VALID TWO-STEP IDENTIFICATION-ROBUST CONFIDENCE SETS FOR GMM

Isaiah Andrews*

Abstract—In models with potentially weak identification, researchers often decide whether to report a robust confidence set based on an initial assessment of model identification. Two-step procedures of this sort can generate large coverage distortions for reported confidence sets, and existing procedures for controlling these distortions are quite limited. This paper introduces a generally applicable approach to detecting weak identification and constructing two-step confidence sets in GMM. This approach controls coverage distortions under weak identification and indicates strong identification, with probability tending to 1 when the model is well identified.

I. Introduction

In contexts where weak identification is a concern, empirical researchers in economics frequently calculate statistics intended to measure identification strength. If these statistics indicate that identification is not "too" weak, researchers proceed as usual and calculate nonrobust confidence sets; if weak identification is detected, researchers may calculate identification-robust confidence sets, look for a different specification, or simply decide not to report results. The latter two approaches can lead to enormous coverage distortions for reported confidence sets, so here I focus on the case where researchers use the first-step identification assessment to decide between reporting robust and nonrobust confidence sets.

We can view such procedures as two-step confidence sets, where the first step assesses identification strength and the second step reports a confidence set chosen based on this assessment. Such two-step procedures underlie many of the applications of identification-robust methods in empirical practice, where robust confidence sets are often computed only after researchers observe evidence suggestive of weak identification.[1] Unless carefully constructed, such procedures can undermine coverage guarantees for robust techniques and result in very poor performance for reported confidence sets.

Received for publication March 4, 2015. Revision accepted for publication March 13, 2017. Editor: Bryan S. Graham.

* MIT.

I am grateful to Anna Mikusheva, Whitney Newey, and Jerry Hausman for their guidance and support and to Josh Angrist, Bruce Hansen, Bryan Graham, and two anonymous referees for extremely helpful feedback. I also thank Victor Chernozhukov, Adam McCloskey, Jose Montiel Olea, Nils Wernerfelt, and the participants of the MIT and Harvard Econometrics Lunches, the Fall 2014 Interactions conference at the University of Chicago, and the Yale Econometrics Workshop for helpful comments. Support from the Silverman (1968) Family Career Development Chair at MIT and the NSF Graduate Research Fellowship (grant number 1122374) are gratefully acknowledged.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00682.

[1] Of the 38 empirical papers calculating robust tests or confidence sets based on Moreira (2003) who themselves have over twenty citations (based on a Google Scholar search on March 4, 2014), for example, 35 report first stage F-statistics, 33 have a first-stage F smaller than fifteen in some specification, and 29 have a first-stage F smaller than ten.

Using the results of Stock and Yogo (2005), one can show that in the linear IV model with homoskedastic errors, two-step confidence sets based on the first stage F-statistic ensure bounded coverage distortions. Stock and Yogo's results do not apply to linear IV models with heteroskedastic, clustered, or serially correlated data, however, much less to nonlinear models. Indeed, even in the linear IV model with heteroskedasticity, two-step confidence sets based on the first-stage F-statistic can exhibit enormous coverage distortions.[2]

To bridge this gap between empirical practice and the theoretical econometric literature, this paper introduces a widely applicable method for constructing two-step confidence sets with controlled coverage distortions. In cases where the model is well identified, and thus nonrobust confidence sets are reliable, the proposed method indicates this and reports nonrobust confidence sets with probability tending to 1. The idea behind this approach is simple: in well-identified models, many different test statistics are asymptotically equivalent local to the true parameter value. Using this equivalence, for any $\gamma > 0$, we can construct identification-robust confidence sets with coverage $1-\alpha-\gamma$, which are contained in the usual nonrobust level $1-\alpha$ confidence set with probability tending to 1 if the model is well identified. A natural way to gauge identification is thus to check if this containment occurs. I show that the resulting two-step confidence sets have coverage at least $1-\alpha-\gamma$. Moreover, I note that rather than picking a bound γ on coverage distortion ex ante, researchers can report robust and nonrobust confidence sets, along with a distortion cutoff $\hat{\gamma}$. Readers whose tolerance for distortion is less than $\hat{\gamma}$ should focus on the robust confidence set, while readers with a higher tolerance for distortion can focus on the nonrobust confidence set.

To implement my approach in (linear or nonlinear) generalized method of moments (GMM) models, I extend the results of Kleibergen (2005) and Chaudhuri and Zivot (2011) and derive identification-robust test statistics that are locally asymptotically equivalent to conventional test statistics in well-identified models for tests of both, the full GMM parameter vector and lower-dimensional parameters. I then construct identification-robust confidence sets by combining these statistics with the S statistic of Stock and Wright (2000) using the linear combination approach discussed by I. Andrews (2016). For lower-dimensional parameters, these confidence sets are based on the projection method, but the choice of test statistic limits the efficiency loss in well-identified models.

[2] For demonstration of this point in simulation, see section C of the supplementary appendix.

The Review of Economics and Statistics, Month Year, 00(0): 000–000

doi:10.1162/REST_a_00682

The next section introduces my approach for combining robust and nonrobust procedures to construct two-step confidence sets with bounded coverage distortions in general models. Section III discusses particular confidence sets, which can be used to implement this approach in GMM, while section IV details the steps needed for implementation and derives results for the nonlinear Euler equation model of Hansen and Singleton (1982). Proofs of all results stated in the paper are given in the appendix; the proof of an auxiliary lemma, details, and additional results for the empirical application and simulation results for the linear IV model are given in the supplementary appendix.

II. Valid Two-Step Confidence Sets

Throughout the paper, I suppose that we observe a sample of size T drawn from distribution $F_T(\beta_0, \psi_0)$, where $\beta \in B \subseteq \mathbb{R}^p$ is finite dimensional while $\psi \in \Psi$ is potentially infinite dimensional. I assume we are interested in constructing a confidence set for the parameter β, treating ψ as a nuisance parameter. The distribution F_T need not be explicitly specified, so this accommodates both parametric and semiparametric models, including moment condition models estimated using GMM. While the primary focus of this paper will be on GMM models, for this section, nothing is gained by limiting attention to GMM, so I do not impose this restriction.

As noted above, when researchers are concerned that conventional inference procedures may be unreliable due to weak identification, they often assess the identification status of the model based on some statistic or collection of statistics. I consider the case where a researcher wants to report an identification-robust confidence set if this initial step indicates weak identification, but will otherwise report a nonrobust confidence set. To formally describe the resulting confidence set, following D. Andrews and Cheng (2012), I represent the first-stage identification diagnostic using an identification category selection (ICS) statistic $\phi_{ICS} \in \{0, 1\}$, where $\phi_{ICS} = 0$ is interpreted as evidence of strong identification and $\phi_{ICS} = 1$ is interpreted as evidence of weak identification. The rule-of-thumb for the first-stage F-statistic in linear IV, for example, indicates weak identification when the first-stage F-statistic is smaller than 10 and so can be represented as $\phi_{ICS} = 1\{F < 10\}$. Denoting the robust and nonrobust confidence sets by CS_R and CS_N, respectively, the procedure described above yields the two-step confidence set CS_2:

$$CS_2 = \begin{cases} CS_N & \text{if } \phi_{ICS} = 0 \\ CS_R & \text{if } \phi_{ICS} = 1 \end{cases}. \tag{1}$$

I will be interested in the probability that this two-step confidence set covers the true parameter value: $Pr_{T,(\beta_0,\psi_0)}\{\beta_0 \in CS_2\}$.

A. Sequential and Asymptotic Coverage Probability

The finite-sample coverage probability $Pr_{T,(\beta_0,\psi_0)}\{\beta_0 \in CS_2\}$ is typically difficult to analyze directly. I thus follow the usual approach and instead consider the limiting coverage probability as the sample size grows. While the traditional justification of nonrobust tests (see, e.g., Newey & McFadden, 1994) considers point-wise asymptotic approximations where we fix (β_0, ψ_0) and take the sample size T to infinity, the weak identification literature following Staiger and Stock (1997) has shown that these approximations may be quite misleading in contexts with potential identification failure. To derive alternative approximations, this literature instead models parameters as drifting with the sample size, so the true parameters in the sample of size T are $(\beta_{0,T}, \psi_{0,T})$. More recently, the literature on robust inference has focused on asymptotic coverage, defined as the lower limit of the minimal finite-sample coverage probability. Formally, the asymptotic coverage probability of CS is

$$ACP(CS) = \liminf_{T\to\infty} \inf_{(\beta_0,\psi_0)\in B\times\Psi} Pr_{T,(\beta_0,\psi_0)}\{\beta_0 \in CS\}.$$

To discuss my results, it is helpful to have compact notation for discussing limiting coverage under particular sequences of parameter values. In particular, let

$$\xi_0 = \left\{(\beta_{0,T}, \psi_{0,T})\right\}_{T=1}^{\infty} \in \Xi = \Pi_{T=1}^{\infty}(B \times \Psi) \tag{2}$$

denote a sequence of true parameter values, with Ξ the space of all such sequences. Define the sequential coverage probability of confidence set CS under the sequence of true parameter values ξ_0 as the lower limit of the coverage probability under ξ_0:

$$\begin{aligned} SCP(CS, \xi_0) &= \liminf_{T\to\infty} Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS\right\} \\ &= \liminf_{T\to\infty} Pr_{T,(\beta_{0,T},\psi_{0,T})}\left\{\beta_{0,T} \in CS\right\}. \end{aligned}$$

Likewise, define the sequential coverage probability of confidence set CS under the set of sequences $\widetilde{\Xi} \subset \Xi$ as the minimal sequential coverage probability under $\xi_0 \in \widetilde{\Xi}$,

$$SCP\left(CS, \widetilde{\Xi}\right) = \inf_{\xi_0\in\widetilde{\Xi}} SCP(CS, \xi_0).$$

Note that sequential coverage probability under Ξ as defined in equation (2) is simply the asymptotic coverage probability

$$SCP(CS, \Xi) = ACP(CS).$$

We can use sequential coverage to formalize what we mean by "robust" and "nonrobust" confidence sets. In particular, I assume we consider two sets of parameter sequences, Ξ_S and Ξ_W, which I will refer to as "strong" and "potentially weak" (or for brevity, simply "weak"), respectively. I

assume that the nonrobust confidence set CS_N has sequential coverage at least $1-\alpha$ under strong identification

$$SCP\left(CS_N, \Xi_S\right) \geq 1-\alpha, \tag{3}$$

but I impose no restriction on the performance of this confidence set under weak identification. By contrast, I assume that the robust confidence set CS_R has coverage at least $1-\alpha$ under both weak and strong identification:

$$\begin{aligned} &SCP\left(CS_R, \Xi_S \cup \Xi_W\right) \\ &\quad = \min\left\{SCP\left(CS_R, \Xi_S\right), SCP\left(CS_R, \Xi_W\right)\right\} \geq 1-\alpha. \end{aligned} \tag{4}$$

Thus, the robust confidence set CS_R is more robust than CS_N in the sense that it has correct sequential coverage for a larger set of sequences.

