Estimation and Inference with Weak Identification

Donald W. K. Andrews*
Cowles Foundation
Yale University

Xu Cheng
Department of Economics
University of Pennsylvania

First Draft: August, 2007
Revised: March 10, 2010

*Andrews gratefully acknowledges the research support of the National Science Foundation via grant number SES-0751517.

Abstract

This paper analyses the properties of standard estimators, tests, and confidence sets (CS's) in a class of models in which the parameters are unidentified or weakly identified in some parts of the parameter space. The paper also introduces a method to make the tests and CS's robust to such identification problems. The results apply to a class of extremum estimators and corresponding tests and CS's, including maximum likelihood (ML), least squares (LS), quantile, generalized method of moments (GMM), generalized empirical likelihood (GEL), and minimum distance (MD) estimators. The consistency/lack-of-consistency and asymptotic distributions of the estimators are established under a full range of drifting sequences of true distributions. The asymptotic size (in a uniform sense) of standard tests and CS's is established. The results are applied to the LS estimator of a nonlinear regression model, a LS estimator of a smooth transition threshold autoregressive model, the instrumental variables estimator of a nonlinear regression model with endogeneity, and the ML estimator of an ARMA (1, 1) model.

Keywords: Asymptotic size, confidence set, estimator, identification, nonlinear models, test, weak identification.

JEL Classification Numbers: C12, C15.

1. Introduction

The literature in econometrics has shown considerable interest in issues related to identification over the last two decades (and, of course, prior to that as well). For example, research has been carried out on models with weak instruments, models with partial identification, models with and without nonparametric identification, tests with nuisance parameters that are unidentified under the null hypothesis, and the finite sample properties of statistics under lack of identification. The present paper is in this line of research, but focuses on a class of models that has not been investigated fully in the literature. It includes models with weak instruments but the focus of the paper is on other models in this class.

We consider a class of models in which lack of identification occurs in part of the parameter space. Specifically, we consider models in which the parameter θ of interest is of the form $\theta = (\beta, \zeta, \pi)$, where π is identified if and only if $\beta \neq 0$, ζ is not related to the identification of π, and $\psi = (\beta, \zeta)$ is always identified. The parameters β, ζ, and π may be scalars or vectors. This a canonical parametrization which may or may not hold in the natural parameterization of the model, but is assumed to hold after suitable reparametrization. For example, the nonlinear regression model, $Y_i = \beta h(X_i, \pi) + Z_i'\zeta + U_i$, where (Y_i, X_i, Z_i) is observed and $h(\cdot, \cdot)$ is known, is of the form just described. So are other models that depend on a nonlinear index of the form $\beta h(X_i, \pi) + Z_i'\zeta$.

Suppose θ is estimated by minimizing a criterion function $Q_n(\theta)$ over a parameter space Θ. Lack of identification of π when $\beta = 0$ leads to $Q_n(\theta)$ being (relatively) flat with respect to (wrt) π when β is close to 0. For example, the LS criterion function in the nonlinear regression example, $n^{-1}\sum_{i=1}^n (Y_i - \beta h(X_i, \pi) + Z_i'\zeta)^2$, has first derivative wrt π equal to $2\beta n^{-1}\sum_{i=1}^n (Y_i - \beta h(X_i, \pi) + Z_i'\zeta)(\partial/\partial\pi)h(X_i, \pi)$, which is close to 0 for β close to 0. Flatness of $Q_n(\theta)$ is well-known to cause numerical difficulties in practice. It also causes difficulties with standard asymptotic approximations because the second derivative matrix of $Q_n(\theta)$ is singular or near singular and standard asymptotic approximations involve the inverse of this matrix.

In addition to the nonlinear regression model, other examples that are considered in the paper (or companion papers) include a smooth transition threshold autoregressive (STAR) model, a smooth transition switching regression model, an ARMA (1, 1) model, a nonlinear regression model with endogenous regressors, limited dependent variable models, including probit and censored regression, with endogeneity and a lin-

ear reduced-form equation for the endogenous variable(s), see Nelson and Olson (1978), Lee (1981), Smith and Blundell (1986), Newey (1987), and Rivers and Vuong (1988), and an endogenous probit model with no exclusion restriction but a nonlinear parametric reduced-form equation for the endogenous regressor, see Dong (2009) for a related model. [These examples are in different stages of completion.] Han (2009) shows that, via reparametrization, a simple bivariate probit model with endogeneity falls into the class of models considered here. Other examples include continuous transition structural change models, continuous transition threshold autoregressive models (e.g., see Chan and Tsay (1998)), seasonal ARMA(1, 1) models (e.g., see Andrews, Liu, and Ploberger (1998)), models with correlated random coefficients (e.g., see Andrews (2001)), GARCH(p, q) models, and time series models with nonlinear deterministic time trends of the form t^π or $(t^\pi - 1)/\pi$. [1]

Not all models with lack of identification at some points in the parameter space fall into the class of models considered here. The models considered here must satisfy a set of criterion function (stochastic) quadratic approximation conditions, as described in more detail below, that do not apply to some models of interest. For example, abrupt transition structural change models, (unobserved) regime switching models, and abrupt transition threshold autoregressive models are not covered by the results of the present paper, e.g., see Picard (1985), Chan (1993), Bai (1997), Hansen (2000), Liu and Shao (2003), Elliott and Müller (2007, 2008), and Drton (2009) for analyses of these models.

The approach of the paper is to consider a general class of extremum estimators that includes ML, LS, quantile, GMM, GEL, and MD estimators. The criterion functions considered may be smooth or non-smooth functions of θ. We place high-level conditions on the behavior of the criterion function $Q_n(\theta)$, provide a variety of more primitive sufficient conditions, and verify the latter in several examples. For example, we provide more primitive sufficient conditions for the case where the criterion function takes the form of a sample average that is a smooth function of θ and is based on i.i.d. or stationary time series observations, which covers ML and LS estimators. These conditions are of a similar nature to standard ML regularity conditions, and indeed cover ML estimators, but allow for non-regularity in terms of a certain type of identification failure. We also provide sufficient conditions for GMM criterion functions. The high-level conditions

[1] Nonlinear time trends can be analyzed asymptotically in the framework considered in this paper via sample size rescaling, i.e., by considering $(t/n)^\pi$ or $((t/n)^\pi - 1)/\pi$, e.g., see Andrews and McDermott (1995).

given here have the attractive features of (i) clarifying precisely which features of the criterion function are essential for the analysis and (ii) covering a wide variety of cases simultaneously.

Given the high-level conditions, we establish the large sample properties of extremum estimators, t and Wald tests, and t and Wald CS's under lack of identification, weak identification, semi-strong identification, and strong identification, as discussed below. We investigate the large sample biases of extremum estimators under weak identification. We determine the asymptotic size of standard tests and CS's, which often deviates from their nominal size in the presence of lack of identification at some points in the parameter space.[2]

We introduce a method of making standard tests and CS's robust to lack of identification, i.e., to have correct asymptotic size. The method is closely related to a method suggested in Andrews (1999, Sec. 6.4; 2000, Sec. 4) for boundary problems and to the generalized moment selection critical value method used in Andrews and Soares (2010) and some other papers for inference in partially-identified models based on moment inequalities. The idea is to use a testing/model selection procedure to determine whether β is close to the non-identification value 0 and, if so, to adjust the critical value to take account of the effect of non-identification or weak identification on the behavior of the test statistic.

The resulting identification-robust tests and CS's are ad hoc in nature and do not have any optimality properties. However, they are generally applicable and often have the advantage of computational ease. In some models with potential identification failure, procedures with explicit asymptotic optimality/admissibility properties are available. For example, see Elliott and Müller (2007, 2008) for some change-point problems.

In the models considered here, weak identification occurs when $\beta \neq 0$ but β is close to 0. As is well-known from the literature on weak instruments, the effect of β of a given magnitude on the behavior of estimators and tests depends on the sample size n. In consequence, to capture asymptotically the finite-sample behavior of estimators, tests, and CS's under near non-identification, one has to consider drifting sequences of true distributions. In the present context, one needs to consider drifting sequences in which β_n drifts to 0 at various rates and β_n drifts to non-zero values.

[2] Asymptotic size is defined to be the limit of exact (i.e., finite-sample) size. For a test, exact size is the maximum rejection probability over distributions in the null hypothesis. For a CI, exact size is the minimum coverage probability over all distributions. Because exact size has uniformity built into its definition, so does asymptotic size as defined here.

Interest in asymptotics with drifting sequences of parameters goes back to Neyman-Pitman drifts, which are used to approximate the power functions of tests, and contiguity results, which are used for asymptotic efficiency calculations among other things. More recently, drifting sequences of parameters have been shown to play a crucial role in the literature on weak instruments, e.g., see Staiger and Stock (1997), and the literature on the (uniform) asymptotic size properties of tests and CS's when the statistics of interest display discontinuities in their pointwise asymptotic distributions, see Andrews and Guggenberger (2009, 2010) and Andrews, Cheng, and Guggenberger (2009). The situation considered here is an example of the latter phenomenon. The latter papers show that to determine asymptotic size, it is both necessary and sufficient to determine the behavior of the relevant statistics under certain drifting sequences of parameters. In this paper, we use the results in those papers and consider a collection of drifting sequences of parameters/distributions that are sufficient to determine the asymptotic size of the tests and CS's considered.

Suppose the true value of the parameter is $\theta_n = (\beta_n, \zeta_n, \pi_n)$ for $n \geq 1$, where n indexes the sample size. The behavior of extremum estimators and tests in the present context depends on the magnitude of $||\beta_n||$. The asymptotic behavior of these statistics varies across the three categories of sequences $\{\beta_n : n \geq 1\}$ defined in Table I.

The asymptotic results of the paper for the extremum estimator $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\zeta}_n, \widehat{\pi}_n)$ are summarized as follows: The estimator $\widehat{\psi}_n = (\widehat{\beta}_n, \widehat{\zeta}_n)$ is $n^{1/2}$-consistent for all categories of sequences $\{\beta_n\}$. The estimator $\widehat{\pi}_n$ is inconsistent for Category I sequences and consistent for Categories II and III. The asymptotic distribution of $n^{1/2}(\widehat{\psi}_n - \psi_n)$ $(= n^{1/2}((\widehat{\beta}_n, \widehat{\zeta}_n) - (\beta_n, \zeta_n)))$ is a functional of a Gaussian process with a mean that is (typically) non-zero for Category I sequences (due to the inconsistency of $\widehat{\pi}_n$) and is normal with mean zero for Categories II and III. The asymptotic distribution of $\widehat{\pi}_n$ is a functional of the same Gaussian process for Category I sequences. These estimation results permit the calculation of the asymptotic biases of $(\widehat{\beta}_n, \widehat{\zeta}_n, \widehat{\pi}_n)$ for Category I sequences as a function of the strength of identification. The asymptotic distribution of $n^{1/2}||\beta_n||(\widehat{\pi}_n - \pi_n)$ is normal with mean zero for Category II sequences. The asymptotic distribution of $n^{1/2}(\widehat{\pi}_n - \pi_n)$ is normal with mean zero for Category III sequences.

Table I. Identification Categories.

Category	$\{\beta_n\}$ Sequence	Identification Property of π
I(a)	$\beta_n = 0 \ \forall n \geq 1$	Unidentified
I(b)	$\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in R^{d_\beta}$ (and, hence, $\|\|\beta_n\|\| = O(n^{-1/2})$)	Weakly identified
II	$\beta_n \to 0$ and $n^{1/2}\|\|\beta_n\|\| \to \infty$	Semi-strongly identified
III	$\beta_n \to \beta_0 \neq 0$	Strongly identified

Similarly, the asymptotic results for tests and CS's vary over the three categories. For Category I sequences, standard tests and CS's have asymptotic rejection/coverage probabilities that may differ, sometimes substantially, from their nominal level. In consequence, the asymptotic size of standard tests and CS's often is substantially different from the desired nominal size. For Category II and III sequences, standard tests and CS's have the desired asymptotic rejection/coverage probability properties. For hypotheses or CS's that involve π, their power/non-coverage properties are standard for Category II and III sequences.

Next, we discuss the literature that is related to this paper. Cheng (2008) considers a nonlinear regression model with multiple nonlinear regressors and, hence, multiple sources of lack of identification. In contrast, the present paper only considers a single source of lack of identification (based on the magnitude of the true value of $||\beta||$), which translates into a single nonlinear regressor in the nonlinear regression example. On the other hand, the present paper covers a much wider variety of models than does Cheng (2008).

