'\right]'$ otherwise (as above, 0 denotes sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding matrix is d_θ).

Based on the above constructions, the local deviation of the null from the truth that we consider is:

$$\sqrt{T}D_{T,R}\Pi'_R(r_0 - \beta^0) = \mu_\beta, \text{ i.e., } r_0 = \beta^0 + \Pi_R \times \text{diag}\left(\lambda^{-1}_{T,j_{n_1}} 1_{c^*_{R,n_1}}, \ldots, \lambda^{-1}_{T,j_{n_{q_R}}} 1_{c^*_{R,n_{q_R}}}\right)\mu_\beta \tag{24}$$

where $D_{T,R}$ and Π_R are as defined in (22) and (21) respectively, r_0 is as defined in (2) and $\beta^0 := R\theta^0$ where θ^0 is the true value of θ, and μ_β is any arbitrary, finite, deterministic, $d_R \times 1$ vector such that $r_0 \in \text{int}(\mathcal{B})$.

Antoine and Renault (2012) note that the one-to-one transformation $\Pi^{-1}_{\rho\theta}\theta$ of θ provides the rate-disentangled directions of θ that are identified under N1-N6. However, since $R\theta$ and not θ is our object of interest, we need to consider the further constructions in (21), (22) and (23) to arrive at a similar rate-disentangled form to characterize in (24) the local deviation of the hypothesized value r_0.

For the hypothesized value r_0 satisfying (24), consider an arbitrary and possibly non-deterministic sequence $\{\gamma_{S,T} : T \geq 1\} \in \Gamma_S$ and, thus, $\theta_T := R^1_S r_0 + S^1_S\gamma_{S,T}$ (i.e., $R\theta_T = r_0$) such that

$$\sqrt{T}D^{-1}_{T,\rho\theta}\Pi^{-1}_{\rho\theta}(\theta_T - \theta^0) = \mu_{T,\theta} \tag{25}$$

where $D_{T,\rho\theta}$ and $\Pi_{\rho\theta}$ are as defined in (18) and (17) respectively, and $\mu_{T,\theta}$ is any arbitrary, $O_p(1)$, $d_\theta \times 1$ vector satisfying $R^*\mu_{T,\theta} \xrightarrow{P} \mu_\beta$, and thus relates (25) to (24). The relationship follows since

$$\frac{1}{\sqrt{T}}\Pi_{\rho\theta}D_{T,\rho\theta}\mu_{T,\theta} = \theta_T - \theta^0 = R^1_S(r_0 - \beta^0) + S^1_S(\gamma_{S,T} - \gamma^0_S) \text{ i.e., } R\Pi_{\rho\theta}D_{T,\rho\theta}\mu_{T,\theta} = \Pi_R^{'-1}D^{-1}_{T,R}\mu_\beta \text{ i.e., } R^*\mu_{T,\theta} \xrightarrow{P} \mu_\beta.$$

The first equality uses the definition in (25), the second one is due to (5), the third one uses a pre-multiplication of both sides by R and then the definition in (24), and the final step uses the definition in (23) and that $\mu_{T,\theta} = O_p(1)$.

Assumption N: (continued)

N7. The version of the tradeoff (see the remark below N6) assumption to be maintained is:

(a) $\rho(\theta)$ is twice continuously differentiable in $\theta \in \text{int}(\Theta)$. $g(z;\theta)$ is twice differentiable in $\theta \in \text{int}(\Theta)$ for each

$z \in \mathbb{R}^{d_z}$ and $\sup_{\theta\in\text{int}(\Theta)} \left\| \frac{\partial}{\partial\theta_i}\left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right]\right\| = o_p(\lambda_{T,l}/\sqrt{T})$ for $i = 1, \ldots, d_\theta$.

(b) λ_{T,j_1} from (18) is such that $\lambda^2_{T,j_1}/\lambda_{T,l} \to \infty$ as $T \to \infty$, and where $\lambda_{T,l}$ is defined in N3.

Remarks:

1. N7(a) imposes additional structure on the convergence in probability similar to assumption 5(i) (also see 3(iv)) of Antoine and Renault (2012), but on the second derivative of the moment vector. However, this structure is still weaker than that in assumption 6(ii) of Antoine and Renault (2012) unless $\lambda_{T,1} = \ldots = \lambda_{T,l}$, i.e., unless there is only a single rate associated with all the rows of the moment vector in (16). N7(b) resembles Antoine and Renault (2012)'s assumption 6*(i), which is a sufficient condition for their high-level orthogonality condition. If $\lim_T \sqrt{T}/\lambda_{T,l} < \infty$, in addition to N3, then N7(b) is equivalent to $\lim_T \lambda^2_{T,j_1}/\sqrt{T} = \infty$, which resembles, in the sense of Antoine and Renault (2012), the well-known condition in equation (4.13) in Andrews (1994).

2. We conjecture that if, when stated in terms of $\partial^a/(\partial\theta_1^{a_1}\ldots\partial\theta_{d_\theta}^{a_{d_\theta}})$ where $a_1, \ldots, a_{d_\theta}$ are non-negative integers and $a := \sum_{j=1}^{d_\theta} a_j$, N7(a) would hold for $a \geq 2$, then in N7(b) we could allow for $\lim_T \lambda^a_{T,j_1}/\lambda_{T,l} = \infty$. (The key steps are sketched for $a = 3$ below the proof of Lemma 3.9(c) in Appendix B.) Thus as $a \to \infty$, an extreme example of whose limiting case is the linear instrumental variables regression with θ as the structural coefficients, we could allow for the slowest rate (i.e. $\lambda_{T,j_1}/\sqrt{T}$) of deviation from zero of the suitably rotated expected moment vector to be anything faster that $T^{-1/2}$. This rate corresponds to just better than the genuinely weak identification of θ.

3. As noted by Antoine and Renault (2012) and also evident from Remark 1, the restriction in N7(b) under our setup is not required if there is only a single rate, i.e., if $\lambda_{T,j_1} = \ldots = \lambda_{T,j_{q_{\rho_\theta}}}$ (see (18)).

Lastly, the following standard assumption similar to M2 completes the characterization of our framework.

Assumption N: (continued)

N8. $\sup_{\theta\in\Theta} \|\widehat{V}_T(\theta) - V(\theta)\| = o_p(1)$ and $\sup_{\theta\in\mathcal{N}(\theta^0)} \|\widehat{V}_{Gg,T}(\theta) - V_{Gg}(\theta)\| = o_p(1)$ where $\mathcal{N}(\theta^0)$ is some open neighborhood of θ^0. $V(\theta)$ and $V_{Gg}(\theta)$ are finite and continuous inside $\mathcal{N}(\theta^0)$. $V(\theta)$ is positive-definite with $\sup_{\theta\in\Theta}$ max[eigen values$(V(\theta))] \leq \bar{c} < \infty$ (thus, finite), $\inf_{\theta\in\Theta}$ min[eigen values$V(\theta))] \geq \underline{c} > 0$. $V(\theta^0) = V$.

Ideally $V_{Gg}(\theta)$ is such that $V_{Gg}(\theta^0) = V_{Gg}$ defined in M1. However, this is not necessary since we do not allow for genuinely weak identification here, and thus do not need to use Kleibergen (2005)'s orthogonalization argument that we needed under the more general characterization when studying the rejection of the true null in Section 3.1.

Lemma 3.3 *Let assumptions O and N hold. Consider a sequence $\{\theta_T = R^1_S r_0 + S^1_S\gamma_{S,T} : T \geq 1\}$ where r_0 satisfies (24) and $\{\gamma_{S,T} : T \geq 1\}$ is such that θ_T satisfies (25). Then the following results hold as $T \to \infty$:*

(a) $LM_T(\theta_T) = LM_T^{infeas}(\theta_0^{infeas}) + o_p(1)$ where $LM_T(\theta_T)$ is as defined in (3), $LM_T^{infeas}(\theta_0^{infeas})$ is as defined in (11) and $\theta_0^{infeas} := A_S^{-1}(r_0', \gamma_S^{0'})' = R^1_S r_0 + S^1_S\gamma^0_S$.

(b) $LM_T(\theta_T) \xrightarrow{d} \chi^2_{d_R}$ with non-centrality parameter $\mu_\beta'\left(R^*(G^{*'}V^{-1}G^*)^{-1}R^{*'}\right)^{-1}\mu_\beta$.

Remarks: Lemma 3.3(a) is a striking result emphasizing that local deviations of the nuisance parameters γ_S from their true value γ^0_S (irrespective of the choice of S) does not matter as long as (25) holds. Under the standard NM-9.2 conditions such a result is expected since $LM_T(\theta)$ is constructed based on the efficient influence function; and Section 2 (in particular, Section 2.5) discusses this. On the other hand, we relax those standard assumptions in this section and, as a result, the ideas behind standard $\sqrt{T}$-consistent estimation, efficiency bound, etc. may no longer hold. Even under these relaxed conditions, that allow for identification failure (albeit not weak), we demonstrate that this asymptotic equivalence result holds. Note that Chaudhuri and Zivot (2011) demonstrate a similar asymptotic equivalence result but without allowing for any identification failure. They consider the classical characterization of local deviations. In our case, however, the characterization of local deviations is nonstandard and this is evident from the (rather long) step-by-step construction that was described before this lemma.

Lemma 3.3(b) specifies the asymptotic distribution of $LM_T(\theta_T)$, which we now use to study the asymptotic properties of the improved projection test defined in (7). (Recall that when the null H_0 is true, we have $\mu_\beta = 0$.)

Proposition 3.4 *Let assumptions O and N hold. Let the hypothesized value r_0 for $\beta^0 := R\theta^0$ satisfy the local deviation from the truth characterized by* (24)*. For the given choice of S, and some $\epsilon > 0$ such that $\epsilon + \alpha < 1$, let $CI_T(\gamma_S;\epsilon)$ be a confidence set for γ_S^0, the true value of γ_S, such that*

$$\sup_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} \sqrt{T}\left\| D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}\left((R_S^1 r_0 + S_S^1\gamma_0) - \theta^0\right)\right\| = O_p(1) \tag{26}$$

where Π_{ρ_θ} and D_{T,ρ_θ} are defined in (17) *and* (18) *respectively. Then*

$$\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) = LM_T^{infeas}(\theta_0^{infeas}) + o_p(1)$$

where $LM_T^{infeas}(\theta_0^{infeas})$ is as defined in (11) *and $\theta_0^{infeas} := A_S^{-1}(r_0', \gamma_S^{0'})' = R_S^1 r_0 + S_S^1\gamma_S^0$.*

Remark: The proposition, therefore, establishes the asymptotic equivalence of the improved projection test and the infeasible test described above (11), provided that the first-step confidence set for γ_S^0 (for the given choice of S) satisfies the condition in (26). Thus, the improved projection test inherits any optimality property (discussed in Section 2) of the infeasible test in such cases.

It is useful to have a closer look at (26). For convenience of future reference, recall that for any $\gamma_0 \in \Gamma_S$:

$$\sqrt{T}\left\| D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}\left((R_S^1 r_0 + S_S^1\gamma_0) - \theta^0\right)\right\| = \sqrt{T}\left\| D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}R_S^1(r_0 - \beta^0) + D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}S_S^1(\gamma_0 - \gamma_S^0)\right\|.$$

If $\Lambda_T = \lambda_T I_{d_g}$ for some $\lambda_T \to \infty$ (but $\lim_T \lambda_T/\sqrt{T} < \infty$), i.e., all the rates are equal, and if our interest lies in testing sub-vectors of θ (e.g., $R = [I_{d_R}, 0]$), then by virtue of (24), the condition in (26) boils down to

$$\sup_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} \lambda_T \left\|\gamma_0 - \gamma_S^0\right\| = O_p(1).$$

If, additionally, we consider the setup of Chaudhuri and Zivot (2011) where (the part related to) asymptotic equivalence similar in spirit to that in Proposition 3.4 is discussed under standard conditions such as the NM-9.2 conditions and those that ensure consistent estimation of γ_S^0 (as in Section 2.5), then again by virtue of (24), the condition in (26) boils down to

$$\sup_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} \sqrt{T}\left\|\gamma_0 - \gamma_S^0\right\| = O_p(1).$$

Both these scenarios are familiar by now. By contrast, in our assumption N3 (and N7(b)) we do not force all the rates to be equal. Nor do we focus only on testing sub-vectors of θ (although it should be noted that such cases are the leading and most common examples of our general null hypotheses on linear restrictions). As a consequence, the general representation (26) in Proposition 3.4 is more involved than the corresponding representations in the two familiar scenarios above. Nevertheless, the intuition behind this general representation is the same.