Example: Linear IV. To illustrate the different notions of limiting coverage described, consider the linear IV model with a single endogenous regressor. The model, written in reduced form, is

$$\begin{aligned} Y &= Z\pi\beta + V_1, \\ X &= Z\pi + V_2, \end{aligned}$$

for Z a $T \times k$ matrix of instruments, X a $T \times 1$ vector of endogenous regressors, Y a $T \times 1$ vector of outcome variables, and V_1 and V_2 both $T \times 1$ vectors of residuals, where I assume that $E\left[V_{1,t} Z_t\right] = E\left[V_{2,t} Z_t\right] = 0$ for Z_t the transpose of row t of Z. For simplicity I assume that either there are no exogenous regressors or that any such regressors have already been partialed out.[3] The nuisance parameter ψ in this context will index both the first-stage parameter π and the joint distribution of (Z, V_1, V_2).

Conventional (strong-instrument, point-wise) asymptotic approximations correspond to fixing $\left(\beta_{0,T}, \psi_{0,T}\right) = (\beta_0, \psi_0)$ at some value with $\pi_0 \neq 0$ and taking T to infinity. Thus, if we define Ξ_S to be a set of such sequences, the usual Wald confidence sets have correct sequential coverage under Ξ_S. By contrast, the weak instrument asymptotics considered by Staiger and Stock (1997) set $\pi_{0,T} = \frac{1}{\sqrt{T}}\pi^*$, while uniform asymptotic results for confidence sets, like those of D. Andrews and Guggenberger (2017), allow arbitrary sequences of values $\left(\beta_{0,T}, \pi_{0,T}\right) \in B \times \Pi \subseteq \mathbb{R}^1 \times \mathbb{R}^k$. Both the results of Staiger and Stock (1997) and those of D. Andrews and Guggenberger (2017) also allow drifting sequences of distributions for (Z, V_1, V_2). If we define Ξ_W to be a set of sequences satisfying the assumptions of Staiger and Stock (1997), then all the identification-robust confidence sets discussed in the weak instruments literature have correct sequential coverage under Ξ_W. If we take Ξ_W to be the set of all sequences Ξ over a base parameter space $B \times \Psi$, then, as noted in D. Andrews and Guggenberger (2017), commonly used robust confidence sets will have correct sequential coverage (and thus correct uniform asymptotic coverage) under appropriate restrictions on Ψ.

[3] That is, for exogenous controls W and initial data $\left(\tilde{Y}, \tilde{X}, \tilde{Z}, W\right)$, $Y = M_W\tilde{Y}$, $X = M_W\tilde{X}$, $Z = M_W\tilde{Z}$, where $M_W = I - W\left(W'W\right)^{-1}W'$.

Defining strong and weak sequences. As the discussion suggests, even in the linear IV model, one may potentially define Ξ_W in a number of ways. Indeed, this reflects the state of the literature, where a number of devices have been used to model weak identification, including the drifting parameter asymptotics considered in Staiger and Stock (1997) and D. Andrews and Cheng (2012), and the drifting moment condition asymptotics considered in Stock and Wright (2000) and Chaudhuri and Zivot (2011). The goal of this paper is to show how, given a definition of weak identification and corresponding robust confidence sets, one may construct a two-step confidence set with bounded coverage distortions. While I will generally take Ξ_S to consist of conventional pointwise asymptotic sequences, with $\left(\beta_{0,T}, \psi_{0,T}\right) = (\beta_0, \psi_0)$ fixed, my construction does not depend on the definition of Ξ_W. Indeed, since results in the literature assume different definitions of Ξ_W, it is helpful to leave the definition of Ξ_W flexible in this section, though in the next section, I impose assumptions on Ξ_W to derive robust confidence sets for GMM models.

B. Coverage Bounds for Two-Step Confidence Sets

The coverage assumptions (3) and (4) for CS_N and CS_R imply an initial bound on the sequential coverage of CS_2:

Lemma 1. *Under equations (3) and (4),*

a. $SCP\left(CS_2, \Xi_W\right) \geq 1-\alpha-\sup_{\xi_0 \in \Xi_W} \limsup_{T\to\infty} \times Pr_{T,\xi_0}\left\{\phi_{ICS} = 0\right\}$.

b. $SCP\left(CS_2, \Xi_S\right) \geq 1-\alpha-\min\left\{\alpha, \sup_{\xi_0 \in \Xi_S} \limsup_{T\to\infty} \times Pr_{T,\xi_0}\left\{\phi_{ICS} = 1\right\}\right\}$.

These bounds are tight in the sense that one cannot obtain a sharper bound without additional conditions on the behavior of (CS_N, CS_R, ϕ_{ICS}). In particular, without further restrictions, the sequential coverage of CS_2 under Ξ_W may be arbitrarily close to 0.

To construct ϕ_{ICS} yielding such additional restrictions, I observe that a number of asymptotic simplifications arise in well-identified models. As I show for GMM models in the next section, we can often construct preliminary robust confidence sets $CS_P(\gamma)$ with coverage $1-\alpha-\gamma$, which are contained in the nonrobust confidence set CS_N with probability tending to 1 when the model is well identified. Formally, I assume:

Assumption 1. We have a preliminary confidence set $CS_P(\gamma)$ such that:

a. $SCP\left(CS_P(\gamma), \Xi_W\right) \geq 1-\alpha-\gamma$.

b. $Pr_{T,\xi_0}\{CS_P(\gamma) \subseteq CS_R\} = 1$ for all T and $\xi_0 \in \Xi$.
c. $\inf_{\xi_0 \in \Xi_S} \liminf_{T\to\infty} Pr_{T,\xi_0}\{CS_P(\gamma) \subseteq CS_N\} = 1$.

This assumption requires the existence of a preliminary confidence set that (a) has sequential coverage at least $1-\alpha-\gamma$ when identification is weak, (b) is contained in CS_R with probability 1, and (c) is contained in CS_N with probability tending to 1 under strong identification. While this might seem quite demanding, in the next section I construct confidence sets $CS_P(\gamma)$ that satisfy these conditions in GMM. Such a preliminary confidence set allows a natural pretest for identification strength; however, since if we see that $CS_P(\gamma)$ is not contained in CS_N, this suggests that the model may not be well identified.

In addition to being intuitively reasonable, this approach to assessing identification implies several good properties for the resulting two-step confidence sets. Formally, this approach corresponds to the ICS statistic:

$$\phi_{ICS}(\gamma) = 1\{CS_P(\gamma) \not\subseteq CS_N\}. \tag{5}$$

For this choice of ICS statistic, the two-step confidence set $CS_2 = CS_2(\gamma)$ as defined in equation (1) contains the preliminary confidence set $CS_P(\gamma)$ by construction, and thus has coverage at least $1-\alpha-\gamma$ under weak identification. Moreover, $CS_2(\gamma)$ coincides with CS_N with probability tending to 1 when the model is well identified. Formally:

Theorem 1. *Under assumption 1, together with equation (3), for ϕ_{ICS} as defined in equation (5), the two-step confidence set $CS_2(\gamma)$ has the following properties:*

a. $SCP(CS_2(\gamma), \Xi_W) \geq 1-\alpha-\gamma$.
b. $SCP(CS_2(\gamma), \Xi_S) \geq 1-\alpha$.
c. $\inf_{\xi_0 \in \Xi_S} \liminf_{T\to\infty} Pr_{T,\xi_0}\{CS_2(\gamma) = CS_N\} = 1$.

Further, $\sup_{\xi_0 \in \Xi_S} \limsup_{T\to\infty} Pr_{T,\xi_0}\{\phi_{ICS}(\gamma) = 1\} = 0$.

Thus, given a preliminary confidence set satisfying assumption 1, we can easily construct two-step confidence sets with coverage at least $1-\alpha-\gamma$. A natural question then is how we ought to choose γ. The next section shows that by reporting results appropriately the choice of γ can be left to the reader.

C. Reporting Results

The discussion has assumed a fixed maximal coverage distortion γ. In practice, however, different readers may be comfortable with different levels of distortion, so it may be preferable to report both robust and nonrobust confidence sets, together with some indication of the reliability of the nonrobust confidence set.