In the models considered in this paper, a test of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, is a test for which π is a nuisance parameter that is unidentified under the null hypothesis. Testing problems of this type have been considered in the literature, see Davies (1977, 1987), Andrews and Ploberger (1994), and Hansen (1996a). In contrast, the hypotheses considered in this paper are of a more general type. Here we consider a full range of nonlinear hypotheses concerning (β, ζ, π)—only special cases are of the type just described. For example, when the null hypothesis concerns ζ, then π is a nuisance parameter that is identified in part of the null hypothesis and unidentified in another part. If the null hypothesis involves all three parameters (β, ζ, π), then the identification scenario is substantially more complicated than when H_0 is $\beta = 0$. Furthermore, even if interest is focussed on β, to obtain asymptotic size results for CS's for β, one needs to consider drifting sequences of null hypotheses of the form $H_0 : \beta = \beta_n^*$ for $n \geq 1$. Such

testing problems are different from those considered in the literature referred to above.

The weak instrumental variable (IV) literature, e.g., see Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2002, 2005), Moreira (2003), and other papers referenced in Andrews and Stock (2007), is related to the present paper. This is especially true of Staiger and Stock (1997). In the weak IV literature, the criterion functions considered do not depend on the parameters that are the source of weak identification. In the present paper, the criterion functions do. In consequence, the present paper and the weak IV literature are complementary—they focus on different criterion functions/models.

However, there is some overlap. For example, in the standard linear IV regression model, the criterion function for the limited information maximum likelihood (LIML) estimator can be written either as (i) a function of the parameters in the structural equation plus the parameters in the accompanying reduced-form equations, which fits the framework of the present paper and yields results that cover both the structural and reduced-form parameters, or (ii) a function of the structural equation parameters only via concentrating out the reduced-form parameters, as in the analysis in Anderson and Rubin (1949) and Staiger and Stock (1997). The same is true of two-stage least squares (2SLS) estimators by writing the criterion function for the structural and reduced-form parameters as a single GMM criterion function without no over-identifying restrictions.

This approach for 2SLS can be extended to two-stage estimators of endogenous limited dependent variable models, see the discussion and references above. Such estimators are not covered by the results of Staiger and Stock (1997). It is not clear whether they fit into the weak instrument framework of Stock and Wright (2000).

The finite-sample results of Dufour (1997) and Gleser and Hwang (1987) for CS's and tests are applicable to the models considered in this paper. This paper considers the case where the potentially unidentified parameter π lies in a bounded set Π. In this case, Cor. 3.4 of Dufour (1997) implies that if the diameter of a CS for π is as large as the diameter of Π with probability less than $1 - 2\alpha$ then the CS has (exact) size less than $1 - \alpha$ (under certain assumptions).

Nelson and Startz (2007) introduces the zero-information-limit condition, which applies to the models considered in this paper, and discuss its implications. Ma and Nelson (2006) considers tests based on linearization for models of the type considered in this paper. Neither of these papers establishes the large sample properties of estimators, tests, and CS's along the lines given in this paper.

Phillips (1989) and Choi and Phillips (2001) provide finite-sample and asymptotic results for linear simultaneous equations and linear spurious regression models in which some parameters are unidentified. Their results do not overlap very much with those in this paper because the present paper is focussed on nonlinear models. Their asymptotic results are pointwise in the parameters, which covers the unidentified and strongly identified categories, but not the weakly identified and semi-strongly identified categories described above.

The results of the present paper apply to the nonlinear regression model estimated by LS. We use this as an example to illustrate the general results of the paper. In the example, the regressors are i.i.d. or stationary and ergodic. One also can apply the approach of this paper to the case where the regressors are integrated. In this case, the general results given below do not apply directly. However, by using the asymptotics for nonlinear and nonstationary processes developed by Park and Phillips (1999, 2001), the approach goes through, as shown recently by Shi and Phillips (2009). With integrated regressors, the nonlinear regression model is a nonlinear cointegration model. Shi and Phillips (2009) employs the same method of computing asymptotic size and of constructing identification-robust CS's as was introduced in an early version of this paper and Cheng (2008).

The remainder of the paper is organized as follows. Section 2 describes the method used in the paper to obtain the asymptotic results. Section 3 introduces the extremum estimators, criterion functions, tests, confidence sets, and drifting sequences of distributions considered in the paper. Section 4 states the high-level assumptions employed. Section 5 provides the asymptotic results for the extremum estimators. Section 6 provides primitive sufficient conditions for the high-level assumptions for the class of estimators based on sample averages, which includes ML and LS estimators, that are smooth functions of the parameter θ. Section 7 establishes the asymptotic distributions of Wald and t statistics, determines the asymptotic size of standard Wald and t CS's, and introduces robust Wald and t tests and Wald and t CS's, whose asymptotic size is equal to their nominal size. Section 8 provides results for two examples: the smooth transition threshold autoregressive model (STAR) and the ARMA(1, 1) model. [This section is still in preparation.] Appendix A gives various sufficient conditions for the high-level conditions stated in Section 4. Appendix B provides proofs of the results given in Sections 5 and 7. Appendix C provides sufficient conditions for the high-level conditions for the class of GMM estimators.

All limits below are taken "as $n \to \infty$." Let $o_{p\pi}(1)$, $O_{p\pi}(1)$, and $o_\pi(1)$ denote terms that are $o_p(1)$, $O_p(1)$, and $o(1)$, respectively, uniformly over a parameter $\pi \in \Pi$. Thus, $X_n(\pi) = o_{p\pi}(1)$ means that $\sup_{\pi\in\Pi} ||X_n(\pi)|| = o_p(1)$, where $||\cdot||$ denotes the Euclidean norm. Let "for all $\delta_n \to 0$" abbreviate "for all sequences of positive scalar constants $\{\delta_n : n \geq 1\}$ for which $\delta_n \to 0$." Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues, respectively, of a matrix A. All vectors are column vectors. For notational simplicity, we often write (a, b) instead of $(a', b')'$ for vectors a and b. Also, for a function $f(c)$ with $c = (a, b)$ $(= (a', b')')$, we often write $f(a, b)$ instead of $f(c)$. Let 0_d denote a d-vector of zeros. Because it arises frequently, we let 0 denote a d_β-vector of zeros, where d_β is the dimension of a parameter β. Let $R_{[\pm\infty]} = R \cup \{\pm\infty\}$. Let $R^p_{[\pm\infty]} = R_{[\pm\infty]} \times ... \times R_{[\pm\infty]}$ with p copies.

Let $\Rightarrow$ denote weak convergence of a sequence of stochastic processes indexed by $\pi \in \Pi$ for some space Π. The definition of weak convergence of R^v-valued functions on Π requires the specification of a metric d on the space $\mathcal{E}_v$ of R^v-valued functions on Π. We take d to be the uniform metric. The literature contains several definitions of weak convergence. We use any of the definitions that is compatible with the use of the uniform metric and for which the continuous mapping theorem (CMT) holds. These include the definitions employed by Pollard (1984, p. 65), Pollard (1990, p. 44), and van der Vaart and Wellner (1996, p. 17). The CMT's that correspond to these definitions are given by Pollard (1984, p. 70), Pollard (1990, p. 46), and van der Vaart and Wellner (1996, Thm. 1.3.6, p. 20). In the event of measurability issues, outer probabilities are used below implicitly in place of probabilities.

2. Description of Approach

The criterion functions/models considered in this paper possess the following characteristics:

(i) the criterion function does not depend on π when $\beta = 0$,

(ii) the criterion function viewed as a function of ψ with π fixed has a (stochastic) quadratic approximation wrt ψ (for ψ close to the true value of ψ) for each $\pi \in \Pi$ when the true β is close to the non-identification value 0 (see Assumption C1 in Section 4.4 below),

(iii) the (generalized) first derivative of this quadratic expansion converges weakly as a process indexed by $\pi \in \Pi$ to a Gaussian process after suitable normalization,

(iv) the (generalized) Hessian of this quadratic expansion is nonsingular asymptotically for all $\pi \in \Pi$ after suitable normalization,
(v) the criterion function viewed as a function of θ has a (stochastic) quadratic approximation wrt θ (for θ close to the true value) whether or not the true β is close to the non-identification value 0 (see Assumption D1 in Section 4.5 below),
(vi) the (generalized) first derivative of this quadratic expansion has an asymptotic normal distribution after a matrix rescaling when β is local to the non-identification value 0, and
(vii) the (generalized) Hessian of this quadratic expansion is nonsingular asymptotically after a matrix rescaling when β is local to the non-identification value 0.

Now, we describe the approach used to establish the asymptotic results discussed in the Introduction. The estimator $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\zeta}_n, \widehat{\pi}_n)$ is defined to minimize a criterion function $Q_n(\theta)$ over $\theta \in \Theta$. Let $\theta_n = (\beta_n, \zeta_n, \pi_n)$ denote the true parameter.

Several steps are employed. The first three steps apply to sequences of true parameters in Categories I and II.

Step 1. We consider the concentrated estimator $\widehat{\psi}_n(\pi)$ that minimizes $Q_n(\theta) = Q_n(\psi, \pi)$ over ψ for fixed $\pi \in \Pi$ and the concentrated criterion function $Q_n^c(\pi) = Q_n(\widehat{\psi}_n(\pi), \pi)$. We show that $\widehat{\psi}_n(\pi)$ is consistent for ψ_n uniformly over $\pi \in \Pi$. The method of proof is a variation of a standard consistency proof for extremum estimators adjusted to yield uniformity over π. The proof is analogous to that used in Andrews (1993) for estimators of structural change models in the situation where no structural change occurs.

Step 2. We employ a stochastic quadratic expansion of $Q_n(\psi, \pi)$ in ψ for given π about the non-identification point $\psi = \psi_{0,n} = (0, \zeta_n)$, rather than the true value ψ_n, which is key. By expanding about $\psi_{0,n}$, the leading term of the expansion, $Q_n(\psi_{0,n}, \pi)$, does not depend on π because $Q_n(\beta, \zeta, \pi)$ does not depend on π when $\beta = 0$. For each $\pi \in \Pi$, we obtain a linear approximation to $\widehat{\psi}_n(\pi)$ after centering around $\psi_{0,n}$ and rescaling. At the same time, we obtain a quadratic approximation of $Q_n^c(\pi)$. Both results hold uniformly in π. The method employed has two steps.

The first step of the two-step method involves establishing a rate of convergence result for $\widehat{\psi}_n(\pi) - \psi_{0,n}$. The second step uses this rate of convergence result to obtain the linear approximation of $\widehat{\psi}_n(\pi) - \psi_{0,n}$ (after rescaling) and the quadratic approximation of $Q_n(\psi, \pi) - Q_n(\psi_{0,n}, \pi)$ (after rescaling) as a function of ψ. Because $Q_n(\psi_{0,n}, \pi)$ does

not depend on π, it does not effect the behavior of $\widehat{\psi}_n(\pi)$ or $\widehat{\pi}_n$. The two-step method used here is like that used by Chernoff (1954), Pakes and Pollard (1989), and Andrews (1999) among others, except that it is carried out for a family of values π, as in Andrews (2001), rather than a single value, and the results hold uniformly over π.

Step 3. We determine the asymptotic behavior of the (generalized) first derivative of $Q_n(\psi, \pi)$ wrt ψ evaluated at $\psi_{0,n}$. Due to the expansion about $\psi_{0,n}$, rather than about the true value ψ_n, a bias is introduced in the first derivative—its mean is not zero. The results here differ between Category I and II sequences. With Category I sequences, one obtains a stochastic term (a mean zero Gaussian process indexed by π) plus a non-stochastic term due to the bias ($K(\pi; \gamma_0)b$ in the notation used below) and the two are of the same order of magnitude. With Category II sequences, the true β_n is farther from the point of expansion 0 than with Category I sequences and, in consequence, the non-stochastic bias term is of a larger order of magnitude than the stochastic term. In this case, the limit is non-stochastic.

We also determine the asymptotic behavior of the (generalized) Hessian matrix of $Q_n(\psi, \pi)$ wrt ψ evaluated at $\psi_{0,n}$. It has a non-stochastic limit. There is no problem here with singularity of the Hessian because it is the Hessian for ψ only, not $\theta = (\psi, \pi)$, and ψ is identified.

For Category I sequences, the results of this step combined with those of Step 2 and the condition $n^{1/2}(\psi_n - \psi_{0,n}) \rightarrow (b, 0)$ gives the asymptotic distribution of (i) the concentrated estimator $\widehat{\psi}_n(\cdot)$ viewed as a stochastic process indexed by $\pi \in \Pi$: $n^{1/2}(\widehat{\psi}_n(\cdot) - \psi_n) \Rightarrow \tau(\cdot)$, where $\tau(\cdot)$ is a Gaussian process indexed by $\pi \in \Pi$ whose mean is non-zero unless $b = 0$, and (ii) the concentrated criterion function $Q_n^c(\cdot)$: $n(Q_n^c(\cdot) - Q_n(\psi_{0,n}, \pi)) \Rightarrow \xi(\cdot)$, where $\xi(\cdot)$ is a quadratic form in $\tau(\cdot)$.