Does there exist such a confidence set $CI_T(\gamma_S;\epsilon)$ satisfying (26)?

The general answer seems to be negative under the framework of this sub-section. For example, consider the choice $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ defined in (13). There are two issues. The first one is specific to this choice, and does not have an adverse effect on the asymptotic power of the improved projection test. The second one may have adverse effects on asymptotic power. This issue is related to the generality of our framework, and does not appear under the special cases that are typically considered in this literature, including that in Chaudhuri and Zivot (2011). Thus, for those special cases and certain generalizations of them, all our results so far go through without any further assumption.

Let us now be more specific about these two issues, and then state the precise condition under which the second issue becomes immaterial and there is no adverse effect on the asymptotic power of the improved projection test.

First, as noted in remark 2 following Proposition 3.2, $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ can be empty with positive probability, a property that we actually deem desirable for the power of the improved projection test.[10] Nevertheless, it means that $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ cannot satisfy (26). This is well-known and has been noted in Chaudhuri and Zivot (2011).

Second, it seems that even without the consideration of the emptiness of the confidence set, it is not possible to conclude that $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ would satisfy (26) without imposing further restrictions. The fundamental problem behind this has nothing to do with the fact that we are testing hypotheses on $R\theta$, but seems to be intrinsic to the framework of this sub-section and our focus on the efficient directions in terms of θ (and thereby in terms of $R\theta$). It is important to highlight this problem since even if we were testing a null hypothesis such as $\theta = \theta_0$ (and thus no projection test is required), this problem affects how the power of a test increases as the hypothesized value θ_0 deviates from the truth along the efficient (in the sense of Antoine and Renault (2012)) direction: $\Pi_{\rho_\theta}^{-1}(\theta^0 - \theta_0)$. Without deviating from the discussion, we note this problem in the context of $CI_T^{SW}(\gamma_S;r_0,\epsilon)$, i.e., the S-test.[11]

Thanks to N1 and subsequently N8, it is straightforward to see that $\inf_{\theta_T:\|\theta_T-\theta^0\|\geq c} T \times Q_T(\theta_T)$ diverges (in probability) to ∞ as $T \to \infty$ for any $c > 0$. Thus, by the definition in (13), such θ_T's (or random sequences θ_T's taking such values) are not contained in $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ with probability approaching one. However, the problem arises when we wish to conclude that sequences $\{\theta_T : T \geq 1\}$ such that $\theta_T - \theta^0 = o(1)$ but $\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta_T - \theta^0) \neq O(1)$ cannot be contained in $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ asymptotically, i.e., when we wish to establish that for such a sequence (or random sequences θ_T's taking such values), $T \times Q_T(\theta_T)$ diverges (in probability) to ∞ as $T \to \infty$.[12] This result is necessary to justify the use of $CI_T^{SW}(\gamma_S;r_0,\epsilon)$, which crucially hinges on (26). The technical problem that we face in establishing it is similar to what necessitated assumption 6(i) in Antoine and Renault (2012).

The following high-level assumption overcomes this problem.

[10] The probability of emptiness decreases as $\epsilon \to 0$. Provided that $CI_T^{SW}(\gamma_S;r_0,\epsilon)$ is non-empty, it contains $\widehat{\gamma}_{S,T}(r_0)$ defined in (15).

[11] This choice helps us to focus on the specific problem since, as noted in Remark 2 following Proposition 3.2, the other tests whose asymptotic size is robust to identification failures suffer from an additional problem of spurious declines in power due to other reasons.

[12] See assumptions 5(i)-(iii) in Andrews and Mikusheva (2016) who also provide further discussion.

Assumption N: (continued)

N9. There exists an open neighborhood $\mathcal{N}(\theta^0)$ of θ^0 such that

$$\sup_{\theta\in\mathcal{N}(\theta^0)} \kappa_T(\theta) = o_p(1) \text{ where } \kappa_T(\theta) := \frac{\left\{\frac{\Lambda_T}{\sqrt{T}}\left[\rho_\theta(\theta)-\rho_\theta(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta}\right\}\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta-\theta^0)}{\|\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta-\theta^0)\|}.$$

Remarks: It is clear that N9 is trivially satisfied when $\rho(\theta)$ is linear in θ, an example of which is the linear instrumental variables regression. Now consider some special cases by allowing for $\rho(\theta)$ to be nonlinear in θ.

First, if $\Lambda_T = \lambda_T I_{d_g}$, i.e., if all the rates are equal, then also N9 holds trivially due to continuity of ρ_θ (see N4). As noted before, this is the scenario considered in Chaudhuri and Zivot (2011) to discuss the asymptotic equivalence of the projection test and the locally optimal infeasible test. Also, as noted in Remark 3 following N7, the assumption N7(b) is redundant under this scenario. Thus, all the results so far automatically hold for the so-called nearly weak identification cases considered in, e.g., Caner (2010).

Second, allow for the $\lambda_{T,j}$'s in Λ_T to be unequal and of different order of magnitude, but let $\rho_\theta(\theta)$ be such that the same Π_{ρ_θ} works for the purpose of (17) for all $\theta\in\mathcal{N}(\theta^0)$. N9 holds in such cases. A further special case of this is where $\Pi_{\rho_\theta} = I_{d_\theta}$, i.e., $\rho_\theta(\theta)$ is already such that the elements of θ are locally identifiable at disentangled rates.[13]

Apart from the above cases, certain restrictions on the $\lambda_{T,j}$'s and $\|\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta-\theta^0)\|$ in addition to N7(b) also make assumption N9 hold. However, in general, assumption N9 is not innocuous and, as a *partial support* to the condition (26), we are only able to provide the following result in Lemma 3.5 by maintaining this assumption.

Although we say *partial support* to acknowledge that $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ can be empty asymptotically with positive probability, we already noted that this can not be bad for the power of the improved projection test. So it is only for technical reasons that we will, in Lemma 3.5, define the sup in (26) to be zero if $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ is empty, as was done in Lemma 2 in Andrews (2016b). In this sense the result below is similar to an intermediate result in the proof of Theorem 3.2(ii) of Chaudhuri and Zivot (2011) and also Theorem 3 of Andrews (2016b) both of whom focus on the case of strong identification, unlike our setup of less than strong and rate-entangled identification.

Lemma 3.5 *Let assumptions O and N hold. Let the hypothesized value r_0 for $\beta^0 := R\theta^0$ satisfy the local deviation from the truth characterized by* (24)*. Then $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ satisfies* (26) *for $\epsilon > 0$, i.e.,*

$$\sup_{\gamma_0\in CI_T^{SW}(\gamma_S;r_0,\epsilon)} \sqrt{T}\left\|D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}\left((R_S^1 r_0 + S_S^1\gamma_0)-\theta^0\right)\right\| = O_p(1)$$

where the left hand side is defined as 0 if $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ is empty.

For completeness, we summarize the result from Proposition 3.4 and Lemma 3.5 in the form of the following corollary. The proof is similar to that of Theorem 3.2(ii) in Chaudhuri and Zivot (2011) and hence is omitted.

Corollary 3.6 *Let assumptions O and N hold. Let the hypothesized value r_0 for $\beta^0 := R\theta^0$ satisfy the local deviation from the truth characterized by* (24)*. Then, for $\epsilon, \alpha > 0$ such that $\epsilon + \alpha < 1$, the asymptotic probability*

[13] While not strictly nested by our model for $E[\bar{G}_T(\theta)] = \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)$, it can be shown that N9 is not required under general cases of rate-disentangled θ, e.g., $E[\bar{G}_T(\theta)] = \rho_\theta(\theta)\text{diag}(\delta_{T,1},\ldots,\delta_{T,d_\theta})/\sqrt{T}$ where $\delta_{T,j}\to\infty$ but $\lim_T \delta_{T,j}/\sqrt{T} < \infty$ for all $j = 1,\ldots,d_\theta$.

of rejection of this hypothesized value by the improved projection test in (7) *based on the choice* $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ *in* (13)*, cannot be smaller than that by the infeasible test in* (11)*.*

3.3 Closely related literature

The notion of optimality in relation to the infeasible test in (11) is less ambitious than that considered in the literature on identification failure inspired by Moreira (2002, 2003). See Andrews et al. (2006), Moreira and Moreira (2013), Andrews (2016b), Andrews and Mikusheva (2016), Montiel-Olea (2016), etc. By contrast, our use of the term is similar to that in Section 9 of Andrews and Guggenberger (2015) and Comment (iii) following Theorem 4.1 of Andrews and Guggenberger (2014). Indeed, the LM-principle generally does not lead to optimality other than in a local sense since it is only based on the slope of the moment vector (slope of log-likelihood function).

Furthermore, as originally noted by Kleibergen (2005), allowing for identification failure necessitates the use of an estimator for the Jacobian matrix that is not simply the sample mean of the derivative of the moment vector, but the sample mean of the residual of the regression of this derivative on the moment vector itself. In certain cases of identification failure, this affects the intended direction along which the LM-principle maximizes local power; see, e.g., Antoine and Renault (2009). Even otherwise, this may lead to a spurious decline in power away from the truth; see Kleibergen (2005). To partially address this problem in the context of testing for sub-vectors, Chaudhuri and Zivot (2011) recommend inverting the S-test to obtain the first-step confidence set for the nuisance parameters; and we follow their approach in this paper. The price to pay is that since this confidence set can be empty with positive probability (even asymptotically), the asymptotic equivalence (as in Section 2) of the improved projection test with the infeasible test no longer holds. However, as we saw in Section 3.1, one can still impose a pre-specified upper bound on the asymptotic size of the improved projection test. And, remarkably, the conventional fixed critical values are sufficient for this purpose even under a very general setup. See McCloskey (2015), Andrews (2016a), etc. for more sophisticated approaches. The results from Section 3.2 indicate that the improved projection test is competitive with the infeasible test in terms of asymptotic power even when we generalize the setup of Section 2.

Lastly, we note that while we generalize the use of the LM and C-alpha principle in Chaudhuri (2008), Zivot and Chaudhuri (2009), Chaudhuri et al. (2010) and Chaudhuri and Zivot (2011); the LM and/or C-alpha tests were originally used in the context of identification failure by Wang and Zivot (1998), Dufour and Jasiak (2001), Kleibergen (2002), Moreira (2003), Kleibergen (2005), Guggenberger and Smith (2005), Antoine and Renault (2009), etc. It has also been considered more recently in Magnusson and Mavroeidis (2010), Guggenberger et al. (2012b), Qu (2014), Dufour et al. (2015), Andrews and Mikusheva (2015), Andrews and Guggenberger (2014), etc.

The equivalence relation established in Section 2 between the alternative constructions of the C-alpha statistics, however, appears to be new. This reconciles the C-alpha statistic in Smith (1987), Dagenais and Dufour (1991), etc. with the efficient score statistic in a re-parameterized model, and thereby makes the latter directly adaptable to our framework. Thus, although we work with the original parameter vector θ and the original linear restrictions $R\theta$ from (1) and (2) respectively to closely adhere to the recent literature, one could alternatively obtain the same results by working with the re-parameterized model and thereby providing a direct generalization of the results in Chaudhuri and Zivot (2011) to the more involved characterizations of identification failures in our paper.

References

Andrews, D. W. K. (1994). Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, 62: 43–72.

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

Andrews, D. W. K. and Cheng, X. (2014). Gmm Estimation and Uniform Subvector Inference with Possible Identification Failure. *Econometric Theory*, 20: 287–333.

Andrews, D. W. K. and Guggenberger, P. (2014). Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models. Working Paper.

Andrews, D. W. K. and Guggenberger, P. (2015). Identification And Singularity Robust Inference For Moment Condition Models. Working Paper.

Andrews, D. W. K., Moreira, M., and Stock, J. H. (2006). Optimal Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, 74: 715–752.

Andrews, I. (2016a). Conditional Linear Combination Tests. Forthcoming: Econometrica.

Andrews, I. (2016b). Robust Two-Step Confidence Sets, and the Trouble with the First Stage F-Statistic. Working Paper.

Andrews, I. and Mikusheva, A. (2015). Maximum Likelihood Inference in Weakly Identified DSGE Models. *Quantitative Economics*, 6: 123–152.

Andrews, I. and Mikusheva, A. (2016). Conditional Inference with a Functional Nuisance Parameter. Forthcoming: Econometrica.