To this end, let us specify some minimal value of γ, $\gamma_{\min} \geq 0$. Suppose that for $\gamma \geq \gamma_{\min}$, we can define a family of preliminary robust confidence sets $CS_P(\gamma)$ that are decreasing in γ in the sense that

$$CS_P(\tilde{\gamma}) \subseteq CS_P(\gamma) \text{ for all } \tilde{\gamma} \geq \gamma.$$

Further, let us assume $CS_P(\gamma_{\min}) \subseteq CS_R$, so that the full family of preliminary confidence sets is contained in our robust confidence set. Define $\hat{\gamma}$ to be the smallest value such that $CS_P(\hat{\gamma}) \subseteq CS_N$,

$$\hat{\gamma} = \min\{\gamma \geq \gamma_{\min} : CS_P(\gamma) \subseteq CS_N\}.$$

$\hat{\gamma}$ is the smallest distortion γ such that $\phi_{ICS}(\gamma)$ will indicate strong identification in this realization of the data.[4] Hence I will refer to $\hat{\gamma}$ as the distortion cutoff. Note that by theorem 1, $\hat{\gamma} \to_p \gamma_{\min}$ under strong identification.[5]

Suppose that rather than reporting the two-step confidence set $CS_2(\gamma)$, we instead report $(CS_N, CS_R, \hat{\gamma})$. A reader who adopts the rule of focusing on CS_N when $\hat{\gamma} \leq \gamma$ and on CS_R when $\hat{\gamma} > \gamma$ is then effectively constructing the two-step confidence set,

$$CS_2(\gamma) = \begin{cases} CS_N & \text{if } \hat{\gamma} \leq \gamma \\ CS_R & \text{if } \hat{\gamma} > \gamma \end{cases} = \begin{cases} CS_N & \text{if } \phi_{ICS}(\gamma) = 0 \\ CS_R & \text{if } \phi_{ICS}(\gamma) = 1 \end{cases},$$

which is the same as $CS_2(\gamma)$ based on $\phi_{ICS}(\gamma)$ as in equation (5). Thus, it follows immediately from theorem 1 that this confidence will have asymptotic coverage at least $1-\alpha-\gamma$ under both weak and strong identification. Thus, by reporting $(CS_N, CS_R, \hat{\gamma})$ we provide the ingredients to construct a variety of two-step confidence sets and supply more information than reporting $CS_2(\gamma)$ alone.

III. Two-Step Confidence Sets for GMM

The two-step procedures described require three inputs: the nonrobust confidence set CS_N, the robust confidence set CS_R, and the family of preliminary confidence sets $CS_P(\gamma)$. To discuss concretely how to construct these confidence sets, I consider models identified by moment equalities and estimated by GMM and provide sufficient conditions to apply theorem 1.

I consider a GMM model with a k-dimensional continuously differentiable moment condition $g_t(\theta)$ that has mean 0 when the m-dimensional parameter θ is equal to its true value $\theta_{0,T}$. In the linear IV model already discussed, for example, $g_t(\theta) = Z_t(Y_t - X_t\theta)$. To reflect the fact that we are frequently interested in inference on a lower-dimensional function of model parameters, I suppose we

[4] If $CS_P(\gamma) \not\subseteq CS_N$ for all $\gamma \in [\gamma_{\min}, 1-\alpha]$, define $\hat{\gamma} = 1-\alpha$.

[5] Taking $\gamma_{\min} = 0$ allows the widest possible range of values for $\hat{\gamma}$. However, this choice may sometimes result in undesirable properties for CS_R, as in the GMM case discussed below, so it is helpful to allow $\gamma_{\min} > 0$.

are interested in inference on a p-dimensional parameter ($p \leq m$) $\beta = f(\theta)$ for f a continuously differentiable function such that $\frac{\partial}{\partial \theta'} f\left(\theta_{0,T}\right)$ has full rank for all T. For example, we may be interested in constructing a confidence set for the ith element of the structural parameter vector and so take $f(\theta) = \theta_i$. Note that we may also take $f(\theta) = \theta$, in which case $\theta = \beta$, and we are conducting inference on the full parameter vector.

Let $g_T(\theta) = \frac{1}{T}\sum_t g_t(\theta)$ be the sample average of $g_t(\theta)$, and let $\widehat{\Sigma}_g$, $\widehat{\Sigma}_{\theta g}$, and $\widehat{\Sigma}_\theta$ be consistent estimators for $\text{Var}\left(\sqrt{T} g_T(\theta)\right)$, $\text{Cov}\left(\sqrt{T} vec\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right), \sqrt{T} g_T(\theta)\right)$, and $\text{Var}\left(\sqrt{T} vec\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right)\right)$, respectively. I assume we have some estimator $\tilde{\theta}$ for θ which under strong identification is first-order equivalent to

$$\widehat{\theta} = \arg\min_{\theta} g_T(\theta)' \widehat{\Omega}(\theta) g_T(\theta) \tag{6}$$

for $\widehat{\Omega}(\theta)$ a symmetric positive-definite weighting matrix, which I assume converges uniformly in probability to a full-rank matrix-valued function $\Omega(\theta)$ under strong identification.[6] Estimators $\tilde{\theta}$ in this class include one-step GMM, efficiently and inefficiently weighted two-step GMM, continuously updating GMM, and many others.

The most common nonrobust confidence set for $\beta_0 = f(\theta_0)$ is based on the Wald statistic,

$$W(\beta) = T \cdot \left(f\left(\tilde{\theta}\right) - \beta\right)' \widehat{\Sigma}_{\tilde{\beta}}^{-1} \left(f\left(\tilde{\theta}\right) - \beta\right), \tag{7}$$

for $\widehat{\Sigma}_{\tilde{\beta}}$, an estimator for the asymptotic variance of $\sqrt{T}\tilde{\beta} = \sqrt{T} f\left(\tilde{\theta}\right)$. Under strong-instrument asymptotics, $\tilde{\beta}$ is $\sqrt{T}$ consistent for θ_0, and the Wald statistic diverges to infinity outside $\sqrt{T}$ neighborhoods of the true parameter value.

Unfortunately, when identification is weak, the distribution of the Wald statistic $W(\beta)$ depends on nuisance parameters, making construction of identification-robust confidence sets based on this statistic challenging in most models. To avoid these issues while constructing CS_R and $CS_P(\gamma)$ satisfying our requirements, I proceed in two steps. First, I seek analytically simple test statistics that are locally asymptotically equivalent to $W(\beta)$ in the well-identified case. Second, I exploit the simple form of these statistics to create identification-robust analogs that remain locally equivalent to $W(\beta)$ when the model is well identified.

To obtain analytically simpler analogs of the Wald statistic $W(\beta)$, note that section 9 of Newey and McFadden (1994) establishes that conventional GMM estimators are asymptotically equivalent to one-step estimators with starting values in a $\sqrt{T}$ neighborhood of the true parameter value. Formally, define the one-step estimator with initial value θ as

$$\bar{\theta}(\theta) = \theta - \left(\frac{\partial}{\partial \theta} g_T(\theta)' \widehat{\Omega}(\theta) \frac{\partial}{\partial \theta} g_T(\theta)\right)^{-1} \times \frac{\partial}{\partial \theta} g_T(\theta)' \widehat{\Omega}(\theta) g_T(\theta).$$

$\bar{\theta}(\theta)$ is first-order asymptotically equivalent to $\tilde{\theta}$ under strong identification provided the initial value θ lies in a $\sqrt{T}$-neighborhood of θ_0.[7] Analogously, we can interpret $\bar{\beta}(\theta) = f(\theta) + \frac{\partial}{\partial \theta} f(\theta)\left(\bar{\theta}(\theta) - \theta\right)$ as a one-step estimator for β, where we have linearized the function f around θ. Thus, in well-identified models, we can construct Wald statistics based on $\bar{\beta}(\theta)$, and they will be first-order asymptotically equivalent to $W(f(\theta))$ local to the true value of θ. Consequently, if we can find identification-robust versions of these one-step Wald statistics, then we can use these to construct CS_R and $CS_P(\gamma)$.

A. *Robust Confidence Sets*

Unfortunately, when identification is weak, even Wald statistics based on $\bar{\beta}(\theta)$ behave irregularly. In particular, as Kleibergen (2005) noted, under weak identification, the Jacobian $\frac{\partial}{\partial \theta} g_T(\theta)$ is asymptotically random and correlated with the moment condition $g_T(\theta)$, with the result that the distribution of $\bar{\beta}\left(\theta_{0,T}\right)$ is nonstandard and depends on unknown parameters.

Happily, the relatively simple structure of $\bar{\beta}\left(\theta_{0,T}\right)$ allows adaptation of the approach of Kleibergen (2005) to address these issues. To eliminate asymptotic dependence between the moment conditions and their Jacobian $\frac{\partial}{\partial \theta} g_T(\theta_0)$, Kleibergen (2005) orthogonalizes the Jacobian with respect to the moment conditions. Define

$$D_T(\theta) = \Bigg[\frac{\partial}{\partial \theta_1} g_T(\theta) - \widehat{\Sigma}_{\theta_1 g}(\theta)\widehat{\Sigma}_g(\theta)^{-1} g_T(\theta), \ldots, \frac{\partial}{\partial \theta_m} g_T(\theta) - \widehat{\Sigma}_{\theta_m g}(\theta)\widehat{\Sigma}_g(\theta)^{-1} g_T(\theta)\Bigg],$$

where $\widehat{\Sigma}_{\theta_i g}(\theta)$ is the $k \times k$ block of $\widehat{\Sigma}_{\theta g}(\theta)$ corresponding to θ_i. One can show that $D_T\left(\theta_{0,T}\right)$ will be asymptotically uncorrelated with $g_T\left(\theta_{0,T}\right)$ even when identification is weak, while $D_T\left(\theta_{0,T}\right)$ is asymptotically equivalent to $\frac{\partial}{\partial \theta} g_T\left(\theta_{0,T}\right)$ when identification is strong. If we then define

$$\theta^*(\theta) = \theta - \left(D_T(\theta)' \widehat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)' \widehat{\Omega}(\theta) g_T(\theta)$$

and $\beta^*(\theta) = f(\theta) + \frac{\partial}{\partial \theta} f(\theta)\left(\theta^*(\theta) - \theta\right)$ to be the analogs of $\bar{\theta}$ and $\bar{\beta}$, which replace $\frac{\partial}{\partial \theta} g_T(\theta)$ by $D_T(\theta)$, then this substitution makes no difference (asymptotically) in the well-identified case, while the Wald statistic based on $\beta^*\left(\theta_{0,T}\right)$ will be robust to weak identification. For

$$M(\theta) = \widehat{\Omega}(\theta) D_T(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta)',$$

[6] By "first-order asymptotic equivalence," I mean that $\sqrt{T}\left(\widehat{\theta} - \tilde{\theta}\right) \to_p 0$ under $\xi_0 \in \Xi_S$.