For Category II sequences, putting the results above together yields: (i) a rate of convergence result for $\widehat{\psi}_n(\pi)$: $\sup_{\pi \in \Pi} ||\widehat{\psi}_n(\pi) - \psi_{0,n}|| = O_p(||\beta_n||)$ that is just fast enough to obtain a rate of convergence result for $\widehat{\psi}_n - \psi_n$ in Step 6 below and (ii) the (non-stochastic) probability limit $\eta(\pi)$ of $Q_n^c(\pi)$ (after normalization): $||\beta_n||^{-1}(Q_n^c(\pi) - Q_n(\psi_{0,n}, \pi)) \rightarrow_p \eta(\pi)$ uniformly over $\pi \in \Pi$.

Step 4. For Category I sequences, we use $\widehat{\pi}_n = \arg\min_{\pi \in \Pi} Q_n^c(\pi)$, $n(Q_n^c(\cdot) - Q_n(\psi_{0,n}, \pi)) \Rightarrow \xi(\cdot)$ from Step 3 (where $Q_n(\psi_{0,n}, \pi)$ does not depend on π), and the continuous mapping theorem (CMT) to obtain $\widehat{\pi}_n \rightarrow_d \pi^* = \arg\min_{\pi \in \Pi} \xi(\pi)$. In this case, $\widehat{\pi}_n$ is not consistent. Given the asymptotic distribution of $\widehat{\pi}_n$, the result $n^{1/2}(\widehat{\psi}_n(\cdot) - \psi_n) \Rightarrow \tau(\cdot)$ from Step 3, and the CMT, we obtain the asymptotic distribution of $\widehat{\psi}_n = \widehat{\psi}_n(\widehat{\pi}_n)$:

$n^{1/2}(\widehat{\psi}_n - \psi_n) \rightarrow_d \tau(\pi^*)$. This completes the asymptotic results for $(\widehat{\psi}_n, \widehat{\pi}_n)$ for Category I sequences of true parameters.

Step 5. For Category II sequences, we obtain the consistency of $\widehat{\pi}_n$ by using the uniform convergence in probability of $Q_n^c(\pi)$ (after normalization) to the non-stochastic quadratic form, $\eta(\pi)$, established in Step 3, combined with the property that $\eta(\pi)$ is uniquely minimized at the limit π_0 of the true values π_n. The vector that appears in the quadratic form $\eta(\pi)$ is the vector of biases of the (generalized) first derivative obtained in Step 3, which appears due to the expansion around $\psi_{0,n}$ rather than around ψ_n.

Step 6. For Category II sequences, we use the rate of convergence result $\sup_{\pi \in \Pi} ||\widehat{\psi}_n(\pi) - \psi_{0,n}|| = O_p(||\beta_n||)$ from Step 3 and a relationship between the bias of the (generalized) first-derivative and the (generalized) Hessian (wrt ψ) to obtain a rate of convergence result for $\widehat{\psi}_n = \widehat{\psi}_n(\widehat{\pi}_n)$ centered at the true value ψ_n: $\widehat{\psi}_n - \psi_n = o_p(||\beta_n||)$.

Step 7. For Category II and III sequences, we carry out stochastic quadratic expansions of $Q_n(\theta)$ about the true value θ_n. The argument proceeds as in Step 2 (but the expansion here is in θ, not in ψ with π fixed, and the expansion is about the true value). First, we obtain a rate of convergence result for $\widehat{\theta}_n - \theta_n$ and then with this rate we obtain the asymptotic distribution of $\widehat{\theta}_n - \theta_n$ (after rescaling) using the quadratic approximation of $Q_n(\theta)$ in a particular neighborhood of θ_n. The result obtained is consistency and asymptotic normality (with mean zero) for $\widehat{\theta}_n$ with rate $n^{1/2}$ for $\widehat{\psi}_n$ for Category II and III sequences, rate $n^{1/2}$ for $\widehat{\pi}_n$ for Category III sequences, and rate $n^{1/2}||\beta_n||$ ($<< n^{1/2}$) for $\widehat{\pi}_n$ for Category II sequences. The last rate result is due to the convergence of β_n to 0 albeit slowly. With Category II sequences, $\widehat{\pi}_n$ is consistent and asymptotically normal but with a slower rate of convergence than is standard.

For Category II sequences, the results in this step are complicated by two issues. First, the (generalized) Hessian matrix for θ with the standard normalization is singular asymptotically because $\beta_n \rightarrow 0$ and the random criterion function $Q_n(\theta)$ becomes more flat wrt π for β in a neighborhood of β_n the closer is β_n to 0. This requires a matrix rescaling of the Hessian based on the magnitude of $||\beta_n||$. Second, the quadratic approximation of the criterion function wrt θ around the true value θ_n only holds for θ close enough to θ_n; specifically, only for $\theta \in \Theta_n(\delta_n) = \{\theta \in \Theta : ||\psi - \psi_n|| \leq \delta_n||\beta_n|| \ \& \ ||\pi - \pi_n|| \leq \delta_n\}$ for constants $\delta_n \rightarrow 0$. Thus, ψ needs to be very close to the true value ψ_n for the quadratic approximation to hold. It is for this reason that the rate of convergence result $\widehat{\psi}_n - \psi_n = o_p(||\beta_n||)$ in Step 6 is a key result. The quadratic approxi-

mation requires $\theta \in \Theta_n(\delta_n)$ because for such $\theta = (\beta, \zeta, \pi)$ we have $||\beta||/||\beta_n|| = 1 + o(1)$ and, hence, the rescaling that enters the Hessian is asymptotically equivalent whether it is based on β or the true value β_n. (For example, see the verification of Assumption Q1(iv) for the LS example in (10.77) to see that the restriction $\theta \in \Theta_n(\delta_n)$ is required for the quadratic approximation to hold in this example.)

Step 8. We obtain the asymptotic null distributions of the Wald and t test statistics for linear and nonlinear restrictions using the asymptotic distributions of the estimators described in Steps 1-7 plus asymptotic results for the variance matrix and standard error estimators upon which the test statistics depend. The latter exhibit non-standard behavior for Category I sequences because $\widehat{\pi}_n$ is random even in the limit. These results yield the asymptotic null rejection probabilities and coverage probabilities of standard Wald and t test for Category I-III sequences.

The analysis of Wald tests for multiple restrictions requires the use of a parameter dependent matrix rotation to separate the effects of randomness in $\widehat{\psi}_n$ and $\widehat{\pi}_n$ because these two estimators have different rates of convergence for sequences $\{\gamma_n\}$ in Categories I and II.

We show that for some multiple nonlinear restrictions the Wald test statistic diverges to infinity in probability under the null hypothesis. Obviously this is not a desirable property and leads to the standard Wald test (for any nominal level $\alpha > 0$) having asymptotic size equal to one.

For Category I sequences, the asymptotic distribution of the t statistic for a scalar linear or nonlinear restriction that involves both π and ψ is found to depend only on the randomness in $\widehat{\pi}_n$ and not on the randomness in $\widehat{\psi}_n$. This occurs because the former is of a larger order of magnitude than the latter. When a scalar restriction does not involve π, then the asymptotic null distribution of the t statistic for Category I sequences usually still depends on the (asymptotically non-standard) randomness of $\widehat{\pi}_n$ through the standard deviation estimator and implicitly through the effect of the randomness of $\widehat{\pi}_n$ on the asymptotic distribution of $\widehat{\psi}_n = \widehat{\psi}_n(\widehat{\pi}_n)$.

Step 9. Using the asymptotic results from Step 8 for Category I-III sequences of true parameters, combined with the argument from Andrews and Guggenberger (2010), as formulated in Andrews, Cheng, and Guggenberger (2009), we obtain a formula for the asymptotic size of standard Wald and t tests and Wald and t CS's. Their behavior under Category I sequences determines whether a test over-rejects asymptotically and whether a CS under-covers asymptotically. Under Category II and III sequences, they

perform asymptotically as desired.

Step 10. We introduce data-dependent critical values that yield Wald and t tests and Wald and t CS's that have correct asymptotic size even in the presence of identification failure and weak identification in part of the parameter space. The adjusted critical values employ the asymptotic formulae derived in Steps 8 and 9.[3]

3. Estimator and Criterion Function

3.1. Extremum Estimators

We consider an estimator $\widehat{\theta}_n$, such as an ML, LS, quantile, GMM, GEL, or MD estimator, that is defined by minimizing a sample criterion function. The sample criterion function, $Q_n(\theta)$, depends on the observations $\{W_i : i \leq n\}$, which may be i.i.d., i.n.i.d., or temporally dependent.

The paper focuses on inference when θ is not identified at some points in the parameter space. Lack of identification occurs when the $Q_n(\theta)$ is flat wrt some sub-vector of θ. To model this identification problem, θ is partitioned into three sub-vectors:

$$\theta = (\beta, \zeta, \pi) = (\psi, \pi), \text{ where } \psi = (\beta, \zeta). \tag{3.1}$$

The parameter $\pi \in R^{d_\pi}$ is unidentified (by the criterion function $Q_n(\theta)$) when $\beta = 0$ $(\in R^{d_\beta})$. The parameter $\psi = (\beta, \zeta) \in R^{d_\psi}$ is always identified. The parameter $\zeta \in R^{d_\zeta}$ does not effect the identification of π. These conditions are stated more precisely in Assumption A below. They allow for a wide range of cases, including cases in which reparametrization is used to convert a model into the framework considered here.

Example 1. This example is a nonlinear regression model estimated by LS. We use it as a running example to illustrate the more general results. The model is

$$Y_i = \beta \cdot h\left(X_i, \pi\right) + Z_i'\zeta + U_i \text{ for } i = 1, ..., n, \tag{3.2}$$

where $h(X_i, \pi) \in R$ is known up to the finite-dimensional parameter $\pi \in R^{d_\pi}$. When the

[3]Steps 1-9 correspond to the following results: Step 1, Lemma 5.1; Step 2, Lemma 10.2; Step 3, Lemmas 10.1 and 10.2; Step 4, Theorem 5.1; Step 5, Lemma 5.3; Step 6, Lemmas 10.3 and 10.4; Step 7, Theorem 5.2; Step 8, Theorems 7.1, 7.2, and 7.3; Step 9, Theorems 7.4 and 7.5; and Step 10, Theorems 7.6 and 7.7.

true value of β is 0, (3.2) becomes a linear model and π is not identified.

Suppose the support of X_i is contained in a set $\mathcal{X}$. We assume here that $h(x,\pi)$ is twice continuously differentiable wrt π, $\forall \pi \in \Pi$, $\forall x \in \mathcal{X}$, although the general theory of the paper allows for non-smooth functions. Let $h_\pi(x,\pi) \in R^{d_\pi}$ and $h_{\pi\pi}(x,\pi) \in R^{d_\pi \times d_\pi}$ denote the first-order and second-order partial derivatives of $h(x,\pi)$ wrt π.

The LS sample criterion function is

$$Q_n(\theta) = n^{-1} \sum_{i=1}^{n} U_i^2(\theta)/2, \text{ where } U_i(\theta) = Y_i - \beta h(X_i,\pi) - Z_i'\zeta. \tag{3.3}$$

When $\beta = 0$, the residual $U_i(\theta)$ and the criterion function $Q_n(\theta)$ do not depend on π. □

The true distribution of the observations $\{W_i : i \leq n\}$ is denoted F_γ for some parameter $\gamma \in \Gamma$. We let P_γ and E_γ denote probability and expectation under F_γ. The parameter space Γ for the true parameter, referred to as the "true parameter space," is compact and is of the form:

$$\Gamma = \{\gamma = (\theta,\phi) : \theta \in \Theta^*, \phi \in \Phi(\theta)\}, \tag{3.4}$$

where Θ^* is a compact subset of R^{d_θ} and $\Phi(\theta) \subset \Phi$ $\forall \theta \in \Theta^*$ for some compact metric space Φ with a metric that induces weak convergence of the bivariate distributions (W_i, W_{i+m}) for all $i, m \geq 1$.[4,5] In unconditional likelihood scenarios, no parameter ϕ appears. In conditional likelihood scenarios, with conditioning variables $\{X_i : i \geq 1\}$, ϕ indexes the distribution of $\{X_i : i \geq 1\}$. In moment condition models, θ is a finite-dimensional parameter that appears in the moment functions and ϕ indexes those aspects of the distribution of the observations that are not determined by θ. In nonlinear regression models estimated by least squares, θ indexes the regression functions and possibly a finite-dimensional feature of the distribution of the errors, such as its variance, and ϕ indexes the remaining characteristics of the distribution of the errors, which may be infinite dimensional.