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

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

Back, K. and Brown, D. P. (1992). GMM, Maximum Likelihood, and Nonparametric Efficiency. *Economics Letters*, 39: 23–28.

Beaulieu, M.-C., Dufour, J.-M., and Khalaf, L. (2013). Identification-robust estimation and testing of the zero-beta CAPM. *Review of Economic Studies*, 83: 892–924.

Caner, M. (2010). Testing, estimation in GMM and CUE with nearly weak identification. *Econometric Reviews*, 29: 330–363.

Chamberlain, G. (1987). Asymptotic Efficiency in Estimation with Conditional Moment Restrictions. *Journal of Econometrics*, 34: 305–334.

Chaudhuri, S. (2008). *Projection-Type Score Tests for Subsets of Paramaters.* PhD thesis, University of Washington.

Chaudhuri, S. and Renault, E. (2011). Finite-sample improvements of score tests by the use of implied probabilities from Generalized Empirical Likelihood. Technical report, University of North Carolina, Chapel Hill.

Chaudhuri, S., Richardson, T., Robins, J., and Zivot, E. (2010). Split-Sample Score Tests in Linear Instrumental Variables Regression. *Econometric Theory*, 26: 1820–1837.

Chaudhuri, S. and Rose, E. (2009). Estimating the Veteran Effect with Endogenous Schooling when Instruments are Potentially Weak. Technical report, University of North Carolina, Chapel Hill and University of Washington.

Chaudhuri, S. and Zivot, E. (2011). A new method of projection-based inference in GMM with weakly identified nuisance parameters. *Journal of Econometrics*, 164: 239–251.

Cheng, X. (2015). Robust inference in nonlinear models with mixed identification strength. *Journal of Ecnometrics*, 189: 207–228.

Choi, I. and Phillips, P. C. B. (1992). Asymptotic and Finite Sample Distribution Theory for IV Estimators and Tests in Partially Identified Structural Equations. *Journal of Econometrics*, 51: 113–150.

Dagenais, M. and Dufour, J. M. (1991). Invariance, Nonlinear Models and Asymptotic Tests. *Econometrica*, 59: 1601–1615.

Davidson, R. and MacKinnon, J. (2014). Confidence sets based on inverting anderson-rubin tests. *Econometrics Journal*, 17: S39–S58.

Dufour, J. M. (1997). Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models. *Econometrica*, 65: 1365–1388.

Dufour, J. M. and Jasiak, J. (2001). Finite Sample Limited Information Inference Methods for Structural Equations and Models with Generated Regressors. *International Economic Review*, 42: 815–843.

Dufour, J.-M., Khalaf, L., and Kichian, M. (2006). Inflation dynamics and the New Keynesian Phillips Curve: An identification robust econometric analysis. *Journal of Economic Dynamics and Control*, 30: 1707–1727.

Dufour, J.-M., Khalaf, L., and Kichian, M. (2013). Identification-robust analysis of dsge and structural macroeconomic models. *Journal of Monetary Economics*, 60: 340–350.

Dufour, J. M. and Taamouti, M. (2005). Further Results on Projection-Based Inference in IV Regressions with Weak, Collinear or Missing Instruments. Discussion Paper.

Dufour, J. M. and Taamouti, M. (2007). Further Results on Projection-Based Inference in IV Regressions with Weak, Collinear or Missing Instruments. *Journal of Econometrics*, 139: 133–153.

Dufour, J.-M., Trognon, A., and Tuvaandorj, P. (2015). Invariant tests based on M-estimators, estimating functions, and the generalized method of moments. Forthcoming in Econometric Reviews.

Guggenberger, P., Kleibergen, F., Mavroeidis, S., and Chen, L. (2012a). On the Asymptotic Sizes of Subset AndersonRubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression. *Economoetrica*, 80: 2649–2666.

Guggenberger, P., Ramalho, J., and Smith, R. (2012b). GEL Statistics under Weak Identification. *Journal of Econometrics*, 170: 331–349.

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

Guggenberger, P. and Smith, R. (2008). Generalized empirical likelihood tests in time series models with potential identification failure. *Journal of Econometrics*, 142: 134–161.

Hahn, J. and Kuersteiner, G. (2002). Discontinuities of Weak Instruments Limiting Distributions. *Economics Letters*, 75: 325–331.

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

Kleibergen, F. (2004). Testing Subsets of Parameters In The Instrumental Variables Regression Model. *The Review of Economics and Statistics*, 86: 418–423.

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

Kleibergen, F. (2007). Subset statistic in the linear IV regression model. Technical report, Brown University Working Paper.

Magnusson, L. and Mavroeidis, S. (2010). Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve. *Journal of Money, Credit and Banking*, 42: 465–481.

McCloskey, A. (2015). Bonferroni-Based Size-Correction for Nonstandard Testing Problems. Brown University.

Mikusheva, A. (2010). Robust Confidence Sets in the Presence of Weak Instruments. *Journal of Econometrics*, 157: 236–247.

Montiel-Olea, J. L. (2016). Admissible, Similar Tests: A Characterization. New York University.

Moreira, H. and Moreira, M. (2013). Contributions to the Theory of Similar Tests. FGV/EPGE.

Moreira, M. J. (2002). *Tests with Correct Size in the Simultaneous Equations Model.* PhD thesis, UC Berkeley.

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

Muller, U. and Norets, A. (2016). Credibility of Confidence Sets in Nonstandard Econometric Problems. *Econometrica*, 84: 2183–2213.

Nelson, C. R. and Startz, R. (1990). The Distribution of the Instrumental Variable Estimator and its t-ratio When the Instrument is a Poor One. *Journal of Business*, 63: 125–140.

Newey, W. K. (1990). Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, 5: 99–135.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

Neyman, J. (1959). Optimal Asymptotic Test of Composite Statistical Hypothesis. In Grenander, U., editor, *Probability and Statistics, the Harald Cramer Volume*, pages 313–334. Almqvist and Wiksell, Uppsala.

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.

Qu, Z. (2014). Inference in DSGE Models with Possible Weak Identification. *Quantitative Economics*, 5: 457–494.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer.

Sargan, D. J. (1983). Identification and lack of identification. *Econometrica*, 51: 1605–1633.

Smith, R. J. (1987). Alternative Asymptotically Optimal Tests and Their Application to Dynamic Specification. *The Review of Economic Studies*, 54: 665–680.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65: 557–586.

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

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

Zivot, E. and Chaudhuri, S. (2009). Comment: Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve by F. Kleibergen and S. Mavroeidis. *Journal of Business and Economic Statistics*, 27: 328–331.

Zivot, E., Startz, R., and Nelson, C. (1998). Valid Confidence Intervals and Inference in the Presence of Weak Instruments. *International Economic Review*, 39: 1119–1144.

Zivot, E., Startz, R., and Nelson, C. (2006). Inference in Weakly Identified Instrumental Variables Regression. In Corbae, D., Durlauf, S. N., and Hansen, B. E., editors, *Frontiers in Analysis and Applied Research: Essays in Honor of Peter C. B. Phillips*, pages 125–166. Cambridge University Press.

Appendix A: For the references from Section 2

A.1 Efficient influence function for $\beta^0 := R\theta^0$ under (1)

It is well-known that under the assumptions that (1) holds, $G(\theta^0)$ is full column-rank, and $V(\theta^0)$ is positive definite: the efficient estimator of $R\theta^0$ has an asymptotically linear representation $-\sqrt{T}l_T(\theta^0)+o_p(1)$. Unfortunately, we could not find a paper to cite the proof of it, and hence we provide a simple proof following Newey (1990) (and maintaining his assumptions) for the sake of completeness. Alternatively, one could follow Sections 2 and 3 of Chamberlain (1987) to use the multinomial approximation, and then appeal to the invariance property of maximum likelihood estimators (MLEs) to establish the same result. Yet otherwise, one could use Back and Brown (1992)'s construction of an exponentially tilted density corresponding to which the MLE of θ is the efficient GMM estimator of θ, and then appeal to the invariance property of MLEs.

Lemma 3.7 *Let $\{Z_t\}_{t=1}^T$ be i.i.d. copies of a random variable Z, and let* (1) *holds. If $G := \frac{\partial}{\partial\theta'}E[g(Z;\theta)]_{\theta=\theta^0}$ is a full column-rank $d_g \times d_\theta$ matrix and $V := E[g(Z;\theta^0)g'(Z;\theta^0)]$ is a positive definite $d_g \times d_g$ matrix, then the asymptotic variance lower bound for any regular estimator of the $d_R \times 1$ parameter vector $\beta^0 := R\theta^0$ where $d_R \le d_\theta$ is $(R(G'V^{-1}G)R')^{-1}$. The regular estimator whose asymptotic variance attains this bound has the asymptotically linear representation $\sqrt{T}(\widehat{\beta^0}-\beta^0) = -\sqrt{T}l_T(\theta^0)+o_p(1)$.*

Proof: Consider a parametric path ξ of the distribution of Z such that for the unique value ξ^0 we have the joint density $f_{\xi^0}(z) = f(z)$. Denote the score with respect to ξ with $s_\xi(Z)$. Without any other restrictions, the tangent space for the model is simply $\mathcal{T} = a(z)$ where $a(z)$ satisfies $E[a(Z)] = 0$, and $E[.]$ equivalently stands for $E_{\xi^0}[.]$. Since $d_g > d_R$, (1) equivalently requires that for any given $d_R \times d_g$ matrix B, the relation $BE[g(Z;\theta^0)] = 0$ holds. Take B as full row-rank without loss of generality. Now, differentiating with respect to ξ under the expectation we obtain $\frac{\partial\theta(\xi^0)}{\partial\xi} = -(BG)^{-1}E[Bg(Z;\theta^0)s_{\xi^0}(Z)]$ and thus $\frac{\partial\beta^0(\xi^0)}{\partial\xi} = -R(BG)^{-1}E[Bg(Z;\theta^0)s_{\xi^0}(Z)]$. Therefore, any regular estimator for β^0 will be asymptotically linear with the influence function $\varphi(B) := -R(BG)^{-1}Bg(Z;\theta^0)$. Given the structure of the tangent space $\mathcal{T}$, (1) implies that the projection of this influence function $\varphi(B)$ onto $\mathcal{T}$ is $\varphi(B)$ itself. For this given B, $Var(\varphi(B)) = \Sigma(B) := R(BG)^{-1}BVB'(BG)^{-1'}R'$. Thus the efficient influence function is obtained by choosing $B^* := \arg\min_B \Sigma(B) = G'V^{-1}$, giving $\Sigma(B^*) = R(G'V^{-1}G)^{-1}R'$ and $\varphi(B^*) = -R(G'V^{-1}G)^{-1}G'V^{-1}g(Z;\theta^0)$. ■

A.2 Proofs of the lemma and the proposition in Section 2:

The following relations that follow from the fact that $A_S = [R', S']'$ and $A_S^{-1} = [R_S^1, S_S^1]$, will be used repeatedly:

$$RR_S^1 = I_{d_R},\ RS_S^1 = 0,\ SR_S^1 = 0,\ SS_S^1 = I_{d_\theta - d_R} \text{ and } R_S^1R + S_S^1S = I_{d_\theta}. \tag{27}$$

We will suppress the dependence of the quantities on θ to avoid notational clutter. We will not consider the negligible set on which the assumptions are allowed to not hold since we only require to show the intended results hold almost surely.

Proof of Lemma 2.1: Consider any $(d_\theta - d_R) \times d_\theta$ full row-rank matrix S in (5), i.e., such that $[R', S']'$ is nonsingular. Let ζ be a $d_\theta \times (d_\theta - d_R)$ matrix whose columns form a basis for the null space of R. Therefore, since $RS_S^1 = 0$ by (27),

$S_S^1 = \zeta B_S$ for some $(d_\theta - d_R) \times (d_\theta - d_R)$ nonsingular matrix B_S. Similarly, if $\widetilde{S}$ is another such $(d_\theta - d_R) \times d_\theta$ matrix in (5), then the corresponding $\widetilde{S}^1_{\widetilde{S}} = \zeta B_{\widetilde{S}}$ for some $(d_\theta - d_R) \times (d_\theta - d_R)$ nonsingular matrix $B_{\widetilde{S}}$. Thus, we have $\widetilde{S}^1_{\widetilde{S}} = S_S^1 B$ where $B = B_S^{-1} B_{\widetilde{S}}$ is a $(d_\theta - d_R) \times (d_\theta - d_R)$ nonsingular matrix.