[7] This is shown formally in the proof of lemma 2 in the supplementary appendix.

this test statistic (which following Kleibergen, 2005, I label a K statistic) is

$$K_{\Omega,f}(\theta) = T \cdot \left(\beta^*(\theta) - f(\theta)\right)' \left(M(\theta)' \widehat{\Sigma}_g(\theta) M(\theta)\right)^{-1} \times \left(\beta^*(\theta) - f(\theta)\right).$$

To derive the limiting distribution of $K_{\Omega,f}\left(\theta_{0,T}\right)$, I make the following assumptions:

Assumption 2. For all $\xi_0 \in \Xi_W \cup \Xi_S$, under ξ_0 we have that for $J_{T,\xi}(\theta) = E_{T,(\beta_T,\psi_T)}\left[\frac{\partial}{\partial\theta'} g_T(\theta)\right]$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \begin{pmatrix} g_t(\theta_{0,T}) \\ vec\left(\frac{\partial}{\partial\theta'} g_t(\theta_{0,T}) - J_{T,\xi_0}\left(\theta_{0,T}\right)\right) \end{pmatrix} \to_d \begin{pmatrix} \psi_g \\ \psi_\theta \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma_g & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_\theta \end{pmatrix}\right)$$

where Σ_g is positive definite and

$$\begin{pmatrix} \Sigma_g & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_\theta \end{pmatrix} = \lim_{T\to\infty} \mathrm{Var}_{T,\xi_0}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \begin{pmatrix} g_t(\theta_{0,T}) \\ vec\left(\frac{\partial}{\partial\theta'} g_t(\theta_{0,T})\right) \end{pmatrix}\right).$$

Assumption 3. We have estimators $\widehat{\Sigma}_g\left(\theta_{0,T}\right), \widehat{\Sigma}_{\theta g}\left(\theta_{0,T}\right)$, and $\widehat{\Sigma}_\theta\left(\theta_{0,T}\right)$, which converge in probability to fixed Σ_g, $\Sigma_{g\theta}$, and Σ_θ under all $\xi_0 \in \Xi_W \cup \Xi_S$. Further, $\widehat{\Omega}\left(\theta_{0,T}\right) \to_p \Omega$ for a nonstochastic symmetric positive-definite limit Ω.

Assumption 4. For all $\xi_0 \in \Xi_W \cup \Xi_S$, there exist sequences of full-rank normalizing matrices $\Lambda_{1,T}$ and $\Lambda_{2,T}$ of dimension $m \times m$ and $p \times p$, respectively, such that

a. $D_T(\theta)\Lambda_{1,T} \to_d D$ for a Gaussian random matrix D that is full rank almost surely but whose variance may be degenerate
b. $\Lambda_{2,T}\frac{\partial}{\partial\theta'} f\left(\theta_{0,T}\right)\Lambda_{1,T} \to F$ for a full-rank matrix F

Further, the elements of $\Lambda_{1,T}$ are of order $O\left(\sqrt{T}\right)$.[8]

Assumption 2 requires that the moment function and its Jacobian be jointly asymptotically normal. Assumption 3 requires that (a) we have consistent estimators for the various terms appearing in the asymptotic variance of $\left(g_T\left(\theta_{0,T}\right), \frac{\partial}{\partial\theta} g_T\left(\theta_{0,T}\right)\right)$ and (b) the weighting matrix $\widehat{\Omega}\left(\theta_{0,T}\right)$ be consistent for some well-behaved limit. Assumption 4 is more opaque but can easily be verified in many leading cases. For example, Kleibergen (2005) considers the case where $\sqrt{T}J_{T,\xi_0}$ converges to a finite matrix J, in which case we can take $\Lambda_{1,T} = \sqrt{T}I_m$ and $\Lambda_{2,T} = \frac{1}{\sqrt{T}}I_p$. More broadly, this assumption holds under the commonly used weakly identified GMM embedding of Stock and Wright (2000). In essence, this assumption requires the existence of a pair of normalizations for D_T and $\frac{\partial}{\partial\theta'} f\left(\theta_{0,T}\right)$ such that both of these terms converge to well-behaved limits.[9]

Given these assumptions, both $K_{\Omega,f}\left(\theta_{0,T}\right)$ and the difference $S\left(\theta_{0,T}\right) - K_{\Omega,f}\left(\theta_{0,T}\right)$ between $K_{\Omega,f}\left(\theta_{0,T}\right)$ and the S statistic of Stock and Wright (2000),

$$S(\theta) = T \cdot g_T(\theta)' \widehat{\Sigma}_g(\theta)^{-1} g_T(\theta), \tag{8}$$

have a well-behaved limiting distribution even under weak identification:

Theorem 2. *Under assumptions 2, 3, and 4, under all* $\xi_0 \in \Xi_W$*,*

$$\left(K_{\Omega,f}\left(\theta_{0,T}\right), S\left(\theta_{0,T}\right) - K_{\Omega,f}\left(\theta_{0,T}\right)\right) \to_d \left(\chi^2_p, \chi^2_{k-p}\right)$$

and $K_{\Omega,f}\left(\theta_{0,T}\right)$ *and* $S\left(\theta_{0,T}\right) - K_{\Omega,f}\left(\theta_{0,T}\right)$ *are asymptotically independent.*

If we take $\widehat{\Omega}(\theta)$ to be the efficient GMM weighting matrix, $K_{\Omega,f}(\theta)$ simplifies to the K statistic of Kleibergen (2005) when we test the full parameter vector, while for $f(\theta)$, which selects a subvector of θ (e.g., the first parameter alone), $K_{\Omega,f}$ is numerically equal to the LM_{eff} statistic proposed by Chaudhuri and Zivot (2011). Thus, this result is a natural generalization of the results of those papers to allow nonlinear functions f of the parameters and inefficient weighting matrices ($\widehat{\Omega}(\theta) \neq \widehat{\Sigma}_g(\theta)^{-1}$).

For the case where we consider hypotheses on the full parameter vector $f(\theta) = \theta$ and use the efficient weighting matrix, the results of D. Andrews and Guggenberger (2017) establish a parameter space on which the conclusion of theorem 2 holds uniformly. It seems likely that an analogous result might be available for the more general case considered here under suitable conditions. Given such uniformity results, one could define Ξ_W to be the set Ξ of all sequences on the appropriate base parameter space and the remainder of the analysis would proceed unchanged. Since my focus is on translating valid robust confidence sets to valid two-step confidence sets rather than on establishing uniform asymptotic validity for robust confidence sets, however, I do not pursue such an extension here.

[8] That is, they are bounded above in absolute value by $C\sqrt{T}$ for some constant C.

[9] These assumptions are stronger than necessary. In particular, using subsequencing arguments as in D. Andrews, Cheng, and Guggenberger (2011) one can relax all of these assumptions to require only that for any subsequence $\left(\beta_{0,T(m)}, \psi_{0,T(m)}\right)$ of a sequence $\xi_0 \in \Xi_W \cup \Xi_S$, there exists a further sub-sequence along which the stated conditions hold.

B. Localizing the Confidence Set

Given the results of theorem 2, we can construct robust confidence sets for $\beta = f(\theta)$. In particular, define

$$\begin{aligned} CS_{K,\theta} &= \left\{\theta : K_{\Omega,f}(\theta) \leq \chi^2_{p,1-\alpha}\right\} \\ CS_K &= \left\{f(\theta) : \theta \in CS_{K,\theta}\right\} \\ &= \left\{\beta : \min_{\theta:\beta=f(\theta)} K_{\Omega,f}(\theta) \leq \chi^2_{p,1-\alpha}\right\}. \end{aligned}$$

$CS_{K,\theta}$ collects the set of values θ where $K_{\Omega,f}(\theta)$ falls below a χ^2_p critical value, and so will cover $\theta_{0,T}$ with probability tending to α by theorem 2. CS_K then takes the image of the initial confidence set $CS_{K,\theta}$ under $f(\cdot)$ to construct a confidence set for $f(\theta)$. This is known as the projection method and ensures correct (albeit potentially conservative) coverage for $f\left(\theta_{0,T}\right)$.

It may seem reasonable to consider CS_K as the basis for CS_R and $CS_P(\gamma)$. In particular, as noted (and established formally in the supplementary appendix), $K_{\Omega,f}(\theta)$ is asymptotically equivalent to $W(f(\theta))$ local to $\theta_{0,T}$ in the well-identified case, and we can construct the nonrobust Wald confidence set in a manner analogous to CS_K:

$$\begin{aligned} CS_{N,\theta} &= \left\{\theta : W(f(\theta)) \leq \chi^2_{p,1-\alpha}\right\}, \\ CS_N &= \left\{f(\theta) : \theta \in CS_{N,\theta}\right\} = \left\{\beta : W(\beta) \leq \chi^2_{p,1-\alpha}\right\}. \end{aligned} \quad (9)$$

Unfortunately, however, the confidence set CS_K is not in general asymptotically equivalent to the confidence set CS_N-based $W(\beta)$, either globally or locally. For global equivalence, Kleibergen (2005) showed that for $\beta = \theta$ and $\widehat{\Omega}(\theta)$, the efficient weighting matrix, $K_{\Omega,f}(\theta)$, can be interpreted as a score statistic based on the continuously updating GMM objective function. In overidentified models, this statistic is thus equal to 0 at any critical point of the continuously updating GMM objective. Similar issues arise more broadly, and even in well-identified models, confidence sets based on $K_{\Omega,f}(\theta)$ are not necessarily consistent for $\theta_{0,T}$. Thus, since Wald confidence sets are consistent when β is well identified, we see that CS_N and CS_K are not globally equivalent. When $\beta = \theta$, one can show that CS_N and CS_K are equivalent on $\sqrt{T}$ neighborhoods of the true parameter value, but when $\beta = f(\theta)$ is of lower dimension, even this local equivalence fails, because while the test statistics $W(f(\theta))$ and $K_{\Omega,f}(\theta)$ are asymptotically equivalent local to $\theta_{0,T}$, the minimization in the definition of CS_K means that this does not suffice to imply local equivalence of CS_N and CS_K.