[4]That is, the metric satisfies: if $\gamma \to \gamma_0$, then (W_i, W_{i+m}) under γ converges in distribution to (W_i, W_{i+m}) under γ_0. Note that Γ is a metric space with metric $d_\Gamma(\gamma_1,\gamma_2) = ||\theta_1 - \theta_2|| + d_\Phi(\phi_1,\phi_2)$, where $\gamma_j = (\theta_j,\phi_j) \in \Gamma$ for $j = 1,2$ and d_Φ is the metric on Φ.

[5]The asymptotic results below give uniformity results over the parameter space Γ. If one is interested in a non-compact parameter space Φ_1 for the parameter ϕ, instead of Φ, then one can apply the results established here to show that the uniformity results hold for all compact subsets Φ of Φ_1 that satisfy the given conditions.

By definition, the extremum estimator $\widehat{\theta}_n$ (approximately) minimizes $Q_n(\theta)$ over an "optimization parameter space" Θ:[6]

$$\widehat{\theta}_n \in \Theta \text{ and } Q_n(\widehat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}). \tag{3.5}$$

We assume that the interior of Θ includes the true parameter space Θ^* (see Assumption B1 below). This ensures that the asymptotic distribution of $\widehat{\theta}_n$ is not effected by boundary constraints for any sequence of true parameters in Θ^*. The focus of this paper is not on the effects of boundary constraints.

3.2. Confidence Sets and Tests

We are interested in the effect of lack of identification or weak identification on the behavior of the extremum estimator $\widehat{\theta}_n$. In addition, we are interested in its effects on CS's for various functions $r(\theta)$ of θ and on tests of null hypotheses of the form $H_0 : r(\theta) = v$.

A CS is obtained by inverting a test. For example, a nominal $1-\alpha$ CS for $r(\theta)$ is

$$CS_n = \{v : T_n(v) \leq c_{n,1-\alpha}(v)\}, \tag{3.6}$$

where $T_n(v)$ is a test statistic and $c_{n,1-\alpha}(v)$ is a critical value for testing $H_0 : r(\theta) = v$. Critical values considered in this paper may depend on the null value v of $r(\theta)$ as well as on the sample size n. The coverage probability of a CS for $r(\theta)$ is

$$P_\gamma(r(\theta) \in CS_n) = P_\gamma(T_n(\theta) \leq c_{n,1-\alpha}(\theta)), \tag{3.7}$$

where $P_\gamma(\cdot)$ denotes the probability when γ is the true value.

The paper focuses on the smallest finite-sample coverage probability of a CS over the parameter space, i.e., the finite-sample size of the CS. It is approximated by the

[6]The $o(n^{-1})$ term in (3.5), and in (5.1) and (5.2) below, is a fixed sequence of constants that does not depend on the true parameter $\gamma \in \Gamma$ and does not depend on π in (5.1). The $o(n^{-1})$ term makes it clear that the infima in these equations need not be achieved exactly. This allows for some numerical inaccuracy in practice and also circumvents the issue of the existence of parameter values that achieve the infima. In contrast to many results in the extremum estimator literature, the $o(n^{-1})$ term is not a random $o_p(n^{-1})$ term here because a quantity is $o_p(n^{-1})$ only for a specific sequence of true distributions and the uniform results given below require properties of the extremum estimators to hold for arbitrary sequences of true distributions.

asymptotic size defined as

$$AsySz = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma\left(r(\theta) \in CS_n\right) = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma\left(T_n\left(\theta\right) \leq c_{n,1-\alpha}\left(\theta\right)\right). \tag{3.8}$$

For a test, we are interested in its null rejection probabilities and in particular its maximum null rejection probability, which is the size of the test. A test's asymptotic size is an approximation to the latter. The null rejection probabilities and asymptotic size of a test are given by

$$\begin{aligned} &P_\gamma\left(T_n\left(\theta\right) > c_{n,1-\alpha}\left(\theta\right)\right) \text{ for } \gamma = (\theta,\phi) \in \Gamma \text{ with } r(\theta) = v \text{ and} \\ &AsySz = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma: r(\theta)=v} P_\gamma\left(T_n\left(v\right) > c_{n,1-\alpha}\left(v\right)\right). \end{aligned} \tag{3.9}$$

3.3. Drifting Sequences of Distributions

In (3.8) and (3.9), the uniformity over $\gamma \in \Gamma$ for any given sample size n is crucial for the asymptotic size to be a good approximation to the finite-sample size. The value of γ at which the finite-sample size of a CS or test is attained may vary with the sample size. Therefore, to determine the asymptotic size we need to derive the asymptotic distribution of the test statistic $T_n(\theta)$ under sequences of true parameters $\gamma_n = (\theta_n, \phi_n)$ that may depend on n.

Similarly, to investigate the finite-sample behavior of the extremum estimator under weak identification, we need to consider its asymptotic behavior under drifting sequences of true distributions—as in Staiger and Stock (1997), Stock and Wright (2000), and numerous other papers that consider weak instruments.

Results in Andrews and Guggenberger (2010) and Andrews, Cheng, and Guggenberger (2009) show that the asymptotic size of CS's and tests are determined by certain drifting sequences of distributions. In this paper, the following sequences $\{\gamma_n\}$ are key:

$$\begin{aligned} \Gamma\left(\gamma_0\right) &= \left\{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \to \gamma_0 \in \Gamma\right\}, \\ \Gamma\left(\gamma_0, 0, b\right) &= \left\{\{\gamma_n\} \in \Gamma\left(\gamma_0\right) : \beta_0 = 0 \text{ and } n^{1/2}\beta_n \to b \in R^{d_\beta}_{[\pm\infty]}\right\}, \text{ and} \\ \Gamma\left(\gamma_0, \infty, \omega_0\right) &= \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : n^{1/2}||\beta_n|| \to \infty \text{ and } \beta_n/||\beta_n|| \to \omega_0 \in R^{d_\beta}\right\}, \end{aligned} \tag{3.10}$$

where $\gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0)$ and $\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n)$. Note that the 0 in $\Gamma(\gamma_0, 0, b)$ and the ∞ in $\Gamma(\gamma_0, \infty, \omega_0)$ stand for different things. In the former, $\beta_0 = 0$, and in the latter

$n^{1/2}||\beta_n|| \to \infty$.

The sequences in $\Gamma(\gamma_0, 0, b)$ are in Categories I and II and are sequences for which $\{\beta_n\}$ is *close* to 0: $\beta_n \to 0$. When $b \in R^{d_\beta}$, $\{\beta_n\}$ is within $O(n^{-1/2})$ of 0 and the sequence is in Category I. The sequences in $\Gamma(\gamma_0, \infty, \omega_0)$ are in Categories II and III and are more *distant* from $\beta = 0$: $n^{1/2}||\beta_n|| \to \infty$. The sets $\Gamma(\gamma_0, 0, b)$ and $\Gamma(\gamma_0, \infty, \omega_0)$ are *not* disjoint. Both contain sequences in Category II.

Throughout the paper we use the terminology: "under $\{\gamma_n\} \in \Gamma(\gamma_0)$" means "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0)$ for any $\gamma_0 \in \Gamma$;" "under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$" means "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for any $\gamma_0 \in \Gamma$ with $\beta_0 = 0$ and any $b \in R^{d_\beta}_{[\pm\infty]}$;" and "under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$" means "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ for any $\gamma_0 \in \Gamma$ and any $\omega_0 \in R^{d_\beta}$ with $||\omega_0|| = 1$."

4. Assumptions

This section provides the high-level conditions under which the results of the paper hold. Various sets of primitive sufficient conditions for the high-level conditions are given for different types of estimators in Section 6 and Appendices A and C. Section 6 considers sample average criterion functions, such as ML and LS, that are smooth functions of θ. Appendix C considers GMM and MD criterion functions. Appendix A contains a variety of other sufficient conditions. Verification of the high-level conditions is illustrated using the running example of LS estimation of a nonlinear regression model. Section 8 verifies the high-level conditions in two other examples [NOT INCLUDED IN THE PAPER CURRENTLY].

4.1. Basic Identification Assumption

The first assumption specifies that θ is not identified (via $Q_n(\theta)$) at some points in the parameter space.

Assumption A. If $\beta = 0$, $Q_n(\theta)$ does not depend on π, $\forall \theta = (\beta, \zeta, \pi) \in \Theta$, $\forall n \geq 1$, for any true parameter $\gamma^* \in \Gamma$.[7]

[7] Assumption A requires the stated condition to hold for all possible realizations of $Q_n(\cdot)$ for any true parameter $\gamma^* \in \Gamma$. Assumption A can be weakened to an a.s. requirement for each $\gamma^* \in \Gamma$, but there seems to be no gain in terms of applications of interest by doing so.

Assumption A specifies that $Q_n(\theta)$ is flat in π when $\beta = 0$. This flatness causes identification failure for π when $\beta = 0$ because $Q_n(\theta)$ cannot distinguish $\theta = (0, \zeta, \pi^*)$ from $\theta' = (0, \zeta, \pi)$ for any $\pi \in \Pi$. The non-identification of π invalidates the standard consistency argument for an extremum estimator based on $Q_n(\theta)$ and causes non-standard asymptotic distributions of extremum estimators and corresponding test statistics. The situation considered in this paper belongs to a broad category of cases where test statistics have discontinuous asymptotic distributions wrt the true parameter value. Here the discontinuity happens when the true value β^* equals 0. It is worth mentioning that the flatness specified in Assumption A does not affect identification of π when $\beta^* \neq 0$ because the extremum estimator $\widehat{\beta}_n$ of β^* is consistent and hence the minimum of the criterion function occurs at values of β where flatness in π does not occur.

Example 1 (cont.). In (3.3), the residual $U_i(\theta)$ does not depend on π when $\beta = 0$. Hence, Assumption A holds for this example. □

4.2. Parameter Space Assumptions

Next, we specify conditions on the parameter spaces Θ and Γ. Without loss of generality (wlog), the optimization parameter space Θ can be written as

$$\begin{aligned} \Theta &= \{\theta = (\psi, \pi) : \psi \in \Psi(\pi), \pi \in \Pi\}, \text{ where} \\ \Pi &= \{\pi : (\psi, \pi) \in \Theta \text{ for some } \psi\} \text{ and} \\ \Psi(\pi) &= \{\psi : (\psi, \pi) \in \Theta\} \text{ for } \pi \in \Pi. \end{aligned} \tag{4.1}$$

We allow $\Psi(\pi)$ to depend on π and, hence, Θ need not be a product space between ψ and π. This is needed in the ARMA(1, 1) example among others.

Define $\Theta_\delta^* = \{\theta \in \Theta^* : ||\beta|| < \delta\}$. The optimization parameter space Θ satisfies:

Assumption B1. (i) $int(\Theta) \supset \Theta^*$,
(ii) for some $\delta > 0$, $\Theta \supset \{\beta \in R^{d_\beta} : ||\beta|| < \delta\} \times \mathcal{Z}^0 \times \Pi \supset \Theta_\delta^*$ for some non-empty open set $\mathcal{Z}^0 \subset R^{d_\zeta}$ and Π as in (4.1), and
(iii) Π is compact.

Because the optimization parameter space is user selected, Assumptions B1(ii)-(iii) can be made to hold by the choice of Θ.[8] Assumption B1(ii) ensures that Θ is compatible with

[8] Assumption B1(iii) is used to show that certain continuous functions on Π introduced in Assump-

(i) a stochastic quadratic approximation of $Q_n(\theta) = Q_n(\psi, \pi)$ wrt ψ around $\psi^* = (0, \zeta^*)$ for each $\pi \in \Pi$, see Assumption C1 below, (ii) the empirical process $\{G_n(\pi) : \pi \in \Pi\}$ defined in Assumption C3 below, and (iii) the definition of $K_n(\theta; \gamma^*)$ in Assumption C5 below.

The true parameter space Γ satisfies:

Assumption B2. (i) Γ is compact and (3.4) holds,
(ii) $\forall \delta > 0$, $\exists \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < ||\beta|| < \delta$, and
(iii) $\forall \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < ||\beta|| < \delta$ for some $\delta > 0$, $\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma$ $\forall a \in [0, 1]$.

Assumption B2(ii) guarantees that Γ is not empty and that there are elements γ of Γ whose β values are non-zero but are arbitrarily close to 0, which is the region of the true parameter space where near lack of identification occurs. Assumption B2(iii) ensures that Γ is compatible with the existence of partial derivatives of certain expectations wrt the true parameter β around $\beta = 0$. These partial derivatives arise in (4.14) and Assumption C5 below.