Now for any $d_\theta \times d_\theta$ nonsingular matrix $M = [M_1, M_2]$, where M_1 is $d_\theta \times d_R$ and M_2 is $d_\theta \times (d_\theta - d_R)$, define:

$$\begin{aligned} \Phi_T(M) &:= T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T M\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right), & (28)\\ \Phi_{1.2,T}(M) &:= T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\left(I_{d_g} - P\left(\widehat{V}_T^{-1/2}\widehat{G}_T M_2\right)\right)\widehat{V}_T^{-1/2}\widehat{G}_T M_1\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right), & (29)\\ \Phi_{2,T}(M_2) &:= T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T M_2\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right), & (30) \end{aligned}$$

and note that, by construction:

(i) $\Phi_T(M) = \Phi_T(I_{d_\theta})$,

(ii) $\Phi_T(M) = \Phi_{1.2,T}(M) + \Phi_{2,T}(M_2)$,

(iii) $\Phi_{2,T}(M_2) = \Phi_{2,T}(M_2 B)$ since B is a $(d_\theta - d_R) \times (d_\theta - d_R)$ nonsingular matrix.

Therefore, considering the choices: $M := [R_S^1, S_S^1]$ and $\widetilde{M} := [R^1_{\widetilde{S}}, \widetilde{S}^1_{\widetilde{S}}]$ corresponding to the two choices S and $\widetilde{S}$, we obtain:

$$\begin{aligned} \Phi_T(M) &= \Phi_T(\widetilde{M}) \quad \text{[by (i)]}\\ \Phi_{1.2,T}(M) + \Phi_{2,T}(M_2) &= \Phi_{1.2,T}(\widetilde{M}) + \Phi_{2,T}(\widetilde{M}_2) \quad \text{[by (ii)]}\\ \Phi_{1.2,T}(M) &= \Phi_{1.2,T}(\widetilde{M}) \quad \text{[by (iii), since } \widetilde{M}_2 := \widetilde{S}^1_{\widetilde{S}} = S_S^1 B =: M_2 B\text{]}. \end{aligned}$$

Thus, (6) implies that $LM^{\text{alt}}_{T,S}(\theta) = \Phi_{1.2,T}(M) = \Phi_{1.2,T}(\widetilde{M}) = LM^{\text{alt}}_{T,\widetilde{S}}(\theta)$, and gives the invariance property. ■

Proof of Proposition 2.2: Consider a $(d_\theta - d_R) \times d_\theta$ matrix S in (5) that satisfies $R\widehat{\Omega}^{-1}S' = 0$, i.e., the $(d_\theta - d_R)$ rows $S_1, \ldots, S_{d_\theta - d_R}$ of S are $(d_\theta - d_R)$ linearly independent elements of the null space of $R\widehat{\Omega}^{-1}$.

Claim 1: With this S, we have a nonsingular $A_S := [R', S']'$.
Proof: Suppose not. Then, the full row-rank of R implies that there has to exist a $(d_\theta - d_R) \times 1$ vector $c := (c_1, \ldots, c_{d_\theta - d_R})' \neq 0$ such that $R_1 = \sum_{j=2}^{d_R} a_j R_j + c'S$ for some scalar coefficients $a_2, \ldots, a_{d_R}$ and where $R = [R_1', \ldots, R_{d_R}']'$. Since $\widehat{\Omega}^{-1}$ is positive definite in except in the negligible set that we are ignoring, it means that for this $c \neq 0$, we have $R_1\widehat{\Omega}^{-1} = \sum_{j=2}^{d_R} a_j R_j \widehat{\Omega}^{-1} + c'S\widehat{\Omega}^{-1}$. Post-multiplying both sides by S' and noting that the rows of S belong in the null space of $R\widehat{\Omega}^{-1}$, it follows that $0 = c'S\widehat{\Omega}^{-1}S'$. Since $S\widehat{\Omega}^{-1}S'$ is positive definite (as $\widehat{\Omega}^{-1}$ is positive definite and as the rows of S are linearly independent), this is only possible if $c = 0$, which contradicts our supposition. Therefore, Claim 1 is true. ■

Claim 2: $R\widehat{\Omega}^{-1}S' = 0$ if and only if $R_S^{1'}\widehat{\Omega}S_S^1 = 0$.
Proof: We use (27) repeatedly in this proof. Post-multiply $R_S^{1'}\widehat{\Omega}S_S^1 = 0$ by S to get $R_S^{1'}\widehat{\Omega}(I_{d_\theta} - R_S^1 R) = 0$ and hence

$$R = (R_S^{1'}\widehat{\Omega}R_S^1)^{-1}R_S^{1'}\widehat{\Omega}. \tag{31}$$

Similarly obtain $S = (S_S^{1'}\widehat{\Omega}S_S^1)^{-1}S_S^{1'}\widehat{\Omega}$. Thus, $R\widehat{\Omega}^{-1}S' = (R_S^{1'}\widehat{\Omega}R_S^1)^{-1}(R_S^{1'}\widehat{\Omega}S_S^1)(S_S^{1'}\widehat{\Omega}S_S^1)^{-1}$ and hence $R\widehat{\Omega}^{-1}S' = 0$ if and only if $R_S^{1'}\widehat{\Omega}S_S^1 = 0$, once again by using the positive definiteness of $\widehat{\Omega}$. ■

Thus (6) implies that $LM^{\text{alt}}_{T,S}(\theta) = T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T R_S^1\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right)$. On the other hand, (4) gives: $LM_T(\theta) = T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T\widehat{\Omega}^{-1}R'\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right) = T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T R_S^1(R_S^{1'}\widehat{\Omega}R_S^1)^{-1}\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right)$ by using (31). But, by the construction of the projection matrix $P(.)$, we have $P\left(\widehat{V}_T^{-1/2}\widehat{G}_T R_S^1(R_S^{1'}\widehat{\Omega}R_S^1)^{-1}\right) = P\left(\widehat{V}_T^{-1/2}\widehat{G}_T R_S^1\right)$ since $(R_S^{1'}\widehat{\Omega}R_S^1)^{-1}$ is nonsingular. Therefore, $LM_T(\theta) = T \times \left(\widehat{V}_T^{-1/2}\bar{g}_T\right)' P\left(\widehat{V}_T^{-1/2}\widehat{G}_T R_S^1\right)\left(\widehat{V}_T^{-1/2}\bar{g}_T\right) = LM^{\text{alt}}_{T,S}(\theta)$. The desired result now follows from Lemma 2.1 for any general choice of S in (5) such that $[R', S']'$ is nonsingular. ■

Remark: The particular choice of S employed to facilitate the proof of Proposition 2.2 has an interesting interpretation. To see it, consider the analogous population version of S, i.e., S such that $R\Omega^{-1}S' = 0$. Similar to the proof of *Claim 1* above, it can be shown that $[R', S']'$ is nonsingular. Similar to the proof of *Claim 2* above, it can be shown that $R\Omega^{-1}S' = 0$ if and only if $R_S^{1'}\Omega S_S^1 = 0$, where the R_S^1 and S_S^1 correspond to this particular choice of S. Now, note from the discussion immediately preceding the statement of Lemma 2.1 that with this particular choice of S, the score for β, i.e., $l_{\beta,S,T}(\theta^0)$ is identical to the efficient score for β, i.e., $l_{\beta.\gamma_S,S,T}(\theta^0)$. In other words, this particular choice of S in the re-parameterization (5) directly makes the scores for β and γ_S uncorrelated (and, by asymptotic normality, asymptotically independent.) In yet other words, this means that the optimal rotation (in the efficient GMM sense) of the moment vector along the directions of β and γ_S are already orthogonal, and thus the subsequent orthogonalization in order to obtain the efficient score is moot.

A.3 $\widetilde{LM}_T(\widetilde{\theta}_T) = LM_T(\widetilde{\theta}_T)$

From (28)-(30) and the definition in (9) it follows that $\widetilde{LM}_T(\theta) = LM_T(\theta) + \Phi_{2,T}(S_S^1, \theta)$ for all θ where the underlying quantities are defined. (Note that by $\Phi_{2,T}(S_S^1, \theta)$ we mean $\Phi_{2,T}(S_S^1)$ with $\bar{g}_T$, $\widehat{G}_T$ and $\widehat{V}_T$ evaluated at θ.) Now, by the definition of the $\widetilde{\theta}_T$, i.e., $(R_S^1 r_o + S_S^1 \widetilde{\gamma}_T)$ where $\widetilde{\gamma}_T$ is the GMM estimator of γ by imposing $\beta = r_0$, it follows from the first order condition of the GMM optimization problem that $\Phi_{2,T}(S_S^1, \widetilde{\theta}_T) = 0$. This is because $\Phi_{2,T}(S_S^1, \theta)$ is simply a quadratic form of the first derivative of the GMM objective function with respect to γ_S, which is zero when evaluated at $\widetilde{\theta}_T$. Thus $\widetilde{LM}_T(\widetilde{\theta}_T) = LM_T(\widetilde{\theta}_T)$. ■

Appendix B: Proofs of the results from Section 3

Since we use (have used) the following result often, let us state it here for reference.

Lemma 3.8 *Let X be an $a \times b$ matrix, and P and Q be $a \times a$ and $b \times b$ nonsingular matrices. Then rank(X) = rank(PX) = rank(XQ).*

Proof: $\text{rank}(X) \geq \text{rank}(PX) \geq \text{rank}(P^{-1}PX) = \text{rank}(X) \geq \text{rank}(XQ) \geq \text{rank}(XQQ^{-1}) = \text{rank}(X)$. ■

Proof of Lemma 3.1: The proof is based on the original work of Antoine and Renault (2012), Andrews and Guggenberger (2014), Andrews and Cheng (2014) and Cheng (2015), with suitable adjustments required by our setup.

Let $\widehat{G}_T := \widehat{G}_T(\theta^0)$, $\widehat{V}_T := \widehat{V}_T(\theta^0)$. By M1 and M2, $\widehat{V}_T$ is positive definite with probability approaching one as $T \to \infty$. Thus, if defined, let $\widehat{V}_T^{-1/2}$ be such that $\widehat{V}_T^{-1/2'} \widehat{V}_T^{-1/2} = \widehat{V}_T^{-1}$ and $\widehat{g}_T := \widehat{V}_T^{-1/2} \bar{g}_T(\theta^0)$. Then, for T sufficiently large, (4) gives

$$LM_T(\theta^0) = T\widehat{g}_T' P\left(H_T\{H_T'H_T\}^- R'\right)\widehat{g}_T = T\widehat{g}_T' P\left(H_T B_T \Upsilon_T \left\{(H_T B_T \Upsilon_T)'(H_T B_T \Upsilon_T)\right\}^- \Upsilon_T B_T' R' \bar{\Pi}_T \bar{D}_T\right)\widehat{g}_T$$

where $H_T := \widehat{V}_T^{-1/2}\widehat{G}_T$, $\Upsilon_T := \text{diag}(1/\delta_{T,1}, \ldots, 1/\delta_{T,p}, \sqrt{T}1_{d_\theta - p})$, a $d_\theta \times d_\theta$ diagonal matrix, nonsingular for any given T. (Recall that by 1_c we mean a $1 \times c$ vector with all elements equal to 1.) Note that Υ_T is $\text{diag}(1/\delta_{T,1}, \ldots, 1/\delta_{T,p})$ if $d_\theta = p$ and is $\text{diag}(\sqrt{T}1_{d_\theta - p})$ if $p = 0$. For a given T, $\bar{\Pi}_T$ and $\bar{D}_T$ are $d_R \times d_R$ nonsingular matrices defined as follows.