To construct robust confidence sets satisfying the requirements of assumption 1, it is thus insufficient to use the statistic $K_{\Omega,f}(\theta)$ alone. Instead, I combine this statistic with the S statistic as defined in equation (8). The S statistic diverges to infinity outside $\sqrt{T}$-neighborhoods of $\theta_{0,T}$ in well-identified models, so considering this statistic limits attention to regions of the parameter space on which $K_{\Omega,f}(\theta)$ is asymptotically equivalent to $W(f(\theta))$. In the case where $\beta = \theta$ and we use the efficient weighting matrix, I. Andrews (2016) establishes a number of desirable properties for tests based on linear combinations,

$$K_{\Omega,f}(\theta) + a \cdot S(\theta), \quad (10)$$

so here I consider test statistics of this form.[10]

Let $H(x; a, k, p)$ be the cumulative distribution function for a $(1+a) \times \chi^2_p + a \times \chi^2_{k-p}$ distribution and $H^{-1}(1-\alpha; a, k, p)$ the $1-\alpha$ quantile of this distribution.[11] For a given value of γ, let $a(\gamma)$ solve

$$H^{-1}(1-\alpha-\gamma; a(\gamma), k, p) = \chi^2_{p,1-\alpha}$$

for $\chi^2_{p,1-\alpha}$ the $1-\alpha$ quantile of a χ^2_p distribution. Define the preliminary robust confidence set through

$$\begin{aligned} CS_{P,\theta}(\gamma) &= \left\{\theta : K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) < \chi^2_{p,1-\alpha}\right\} \\ CS_P(\gamma) &= \left\{f(\theta) : \theta \in CS_{P,\theta}(\gamma)\right\} \\ &= \left\{\beta : \min_{\theta:\beta=f(\theta)} \left(K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta)\right) \right. \\ &\qquad \left. < \chi^2_{p,1-\alpha}\right\}. \end{aligned} \quad (11)$$

Analogously, define the robust confidence set

$$\begin{aligned} CS_{R,\theta} &= \left\{\theta : K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) \right. \\ &\qquad \left. \leq H^{-1}(1-\alpha; a(\gamma), k, p)\right\}, \\ CS_R &= \left\{f(\theta) : \theta \in CS_{R,\theta}\right\} \\ &= \left\{\beta : \min_{\theta:\beta=f(\theta)} \left(K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta)\right) \right. \\ &\qquad \left. \leq H^{-1}(1-\alpha; a(\gamma), k, p)\right\}. \end{aligned} \quad (12)$$

Theorem 2 implies that $CS_P(\gamma)$ has sequential coverage probability at least $1-\alpha-\gamma$ under both Ξ_W and Ξ_S, while CS_R has sequential coverage at least $1-\alpha$, as desired.

Corollary 1. *Under the conditions of theorem 2,*

$$\begin{aligned} SCP(CS_P(\gamma), \Xi_W \cup \Xi_S) &\geq 1-\alpha-\gamma, \\ SCP(CS_R, \Xi_W \cup \Xi_S) &\geq 1-\alpha. \end{aligned}$$

Thus, since $CS_P(\gamma) \subseteq CS_R$ by construction, these choices satisfy assumptions 1.1 and 2. Hence, to apply theorem 1, all that remains is to give sufficient conditions for assumption 1.3.

[10] Note that here I consider linear combination statistics of the form $K(\beta) + a \cdot S(\beta)$ while I. Andrews (2016) considers statistics of the form $(1-\tilde{a}) \cdot K(\beta) + \tilde{a} \cdot S(\beta)$. For $a = \tilde{a}/(1-\tilde{a})$, the level $1-\alpha$ confidence sets based on these two definitions are equivalent. I use the formulation in this paper rather than that in I. Andrews (2016) to simplify the expression for $CS_P(\gamma)$ below.

[11] Note that by $(1+a) \times \chi^2_p + a \times \chi^2_{k-p}$, I mean the distribution for the linear combination of χ^2 variables.

The role of $a \times S(\theta)$**.** The term $a \times S(\theta)$ in equation (10) serves two conceptually distinct purposes. First, it overcomes the issue discussed at the start of this section and ensures that confidence sets based on the linear combination statistic (10), including both $CS_P(\gamma)$ and CS_R, will be consistent in the strongly identified case. Second, since $CS_P(\gamma)$ is formed by comparing these linear combination statistics to a χ^2 critical value, the value a is also tied to the coverage distortion γ of this preliminary confidence set. There are alternative ways to construct $CS_P(\gamma)$ that avoid this one-to-one link between a and γ, but since in constructing $CS_P(\gamma)$, we want both $a > 0$ and $\gamma > 0$, setting $a = a(\gamma)$ avoids introducing additional free parameters and so is a natural choice.

C. *Asymptotic Results under Strong Identification*

We next establish conditions under which the confidence sets CS_R and $CS_P(\gamma)$, along with the Wald confidence set (9), satisfy assumption 1. To this end, I impose standard conditions for the consistency of $\hat{\theta}$:

Assumption 5. For all $\xi_0 \in \Xi_S$ the following conditions hold:

a. $g_T(\theta) \to_p \lim_{T\to\infty} E_{T,\xi_0}\left[g_T(\theta)\right]$ uniformly over the compact parameter space Θ for θ, and $\lim_{T\to\infty} \left\|E_{T,\xi_0}\left[g_T(\theta)\right]\right\|$ is uniformly bounded.
b. $E_{T,\xi_0}\left[g_T(\theta_0)\right] = 0\ \forall T$.
c. $\hat{\Omega}(\theta) \to_p \Omega(\theta)$ uniformly over Θ for $\Omega(\theta)$ continuous and everywhere positive definite with a uniformly bounded maximal eigenvalue and minimal eigenvalue bounded away from 0.
d. For all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left(\lim_{T\to\infty} E_{T,\xi_0}\left[g_T(\theta)\right]\right)' \Omega(\theta) \times \left(\lim_{T\to\infty} E_{T,\xi_0}\left[g_T(\theta)\right]\right) < \delta$$

only if $\|\theta - \theta_0\| < \varepsilon$.

Assumption 5a requires that the sample average of the moment condition $g_T(\theta)$ be uniformly close to its mean in large samples, while assumption 5c requires that the weighting matrix be well behaved. Assumptions 5b and 5d are identification conditions, which ensure that the population objective function is small if and only if evaluated in a neighborhood of the true parameter value, and assumption 5d will fail in contexts where weak or partial identification issues arise. Provided these conditions hold, standard arguments yield the consistency of $\hat{\theta}$. Next, I consider an assumption yielding asymptotic normality of $\hat{\theta}$.

Assumption 6. The following conditions hold for all $\xi_0 \in \Xi_S$:

a. θ_0 belongs to the interior of Θ.
b. $g_T(\theta)$ and $\hat{\Omega}(\theta)$ are almost surely continuously differentiable on some open ball $B(\theta_0)$ around θ_0.
c. For

$$J(\theta) = \lim_{T\to\infty} J_{T,\xi_0}(\theta) = \lim_{T\to\infty} E_{T,\xi_0}\left[\frac{\partial}{\partial\theta'} g_T(\theta)\right],$$

$J(\theta)$ is continuous at θ_0, $G_T(\theta) = \frac{\partial}{\partial\theta'} g_T(\theta) \to_p J(\theta)$ uniformly on $B(\theta_0)$, and $J(\theta_0)$ is full rank.
d. $\sup_{\theta\in B(\theta_0)} \left\| \frac{\partial vec\left(\hat{\Omega}(\theta)\right)}{\partial\theta'} \right\| = O_p(1)$.
e. $\hat{\Sigma}_g(\theta) \to_p \Sigma_g(\theta)$ uniformly on $B(\theta_0)$, and $\Sigma_g(\theta) = \lim_{T\to\infty} Var_{T,\xi_0}(\sqrt{T} g_T(\theta))$ is continuous in θ and everywhere positive-definite on $B(\theta_0)$.

Assumption 6a rules out cases where the true parameter value lies near the boundary of the parameter space. Assumption 6b requires that the moment condition and weight function both be smooth, while assumptions 6.3 and 6.4 require that their derivatives be well behaved. Finally Assumption 6e requires that we have a uniformly consistent estimator for $\Sigma_g(\theta)$ on a neighborhood of θ_0.

Assumptions 2, 5, and 6 together establish assumption 1.3. Stated formally:

Theorem 3. *Under assumptions 2, 5, and 6, for* $CS_P(\gamma)$ *as defined in equation (11),* CS_N *as in equation (9), and* $\gamma > 0$*,*

$$\inf_{\xi\in\Xi_S} \liminf_{T\to\infty} Pr_{T,\xi}\left\{CS_P(\gamma) \subseteq CS_N\right\} = 1.$$

Thus, we see that for the proposed $(CS_R, CS_P(\gamma), CS_N)$, assumptions 2 to 6 provide sufficient conditions for assumption 1 and allow us to apply theorem 1. Thus, we can construct two-step confidence sets with bounded sequential coverage distortions in potentially nonlinear GMM models. Further, as in section IIC, rather than picking a value γ, we can instead report $\left(CS_R, CS_{N,\hat{\gamma}}\right)$ for CS_R based on $K_{\Omega f}(\theta) + a(\gamma_{\min}) \times S(\theta)$.