Example 1 (cont.). In this example, the random variables $\{(X_i, Z_i, U_i) : i = 1, ..., n\}$ are i.i.d. with distribution $F \in \mathcal{F}$, where $\mathcal{F}$ is a compact metric space with some metric that induces weak convergence. The parameter of interest is $\theta = (\beta, \zeta, \pi)$ and the nuisance parameter is $\phi = F$, which is infinite dimensional. The true parameter space for θ is

$$\Theta^* = \mathcal{B}^* \times \mathcal{Z}^* \times \Pi^*, \text{ where } \mathcal{B}^* = [-b_1^*, b_2^*] \subset R \tag{4.2}$$

with $b_1^* \geq 0$, $b_2^* \geq 0$, b_1^* and b_2^* are not both equal to 0, $\mathcal{Z}^*$ ($\subset R^{d_\zeta}$) is compact, and Π^*

tions C6 and C7 below, which have unique minima on Π, satisfy "identifiable uniqueness" properties. Assumption B1(iii) could be avoided by imposing "identifiable uniqueness" properties directly in Assumptions C6 and C7.

($\subset R^{d_\pi}$) is compact. For any $\theta^* \in \Theta^*$, the true parameter space for ϕ is

$$\begin{aligned}
&\Phi(\theta^*) = \{F \in \mathcal{F} : E_F(U_i|X_i, Z_i) = 0 \text{ a.s.},\ E_F(U_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i) > 0 \text{ a.s.},\\
&E_F\left(\sup_{\pi\in\Pi} ||h\left(X_i,\pi\right)||^{4+\varepsilon} + \sup_{\pi\in\Pi} ||h_\pi\left(X_i,\pi\right)||^{4+\varepsilon} + \sup_{\pi\in\Pi} ||h_{\pi\pi}\left(X_i,\pi\right)||^{2+\varepsilon}\right) \leq C,\\
&||h_{\pi\pi}(X_i,\pi_1) - h_{\pi\pi}(X_i,\pi_2)|| \leq M(X_i)||\pi_1 - \pi_2||\ \forall \pi_1,\pi_2 \in \Pi \text{ for some function}\\
&M(X_i),\ E_F M(X_i)^{2+\varepsilon} \leq C,\ E_F|U_i|^{4+\varepsilon} \leq C,\ E_F\left\|Z_i\right\|^{4+\varepsilon} \leq C,\\
&P_F(a'(h(X_i,\pi_1), h\left(X_i,\pi_2\right), Z_i) = 0) < 1,\ \forall \pi_1,\pi_2 \in \Pi \text{ with } \pi_1 \neq \pi_2,\ \forall a \in R^{d_\zeta+2}\\
&\text{with } a \neq 0,\ \lambda_{\min}(E_F(h(X_i,\pi), Z_i')'(h(X_i,\pi), Z_i')) \geq \varepsilon\ \forall \pi \in \Pi,\ \text{and}\\
&\lambda_{\min}(E_F d_i\left(\pi\right) d_i(\pi)') \geq \varepsilon\ \forall \pi \in \Pi\}
\end{aligned} \tag{4.3}$$

for some constants $C < \infty$ and $\varepsilon > 0$, and by definition $d_i(\pi) = (h\left(X_i,\pi\right), Z_i, h_\pi\left(X_i,\pi\right))'$. The moment conditions are needed to ensure the uniform convergence of various sample averages. The other conditions are for the identification of β and ζ and the identification of π when $\beta \neq 0$.

Given the definitions above, the true parameter space Γ is of the form in (3.4) with $\Phi = \mathcal{F}$. Thus, Assumption B2(i) holds immediately. Assumption B2(ii) follows from the form of $\mathcal{B}^*$ given in (4.2). Assumption B2(iii) follows from the form of $\mathcal{B}^*$ and the fact that Θ^* is a product space and $\Phi(\theta^*)$ does not depend on β^*. Hence, the true parameter space Γ satisfies Assumption B2.

The LS estimator of θ minimizes $Q_n(\theta)$ over $\theta \in \Theta$. The optimization parameter space Θ takes the form

$$\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi,\ \text{where } \mathcal{B} = [-b_1, b_2] \subset R \tag{4.4}$$

with $b_1 > b_1^*$, $b_2 > b_2^*$, $\mathcal{Z}$ ($\subset R^{d_\zeta}$) is compact, Π ($\subset R^{d_\pi}$) is compact, $\mathcal{Z}^* \in int(\mathcal{Z})$, and $\mathcal{B}^* \in int(\mathcal{B})$. Given these conditions, Assumptions B1(i) and B1(iii) follow immediately. Assumption B1(ii) holds by taking $\delta < \min\{b_1^*, b_2^*\}$ and $\mathcal{Z}^0 = int(\mathcal{Z})$. □

4.3. Criterion Function Limit Assumption

Here we specify the limit of the sample criterion function $Q_n(\theta)$ along drifting sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0)$ whose limit is $\gamma_0 \in \Gamma$.

Assumption B3. (i) For some non-stochastic real-valued function $Q(\theta;\gamma_0)$ on $\Theta \times \Gamma$,

$$\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta;\gamma_0)| \to_p 0$$

under $\{\gamma_n\} \in \Gamma(\gamma_0)$, $\forall \gamma_0 \in \Gamma$.
(ii) When $\beta_0 = 0$, for every neighborhood Ψ_0 $(\subset R^{d_\psi})$ of $\psi_0 = (\beta_0, \zeta_0)$,

$$\inf_{\pi\in\Pi} \left(\inf_{\psi\in\Psi(\pi)/\Psi_0} Q(\psi,\pi;\gamma_0) - Q(\psi_0,\pi;\gamma_0) \right) > 0, \ \forall \gamma_0 = (\psi_0,\pi_0,\phi_0) \in \Gamma.$$

(iii) When $\beta_0 \neq 0$, for every neighborhood Θ_0 $(\subset \Theta)$ of $\theta_0 = (\beta_0, \zeta_0, \pi_0)$,

$$\inf_{\theta\in\Theta/\Theta_0} Q(\theta;\gamma_0) - Q(\theta_0;\gamma_0) > 0, \ \forall \gamma_0 = (\theta_0,\phi_0) \in \Gamma.$$

Assumption B3(i) defines the (asymptotic) population criterion function $Q(\theta;\gamma_0)$. Assumption B3(ii) provides a condition for the identification of β and ζ despite the non-identification of π when $\beta_0 = 0$. Uniformity over Π is required due to the non-identification of π. A condition of this type also is used in Andrews (1993) for the uniform consistency of a family of estimators. A necessary condition for Assumption B3(ii) is that for any given $\pi \in \Pi$ and $\gamma_0 \in \Gamma$ with $\beta_0 = 0$, $Q(\psi,\pi;\gamma_0)$ is uniquely minimized by ψ_0. Assumption B3(iii) is a standard identification condition for θ when $\beta_0 \neq 0$. A condition of this sort is verified for various extremum estimators in Newey and McFadden (1994).

A set of primitive sufficient conditions for Assumptions B3(ii) and B3(iii) is given in Assumption B3* in Appendix A.

Example 1 (cont.). In this example, the function $Q(\theta;\gamma_0)$ in Assumption B3(i) is

$$Q(\theta;\gamma_0) = E_{F_0} U_i^2/2 + E_{F_0}(\beta_0 h(X_i,\pi_0) + Z_i'\zeta_0 - \beta h(X_i,\pi) - Z_i'\zeta)^2/2, \tag{4.5}$$

where $\gamma_0 = (\beta_0, \zeta_0, \pi_0, F_0)$. The uniform convergence in Assumption B3(i) holds by Lemma 10.7, which in turn uses the pointwise WLLN for a triangular array of row-wise i.i.d. random variables and a uniformity result in Andrews (1992). Assumptions B3(ii) and B3(iii) are verified by verifying the sufficient condition Assumption B3* given in Appendix A. For brevity, the details are given in Appendix B. □

4.4. Close to $\beta = 0$ Assumptions

The following Assumptions C1-C8 are used to determine the asymptotic distributions of estimators and test statistics under sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$ and to establish the consistency of $\widehat{\pi}_n$ under sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| = \infty$. The "C" denotes that the sequences of parameters $\{\gamma_n\}$ considered are *close* to the point of non-identification.

The first assumption, Assumption C1, requires that the criterion function $Q_n(\theta)$ has a stochastic quadratic expansion in ψ around the non-identification point $\psi_{0,n} = (0, \zeta_n)$ uniformly in $\pi \in \Pi$. Assumptions C2, C3, and C8 concern the behavior of the (generalized) first derivative in the expansion. Assumption C4 concerns the behavior of the (generalized) second derivative. Assumptions C5 and C7 arise because the quadratic expansion is about the non-identification point $\psi_{0,n}$, rather than the true value ψ_n. Assumptions C6 and C7 are used when determining the asymptotic behavior of $\widehat{\pi}_n$.

We now define a sequence of scalar constants $\{a_n(\gamma_n) : n \geq 1\}$ that provides the normalization required so that the (generalized) first derivative in the quadratic expansion in Assumption C1 is non-degenerate asymptotically (see Lemma 10.1 in Appendix B). These constants arise in the conditions on the remainder term of the approximation in Assumption C1. Define

$$a_n(\gamma_n) = \begin{cases} n^{1/2} & \text{if } \{\gamma_n\} \in \Gamma(\gamma_0, 0, b) \text{ and } b \in R^{d_\beta} \\ ||\beta_n||^{-1} & \text{if } \{\gamma_n\} \in \Gamma(\gamma_0, 0, b) \text{ and } ||b|| = \infty. \end{cases} \tag{4.6}$$

Note that $||\beta_n||^{-1} < n^{1/2}$ for n large when $||b|| = \infty$, because $n^{1/2}||\beta_n|| \to \infty$.[9] Hence, $a_n(\gamma_n) \leq n^{1/2}$ for n large.

Assumption C1. Under $\{\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n)\} \in \Gamma(\gamma_0, 0, b)$, for some $\delta > 0$, $\forall \theta = (\psi, \pi) \in \Theta_\delta = \{\theta \in \Theta : ||\beta|| < \delta\}$,
(i) the sample criterion function $Q_n(\psi, \pi)$ has a quadratic expansion in ψ around $\psi_{0,n} = (0, \zeta_n)$ for given π:

$$\begin{aligned} Q_n(\psi, \pi) = {} & Q_n(\psi_{0,n}, \pi) + D_\psi Q_n(\psi_{0,n}, \pi)'(\psi - \psi_{0,n}) + \\ & \frac{1}{2}(\psi - \psi_{0,n})' D_{\psi\psi} Q_n(\psi_{0,n}, \pi)(\psi - \psi_{0,n}) + R_n(\psi, \pi), \end{aligned}$$

[9] The nth term $a_n(\gamma_n)$ in the sequence of constants $\{a_n(\gamma_n)\}$ actually depends on the entire sequence $\{\gamma_n\}$ because b depends on $\{\gamma_n\}$. For notational simplicity, however, this is not reflected in the notation $a_n(\gamma_n)$.

where $D_\psi Q_n(\psi_{0,n}, \pi) \in R^{d_\psi}$ is a stochastic generalized first partial-derivative vector and $D_{\psi\psi} Q_n(\psi_{0,n}, \pi) \in R^{d_\psi \times d_\psi}$ is a generalized second partial-derivative matrix that is symmetric and may be stochastic or non-stochastic,
(ii) the remainder, $R_n(\psi, \pi)$, satisfies

$$\sup_{\psi \in \Psi(\pi): ||\psi - \psi_{0,n}|| \leq \delta_n} \frac{|a_n^2(\gamma_n) R_n(\psi, \pi)|}{(1 + ||a_n(\gamma_n)(\psi - \psi_{0,n})||)^2} = o_{p\pi}(1)$$

for all constants $\delta_n \to 0$, and
(iii) $D_\zeta Q_n(\theta)$ and $D_{\zeta\zeta} Q_n(\theta)$ do not depend on π when $\beta = 0$, where $\theta = (\beta, \zeta, \pi) \in \Theta$, $D_\zeta Q_n(\theta)$ denotes the last d_ζ elements of $D_\psi Q_n(\theta)$, and $D_{\zeta\zeta} Q_n(\theta)$ is the lower $d_\zeta \times d_\zeta$ block of $D_{\psi\psi} Q_n(\theta)$.