Step 1: Definition of $\bar{\Pi}_T$ and $\bar{D}_T$, and the asymptotic behavior of $\Upsilon_T B_T' R' \bar{\Pi}_T \bar{D}_T$

Note that under assumption M3(a) we can, without loss of generality, partition the set of elements $\delta_{T,1}, \ldots, \delta_{T,p}$ into $m-1$ groups containing $p_1, p_2, \ldots, p_{m-1}$ elements respectively as $(\delta_{T,1}, \ldots, \delta_{T,p_1})$, $(\delta_{T,\bar{p}_1+1}, \ldots, \delta_{T,\bar{p}_2})$,..., $(\delta_{T,\bar{p}_{m-2}+1}, \ldots, \delta_{T,\bar{p}_{m-1}})$ where $p_j \geq 0$ and $\bar{p}_j := \sum_{k=1}^j p_k$ for $j = 1, \ldots, m-1$ and $m \in \{1, \ldots, p+1\}$ (let $p_m := d_\theta - p$; and when $p = 0$ let $m = 1$; and also, by construction, $\bar{p}_{m-1} = p$ and $\bar{p}_m = d_\theta$), such that:

$$\delta_{T,\bar{p}_j} \neq o(\delta_{T,\bar{p}_j - p_j + 1}) \text{ for } j = 1, \ldots, m-1, \text{ and } \delta_{T,\bar{p}_j+1} = o(\delta_{T,\bar{p}_j}) \text{ for } j = 1, \ldots, m-2. \tag{32}$$

Now, define $\bar{\Pi}_T$ as the Π_T matrix from the UBT-Construction in Section 3.2 and with $W_T := RB_T = [W_{T,1}, \ldots, W_{T,m}]$ where $W_{T,j} := RB_{T,(\bar{p}_j - p_j + 1:\bar{p}_j)}$ for $j = 1, \ldots, m$. Since B_T is orthogonal for each T and also $B_T \to B$, which is nonsingular by M3(c), the $q_T = q$ and $c^*_{T,j_i,T} = c^*_{j_i}$, i.e., these quantities in the UBT-Construction do not depend on T.[14] $\sum_{i=1}^q c^*_{j_i} = d_R$.

Define $\bar{D}_T = \text{diag}(\delta_{T,\bar{p}_{j_1}} 1_{c^*_{j_1}}, \ldots, \delta_{T,\bar{p}_{j_q}} 1_{c^*_{j_q}})$ where, for simplicity, we use the notation $\delta_{T,\bar{p}_{m-1}+1} = \ldots = \delta_{T,\bar{p}_m} = T^{-1/2}$ to accommodate for the possible case that $j_q = m$. $\bar{D}_T$ is a $d_R \times d_R$ nonsingular diagonal matrix for each T.

Therefore, as $T \to \infty$, it follows by M3(a) and (32), and then again using Lemma 3.8, that:

$$\Upsilon_T B_T' R' \bar{\Pi}_T \bar{D}_T \to W^{*'}, \text{ say, where } W^{*'} \text{ is a finite, non-random, } d_\theta \times d_R \text{ matrix with full column-rank } d_R. \tag{33}$$

In particular, by using arguments similar to those below (18) along with M3(a), we obtain for the matrix $W^{*'}$ that its columns from $(d_R - \sum_{i'=i}^q c^*_{j_{i'}})$ to $(d_R - \sum_{i'=i}^q c^*_{j_{i'}} + c^*_{j_i})$ for $i = 1, \ldots, q$ are represented by the $d_g \times c^*_{j_i}$ matrix: $\left[(\delta_{\bar{p}_1}\text{diag}(\delta_1^{-1}, \ldots, \delta_{\bar{p}_1}^{-1})B'_{(1:p_1)}R'\bar{\Pi}_1)', 0'\right]'$ if $j_i = 1$, and $\left[0', (\delta_{\bar{p}_{j_i}}\text{diag}(\delta^{-1}_{\bar{p}_{j_i} - p_{j_i}+1}, \ldots, \delta^{-1}_{\bar{p}_{j_i}})B'_{(\bar{p}_{j_i} - p_{j_i}+1:\bar{p}_{j_i})}R'\bar{\Pi}_{j_i})', 0'\right]'$ otherwise (as it was below (18), 0 denotes sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding big matrix is d_θ). Thus the non-zero blocks in such sets of columns (one block per set of columns) are: (a) at mutually non-overlapping positions (sets of rows); (b) are finite by M1, M3(a); (c) of full column-rank by Lemma 3.8, which tells that pre-multiplication by the nonsingular matrix $\delta_{\bar{p}_{j_i}}\text{diag}(\delta^{-1}_{\bar{p}_{j_i} - p_{j_i}+1}, \ldots, \delta^{-1}_{\bar{p}_{j_i}})$ does not change the rank of $B'_{(\bar{p}_{j_i} - p_{j_i}+1:\bar{p}_{j_i})}R'\bar{\Pi}_{j_i}$. The latter has full column-rank $c^*_{j_i}$ for $i = 1, \ldots, q$ by (i) in the UBT-Construction. Therefore, full column-rank d_R of $W^{*'}$ follows by noting that $\sum_{i=1}^q c^*_{j_1} = d_R$.

The rest of the proof is completely based on Andrews and Guggenberger (2014).

Step 2: Asymptotic behavior of $H_T B_T \Upsilon_T$

Under (12), $\|\Delta_T\| \leq c \times \bar{c}$ for some $c > 0$ by M2. Then it follows that

$$\begin{aligned}
V_T^{-1/2}\widehat{G}_T B_T \Upsilon_T &= V_T^{-1/2}\widehat{G}_T\left[B_{T,(1:p)}\Delta^{-1}_{T,(1:p)}, \sqrt{T}B_{T,(p+1:d_\theta)}\right] \\
&= V_T^{-1/2}G_T\left[B_{T,(1:p)}\Delta^{-1}_{T,(1:p)}, \sqrt{T}B_{T,(p+1:d_\theta)}\right] \\
&\quad + V_T^{-1/2}\sqrt{T}\left(\widehat{G}_T - G_T\right)\left[B_{T,(1:p)}(\sqrt{T}\Delta_{T,(1:p)})^{-1}, B_{T,(p+1:d_\theta)}\right].
\end{aligned}$$

[14] We use B_T instead of B in the UBT-Construction to avoid strong conditions on the rate of convergence of $B_T \to B$ as $T \to \infty$.

By the orthogonality of B_T it follows from the relation $V_T^{-1/2}G_T = C_{T,(1:d_R)}\Delta_T B_T'$ (obtained from (12)) and M3, that the first term on the right hand side of the above equation converges to $[C_{(1:p)}, C_{(p+1:d_\theta)}L]$. On the other hand, M1 and M2 give $\sqrt{T}\left(\widehat{G}_T - G_T\right) \xrightarrow{d} devec_{d_g}(\psi_G - V_{Gg}V^{-1}\psi) = O_p(1)$ which, crucially, is independent of ψ. Also M3 implies that $[B_{T,(1:p)}(\sqrt{T}\Delta_{T,(1:p)})^{-1}, B_{T,(p+1:d_\theta)}] \to [0, B_{(p+1:d_\theta)}]$ as $T \to \infty$. Thus, by M1, the second term on the right hand side of the above equation (i.e., $V_T^{-1/2}\sqrt{T}\left(\widehat{G}_T - G_T\right)\left[B_{T,(1:p)}(\sqrt{T}\Delta_{T,(1:p)})^{-1}, B_{T,(p+1:d_\theta)}\right]$) converges in distribution to $[0, V^{-1/2}devec_{d_g}(\psi_G - V_{Gg}V^{-1}\psi)R^1 B_{(p+1:d_\theta)}]$. Since M2 implies that $\widehat{V}_T^{-1/2}V_T^{1/2} \xrightarrow{p} I_{d_g}$, it follows that

$$H_T B_T \Upsilon_T = \widehat{V}_T^{-1/2}\widehat{G}_T B_T \Upsilon_T = \left(\widehat{V}_T^{-1/2}V_T^{1/2}\right)V_T^{-1/2}\widehat{G}_T B_T \Upsilon_T \xrightarrow{d} G^* \tag{34}$$

where $G^* := [C_{(1:p)}, C_{(p+1:d_\theta)}L + V^{-1/2}devec_{d_g}(\psi_G - V_{Gg}V^{-1}\psi)B_{(p+1:d_\theta)}]$, as defined in M3(d).

Step 3: Asymptotic behavior of $LM_T(\theta^0)$

Therefore, $P(H_T B_T \Upsilon_T\{(H_T B_T \Upsilon_T)'(H_T B_T \Upsilon_T)\}^{-}\Upsilon_T B_T' R' \bar{\Pi}_T \bar{D}_T) \xrightarrow{d} P(G^*(G^{*'}G^*)^{-1}W^{*'})$, a finite matrix with full column-rank d_R almost surely by (33), (34) and Lemma 3.8. Now, since M1 and M2 imply that $\sqrt{T}\widehat{g}_T \xrightarrow{d} V^{-1/2}\psi \sim N(0, I_{d_g})$, and since we already noted the independence between ψ and G^*, it follows that $LM_T(\theta^0) \xrightarrow{d} \chi^2_{d_R}$. ■

Proof of Proposition 3.2: Recall that our definition of the improved test accommodates for the convention that $\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) = \infty$ if $CI_T(\gamma_S;\epsilon)$ is empty. Let $\{\phi_{\gamma_S,T} : T \geq 1\}$ denote the sequence of indicator variables where $\phi_{\gamma_S,T} = 0$ if $CI_T(\gamma_S;\epsilon)$ contains γ_S^0, and $\phi_{\gamma_S,T} = 1$ otherwise. Since it is given that $CI_T(\gamma_S;\epsilon)$ has asymptotic coverage $(1-\epsilon)$ when H_0 is true, naturally, $\lim_{T\to\infty} Pr_T(\phi_{\gamma_S,T} = 0) \geq (1-\epsilon)$ where $Pr_T(.)$ denotes the probability of an event under F_T constrained by assumptions O and M1-M3 and when $\beta^0 = r_0$ (equivalently, $R\theta^0 = r_0$). Therefore, by construction:

$$\lim_{T\to\infty} Pr_T\left(\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) \leq LM_T(\theta^0)\right) \geq \lim_{T\to\infty} Pr_T(\phi_{\gamma_S,T} = 0) \geq 1 - \epsilon, \tag{35}$$

since for any $T \geq 1$, the event $\{\phi_{\gamma_S,T} = 0\}$ implies the event $\{\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) \leq LM_T(\theta^0)\}$.

Let $\{\phi_{\beta,T} : T \geq 1\}$ denote the sequence of indicator variables where $\phi_{\beta,T} = 1$ if $\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > \chi^2_{d_R}(1-\alpha)$, and $\phi_{\beta,T} = 0$ otherwise. Therefore,

$$\begin{aligned}
Pr_T\left(\phi_{\beta,T} = 0\right) &= Pr_T\left(\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) \leq \chi^2_{d_R}(1-\alpha)\right) \\
&\geq Pr_T\left(\left\{\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) \leq LM_T(\theta^0)\right\}\bigcap\left\{LM_T(\theta^0) \leq \chi^2_{d_R}(1-\alpha)\right\}\right) \\
&= 1 - Pr_T\left(\left\{\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > LM_T(\theta^0)\right\}\bigcup\left\{LM_T(\theta^0) > \chi^2_{d_R}(1-\alpha)\right\}\right) \\
&\geq 1 - \left(Pr_T\left(\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > LM_T(\theta^0)\right) + Pr_T\left(LM_T(\theta^0) > \chi^2_{d_R}(1-\alpha)\right)\right)
\end{aligned}$$

where the first line follows by the definition of $\phi_{\beta,T}$, the second line by the construction of the improved projection test, the third line by De Morgan's law, and the fourth line by Bonferroni's inequality. Taking limits on both sides gives:

$$\begin{aligned}
\lim_{T\to\infty} Pr_T\left(\phi_{\beta,T} = 0\right) &\geq 1 - \lim_{T\to\infty}\left(Pr_T\left(\inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1}(r_0', \gamma_0')'\right) > LM_T(\theta^0)\right) + Pr_T\left(LM_T(\theta^0) > \chi^2_{d_R}(1-\alpha)\right)\right) \\
&\geq 1 - (\epsilon + \alpha)
\end{aligned}$$

where the last line follows by (35) and Lemma 3.1. ■

Remark: Since the way it is stated in the statement of the proposition, the coverage probability of $CI_T(\gamma_S;\epsilon)$ is $(1-\epsilon)$ possibly under a larger class of distributions than F_T constrained by assumptions O and M1-M3. This is the reason behind the inequality $\lim_{T\to\infty} Pr_T(\phi_{\gamma_S,T} = 0) \geq (1-\epsilon)$. However, the confidence sets $CI_T(\gamma_S;\epsilon)$, e.g., $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ defined in (13), that we actually specify are asymptotically similar and hence for them the above inequality will hold as an equality.

Lemma 3.9 *Let assumptions O and N hold. Consider a sequence $\{\theta_T = R_S^1 r_0 + S_S^1 \gamma_{S,T} : T \geq 1\}$ where r_0 satisfies* (24) *and $\{\gamma_{S,T} : T \geq 1\}$ is such that θ_T satisfies* (25)*. Then the following results hold as $T \to \infty$:*

(a) $\widehat{V}_T(\theta_T) \xrightarrow{P} V(\theta^0) \equiv V$.