Linear IV simulations. As a complement to these theoretical results, section C of the supplementary appendix simulates the performance of CS_R, $CS_P(\gamma)$, and $CS_2(\gamma)$ in the linear IV model with a single endogenous regressor. These simulations confirm the good coverage properties of these confidence sets in models with both weak and strong identification. Moreover, in linear IV models with homoskedastic errors where one can use the results of Stock and Yogo (2005) to construct two-step confidence sets based on the first-stage F-statistic, the approach developed here is found to be competitive and indicates weak identification substantially less often in some contexts.

IV. Empirical Illustration and User's Guide

To illustrate the application of the two-step confidence sets, I revisit the nonlinear Euler equation model of Hansen

and Singleton (1982). As noted by Hansen, Heaton, and Yaron (1996) and Stock and Wright (2000), there is evidence of weak identification in this context, so it is a natural setting in which to examine the performance of the procedures proposed here. I also detail the steps needed to calculate CS_R, CS_N, and $\hat{\gamma}$ in practice.

The parameters in this model are $\theta = (\delta, \eta)$, which represent the discount factor and the coefficient of relative risk aversion, respectively. The moments are

$$g_t(\theta) = \left(\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\eta} R_t - 1\right) Z_t$$

for C_t aggregate consumption in period t, R_t an aggregate stock return from $t-1$ to t, and Z_t a vector of instruments. Following Stock and Wright (2000), I use an extension of the long annual data set of Campbell and Shiller (1987) and take the vector Z_t to contain a constant, C_{t-1}/C_{t-2} and R_{t-1}. (For further discussion of the data, see section B of the supplementary appendix.) As Stock and Wright noted, results in this context are quite sensitive to the details of the specification, and there is evidence of model misspecification. Here, I follow the CRRA-1 specification of Stock and Wright (2000) except for covariance matrix estimation, where, unlike Stock and Wright (2000), I use the Newey and West (1987) covariance estimator with four lags to allow for serial dependence in $\frac{\partial}{\partial\theta'} g_t(\theta)$.[12]

I compute CS_R, CS_N, and $\hat{\gamma}$ for both the full parameter vector θ and for each parameter separately, corresponding to three different choices of $f(\theta)$: $f(\theta) = \theta$, $f(\theta) = \delta$, and $f(\theta) = \eta$. In all cases, I set $\alpha = 5\%$ and $\gamma_{\min} = 5\%$, so robust confidence sets have coverage at least 95%. The next section walks through the steps required to implement my suggested approach for a given $f(\theta)$ in detail, while the following section presents results.[13]

A. *Calculating CS_R, CS_N, and $\hat{\gamma}$*

This section details the steps needed to implement the approach developed above for a given choice of f and discusses my particular implementation choices in this application. Note that when one considers multiple choices of $f(\theta)$, as I do in the nonlinear Euler equation application, one can economize on computation by running steps 1 to 3 below for all choices of $f(\theta)$ at the same time. For expositional simplicity, however, I assume a fixed choice of $f(\theta)$ in this discussion.

Step 1: Choose weighting matrix and estimator. To implement this approach, we first need to choose a weighting matrix $\Omega(\theta)$ to use in estimation, since this choice affects both the robust and nonrobust confidence sets. In this application, I use the continuously updating GMM estimator of Hansen et al. (1996), which is given by equation (6), with $\widehat{\Omega}(\theta) = \widehat{\Sigma}(\theta)^{-1}$ the efficient weighting matrix. I then define the Wald statistic, equation (7), where $\widehat{\Sigma}_{\hat{\beta}}$ is the usual GMM variance estimator for $f\left(\hat{\theta}\right)$:

$$\widehat{\Sigma}_{\hat{\beta}} = \left(\frac{\partial}{\partial\theta} f\left(\hat{\theta}\right) \frac{\partial}{\partial\theta'} g_T(\theta)' \widehat{\Sigma}\left(\hat{\theta}\right)^{-1} \frac{\partial}{\partial\theta'} g_T(\theta) \frac{\partial}{\partial\theta} f\left(\hat{\theta}\right)'\right)^{-1}.$$

Step 2: Choose grid of parameter values. To calculate robust confidence sets, we need to collect the set of all parameter values where the identification-robust test statistics fall below given thresholds. To facilitate these computations, as is common in the identification-robust inference literature, we can take a discrete approximation Θ_D to the parameter space. In this application, I consider

$$\theta = (\delta, \eta) \in \Theta_D = \{0.6, 0.6025, ..., 1.1\} \times \{-6 : -5.975, ..., 60\}.$$

Let us label the elements of Θ_D as $\left\{\theta_1, \theta_2, ..., \theta_{|\Theta_D|}\right\}$.[14]

Step 3: Calculate test statistics. Given a discrete approximation to the parameter space, we next need to calculate our test statistics at each point in Θ_D. For each $\theta_i \in \Theta_D$, we can first calculate $g_T(\theta_i)$, $\widehat{\Sigma}(\theta_i)$, and $D_T(\theta_i)$. This suffices to let us calculate $S(\theta_i)$, as well as $K_{\Omega,f}(\theta_i)$, while we can calculate the Wald statistic $W(f(\theta_i))$ based on equation (7). Let us store the values $\left\{S(\theta_i), K_{\Omega,f}(\theta_i), W(\theta_i) : \theta_i \in \Theta_D\right\}$.

Step 4: Calculate $a(\gamma_{\min})$. Next, we need to determine the value $a(\gamma_{\min})$ to use in the construction of the robust confidence set CS_R. By definition, $a(\gamma_{\min})$ solves

$$Pr\left\{(1 + a(\gamma_{\min})) \times \chi^2_p + a(\gamma_{\min}) \cdot \chi^2_{k-p} \leq \chi^2_{p,1-\alpha}\right\} = 1 - \alpha - \gamma_{\min}.$$

To find this value in practice, we can take independent draws from χ^2_p and χ^2_{k-p} distributions and solve numerically for the value a, which sets the $1-\alpha-\gamma$ quantile of the corresponding linear combination of these draws to $\chi^2_{p,1-\alpha}$.[15]

Step 5: Calculate CS_R, CS_N. Now that we have $a(\gamma_{\min})$, we are ready to calculate the confidence sets CS_R and CS_N. In particular, we can first calculate the critical value used to construct CS_R, $H^{-1}(1-\alpha; a(\gamma_{\min}), k, p)$, by taking the

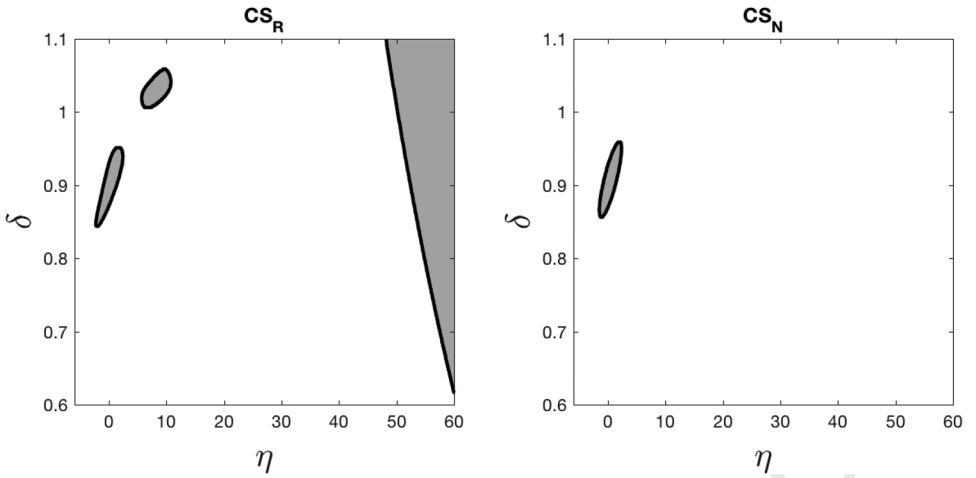
[12] See Kleibergen (2005) on the importance of allowing for serial correlation in this setting.

[13] Matlab code for performing these calculations with user-specified moment functions and weighting matrices, as well as for replicating the results, is available on my website: http://economics.mit.edu/faculty/iandrews

[14] Rather than considering grids in the parameter space, which will become computationally daunting when the dimension of the parameter is moderate or high, one could instead use Markov chain Monte Carlo methods based on the identification-robust test statistics, as suggested by Chernozhukov, Hansen, and Jansson (2009). Given the low dimension of the parameter space in the present application, however, I focus on the discrete approximation.

[15] All results reported here are based on 1 million simulation draws.

FIGURE 1.—ROBUST AND NONROBUST CONFIDENCE SETS FOR FULL PARAMETER VECTOR θ



The distortion cutoff $\hat{\gamma}$ is 10.42%.

$1-\alpha$ quantile of a $(1+a(\gamma_{\min}))\times\chi^2_p + a(\gamma_{\min})\times\chi^2_{k-p}$ distribution. The robust confidence set for $f(\theta)$ is then

$$CS_R = \left\{ f(\theta_i) : \theta_i \in \Theta_D, K_{\Omega,f}(\theta_i) + a \times S(\theta_i) \leq H^{-1}(1-\alpha; a(\gamma_{\min}), k, p) \right\}.$$

Likewise, the nonrobust confidence set is

$$CS_N = \left\{ f(\theta_i) : \theta_i \in \Theta_D, W(f(\theta_i)) \leq \chi^2_{p,1-\alpha} \right\}.$$

Step 6: Calculate $\hat{\gamma}$. Finally, we calculate the distortion cutoff $\hat{\gamma}$. Note that if $\gamma_{\min} = 0$, then for the discretized problem we consider here, $\hat{\gamma}$ solves

$$\min_{\theta_i \in \Theta_D : W(f(\theta_i)) > \chi^2_{p,1-\alpha}} K_{\Omega,f}(\theta_i) + a(\hat{\gamma}) \times S(\theta_i) = \chi^2_{p,1-\alpha},$$

since for any γ larger than this,

$$\left\{ \theta_i \in \Theta_D : K_{\Omega,f}(\theta_i) + a(\gamma) \times S(\theta_i) \leq \chi^2_{p,1-\alpha} \right\} \subseteq \left\{ \theta_i \in \Theta_D : W(f(\theta_i)) \leq \chi^2_{p,1-\alpha} \right\}.$$

Thus, if for any value $\gamma_{\min}$ we define

$$\tilde{a} = \max_{\theta_i \in \Theta_D} \frac{\chi^2_{p,1-\alpha} - K_{\Omega,f}(\theta_i)}{S(\theta_i)} 1\left\{ W(f(\theta_i)) > \chi^2_{p,1-\alpha} \right\},$$

then for

$$\tilde{\gamma} = 1 - \alpha - Pr\left\{ (1+\tilde{a}) \times \chi^2_p + \tilde{a} \cdot \chi^2_{k-p} \leq \chi^2_{p,1-\alpha} \right\},$$

we see that $\hat{\gamma} = \max\{\tilde{\gamma}, \gamma_{\min}\}$. Hence, given the discretization of the parameter space we can easily determine $\hat{\gamma}$ from the quantities calculated above.