Sufficient conditions for Assumption C1 when $Q_n(\theta)$ is a sample average that is smooth in θ are given in Lemma 9.5 in Appendix A. In this case, $D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are the pointwise partial and second partial derivatives of $Q_n(\theta)$. For the non-smooth sample average case, sufficient conditions are given in Lemma 9.6 in Appendix A. In this case, $D_\psi Q_n(\theta)$ is a "stochastic derivative" of $Q_n(\theta)$, which typically equals the pointwise derivative for points where the latter exists, and $D_{\psi\psi} Q_n(\theta)$ is the (non-stochastic) second partial derivative of the expected value of $Q_n(\theta)$. For example, this case covers quantile estimators and ML and LS estimators in continuous, but not smooth, threshold autoregressive models, as in Chan and Tsay (1998).

Sufficient conditions for Assumption C1 when $Q_n(\theta)$ is a GMM or MD criterion function, smooth or non-smooth in θ, are given in Appendix C. In the GMM case, $D_\psi Q_n(\theta)$ is the product of two matrices and a vector: (i) the derivative wrt ψ of the expected value of the moment conditions, (ii) the limit of the GMM weight matrix, and (iii) the sample moment vector. The non-stochastic matrix $D_{\psi\psi} Q_n(\theta)$ is the same as $D_\psi Q_n(\theta)$ except the sample moment vector is replaced by the transpose of the matrix in (i).

If $D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are the pointwise partial and second partial derivatives of $Q_n(\theta)$, then Assumption C1(iii) is implied by Assumption A. When $D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are generalized derivatives, then Assumption C1(iii) is not necessarily implied by Assumption A (because generalized derivatives are not uniquely defined), but in the presence of Assumption A the condition is not restrictive.

Example 1 (cont.). The sample criterion function $Q_n(\theta)$ is a sample average:

$$Q_n(\theta) = n^{-1}\sum_{i=1}^{n}\rho(W_i,\theta), \text{ where } \rho(W_i,\theta) = U_i^2(\theta)/2 \text{ and } W_i = (Y_i, X_i, Z_i')'. \tag{4.7}$$

The first- and second-order partial derivatives of $\rho(W_i,\theta)$ wrt to ψ are

$$\begin{aligned}\rho_\psi(W_i,\theta) &= -U_i(\theta)d_{\psi,i}(\pi) \text{ and } \rho_{\psi\psi}(W_i,\theta) = d_{\psi,i}(\pi)d_{\psi,i}(\pi)', \text{ where}\\ d_{\psi,i}(\pi) &= (h(X_i,\pi), Z_i')'.\end{aligned} \tag{4.8}$$

We verify Assumption C1 with

$$D_\psi Q_n(\theta) = -n^{-1}\sum_{i=1}^{n}U_i(\theta)d_{\psi,i}(\pi) \text{ and } D_{\psi\psi}Q_n(\theta) = n^{-1}\sum_{i=1}^{n}d_{\psi,i}(\pi)d_{\psi,i}(\pi)' \tag{4.9}$$

using Lemma 9.5 in Appendix A. For brevity, the verification is given in Appendix B. □

The (generalized) first derivative of $Q_n(\theta)$ wrt ψ is assumed to satisfy:

Assumption C2. (i) $D_\psi Q_n(\theta)$ takes the form

$$D_\psi Q_n(\theta) = n^{-1}\sum_{i=1}^{n}m(W_i,\theta)$$

for some function $m(W_i,\theta) \in R^{d_\psi}\ \forall\theta \in \Theta_\delta$, for any true parameter $\gamma^* \in \Gamma$.
(ii) $E_{\gamma^*}m(W_i,\theta^*) = 0\ \forall\pi \in \Pi, \forall i \geq 1$ when the true parameter is $\gamma^*\ \forall\gamma^* = (\psi^*,\pi^*,\phi^*) \in \Gamma$ with $\beta^* = 0$.[10]

Example 1 (cont). Assumption C2(i) holds in this example with

$$m(W_i,\theta) = -U_i(\theta)d_{\psi,i}(\pi). \tag{4.10}$$

Assumption C2(ii) holds because $E_{\gamma^*}m(W_i,\theta^*) = -E_{\gamma^*}U_i(h(X_i,\pi^*),Z_i')' = 0\ \forall\gamma^* \in \Gamma$. Assumption C2(iii) holds because $E_{\gamma^*}m(W_i,\psi^*,\pi) = -E_{\gamma^*}(U_i+\beta^*h(X_i,\pi^*)-\beta^*h(X_i,\pi))$ $\times\ (h(X_i,\pi),Z_i')' = 0\ \forall\pi \in \Pi$ when $\beta^* = 0$. □

[10] In some time series examples $D_\psi Q_n(\theta)$ is of the form $n^{-1}\sum_{i=1}^{n}m_i(\theta)$, where $m_i(\theta)$ depends on $\{W_j : \forall 1 \leq j \leq i\}$. Assumption C2 can be relaxed to cover such cases without any changes to the results of the paper. In such cases, Assumption C3 below still can hold provided $\{m_i(\theta) : i \leq n\}$ satisfies a suitable "asymptotic weak dependence" condition, such as near epoch dependence.

For simplicity, $m(W_i, \theta)$ is abbreviated as $m_i(\theta)$. Define an empirical process $\{G_n(\pi) : \pi \in \Pi\}$ by

$$G_n(\pi) = n^{-1/2} \sum_{i=1}^{n} \left(m_i(\psi_{0,n}, \pi) - E_{\gamma_n} m_i(\psi_{0,n}, \pi) \right). \tag{4.11}$$

The recentered and rescaled (generalized) first derivative of $Q_n(\theta)$ wrt ψ is assumed to satisfy an empirical process CLT:

Assumption C3. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $G_n(\cdot) \Rightarrow G(\cdot; \gamma_0)$, where $G(\cdot; \gamma_0)$ is a mean zero Gaussian process indexed by $\pi \in \Pi$ with bounded continuous sample paths and some covariance kernel $\Omega(\pi_1, \pi_2; \gamma_0)$ for $\pi_1, \pi_2 \in \Pi$.

Numerous empirical process results in the literature can be used to verify this assumption, including results in Pollard (1984, 1990), Andrews (1994), and van der Vaart and Wellner (1996).

Example 1 (cont). To verify Assumption C3, we have

$$\begin{aligned} U_i(\psi_{0,n}, \pi) &= Y_i - Z_i'\zeta_n = U_i + \beta_n h(X_i, \pi_n) \text{ and} \\ G_n(\pi) &= -n^{-1/2} \sum_{i=1}^{n} (U_i d_{\psi,i}(\pi) + \beta_n [h(X_i, \pi_n) d_{\psi,i}(\pi) - E_{\gamma_n} h(X_i, \pi_n) d_{\psi,i}(\pi)]). \end{aligned} \tag{4.12}$$

Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $G_n(\pi) \Rightarrow G(\pi; \gamma_0)$, where $G(\pi; \gamma_0)$ is a Gaussian process with bounded continuous sample paths and covariance kernel $\Omega(\pi_1, \pi_2; \gamma_0) = E_{F_0} U_i^2 d_{\psi,i}(\pi_1)$ $d_{\psi,i}(\pi_2)'$. This weak convergence follows from Andrews (1994, p. 2251) because (i) Π is compact, (ii) the finite-dimensional convergence holds by the CLT for a triangular array of row-wise i.i.d. random variables, where the Lindeberg condition holds by the $L^{2+\delta}$-boundedness of its summands, and $\beta_n \to 0$, and (iii) the stochastic equicontinuity (SE) holds by applying the type II class (Lipschitz functions) using the differentiability of $h(x, \pi)$ in π. □

The (generalized) second derivative of $Q_n(\theta)$ wrt ψ is assumed to satisfy:

Assumption C4. (i) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\sup_{\pi \in \Pi} ||D_{\psi\psi} Q_n(\psi_{0,n}, \pi) - H(\pi; \gamma_0)|| \to_p 0$ for some non-stochastic symmetric $d_\psi \times d_\psi$-matrix-valued function $H(\pi; \gamma_0)$ on $\Pi \times \Gamma$ that is continuous on Π $\forall \gamma_0 \in \Gamma$, and
(ii) $0 < \inf_{\pi \in \Pi} \lambda_{\min}(H(\pi; \gamma_0)) \leq \sup_{\pi \in \Pi} \lambda_{\max}(H(\pi; \gamma_0)) < \infty$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

Example 1 (cont). Assumption C4(i) holds in this example with

$$H(\pi;\gamma_0) = E_{F_0} d_{\psi,i}(\pi) d_{\psi,i}(\pi)' \tag{4.13}$$

by applying the uniform LLN for drifting true distributions given in Lemma 10.7 in Appendix B to $n^{-1}\sum_{i=1}^{n} d_{\psi,i}(\pi)d_{\psi,i}(\pi)$. The continuity of $H(\pi;\gamma_0)$ is implied by the continuity of $h(X_i,\pi)$ in π, $E_{F_0}\sup_{\pi\in\Pi}||d_{\psi,i}(\pi)d_{\psi,i}(\pi)'|| < \infty$, and the dominated convergence theorem (DCT). Assumption C4(ii) follows immediately from the conditions in (4.3). □

Define the $d_\psi \times d_\beta$-matrix of partial derivatives of the average population moment function wrt the true β value, β^*, to be

$$K_n(\theta;\gamma^*) = n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial \beta^{*\prime}} E_{\gamma^*} m(W_i,\theta). \tag{4.14}$$

The domain of the function $K_n(\theta;\gamma^*)$ is $\Theta_\delta \times \Gamma_0$, where $\Gamma_0 = \{\gamma_a = (a\beta,\zeta,\pi,\phi) \in \Gamma :$ $\gamma = (\beta,\zeta,\pi,\phi) \in \Gamma$ with $||\beta|| < \delta$ and $a \in [0,1]\}$ and $\delta > 0$ is as in Assumption B2(iii). The set Γ_0 is not empty by Assumption B2(ii).

Assumption C5. (i) $K_n(\theta;\gamma^*)$ exists $\forall(\theta,\gamma^*) \in \Theta_\delta \times \Gamma_0$, $\forall n \geq 1$,
(ii) for some non-stochastic $d_\psi \times d_\beta$-matrix-valued function $K(\psi_0,\pi;\gamma_0)$, $K_n(\psi_n,\pi;\widetilde{\gamma}_n) \to K(\psi_0,\pi;\gamma_0)$ uniformly over $\pi \in \Pi$ for all non-stochastic sequences $\{\psi_n\}$ and $\{\widetilde{\gamma}_n\}$ such that $\widetilde{\gamma}_n \in \Gamma$, $\widetilde{\gamma}_n \to \gamma_0 = (0,\zeta_0,\pi_0,\phi_0)$ for some $\gamma_0 \in \Gamma$, $(\psi_n,\pi) \in \Theta$, and $\psi_n \to \psi_0 = (0,\zeta_0)$, and
(iii) $K(\psi_0,\pi;\gamma_0)$ is continuous on Π $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.

Assumption C5 is not restrictive. A set of primitive sufficient conditions for Assumption C5 is given in Appendix A.

For simplicity, $K(\psi_0,\pi;\gamma_0)$ is abbreviated as $K(\pi;\gamma_0)$. Note that $(\psi_n,\widetilde{\gamma}_n)$ in Assumption C5(ii) is in $\Theta_\delta \times \Gamma_0$ for n large.

Example 1 (cont). To verify Assumption C5(i), we have

$$\begin{aligned} K_n(\theta,\gamma^*) &= \frac{\partial}{\partial\beta^*} E_{F^*} m(W_i,\theta) = -\frac{\partial}{\partial\beta^*} E_{F^*}(Y_i - \beta h(X_i,\pi) - Z_i'\zeta) d_{\psi,i}(\pi) \\ &= -\frac{\partial}{\partial\beta^*} E_{F^*}(U_i + \beta^* h(X_i,\pi^*) - \beta h(X_i,\pi) - Z_i'(\zeta - \zeta^*)) d_{\psi,i}(\pi) \\ &= -E_{F^*} h(X_i,\pi^*) d_{\psi,i}(\pi). \end{aligned} \tag{4.15}$$

Assumptions C5(ii) and C5(iii) hold with

$$K(\pi;\gamma_0) = K(\psi_0,\pi;\gamma_0) = -E_{F_0} h(X_i,\pi_0) d_{\psi,i}(\pi), \tag{4.16}$$

see Appendix B for the proof. □

Next, we introduce the limits of the concentrated criterion function after suitable normalization. Define a "weighted non-central chi-square" process $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ and a non-stochastic function $\eta(\pi;\gamma_0,\omega_0)$ by

$$\begin{aligned} \xi(\pi;\gamma_0,b) &= -\frac{1}{2}\left(G(\pi;\gamma_0) + K(\pi;\gamma_0)\, b\right)' H^{-1}(\pi;\gamma_0)\left(G(\pi;\gamma_0) + K(\pi;\gamma_0) b\right) \text{ and} \\ \eta(\pi;\gamma_0,\omega_0) &= -\omega_0' K(\pi;\gamma_0)' H^{-1}(\pi;\gamma_0) K(\pi;\gamma_0)\omega_0. \end{aligned} \tag{4.17}$$

The process $\xi(\pi;\gamma_0,b)$ is the limit under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ for $b \in R^{d_\beta}$ and the function $\eta(\pi;\gamma_0,b)$ is the limit under $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$. Under Assumptions C3, C4, and C5(iii), $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ has bounded continuous sample paths a.s.