(b) $\widehat{V}_{Gg,T}(\theta_T) \xrightarrow{P} V_{Gg}(\theta^0) \equiv V_{Gg}$.

(c) $\bar{G}_T(\theta_T)\Pi_{\rho_\theta}D_{T,\rho_\theta} \xrightarrow{P} G^*$ *where* Π_{ρ_θ}, D_{T,ρ_θ} *and* G^* *are as defined in* (17), (18) *and* (20) *respectively.*

(d) $\sqrt{T}\bar{g}_T(\theta_T) = \sqrt{T}\bar{g}_T(\theta^0) + G^*\mu_{T,\theta} + o_p(1)$ *where* G^* *and* $\mu_{T,\theta}$ *are as defined in* (20) *and* (25) *respectively.*

(e) $\left[\widehat{V}_{1,g,T}(\theta_T)\widehat{V}_T^{-1}(\theta_T)\bar{g}_T(\theta_T), \ldots, \widehat{V}_{d_\theta,g,T}(\theta_T)\widehat{V}_T^{-1}(\theta_T)\bar{g}_T(\theta_T)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} = o_p(1)$ *(a $d_g \times d_\theta$ matrix).*

(f) $\widehat{G}_T(\theta_T)\Pi_{\rho_\theta}D_{T,\rho_\theta} \xrightarrow{P} G^*$ *where* Π_{ρ_θ}, D_{T,ρ_θ} *and* G^* *are as defined in* (17), (18) *and* (20) *respectively.*

Proof: (a) and (b) follow by assumption N8 since $\theta_T = \theta^0 + o_p(1)$.

(c) We prove it working term-by-term in the following decomposition:

$$\bar{G}_T(\theta_T)\Pi_{\rho_\theta}D_{T,\rho_\theta} = \left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} + \sqrt{T}\left[\bar{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta^0)\right]\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta^0)\Pi_{\rho_\theta}D_{T,\rho_\theta}. \quad (36)$$

From the definitions in (17) and (18) it follows that $\Pi_{\rho_\theta}D_{T,\rho_\theta} = o(\sqrt{T})$ by N3, and hence using N6 it follows that the second term on the right hand side (RHS) of (36) is $o_p(1)$. On the other hand, (19) and (20) imply that the third term on the RHS of (36) converges to G^* by construction.

To complete the proof, now let us show that the first term on the RHS of (36) is $o_p(1)$. It is the treatment of this term where we deviate from Antoine and Renault (2012), and the result thus obtained has substantive implications in terms of the allowable weakness of identification in the system. Let $\bar{G}_{T,i}(\theta) := \frac{\partial}{\partial\theta_i}\bar{g}_T(\theta)$ denote the i-th column of $\bar{G}_T(\theta)$ for $i = 1, \ldots, d_\theta$ (recall that $\theta = (\theta_1, \ldots, \theta_{d_\theta})'$). Therefore, with a bad but common abuse of notation in denoting the mean values element by element, we obtain by a mean value expansion of $\bar{G}_{T,i}(\theta_T)$ around $\bar{G}_{T,i}(\theta^0)$ for $i = 1, \ldots, d_\theta$ that:

$$\begin{aligned}\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} &= \left[\left\{\frac{\partial}{\partial\theta'}\bar{G}_{T,1}(\theta_T(\theta_1))\right\}(\theta_T - \theta^0), \ldots, \left\{\frac{\partial}{\partial\theta'}\bar{G}_{T,d_\theta}(\theta_T(\theta_{d_\theta}))\right\}(\theta_T - \theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} \\ &= \left[\left\{\frac{\partial}{\partial\theta_1}\bar{G}_T(\theta_T(\theta_1))\right\}(\theta_T - \theta^0), \ldots, \left\{\frac{\partial}{\partial\theta_{d_\theta}}\bar{G}_T(\theta_T(\theta_{d_\theta}))\right\}(\theta_T - \theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} \quad (37)\end{aligned}$$

by twice interchanging the order in which the derivatives are taken in each of the d_θ columns. Note that, for $i = 1, \ldots, d_\theta$, we used $\theta_T(\theta_i)$ (such that $\|\theta_T(\theta_i) - \theta^0\| \leq \|\theta_T - \theta^0\|$) to denote the mean value, row by row, on the first line of the above equation. Recalling that $\mu_{T,\theta} = \sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta_T - \theta^0)$ by (25), define $U_{T,i}$ for $i = 1, \ldots, d_\theta$ as the $d_g \times d_\theta$ matrix with

$$\left\{\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta_T(\theta_i))\right\}(\theta_T - \theta^0) = \left\{\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta_T(\theta_i))\frac{\sqrt{T}}{\lambda_{T,l}}\right\}\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\mu_{T,\theta}\frac{\lambda_{T,l}}{\sqrt{T}\lambda_{T,j_1}}$$

in the i-th column and zero everywhere else. (See N7(b) for more on λ_{T,j_1}.) Therefore, (37) implies that

$$\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta} = \sum_{i=1}^{d_\theta}U_{T,i}\Pi_{\rho_\theta}D_{T,\rho_\theta}$$

and thus

$$\begin{aligned}&\left\|\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta}\right\| \\ \leq\ & \sum_{i=1}^{d_\theta}\|U_{T,i}\| \times \|\Pi_{\rho_\theta}D_{T,\rho_\theta}\| \\ \leq\ & \sum_{i=1}^{d_\theta}\left\|\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta_T(\theta_i))\frac{\sqrt{T}}{\lambda_{T,l}}\right\| \times \left\|\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\right\| \times \|\mu_{T,\theta}\| \times \left\|\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\right\|\frac{\sqrt{T}\lambda_{T,l}}{\lambda_{T,j_1}^2\sqrt{T}} \\ \leq\ & \sum_{i=1}^{d_\theta}\sup_\theta\left\{\left\|\frac{\sqrt{T}}{\lambda_{T,l}}\frac{\partial}{\partial\theta_i}\frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right\| + \left\|\frac{\sqrt{T}}{\lambda_{T,l}}\frac{\partial}{\partial\theta_i}\left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right]\right\|\right\} \times \|\mu_{T,\theta}\| \times \left\|\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\right\|^2\frac{\lambda_{T,l}}{\lambda_{T,j_1}^2} \\ =\ & o_p(1)\end{aligned}$$

since, on the third line from above, the order of magnitude of the terms (from left to right) inside the sum is respectively: (i) $\sup_\theta\left\|\frac{\sqrt{T}}{\lambda_{T,l}}\frac{\partial}{\partial\theta_i}\frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right\| = O(1)$ by N3 and N4, (ii) $\sup_\theta\left\|\frac{\sqrt{T}}{\lambda_{T,l}}\frac{\partial}{\partial\theta_i}\left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right]\right\| = o_p(1)$ by N3 and N7(a), (iii) $\|\mu_{T,\theta}\| = O_p(1)$ by (25), (iv) $\left\|\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\right\| = O(1)$ by N3, (17) and (18), and (v) $\frac{\lambda_{T,l}}{\lambda_{T,j_1}^2} = o(1)$ by N7(b).

(d) A mean value expansion (with similar abuse of notation as above to denote the mean value, this time, $\bar{\theta}_T$) gives $\sqrt{T}\bar{g}_T(\theta_T) = \sqrt{T}\bar{g}_T(\theta^0) + \bar{G}_T(\bar{\theta}_T)\sqrt{T}(\theta_T - \theta^0) = \sqrt{T}\bar{g}_T(\theta^0) + \bar{G}_T(\bar{\theta}_T)\Pi_{\rho_\theta}D_{T,\rho_\theta}\mu_{T,\theta} = \sqrt{T}\bar{g}_T(\theta^0) + G^*\mu_{T,\theta} + o_p(1)$ where the second equality uses (25) and the last equality uses the result from (c).

(e) The result follows by using (a), (b), (d) and since $\Pi_{\rho_\theta}D_{T,\rho_\theta} = o(\sqrt{T})$ by N3.

(f) The result follows by (c) and (e). ■

Remark:

Let us briefly elaborate on Remark 2 following assumption N7 from the main text. Take $a = 3$ there. An argument by induction can be used to extend it to a general a, to essentially demonstrate that more smoothness in the moment vector helps to weaken the restrictions on the relative order of magnitude of the $\lambda_{\lambda_{T,j}}$'s, and in the limit ($a \to \infty$), an example of which is the linear instrumental variables regression, we would not need restrictions beyond N3.

Since the main idea remains the same for all a, we focus on $a = 3$ to avoid further clutter in notation. From the above lemma, which is key to all the results under the local deviation from the truth, it is clear that Remark 2 is pertinent only to part (c) of this lemma. Indeed the only part of (c) that needs attention is where we show that $\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta}$, i.e., the first term on the RHS of (36), is $o_p(1)$.

To accommodate for $a = 3$ we extend assumption N6 as N6' to the second derivative, and replace N7 by N7' as follows. Naturally, the existence of the derivatives of appropriate order are assumed. (Assumptions N1-N5 and N8 remain the same.)

Assumption N6': (a one-time assumption for this remark only)

(a) $\frac{\partial}{\partial\theta'}\psi_T(\theta^0) = \sqrt{T}\left[\bar{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta^0)\right] = O_p(1)$. (This was the original N6.)

(b) For $i = 1,\dots,d_\theta$: $\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta'}\psi_T(\theta^0) = \sqrt{T}\frac{\partial}{\partial\theta_i}\left[\bar{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta^0)\right] = O_p(1)$. (This is the extension.)

Assumption N7': (a one-time assumption for this remark only)

(a) $\rho(\theta)$ is thrice continuously differentiable in $\theta \in \mathrm{int}(\Theta)$. $g(z;\theta)$ is thrice differentiable in $\theta \in \mathrm{int}(\Theta)$ for each $z \in \mathbb{R}^{d_z}$ and $\sup_{\theta\in\mathrm{int}(\Theta)}\left\|\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_k}\left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta)\right]\right\| = o_p(\lambda_{T,l}/\sqrt{T})$ for $i,k = 1,\dots,d_\theta$.

(b) λ_{T,j_1} from (18) satisfies $\lambda^3_{T,j_1}/\lambda_{T,l} \to \infty$ as $T \to \infty$.

Comparing assumption N7 with N7' reveals the tradeoff in terms of parts (a) and (b) of these assumptions. We note that for $a = 4, 5, \dots$, similar tradeoffs would result in the same result as we obtain below for $a = 3$.

For clarity, introduce further structure but without loss of generality. First, for $i = 1,\dots,d_\theta$, define $\Pi_{\rho_\theta,i}$ and $D_{T,\rho_\theta,i}$ by the UBT-Construction in a similar to that in (17) and (18), but this time, by taking

$$W_T = \left[\frac{\partial}{\partial\theta_i}\rho'_{\theta,1}(\theta^0),\dots,\frac{\partial}{\partial\theta_i}\rho'_{\theta,l}(\theta^0)\right] = \left(I^*\frac{\partial}{\partial\theta_i}\rho_\theta(\theta^0)\right)'$$

(instead of $W_T = \left[\rho'_{\theta,1}(\theta^0),\dots,\rho'_{\theta,l}(\theta^0)\right] = \left(I^*\rho_\theta(\theta^0)\right)'$) not depending on T in the UBT-Construction. The corresponding quantities with full column-rank, and thus also the elements of $D_{T,\rho_\theta,i}$ will change. Indeed no full-rank conditions are required, and instead, for the purpose of this proof, the only properties we will require are: For $i = 1,\dots d_\theta$,

$$\begin{aligned}\left(I^{*'}\left\{\frac{\partial}{\partial\theta_i}I^*\frac{\Lambda_T}{\sqrt{T}}I^{*'}I^*\rho_\theta(\theta^0)\right\}\Pi_{\rho_\theta,i}D_{T,\rho_\theta,i}\right) &= O(1), &(38)\\ \Pi_{\rho_\theta,i}D_{T,\rho_\theta,i} &= o(\sqrt{T}) &(39)\end{aligned}$$

and these will not change since (38) holds by the construction of $D_{T,\rho_\theta,i}$, while (39) follows from N3.