TABLE 1.—CONFIDENCE SETS AND DISTORTION CUTOFFS $\hat{\gamma}$ FOR PARAMETERS δ AND η

Parameter	CS_R	CS_N	$\hat{\gamma}$
δ	[0.6, 1.1]	[0.867, 0.948]	6.64%
η	[−6, −5.3]∪[−1.45, 1.95]∪ [5.25, 35.7] ∪ [54, 60]	[−1.1, 1.95]	6.64%

B. *Empirical Results*

Figure 1 reports joint confidence sets for the full parameter vector θ in this application, and table 1 reports marginal confidence sets for the parameters δ and η separately. In all cases, the robust confidence sets have larger volume than the nonrobust ones. Nonetheless, we see that the distortion cutoff $\hat{\gamma}$ is 10.42% for the joint confidence set and just 6.64% for both marginal confidence sets. Thus, while readers interested in two-step confidence sets and willing to accept at most a 5% coverage distortion should focus on the robust confidence sets in all cases, readers willing to accept a 10% coverage distortion could use the nonrobust marginal confidence sets for δ and η.

A notable feature of these results is that the distortion cutoff $\hat{\gamma}$ is the same for the parameters δ and η in this application. This results from the fact that the continuously updating the GMM objective function in this application has a saddle point. Using the results of Kleibergen (2005), one can show that all optimally weighted $K_{\Omega,f}$ statistics are equal to 0 at this saddle point by construction, and $\hat{\gamma}$ must be large enough to ensure that $CS_P(\hat{\gamma})$ excludes this point. The minimal value γ required for this purpose is the same for both δ and η, however, and in both cases, this value also suffices to ensure that $CS_P(\gamma)$ is also contained in CS_N. Thus, in this application, $\hat{\gamma}$ is the same for both δ and η.

As suggested above, results in this setting are quite sensitive to the specification considered. While our baseline moments take R_t to be an equity return, if we add moments

that take R_t to be an interest rate, then as elsewhere in the consumption-based asset pricing literature (e.g., Lettau & Ludvigson, 2009) we obtain a much larger estimate of risk aversion. Moreover, in these specifications, we obtain larger $\hat{\gamma}$, equal to 65.72% and 67.49% for δ and η, respectively. Details of these results are provided in section B of the supplementary appendix.

V. Conclusion

This paper develops two-step confidence sets with controlled coverage distortions in GMM models. The particular implementation I propose is based on generalizations of the statistics studied by Kleibergen (2005) and Chaudhuri and Zivot (2011), but there are many other ways one could construct confidence sets CS_R and $CS_P(\gamma)$ satisfying the requirements of theorem 1, and the comparative performance of different choices is an interesting question for future research. While I have established the validity of the confidence sets I construct under particular sequences of parameter values, conditions for uniform asymptotic validity are an interesting open question.

REFERENCES

Andrews, D., and X. Cheng, "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," *Econometrica* 80 (2012), 2153–2211.

Andrews, D., X. Cheng, and P. Guggenberger, "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Cowles Foundation working paper (2011).

Andrews, D., and P. Guggenberger, "Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models," *Econometric Theory* 83 (2017), 1046–1080.

Andrews, I. "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica* 84 (2016), 2155–2182.

Campbell, J. Y., and R. J. Shiller, "Cointegration Tests of Present Value Models," *Journal of Political Economy* 95 (1987), 1062–1088.

Chaudhuri, S., and E. Zivot, "A New Method of Projection-Based Inference in GMM with Weakly Identified Nuisance Parameters," *Journal of Econometrics* 164 (2011), 239–251.

Chernozhukov, V., C. Hansen, and M. Jansson, "Finite Sample Inference for Quantile Regression Models," *Journal of Econometrics* 152 (2009), 93–103.

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

Hansen, L. P., and K. J. Singleton, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50 (1982), 1269–1286.

Kleibergen, F. "Testing Parameters in GMM without Assuming They Are Identified," *Econometrica* 73 (2005), 1103–1123.

Lettau, M., and S. C. Ludvigson, "Euler Equation Errors," *Review of Economic Dynamics* 12 (2009), 255–283.

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

Newey, W., and D. McFadden, "Large Sample Estimation and Hypothesis Testing," Robert Engle and Daniel McFadden, eds, *Handbook of Econometrics* (Dordrecht: Elsevier, 1994).

Newey, W., and K. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55 (1987), 703–708.

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

Stock, J., and J. Wright, "Gmm with Weak Identification," *Econometrica* 68 (2000), 1055–1096.

Stock, J., and M. Yogo, "Testing for Weak Instruments in Linear IV Regression" (pp. 80–108) in Donald Andrews and James Stock; eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (Cambridge: Cambridge University Press, 2005).

APPENDIX

This appendix contains proofs for results stated in the paper. Proofs for an auxiliary lemma, additional details on the empirical application, and simulation results for the linear IV model are given in the supplementary appendix.[16]

Proof of Lemma 1. To prove equation (1), note that for any $\xi_0 \in \Xi$ and any T,

$$Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_2\right\} \geq Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_R\right\} - Pr_{T,\xi_0}\left\{\phi_{ICS} = 0\right\}.$$

By equation (4), $SCP\left(CS_R, \Xi_W\right) \geq 1-\alpha$, so lemma 1.1 follows immediately from the definition of sequential coverage probability.

To prove equation (2), note that

$$\begin{aligned} &Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_2\right\} \\ &\quad \geq Pr_{T,\xi_0}\left\{\left\{\beta_{0,T} \in CS_N\right\} \cap \left\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\right\}^C\right\} \\ &\quad \geq Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_N\right\} - Pr_{T,\xi_0}\left\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\right\} \end{aligned}$$

and

$$\begin{aligned} &Pr_{T,\xi_0}\left\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\right\} \\ &\quad \leq \min\left\{Pr_{T,\xi_0}\left\{\beta_{0,T} \notin CS_R\right\}, Pr_{T,\xi_0}\left\{\phi_{ICS} = 1\right\}\right\}. \end{aligned}$$

By equation 3, $SCP\left(CS_N, \Xi_S\right) \geq 1 - \alpha$ so

$$\begin{aligned} SCP\left(CS_2, \Xi_S\right) \geq 1 - \alpha - \sup_{\xi_0 \in \Xi_S} \limsup_{T\to\infty} \min\Big\{&Pr_{T,\xi_0}\left\{\beta_{0,T} \notin CS_R\right\}, \\ &Pr_{T,\xi_0}\left\{\phi_{ICS} = 1\right\}\Big\}, \end{aligned}$$

but $\sup_{\xi_0 \in \Xi_S} \limsup_{T\to\infty} Pr_{T,\xi_0}\left\{\beta_{0,T} \notin CS_R\right\} \leq \alpha$ by assumption, implying the result.

Proof of Theorem 1. To establish equation (1), note that by assumption 1b, $Pr_{T,\xi_0}\left\{CS_P(\gamma) \subseteq CS_R\right\} = 1$ for all T and $\xi_0 \in \Xi$. Thus, by the definition of CS_2, $Pr_{T,\xi_0}\left\{CS_P(\gamma) \subseteq CS_2\right\} = 1$ for all T and $\xi_0 \in \Xi$. Consequently, $Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_P(\gamma)\right\} \leq Pr_{T,\xi_0}\left\{\beta_{0,T} \in CS_2(\gamma)\right\}$, so equation (1) follows immediately from assumption 1a. Equation (2) follows immediately from lemma lb and assumption 1c. Equation (3) is implied by

$$\sup_{\xi_0 \in \Xi_S} \limsup_{T\to\infty} Pr_{T,\xi_0}\left\{\phi_{ICS} = 1\right\} = 0,$$

which is an immediate consequence of assumption 1c.

Proof of Theorem 2. We can rewrite $K_{\Omega,f}$ as

$$\begin{aligned} &K_{\Omega,f}(\theta) = \\ &\quad T \cdot g_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1} \\ &\qquad \times \Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}' \times \Bigg(\Lambda_{2,T}\frac{\partial}{\partial\theta'}f(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1} \\ &\qquad \times \Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}(\theta)\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1} \\ &\qquad \times \Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\Bigg)^{-1}\Lambda_{2,T}\frac{\partial}{\partial\theta'}f(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1} \\ &\qquad \times \Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta). \end{aligned}$$

[16] Available at http://economics.mit.edu/faculty/iandrews.

By Lemma 1 of Kleibergen (2005), $\left(\sqrt{T}g_T\left(\theta_{0,T}\right), \sqrt{T}\text{vec}\left(D_T\left(\theta_{0,T}\right) - J_{T,\xi}\right)\right)$ converges to $\left(\psi_g, \psi_D\right)$, which are mutually independent. By assumption, the elements of $\Lambda_{1,T}$ are of order $\sqrt{T}$, so $\frac{1}{\sqrt{T}}\Lambda_{1,T} = O(1)$ and $\left(\sqrt{T}g_T\left(\theta_{0,T}\right), D_T(\theta_{0,T})\Lambda_{1,T}\right)$ are asymptotically independent as well. In particular, $\left(\sqrt{T}g_T\left(\theta_{0,T}\right), D_T(\theta_{0,T})\Lambda_{1,T}\right) \to_d \left(\psi_g, D\right)$ where $\psi_g|D \sim N\left(0, \Sigma_g\right)$.