To obtain the asymptotic distribution of $\widehat{\pi}_n$ when $\beta_n = O(n^{-1/2})$ via the continuous mapping theorem, we use the following assumption.

Assumption C6. Each sample path of the stochastic process $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ in some set $A(\gamma_0,b)$ with $P_{\gamma_0}(A(\gamma_0,b)) = 1$ is minimized over Π at a unique point (which may depend on the sample path), denoted $\pi^*(\gamma_0,b)$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$, $\forall b \in R^{d_\beta}$.

In Assumption C6, $\pi^*(\gamma_0,b)$ is random.

We now provide a primitive sufficient condition for Assumption C6 for the case when β is a scalar, i.e., $d_\beta = 1$, which covers many cases of interest. Assumptions C1(iii) and C2 and (4.11) imply that $G(\pi;\gamma_0)$ can be partitioned as $(G_1(\pi)', G_2')'$, where $G_1(\pi) \in R^{d_\beta}$, $G_2 \in R^{d_\zeta}$, and G_2 does not depend on π. We partition the covariance kernel $\Omega(\pi_1,\pi_2;\gamma_0)$ in Assumption C3 analogously to $G(\pi;\gamma_0)$ and obtain

$$\Omega(\pi_1,\pi_2;\gamma_0) = \begin{bmatrix} \Omega_{11}(\pi_1,\pi_2;\gamma_0) & \Omega_{12}(\pi_1;\gamma_0) \\ \Omega_{12}(\pi_2;\gamma_0)' & \Omega_{22}(\gamma_0) \end{bmatrix}, \tag{4.18}$$

where $\Omega_{22}(\gamma_0) \in R^{d_\zeta \times d_\zeta}$ does not depend on π. For any $\pi_1, \pi_2 \in \Pi$ and $\pi_1 \neq \pi_2$,

$(G_1(\pi_1), G_1(\pi_2), G_2)'$ is normally distributed with mean zero and covariance matrix

$$\Omega_G(\pi_1, \pi_2; \gamma_0) = \begin{bmatrix} \Omega_{11}(\pi_1, \pi_1; \gamma_0) & \Omega_{11}(\pi_1, \pi_2; \gamma_0) & \Omega_{12}(\pi_1; \gamma_0) \\ \Omega_{11}(\pi_2, \pi_1; \gamma_0) & \Omega_{11}(\pi_2, \pi_2; \gamma_0) & \Omega_{12}(\pi_2; \gamma_0) \\ \Omega_{12}(\pi_1; \gamma_0)' & \Omega_{12}(\pi_2; \gamma_0)' & \Omega_{22}(\gamma_0) \end{bmatrix}. \tag{4.19}$$

Typically, the covariance matrix $\Omega_G(\pi_1, \pi_2; \gamma_0)$ takes the form of an outer product, which facilitates the verification of Assumption C6**, as shown in the examples.

Assumption C6.** (i) $d_\beta = 1$ (i.e., β is a scalar).
(ii) $\Omega_G(\pi_1, \pi_2; \gamma_0)$ is positive definite, $\forall \pi_1, \pi_2 \in \Pi$ with $\pi_1 \neq \pi_2$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

Lemma 4.1. *Assumption* C6** *implies Assumption* C6.

Comment. A slightly more general sufficient condition, Assumption C6*, for Assumption C6 is given in Appendix A.

Example 1 (cont). We verify Assumption C6**. Assumption C6**(i) holds because β is a scalar. By the discussion following (4.12), $a'(G_1(\pi_1), G_1(\pi_2), G_2)$ has variance $E_{F_0} U_i^2 d_a^2(\pi_1, \pi_2)$, where $d_a(\pi_1, \pi_2) = a'(h(X_i, \pi_1), h(X_i, \pi_2), Z_i)$. By the conditions in (4.3), $P_{F_0}(d_a(\pi_1, \pi_2) = 0) < 1$ $\forall a \in R^{d_\zeta + 2}$ with $a \neq 0$, $\forall \pi_1 \neq \pi_2$, $\forall F_0 \in \Phi(\theta_0)$, and $E_{F_0}(U_i^2 | X_i, Z_i) > 0$ a.s. Hence, $E_{F_0} U_i^2 d_a^2(\pi_1, \pi_2) > 0$ $\forall a \neq 0$ and Assumption C6**(ii) holds. □

The following assumption is used in the proof of consistency of $\widehat{\pi}_n$ in the "distant local to $\beta = 0$" case in which $\beta_n \to 0$ and $n^{1/2} ||\beta_n|| \to \infty$.

Assumption C7. The non-stochastic function $\eta(\pi; \gamma_0, \omega_0)$ is uniquely minimized over $\pi \in \Pi$ at π_0 $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

In Assumption C7, π_0 is non-random. Assumption C7 can be verified using the Cauchy-Schwarz inequality or a matrix version of it, see Tripathi (1999), when $K(\pi; \gamma_0)$ and $H(\pi; \gamma_0)$ take proper forms, as in our examples. [ADD REFERENCE TO SUFFICIENT CONDITIONS AS. C7 FOR ML ESTIMATORS IN APPENDIX A AFTER THAT HAS BEEN ADDED.]

Lemma 10.3 in Appendix B shows that when $\pi = \pi_0$, $K(\pi; \gamma_0) = -H(\pi; \gamma_0) S_\beta'$, where $S_\beta = [I_{d_\beta} : 0] \in R^{d_\beta \times d_\psi}$, whereas this relationship does not hold for $\pi \neq \pi_0$ in general.

Example 1 (cont). We verify Assumption C7 as follows. Given the form of $H(\pi;\gamma_0)$ and $K(\pi;\gamma_0)$ in (4.13) and (4.16), respectively, we have

$$\begin{aligned} &K(\pi;\gamma_0)'H^{-1}(\pi;\gamma_0)K(\pi;\gamma_0) \qquad (4.20)\\ &= [E_{F_0}h(X_i,\pi_0)d_{\psi,i}(\pi)]'[E_{F_0}d_{\psi,i}(\pi)d_{\psi,i}(\pi)']^{-1}[E_{F_0}d_{\psi,i}(\pi)h(X_i,\pi_0)] \leq E_{F_0}h^2(X_i,\pi_0), \end{aligned}$$

where the inequality holds by the matrix Cauchy-Schwarz inequality in Tripathi (1999). The "$\leq$" holds as an equality if and only if $h(X_i,\pi_0)a + d_{\psi,i}(\pi)'b = 0$ with probability 1 for some $a \in R$, $b \in R^{d_\zeta+1}$, and $(a,b') \neq 0$. The "$\leq$" holds as an equality uniquely at $\pi = \pi_0$ because for any $\pi \neq \pi_0$, $P_{F_0}(c'(h(X_i,\pi_0), h(X_i,\pi), Z_i) = 0) < 1$ for any $c \neq 0$ by (4.3). This completes the verification of Assumption C7. □

The following technical assumption is used when obtaining a rate of convergence result for $\widehat{\psi}_n$ for sequences $\{\gamma_n\}$ for which $\beta_n \to 0$ and $n^{1/2}||\beta_n|| \to \infty$.

Assumption C8. Under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$, $\frac{\partial}{\partial\psi'}E_{\gamma_n}D_\psi Q_n(\psi,\pi_n)|_{\psi=\psi_n} \to H(\pi_0;\gamma_0)$.

By Assumption C4(i), $H(\pi;\gamma_0)$ is the probability limit of $D_{\psi\psi}Q_n(\psi_{0,n},\pi_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$. When $Q_n(\theta)$ is a twice differentiable sample average, $D_\psi Q_n(\theta)$ and $D_{\psi\psi}Q_n(\theta)$ are its first and second-order partial derivatives wrt ψ, respectively. One can switch E and ∂ under certain regularity conditions, so that $(\partial/\partial\psi')E_{\gamma_n}D_\psi Q_n(\psi_n,\pi_n)$ is the expectation of $D_{\psi\psi}Q_n(\psi_n,\pi_n)$ in this case. Hence, Assumption C8 can be verified by a uniform LLN and the continuity of $D_{\psi\psi}Q_n(\psi,\pi)$ in ψ. When $Q_n(\theta)$ is non-smooth, one can show that $E_{\gamma_n}D_\psi Q_n(\theta)$ is close to the first-order partial derivative of $Q(\theta;\gamma_0)$ wrt ψ, roughly by switching E_{γ_n} and D_ψ under some regularity conditions, and $D_{\psi\psi}Q_n(\theta)$ is typically taken to be the second-order partial derivative of $Q(\theta;\gamma_0)$ wrt ψ in this case.

Example 1 (cont). To verify Assumption C8, we have

$$(\partial/\partial\psi')E_{\gamma_n}D_\psi Q_n(\psi,\pi_n)|_{\psi=\psi_n} = E_{F_n}d_{\psi,i}(\pi_n)d_{\psi,i}(\pi_n)' \qquad (4.21)$$

by the form of $D_\psi Q_n(\theta_n)$ given in (4.9). Assumption C8 holds provided $E_{F_n}d_{\psi,i}(\pi)d_{\psi,i}(\pi)'$ converges to $E_{F_0}d_{\psi,i}(\pi)d_{\psi,i}(\pi)'$ uniformly over $\pi \in \Pi$ and $E_{F_0}d_{\psi,i}(\pi)d_{\psi,i}(\pi)'$ is continuous in π. This holds by the same argument as in the verification of Assumption C5, see Appendix B. The smoothness and moment conditions are satisfied by the conditions in (4.3). □

4.5. Distant from $\beta = 0$ Assumptions

Assumptions D1-D3 below are used to derive asymptotic distributions under sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The "D" denotes that the sequences of true parameters considered are more *distant* from the point of non-identification than are the sequences in the "C" assumptions.

Next we define a matrix $B(\beta)$ that is used to normalize the (generalized) second-derivative matrix $D^2Q_n(\theta_n)$ of $Q_n(\theta_n)$ (which is introduced in Assumption D1 below) so that it is nonsingular asymptotically, as specified in Assumption D2. Let

$$\begin{aligned} B(\beta) &= \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & \iota(\beta) I_{d_\pi} \end{bmatrix} \in R^{d_\theta \times d_\theta}, \text{ where} \\ \iota(\beta) &= \begin{cases} \beta & \text{if } \beta \text{ is a scalar} \\ ||\beta|| & \text{if } \beta \text{ is a vector} \end{cases}. \end{aligned} \tag{4.22}$$

We use a different definition of $B(\beta)$ in the scalar and vector β cases because in the scalar case the use of β, rather than $||\beta||$, produces noticeably simpler (but equivalent) formulae, but in the vector case $||\beta||$ is required.

Assumption D1. When the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,
(i) the sample criterion function $Q_n(\theta)$ has a quadratic expansion in θ around θ_n:

$$Q_n(\theta) = Q_n(\theta_n) + DQ_n(\theta_n)'(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)D^2Q_n(\theta_n)(\theta - \theta_n) + R_n^*(\theta),$$

where $DQ_n(\theta_n) \in R^{d_\theta}$ is a stochastic generalized first derivative vector and $D^2Q_n(\theta_n) \in R^{d_\theta \times d_\theta}$ is a generalized second derivative matrix that is symmetric and may be stochastic or non-stochastic, and
(ii) the remainder, $R_n^*(\theta)$, satisfies

$$\sup_{\theta \in \Theta_n(\delta_n)} \frac{|nR_n^*(\theta)|}{(1 + ||n^{1/2}B(\beta_n)(\theta - \theta_n)||)^2} = o_p(1)$$

for all $\delta_n \to 0$, where $\Theta_n(\delta_n) = \{\theta \in \Theta : \|\psi - \psi_n\| \leq \delta_n \|\beta_n\| \text{ and } \|\pi - \pi_n\| \leq \delta_n\}$.

The set $\Theta_n(\delta_n)$ in Assumption D1(ii) is a neighborhood of θ_n whose radius shrinks as the sample size gets larger. In particular, the distance between ψ and ψ_n shrinks faster than $||\beta_n||$ when $\beta_n \to 0$. It is shown below that, under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\widehat{\theta}_n \in \Theta_n(\delta_n)$ with probability that goes to one as $n \to \infty$ for some $\delta_n \to 0$. (This holds

because $\widehat{\theta}_n$ is consistent by Lemma 5.3 below and $\widehat{\psi}_n - \psi_n = o_p(\|\beta_n\|)$ when $\beta_n \to 0$ by Lemma 10.4 in Appendix B.)