Start from (37). All we do below is to tease out further structure in the non-zero (i.e.,the i-th) column of $U_{T,i}$ (defined below (37)) so that assumption N7' could be effectively used to show that $\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right]\Pi_{\rho_\theta}D_{T,\rho_\theta}$, i.e., the first term on the RHS of (36) is $o_p(1)$. With this purpose in mind, for each $i = 1,\dots,d_\theta$, consider a further mean value expansion (with similar abuse of notation, and this time using $\theta_T(\theta_i^k)$ to denote the mean value such that $\|\theta_T(\theta_i^k) - \theta^0\| \le \|\theta_T(\theta_i) - \theta^0\| \le \|\theta_T - \theta^0\|$ for $k = 1,\dots,d_\theta$):

$$\begin{aligned}\left\{\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta_T(\theta_i))\right\}(\theta_T - \theta^0) &= \left\{\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta^0)\right\}(\theta_T - \theta^0)\\ &+\left[\left\{\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_1}\bar{G}_T(\theta_T(\theta_i^1))\right\}(\theta_T(\theta_i) - \theta^0),\dots,\left\{\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_{d_\theta}}\bar{G}_T(\theta_T(\theta_i^{d_\theta}))\right\}(\theta_T(\theta_i) - \theta^0)\right](\theta_T - \theta^0)\end{aligned}$$

by similar (to above) interchange in the order of the derivatives. Since $\mu_{T,\theta} = \sqrt{T}D^{-1}_{T,\rho_\theta}\Pi^{-1}_{\rho_\theta}(\theta_T - \theta^0)$ by (25), it follows that

$$\begin{aligned}\left\{\frac{\partial}{\partial\theta_i}\bar{G}_T(\theta^0)\right\}(\theta_T - \theta^0) &= \underbrace{\left(I^{*'}\left\{\frac{\partial}{\partial\theta_i}I^*\frac{\Lambda_T}{\sqrt{T}}I^{*'}I^*\rho_\theta(\theta^0)\right\}\Pi_{\rho_\theta,i}D_{T,\rho_\theta,i}\right)\mu_{T,\theta}\frac{1}{\sqrt{T}}}_{=\,u_{a,T,i}\ \text{(say)}}\\ &+\underbrace{\left(\frac{\partial}{\partial\theta_i}\sqrt{T}\left(\bar{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}}\rho_\theta(\theta^0)\right)\right)\left(\frac{\Pi_{\rho_\theta,i}D_{T,\rho_\theta,i}}{\sqrt{T}}\right)\mu_{T,\theta}\frac{1}{\sqrt{T}}}_{=\,u_{b,T,i}\ \text{(say)}}\end{aligned}$$

for $i = 1,\dots,d_\theta$. Define the $d_g \times d_\theta$ matrices $U_{a,T,i}$ and $U_{b,T,i}$ such that all their columns are zeros, except for the i-th column, which for them is $u_{a,T,i}$ and $u_{b,T,i}$ respectively. Do this for all $i = 1,\dots,d_\theta$.

For the notation-abused quantity $\theta_T(\theta_i)$, define $\mu_{T,\theta(i)} := \sqrt{T}D^{-1}_{T,\rho_\theta}\Pi^{-1}_{\rho_\theta}(\theta_T(\theta_i) - \theta^0)$ where $\|\mu_{T,\theta(i)}\| \le \|\mu_{T,\theta}\|$ by construction for $i = 1,\dots,d_\theta$ (recall that $\mu_{T,\theta} = \sqrt{T}D^{-1}_{T,\rho_\theta}\Pi^{-1}_{\rho_\theta}(\theta_T - \theta^0)$ by (25)). Now for each $i = 1,\dots,d_\theta$ define the $d_g \times d_\theta$ matrices $U_{c,T,i,k}$ for $k = 1,\dots,d_\theta$ such that all columns of $U_{c,T,i,k}$ are zeros, except for the k-th column which is

$$\left\{\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_k}\bar{G}_T(\theta_T(\theta_i^k))\right\}(\theta_T(\theta_i) - \theta^0) = \left\{\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_k}\bar{G}_T(\theta_T(\theta_i^k))\frac{\sqrt{T}}{\lambda_{T,l}}\right\}\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\mu_{T,\theta(i)}\frac{\lambda_{T,l}}{\sqrt{T}\lambda_{T,j_1}}.$$

Therefore, it follows that $U_{T,i}$ (defined below (37)) can be written as:

$$U_{T,i} = U_{a,T,i} + U_{b,T,i} + \left(\sum_{k=1}^{d_\theta}U_{c,T,i,k}\right)(\theta_T - \theta^0) = U_{a,T,i} + U_{b,T,i} + \left(\sum_{k=1}^{d_\theta}U_{c,T,i,k}\right)\frac{\Pi_{\rho_\theta}D_{T,\rho_\theta}\lambda_{T,j_1}}{\sqrt{T}}\mu_{T,\theta}\frac{1}{\lambda_{T,j_1}}.$$

And, therefore,

$$\begin{aligned}
& \left\| \left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right] \Pi_{\rho_\theta} D_{T,\rho_\theta} \right\| \\
\leq \; & \sum_{i=1}^{d_\theta} \|U_{T,i}\| \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| \\
\leq \; & \sum_{i=1}^{d_\theta} \|U_{a,T,i}\| \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| + \sum_{i=1}^{d_\theta} \|U_{b,T,i}\| \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| + \sum_{i=1}^{d_\theta}\sum_{k=1}^{d_\theta} \|U_{c,T,i,k}\| \times \left\| \frac{\Pi_{\rho_\theta} D_{T,\rho_\theta} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \frac{\|\mu_{T,\theta}\|}{\lambda_{T,j_1}} \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\|.
\end{aligned}$$

Since $\|u_{a,T,i}\| = O(1/\sqrt{T})$ by its definition and using (38), it follows that $\sum_{i=1}^{d_\theta} \|U_{a,T,i}\| \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| = o_p(1)$ by then using (39). Since $\|u_{b,T,i}\| = O(1/\sqrt{T})$ by its definition and using N6' and (39), it follows that $\sum_{i=1}^{d_\theta} \|U_{b,T,i}\| \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| = o_p(1)$ by then using (39). Finally, note that $\sum_{i=1}^{d_\theta}\sum_{k=1}^{d_\theta} \|U_{c,T,i,k}\| \times \left\| \frac{\Pi_{\rho_\theta} D_{T,\rho_\theta} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \frac{\|\mu_{T,\theta}\|}{\lambda_{T,j_1}} \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| = o_p(1)$ since (collecting similar terms together)

$$\begin{aligned}
& \|U_{c,T,i,k}\| \times \left\| \frac{\Pi_{\rho_\theta} D_{T,\rho_\theta} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \frac{\|\mu_{T,\theta}\|}{\lambda_{T,j_1}} \times \|\Pi_{\rho_\theta} D_{T,\rho_\theta}\| \\
\leq \; & \sup_\theta \left\{ \left\| \frac{\Lambda_T}{\sqrt{T}} \frac{\partial}{\partial\theta_i} \frac{\partial}{\partial\theta_k} \rho_\theta(\theta) \times \frac{\sqrt{T}}{\lambda_{T,l}} \right\| + \left\| \frac{\partial}{\partial\theta_i} \frac{\partial}{\partial\theta_k} \left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}} \rho_\theta(\theta) \right] \frac{\sqrt{T}}{\lambda_{T,l}} \right\| \right\} \times \left\| \frac{\Pi_{\rho_\theta} D_{T,\rho_\theta} \lambda_{T,j_1}}{\sqrt{T}} \right\|^3 \times \|\mu_{T,\theta(i)}\| \times \|\mu_{T,\theta}\| \times \frac{\lambda_{T,l}}{\lambda_{T,j_1}^3} \\
= \; & O_p(1) \times O(1) \times O_p(1) \times O_p(1) \times o(1)
\end{aligned}$$

term by term: (i) $\sup_\theta \left\| \frac{\Lambda_T}{\sqrt{T}} \frac{\partial}{\partial\theta_i} \frac{\partial}{\partial\theta_k} \rho_\theta(\theta) \times \frac{\sqrt{T}}{\lambda_{T,l}} \right\| + \left\| \frac{\partial}{\partial\theta_i} \frac{\partial}{\partial\theta_k} \left[\bar{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}} \rho_\theta(\theta) \right] \frac{\sqrt{T}}{\lambda_{T,l}} \right\| = O(1) + o_p(1)$ by using N3 and N7'(a), (ii) $\left\| \frac{\Pi_{\rho_\theta} D_{T,\rho_\theta} \lambda_{T,j_1}}{\sqrt{T}} \right\|^3 = O(1)$ by using N3, (17) and (18), (iii) $\|\mu_{T,\theta(i)}\| = O_p(1)$ by using (25) and the definition of $\mu_{T,\theta(i)}$, (iv) $\|\mu_{T,\theta}\| = O_p(1)$ by using (25), and (v) $\frac{\lambda_{T,l}}{\lambda_{T,j_1}^3} = o(1)$ by using N7'(b). Thus $\left[\bar{G}_T(\theta_T) - \bar{G}_T(\theta^0)\right] \Pi_{\rho_\theta} D_{T,\rho_\theta} = o_p(1)$. ■

Proof of Lemma 3.3: (a) Utilizing the constructions of the nonsingular matrices Π_{ρ_θ}, D_{T,ρ_θ}, Π_R and $D_{T,R}$ in (17),(18), (21) and (22) respectively, recall from (4) that $LM_T(\theta)$ and $LM_T^{\text{infeas}}(\theta)$ can be written as

$$\begin{aligned}
LM_T(\theta) &= T \times \left(\widehat{V}_T^{-1/2}(\theta) \bar{g}_T(\theta) \right)' P \left(\widehat{H}_T (\widehat{H}_T' \widehat{H}_T)^- D_{T,\rho_\theta} \Pi_{\rho_\theta}' R' \Pi_R D_{T,R} \right) \left(\widehat{V}_T^{-1/2}(\theta) \bar{g}_T(\theta) \right), \\
LM_T^{\text{infeas}}(\theta) &= T \times \left(V_T^{-1/2}(\theta) \bar{g}_T(\theta) \right)' P \left(H_T (H_T' H_T)^- D_{T,\rho_\theta} \Pi_{\rho_\theta}' R' \Pi_R D_{T,R} \right) \left(V_T^{-1/2}(\theta) \bar{g}_T(\theta) \right),
\end{aligned}$$

where $\widehat{H}_T(\theta) := \widehat{V}_T^{-1/2}(\theta) \widehat{G}_T(\theta) \Pi_{\rho_\theta} D_{T,\rho_\theta}$ and $H_T := V_T^{-1/2}(\theta) E_T[\bar{G}_T(\theta)] \Pi_{\rho_\theta} D_{T,\rho_\theta}$ respectively. Essentially $LM_T^{\text{infeas}}(\theta)$ is $LM_T^{\text{infeas}}(\theta^{\text{infeas}})$, but without plugging in θ^{infeas} in place of the general θ. Now recall that for θ_T defined in (25), we have:

(i) $\widehat{V}_T^{-1/2}(\theta_T) \xrightarrow{P} V^{-1/2}$ by N8 and $V_T^{-1/2}(\theta_T) \to V^{-1/2}$ by definition;

(ii) $\widehat{V}_T^{-1/2}(\theta_T) \sqrt{T} \bar{g}_T(\theta_T) = V_T^{-1/2}(\theta_T) \sqrt{T} \bar{g}_T(\theta_T) + o_p(1) = V^{-1/2} [\sqrt{T} \bar{g}_T(\theta^0) + G^* \mu_{T,\theta}] + o_p(1)$ by (i) and Lemma 3.9(c); and this is $O_p(1)$ by N2 and N8;

(iii) $D_{T,\rho_\theta} \Pi_{\rho_\theta}' R' \Pi_R D_{T,R} \to R^{*'}$ by (23).

Therefore, to show that $LM_T(\theta_T) = LM_T^{\text{infeas}}(\theta_T) + o_p(1)$, it suffices to show that $\widehat{H}_T - H_T = o_p(1)$. Thus, by virtue of (i) and Lemma 3.9(f), it suffices to show that $E_T[\bar{G}_T(\theta_T)] \Pi_{\rho_\theta} D_{T,\rho_\theta} \to G^*$. Further, by virtue of (20), it is now sufficient to show that $E_T[(G_T(\theta_T) - G_T(\theta^0)] \Pi_{\rho_\theta} D_{T,\rho_\theta} = o(1)$. This follows exactly by proceeding from (37) onward in the proof of Lemma 3.9 simply by replacing $G_T(.)$ in that proof with $E_T[G_T(.)]$. Thus $LM_T(\theta_T) = LM_T^{\text{infeas}}(\theta) + o_p(1)$.