We can further rewrite $K_{\Omega,f}(\theta)$ as

$$T \cdot g_T(\theta)'\hat{\Sigma}_g(\theta)^{-\frac{1}{2}} P\left(\hat{\Sigma}_g(\theta)^{\frac{1}{2}} \hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)' \times \hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\right)\hat{\Sigma}_g(\theta)^{-\frac{1}{2}} g_T(\theta).$$

where $P(X) = X(X'X)^{-1}X'$ denotes the projection matrix onto X. By assumptions 3 and 4 and the continuous mapping theorem,

$$\begin{aligned}&\hat{\Sigma}_g\left(\theta_{0,T}\right)^{\frac{1}{2}} \hat{\Omega}\left(\theta_{0,T}\right) D_T(\theta_{0,T})\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta_{0,T})'\hat{\Omega}\left(\theta_{0,T}\right)\hat{\Sigma}_g\left(\theta_{0,T}\right)\right.\\ &\left.\times \hat{\Omega}\left(\theta_{0,T}\right) D_T(\theta_{0,T})\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta_{0,T})'\Lambda_{2,T}\\ &\to_d \Sigma_g^{\frac{1}{2}}\Omega D\left(D'\Omega\Sigma_g\Omega D\right)^{-1}F'\end{aligned}$$

where the sole random component on the right-hand side is D and the right-hand side has rank p almost surely. Together with the fact that $\Sigma_g^{-\frac{1}{2}}\psi_g|D \sim N(0, I_k)$, this implies by the continuous mapping theorem that $\left(K_{\Omega,f}\left(\theta_{0,T}\right), D_T(\theta_{0,T})\Lambda_{1,T}\right) \to_d \left(\tilde{K}_{\Omega,f}, D\right)$, where $\tilde{K}_{\Omega,f}|D \sim \chi_p^2$, since conditional on D, $\tilde{K}_{\Omega,f}$ is a quadratic form in a standard-normal random vector and a rank-p projection matrix.

One can handle $S\left(\theta_{0,T}\right) - K_{\Omega,f}\left(\theta_{0,T}\right)$ in a similar manner. In particular, note that

$$\begin{aligned}S(\theta) - K_{\Omega,f}(\theta) = {} & Tg_T(\theta)'\hat{\Sigma}_g(\theta)^{-\frac{1}{2}}\left(I - P\left(\hat{\Sigma}_g(\theta)^{\frac{1}{2}} \hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right.\right.\\ &\left.\left.\times\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)\hat{\Omega}(\theta) D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\right)\right)\\ &\times \hat{\Sigma}_g(\theta)^{-\frac{1}{2}} g_T(\theta).\end{aligned}$$

so

$$\begin{aligned}&\left(K_{\Omega,f}\left(\theta_{0,T}\right), S\left(\theta_{0,T}\right) - K_{\Omega,f}\left(\theta_{0,T}\right), D_T(\theta_{0,T})\Lambda_{1,T}\right)\\ &\quad\to_d \left(\tilde{K}_{\Omega,f}, \tilde{S} - \tilde{K}_{\Omega,f}, D\right),\end{aligned}$$

where $\left(\tilde{K}_{\Omega,f}, \tilde{S} - \tilde{K}_{\Omega,f}\right)|D \sim \left(\chi_p^2, \chi_{k-p}^2\right)$ and $\left(\tilde{K}_{\Omega,f}, \tilde{S} - \tilde{K}_{\Omega,f}\right)$ are independent conditional on D. Thus $\left(\tilde{K}_{\Omega,f}, \tilde{S} - \tilde{K}_{\Omega,f}\right)$ are independent and distributed $\left(\chi_p^2, \chi_{k-p}^2\right)$ unconditionally as well, which establishes the result.

Proof of Corollary 1. I prove the statement for $CS_P(\gamma)$, since the statement for CS_R follows by the same argument. Define

$$CS_{P,\theta}(\gamma) = \left\{\theta : K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) \leq \chi_{p,1-\alpha}^2\right\},$$

and note that since linear combinations of χ^2 random variables are continuously distributed and $K_{\Omega,f}(\theta) + a \times S(\theta) = (1+a) \times K_{\Omega,f}(\theta) + a \times \left(S(\theta) - K_{\Omega,f}(\theta)\right)$, theorem 2 implies that

$$\lim_{T\to\infty} Pr_{T,\xi_0}\left\{K_{\Omega,f}\left(\theta_{0,T}\right) + a(\gamma) \times S\left(\theta_{0,T}\right) \leq \chi_{p,1-\alpha}^2\right\} = 1 - \alpha - \gamma.$$

Thus,

$$\lim_{T\to\infty} Pr_{T,\xi_0}\left\{\theta_{0,T} \in CS_{P,\theta}(\gamma)\right\} = 1 - \alpha - \gamma.$$

Note, however, that $\theta_{0,T} \in CS_{P,\theta}$ implies that $f\left(\theta_{0,T}\right) \in CS_P(\gamma)$. Thus, we obtain

$$\liminf_{T\to\infty} Pr_{T,\xi_0}\left\{\theta_{0,T} \in CS_{P,\theta}(\gamma)\right\} \geq 1 - \alpha - \gamma,$$

as desired.

The proof of theorem 3 uses the following lemma, which is proved in the supplementary appendix.

Lemma 2. *Let $\left\{A_{\theta,T}\right\}$ be a sequence of random sets such that $\limsup_{T\to\infty} Pr_{\xi_0,T}\left\{A_{\theta,T} = \emptyset\right\} < 1$ and $\sup_{\theta\in A_{\theta,T}}\|\theta - \theta_0\| = O_p\left(\frac{1}{\sqrt{T}}\right)$ (where I define the sup to be 0 if $A_{\theta,T}$ is empty). Under assumptions 2, 5, and 6, under all $\xi_0 \in \Xi_S$,*

$$\sup_{\theta\in A_{\theta,T}}\left\|W(f(\theta)) - K_{\Omega,f}(\theta)\right\| = o_p(1).$$

Proof of Theorem 3. For $S(\theta)$ as in equation (8), note that Assumption 5 implies that for any $\varepsilon > 0$,

$$\inf_{\|\theta-\theta_0\|\geq\varepsilon} S(\theta) \to_p \infty.$$

Thus, if we define $A_{\theta,T} = \left\{\theta : a(\gamma) \cdot S(\theta) \leq \chi_{p,1-\alpha}^2\right\}$ then $\sup_{\theta\in A_{\theta,T}}\|\theta - \theta_0\| = o_p(1)$. A mean-value expansion yields that $g_T(\theta) = g_T(\theta_0) + G_T(\theta^*)(\theta - \theta_0)$. Since $\sup_{\theta\in B(\theta_0)}\|G_T(\theta) - J(\theta)\| = o_p(1)$, and $\sup_{\theta\in B(\theta_0)}\left\|\hat{\Sigma}_g(\theta) - \Sigma_g(\theta)\right\| = o_p(1)$ for an open ball $B(\theta_0)$ around θ_0 as in assumption 6 and $J(\theta)$ and $\Sigma_g(\theta)$ are continuous in θ,

$$\begin{aligned}\sup_{\theta\in A_{\theta,T}}&\left|S(\theta) - T\left(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\right)'\right.\\ &\left.\times\Sigma_g(\theta_0)^{-1}\left(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\right)\right| = o_p(1).\end{aligned}$$

Thus, for any $\varepsilon > 0$ and for $\underline{\lambda}$ the minimal eigenvalue of $\Sigma_g(\theta_0)^{-1}$,

$$Pr_{T,\xi_0}\left\{\inf_{\theta\in A_{\theta,T}}\left(S(\theta) - \underline{\lambda}T\|g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\|^2\right) > -\varepsilon\right\} \to 1.$$

Since $\sqrt{T}g_T(\theta_0) = O_p(1)$ by assumption 2, this implies that $\sup_{\theta\in A_{\theta,T}}\|\theta - \theta_0\| = O_p\left(\frac{1}{\sqrt{T}}\right)$. Thus $A_{\theta,T} = \left\{\theta : a(\gamma) \times S(\theta) \leq \chi_{p,1-\alpha}^2\right\}$ shrinks toward θ_0 at rate $\sqrt{T}$.

Next, note that $K_{\Omega,f}(\theta) \geq 0$ by construction, so $K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) \geq a(\gamma) \times S(\theta)$ and $CS_P(\gamma) \subseteq A_{\theta,T}$. By standard results on the distribution of tests for overidentifying restrictions $\inf_\theta S(\theta) \to_d \chi_{k-p}^2$, so since

$$K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) \geq K_{\Omega,f}(\theta) + a(\gamma) \cdot \inf_\theta S(\theta)$$

and by lemma 2 $\sup_{\theta\in A_{\theta,T}}\left|K_{\Omega,f}(\theta) - W(f(\theta))\right| = o_p(1)$, we obtain that if $k > p$, then

$$Pr_{T,\xi_0}\left\{\inf_{\theta\in A_{\theta,T}}\left(K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) - W(f(\theta))\right) > 0\right\} \to 1,$$

with the consequence that $Pr_{T,\xi}\{CS_P(\gamma) \subseteq CS_N\} \to 1$, as we wanted to show. If $k = p$, then $K_{\Omega,f}(\theta) + a(\gamma) \times S(\theta) = (1 + a(\gamma))K_{\Omega,f}(\theta)$, and the same conclusion follows from the fact that $\sup_{\theta\in A_{\theta,T}}\left|K_{\Omega,f}(\theta) - W(f(\theta))\right| = o_p(1)$.