Now, using (ii), (iii), N4, N8 and the fact that we just established $H_T \to V^{-1/2} G^*$, note that:

$$\begin{aligned}
\left(H_T (H_T' H_T)^- D_{T,\rho_\theta} \Pi_{\rho_\theta}' R' \Pi_R D_{T,R} \right)' \left(V_T^{-1/2}(\theta) \sqrt{T} \bar{g}_T(\theta) \right) &= R^* (G^{*'} V^{-1} G^*)^{-1} G^{*'} V^{-1} [\sqrt{T} \bar{g}_T(\theta^0) + G^* \mu_{T,\theta}] + o_p(1) \\
&= R^* (G^{*'} V^{-1} G^*)^{-1} G^{*'} V^{-1} \sqrt{T} \bar{g}_T(\theta^0) + R^* \mu_{T,\theta} + o_p(1) \\
&= R^* (G^{*'} V^{-1} G^*)^{-1} G^{*'} V^{-1} \sqrt{T} \bar{g}_T(\theta^0) + \mu_\beta + o_p(1).
\end{aligned}$$

by using the relation $R^* \mu_{T,\theta} \xrightarrow{P} \mu_\beta$ (see below (25)). It therefore follows that the RHS on the last line does not depend on $\gamma_{S,T}$ at all as long as the latter is such that (25) holds. Note that γ_S^0, a constant for all T, is trivially such a choice of the sequence $\{\gamma_{S,T} : T \geq 1\}$. Thus $LM_T^{\text{infeas}}(\theta_0^{\text{infeas}}) = LM_T^{\text{infeas}}(\theta_T) + o_p(1) = LM_T(\theta_T) + o_p(1)$.

(b) From (a) it now follows that $LM_T(\theta_T) \xrightarrow{d} \chi^2_{d_R}$ with non-centrality parameter $\mu_\beta' \left(R^* (G^{*'} V^{-1} G^*)^{-1} R^{*'} \right)^{-1} \mu_\beta$. ■

Proof of Proposition 3.4: Define the sequence $\{\gamma_T^\dagger : T \geq 1\}$ such that

$$\gamma_T^\dagger := \arg \inf_{\gamma_0 \in CI_T(\gamma_S;\epsilon)} LM_T\left(A_S^{-1} (r_0', \gamma_0')' \right).$$

By condition (26) on $CI_T(\gamma_S;\epsilon)$, it then follows that $\gamma_T^\dagger$ is such that $\mu_{T,\theta}^\dagger = O_p(1)$ where $\mu_{T,\theta}^\dagger := \sqrt{T} D_{T,\rho_\theta}^{-1} \Pi_{\rho_\theta}^{-1} R_S^1 (\theta_T^\dagger - \theta^0)$ and $\theta_T^\dagger := R_S^1 r_0 + S_S^1 \gamma_T^\dagger$. Therefore, the final result follows by Lemma 3.3. ■

Proof of Lemma 3.5: The proof partly follows the intermediate steps in the proof of Theorem 3 of Andrews (2016b) after adapting it to our setup. Take any $\varpi > 0$ and note that N1 implies that $\inf_{\beta\in\mathcal{B},\gamma:\|\gamma-\gamma_S^0\|\geq\varpi}\|\rho(R_S^1\beta + S_S^1\gamma)\| > 0$ (see equation (2.1) in Antoine and Renault (2012)). Let the infimum occur at θ^* (not necessarily unique). Using N1, let λ_T^* be the largest in order of magnitude diagonal element of Λ_T whose corresponding element of $\rho(\theta^*)$ is not zero. Now note that

$$V^{-1/2}(\theta)\frac{\sqrt{T}}{\lambda_T^*}\bar{g}_T(\theta) = V^{-1/2}(\theta)\frac{\sqrt{T}}{\lambda_T^*}\left[\bar{g}_T(\theta) - E_T[\bar{g}_T(\theta)]\right] + V^{-1/2}(\theta)\frac{\sqrt{T}}{\lambda_T^*}E_T[\bar{g}_T(\theta)].$$

By N2, N3 and N8, the first term on the RHS is $o_p(1)$ uniformly in $\theta \in \Theta$. By (16), N2, N3, N8 and the definition of λ_T^*, for the second term on the RHS we have:

$$\inf_{\beta\in\mathcal{B},\gamma:\|\gamma-\gamma_S^0\|\geq\varpi}\left\|V^{-1/2}(\theta)\frac{\sqrt{T}}{\lambda_T^*}E_T[\bar{g}_T(R_S^1\beta + S_S^1\gamma)]\right\| = \inf_{\beta\in\mathcal{B},\gamma:\|\gamma-\gamma_S^0\|\geq\varpi}\left\|V^{-1/2}(\theta)\frac{1}{\lambda_T^*}\Lambda_T\rho(R_S^1\beta + S_S^1\gamma)\right\| \to c > 0$$

as $T \to \infty$ for some $c > 0$. Therefore, by using the uniform consistency of $\widehat{V}_T^{-1}(\theta)$ for $V^{-1}(\theta)$ from N8, and the definition of $Q_T(\theta)$ from (14), it follows that $\inf_{\beta\in\mathcal{B},\gamma:\|\gamma-\gamma_S^0\|\geq\varpi}(\lambda_T^*)^{-2} \times T \times Q_T(R_S^1\beta + S_S^1\gamma') \xrightarrow{P} c^*$ for some $c^* > 0$. Therefore, by definition (and since $\lambda_T^* \to \infty$ by N3) it follows that $\inf_{\gamma:\|\gamma-\gamma_S^0\|\geq\varpi} T \times Q_T(R_S^1 r_0 + S_S^1\gamma') \xrightarrow{P} \infty$. Since $\varpi > 0$ is arbitrary, by the definition in (13) where the critical value is a fixed, finite positive number for a given $\epsilon < 1$, it follows that

$$\sup_{\gamma_0\in CI_T^{SW}(\gamma_S;r_0,\epsilon)}\|\gamma_0 - \gamma_S^0\| = o_p(1).$$

This is an intermediate result since we actually need to show more. For that purpose let us proceed as follows. Since $r_0 - \beta^0 = o(1)$ by (24), we appeal to the above result and now focus on θ such that $\theta - \theta^0 = o_p(1)$. Define

$$\vartheta_T := \sup_{\gamma_0\in CI_T^{SW}(\gamma_S;r_0,\epsilon)}\|\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(R_S^1(r_0 - \beta^0) + S_S^1(\gamma_0 - \gamma_S^0))\|.$$

Since $CI_T^{SW}(\gamma_S; r_0, \epsilon)$ is closed by construction (see (13)), we have

$$\gamma_T := \arg\sup_{\gamma_0\in CI_T^{SW}(\gamma_S;r_0,\epsilon)}\|\sqrt{T}D_{T,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(R_S^1(r_0 - \beta^0) + S_S^1(\gamma_0 - \gamma_S^0))\| \in CI_T^{SW}(\gamma_S; r_0, \epsilon)$$

for all T. We know from the above result that $\|\gamma_T - \gamma_S^0\| = o_p(1)$ and hence for $\theta_T := R_S^1 r_o + S_S^1\gamma_T$, it follows that $\|\theta_T - \theta^0\| = o_p(1)$. Hence $\theta_T \in \mathcal{N}(\theta^0)$ (defined in N9) with probability approaching one.

Note that if $\vartheta_T = o_p(1)$, the required result for the lemma is already proved. So let us consider the case that $\vartheta_T \neq o_p(1)$. Now, we wish to prove that $\vartheta_T = O_p(1)$, and we do it by contradiction.

Suppose that $\vartheta_T \neq O_p(1)$. Then there exist a $\varepsilon > 0$ and a subsequence $\{T_n\}$ such that $Pr_{T_n}(\vartheta_{T_n} > n) \geq \varepsilon$ for all n. The required proof follows by contradiction if we show that this is not possible.

First, using (16) and N1, we obtain by a mean value expansion of $\rho(\theta_T)$ that

$$\sqrt{T}\bar{g}_T(\theta_T) = \sqrt{T}\left(\bar{g}_T(\theta_T) - E_T[\bar{g}_T(\theta_T)]\right) + \sqrt{T}\left[\frac{\Lambda_T}{\sqrt{T}}\left\{\rho(\theta^0) + \rho_\theta(\theta^0)(\theta_T - \theta^0) + [\rho_\theta(\bar{\theta}_T) - \rho_\theta(\theta^0)](\theta_T - \theta^0)\right\}\right]$$

where, as before, $\rho_\theta(.) := \frac{\partial}{\partial\theta'}\rho(.)$. $\psi_T(\theta) := \sqrt{T}\left(\bar{g}_T(\theta_T) - E_T[\bar{g}_T(\theta_T)]\right)$ (be its definition in N2) and $\rho(\theta^0) = 0$ by N1. Hence, focusing on the subsequence $\{T_n\}$, we obtain that

$$\left\|\sqrt{T_n}\bar{g}_{T_n}(\theta_{T_n}) - \left\{\psi_{T_n}(\theta_{T_n}) + \left(\frac{\Lambda_{T_n}}{\sqrt{T_n}}\rho_\theta(\theta^0)\Pi_{\rho_\theta}D_{T_n,\rho_\theta}\right)\sqrt{T_n}D_{T_n,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta_{T_n} - \theta^0)\right\}\right\| \leq \vartheta_{T_n} \times \sup_{\theta\in\mathcal{N}(\theta^0)}\kappa_{T_n}(\theta) = \vartheta_{T_n} \times o_p(1)$$

where $\psi_T(\theta)$ and $\kappa_T(\theta)$ are defined in N2 and N9 respectively. For notational simplicity define

$$b_{T_n}(\theta_{T_n}) := \psi_{T_n}(\theta_{T_n}) + \left(\frac{\Lambda_{T_n}}{\sqrt{T_n}}\rho_\theta(\theta^0)\Pi_{\rho_\theta}D_{T_n,\rho_\theta}\right)\sqrt{T_n}D_{T_n,\rho_\theta}^{-1}\Pi_{\rho_\theta}^{-1}(\theta_{T_n} - \theta^0),$$

which is $O_p(1) + O_p(\vartheta_{T_n})$ by virtue of N2, (20) (using the full column rank of G^*).

Define $\varpi_T := 1 + \vartheta_T^2 \geq 1$. Now, by the definition of $Q_T(\theta)$, it follows from using N8 that

$$\frac{1}{\varpi_{T_n}}\left|T_n \times Q_{T_n}(R_S^1 r_0 + S_S^1\gamma_{T_n}) - b_{T_n}'(\theta_{T_n})V^{-1}b_{T_n}(\theta_{T_n})\right| = o_p(1)$$

since the minimum eigen value of $V(\theta)$ is uniformly bounded away from zero. Further noting the full column rank of G^* and the fact that $b_{T_n}(\theta_{T_n}) = O_p(1) + O_p(\vartheta_{T_n})$ (please note the two additive components of $b_{T_n}(\theta_{T_n})$ from above) will imply that $\frac{1}{\varpi_{T_n}}b_{T_n}'(\theta_{T_n})V^{-1}b_{T_n}(\theta_{T_n}) \neq o_p(1)$ but $O_p(1)$. Therefore, $\frac{T_n}{\varpi_{T_n}} \times Q_{T_n}(R_S^1 r_0 + S_S^1\gamma_{T_n}) \neq o_p(1)$ but $O_p(1)$. On the other hand, since $\gamma_{T_n} \in CI_{T_n}^{SW}(\gamma_S; r_0, \epsilon)$, we also know that $T_n \times Q_{T_n}(R_S^1 r_0 + S_S^1\gamma_{T_n}) = O_p(1)$ by the definition of $CI_{T_n}^{SW}(\gamma_S; r_0, \epsilon)$ in (13). Therefore, there cannot exist a $\varepsilon > 0$ such that $Pr_{T_n}(\varpi_{T_n} > 1 + n^2) \geq \varepsilon$ for all n. Equivalently, since $\varpi_T := 1 + \vartheta_T^2 \geq 1$, there cannot exist a $\varepsilon > 0$ such that $Pr_{T_n}(\vartheta_{T_n}^2 > n^2) \geq \varepsilon$ for all n. Equivalently, there cannot exist a $\varepsilon > 0$ such that $Pr_{T_n}(\vartheta_{T_n} > n) \geq \varepsilon$ for all n, which is a contradiction to our supposition that $\vartheta_T \neq O_p(1)$. Hence, $\vartheta_T = O_p(1)$. ■