The sufficient conditions for Assumption C1 referenced in the previous sub-section also are sufficient for Assumption D1. The quantities $DQ_n(\theta_n)$ and $D^2Q_n(\theta_n)$ take similar forms to $D_\psi Q_n(\psi_{0,n}, \pi)$ and $D_{\psi\psi} Q_n(\psi_{0,n}, \pi)$ (see the discussion following Assumption C1), but involve derivatives wrt θ, not ψ, and hence are not functions of π.

Example 1 (cont). The first- and second-order partial derivatives of $\rho(W_i, \theta)$ wrt to θ are

$$\begin{aligned}
\rho_\theta(W_i, \theta) &= -U_i(\theta) B(\beta) d_i(\pi) \text{ and} \\
\rho_{\theta\theta}(W_i, \theta) &= -U_i(\theta) D_i(\theta) + B(\beta) d_i(\pi) d_i(\pi)' B(\beta), \text{ where} \\
d_i(\pi) &= (h(X_i, \pi), Z_i', h_\pi(X_i, \pi)')', \\
D_i(\theta) &= \begin{bmatrix} 0 & 0_{1\times d_\zeta} & h_\pi(X_i, \pi)' \\ 0_{d_\zeta \times 1} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ h_\pi(X_i, \pi) & 0_{d_\pi \times d_\zeta} & h_{\pi\pi}(X_i, \pi)\beta \end{bmatrix},
\end{aligned} \tag{4.23}$$

and $B(\beta)$ depends on β, not $||\beta||$, because β is a scalar. Assumption D1 holds with

$$\begin{aligned}
DQ_n(\theta) &= -n^{-1} \sum_{i=1}^n U_i(\theta) B(\beta) d_i(\pi) \text{ and} \\
D^2 Q_n(\theta) &= n^{-1} \sum_{i=1}^n \left(B(\beta) d_i(\pi) d_i(\pi)' B(\beta) - U_i(\theta) D_i(\theta) \right)
\end{aligned} \tag{4.24}$$

by Lemma 9.5 in Appendix A. The verification is given in Appendix B.[11] □

The next assumption requires good behavior of the (generalized) second derivative of $Q_n(\theta_n)$ after it has been rescaled to eliminate its singularity when β_n converges to zero.

Assumption D2. Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$J_n = B^{-1}(\beta_n) D^2 Q_n(\theta_n) B^{-1}(\beta_n) \to_p J(\gamma_0) \in R^{d_\theta \times d_\theta},$$

[11] This example illustrates why defining $B(\beta)$ using β, not $||\beta||$, is preferred in the scalar β case. If $B(\beta)$ is defined with $||\beta||$ in place of β, then $d_i(\pi)$ needs to be replaced by $d_i(\beta, \pi) = (h(X_i, \pi), Z_i', sgn(\beta) h_\pi(X_i, \pi)')'$. The appearance of $sgn(\beta)$ complicates matters because it introduces a dependence of $d_i(\beta, \pi)$ on β, which otherwise does not appear, and it is a discontinuous function of β.

where $J(\gamma_0)$ is nonsingular and symmetric, $\forall \gamma_0 \in \Gamma$.[12]

Example 1 (cont). To verify Assumption D2 with $D^2 Q_n(\theta)$ given in (4.24), we have

$$J_n = n^{-1} \sum_{i=1}^{n} d_i(\pi_n) d_i(\pi_n)' - \tag{4.25}$$
$$(n^{1/2}\beta_n)^{-1} \begin{bmatrix} 0 & 0_{1\times d_\zeta} & n^{-1/2} \sum_{i=1}^n U_i h_\pi(X_i, \pi_n)' \\ 0_{d_\zeta \times 1} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ n^{-1/2} \sum_{i=1}^n U_i h_\pi(X_i, \pi_n) & 0_{d_\pi \times d_\zeta} & n^{-1/2} \sum_{i=1}^n U_i h_{\pi\pi}(X_i, \pi) \end{bmatrix}.$$

Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $n^{-1} \sum_{i=1}^n d_i(\pi_n) d_i(\pi_n)' \to_p E_{\gamma_0} d_i(\pi_0) d_i(\pi_0)'$ because $n^{-1} \sum_{i=1}^n d_i(\pi) d_i(\pi)' \to_p E_{\gamma_0} d_i(\pi) d_i(\pi)'$ uniformly over $\pi \in \Pi$ by Lemma 10.7 in Appendix B and the continuity of $E_{\gamma_0} d_i(\pi) d_i(\pi)'$ in π. The second line of (4.25) is $o_p(1)$ because $n^{1/2}|\beta_n| \to \infty$, $n^{-1/2} \sum_{i=1}^n U_i h_\pi(X_i, \pi_n)' = O_p(1)$, and $n^{-1/2} \sum_{i=1}^n U_i h_{\pi\pi}(X_i, \pi_n) = O_p(1)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The latter two terms are $O_p(1)$ by the CLT for a triangular array of row-wise i.i.d. random variables under the moment conditions in (4.3). Hence, Assumption D2 holds with the matrix

$$J(\gamma_0) = E_{\gamma_0} d_i(\pi_0) d_i(\pi_0)', \tag{4.26}$$

which is nonsingular by the conditions in (4.3).

The following assumption requires the rescaled (generalized) first derivative to satisfy a CLT.

Assumption D3. (i) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$n^{1/2} B^{-1}(\beta_n) D Q_n(\theta_n) \to_d G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0)),$$

for some symmetric $d_\theta \times d_\theta$-matrix $V(\gamma_0)$.[13]
(ii) $V(\gamma_0)$ is positive definite $\forall \gamma_0 \in \Gamma$.

Example 1 (cont). To verify Assumption D3 in this example, we have

$$n^{1/2} B^{-1}(\beta_n) D Q_n(\theta_n) = -n^{-1/2} \sum_{i=1}^{n} U_i d_i(\pi_n) \to_d N(0_{d_\theta}, V(\gamma_0)), \text{ where}$$
$$V(\gamma_0) = E_{F_0} U_i^2 d_i(\pi_0) d_i(\pi_0)'. \tag{4.27}$$

[12] In the vector β case, $J(\gamma_0)$ may depend on ω_0 as well as γ_0.
[13] In the vector β case, $V(\gamma_0)$ may depend on ω_0 as well as γ_0.

The convergence in distribution holds by the CLT for a triangular array of row-wise i.i.d. random variables. Assumption D3(ii) holds because $E_{F_0} d_i(\pi_0) d_i(\pi_0)'$ is non-singular and $E_{F_0}(U_i^2|X_i, Z_i) > 0$ a.s. by (4.3). □

5. Estimation Results

This section provides the asymptotic results of the paper for the extremum estimator $\widehat{\theta}_n$. Define a concentrated extremum estimator $\widehat{\psi}_n(\pi)$ $(\in \Psi(\pi))$ of ψ for given $\pi \in \Pi$ by

$$Q_n(\widehat{\psi}_n(\pi), \pi) = \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o(n^{-1}). \tag{5.1}$$

Let $Q_n^c(\pi)$ denote the concentrated sample criterion function $Q_n(\widehat{\psi}_n(\pi), \pi)$. Define an extremum estimator $\widehat{\pi}_n$ $(\in \Pi)$ by

$$Q_n^c(\widehat{\pi}_n) = \inf_{\pi \in \Pi} Q_n^c(\pi) + o(n^{-1}). \tag{5.2}$$

We assume that the extremum estimator $\widehat{\theta}_n$ in (3.5) can be written as $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n), \widehat{\pi}_n)$. Note that if (5.1) and (5.2) hold and $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n), \widehat{\pi}_n)$, then (3.5) automatically holds.

Lemma 5.1. *Suppose Assumptions* A *and* B3 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0)$, *where* $\gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0)$,
(a) *when* $\beta_0 = 0$, $\sup_{\pi \in \Pi} ||\widehat{\psi}_n(\pi) - \psi_n|| \to_p 0$, *and*
(b) *when* $\beta_0 \neq 0$, $\widehat{\theta}_n - \theta_n \to_p 0$.

For $\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n) \in \Gamma$, let $Q_{0,n} = Q_n(\psi_{0,n}, \pi)$, where $\psi_{0,n} = (0, \zeta_n)$ as in Assumption C1. Note that $Q_{0,n}$ does not depend on π by Assumption A.

Lemma 5.2. *Suppose Assumptions* A, B1-B3, *and* C1-C5 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$,
(a) *when* $b \in R^{d_\beta}$, $n(Q_n^c(\cdot) - Q_{0,n}) \Rightarrow \xi(\cdot; \gamma_0, b)$,
(b) *when* $||b|| = \infty$ *and* $\beta_n / ||\beta_n|| \to \omega_0$ *for some* $\omega_0 \in R^{d_\beta}$ *with* $||\omega_0|| = 1$,
$||\beta_n||^{-2}(Q_n^c(\pi) - Q_{0,n}) \to_p \eta(\pi; \gamma_0, \omega_0)$ *uniformly over* $\pi \in \Pi$.

Define the Gaussian process $\{\tau(\pi; \gamma_0, b) : \pi \in \Pi\}$ by

$$\tau(\pi; \gamma_0, b) = -H^{-1}(\pi; \gamma_0)(G(\pi; \gamma_0) + K(\pi; \gamma_0) b) - (b, 0_{d_\zeta}), \tag{5.3}$$

where $(b, 0_{d_\zeta}) \in R^{d_\psi}$. Note that, by (4.17) and (5.3), $\xi(\pi; \gamma_0, b) = -(1/2)(\tau(\pi; \gamma_0, b) + (b, 0_{d_\zeta}))' H(\pi; \gamma_0)(\tau(\pi; \gamma_0, b) + (b, 0_{d_\zeta}))$. Let

$$\pi^*(\gamma_0, b) = \underset{\pi \in \Pi}{\arg\min}\ \xi(\pi; \gamma_0, b). \tag{5.4}$$

Theorem 5.1. *Suppose Assumptions* A, B1-B3, *and* C1-C6 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $b \in R^{d_\beta}$,
(a) $\begin{pmatrix} n^{1/2}(\widehat{\psi}_n - \psi_n) \\ \widehat{\pi}_n \end{pmatrix} \to_d \begin{pmatrix} \tau(\pi^*(\gamma_0, b); \gamma_0, b) \\ \pi^*(\gamma_0, b) \end{pmatrix}$, *and*
(b) $n\left(Q_n(\widehat{\theta}_n) - Q_{0,n}\right) \to_d \xi(\pi^*(\gamma_0, b); \gamma_0, b)$.

Comment. Define the Gaussian process $\{\tau_\beta(\pi; \gamma_0, b) : \pi \in \Pi\}$ by

$$\tau_\beta(\pi; \gamma_0, b) = S_\beta \tau(\pi; \gamma_0, b) + b, \tag{5.5}$$

where $S_\beta = [I_{d_\beta} : 0_{d_\beta \times d_\zeta}]$ is the $d_\beta \times d_\psi$ selector matrix that selects β out of ψ. The asymptotic distribution of $n^{1/2}\widehat{\beta}_n$ (without centering at β_n) under $\Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$ is given by $\tau_\beta(\pi^*(\gamma_0, b); \gamma_0, b)$. This quantity appears in the asymptotic distributions of Wald and t statistics below.

Lemma 5.3. *Suppose Assumptions* A, B1-B3, *and* C1-C7 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,
(a) $\widehat{\pi}_n - \pi_n \to_p 0$ *and* (b) $\widehat{\psi}_n - \psi_n \to_p 0$.

Theorem 5.2. *Suppose Assumptions* A, B1-B3, C1-C8, *and* D1-D3 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,
(a) $n^{1/2} B(\beta_n)(\widehat{\theta}_n - \theta_n) \to_d J^{-1}(\gamma_0) G^*(\gamma_0) \sim N(0_{d_\theta}, J^{-1}(\gamma_0) V(\gamma_0) J^{-1}(\gamma_0))$, *and*
(b) $n(Q_n(\widehat{\theta}_n) - Q_n(\theta_n)) \to_d -\frac{1}{2} G^*(\gamma_0)' J^{-1}(\gamma_0) G^*(\gamma_0)$.

Example 1 (cont.). In this example, the components of the stochastic processes $\tau(\pi; \gamma_0, b)$ and $\xi(\pi; \gamma_0, b)$, the function $\eta(\pi; \gamma_0, \omega_0)$, and matrices $J(\gamma_0)$ and $V(\gamma_0)$ that

appear in the asymptotic results above are

$$\begin{aligned} H(\pi;\gamma_0) &= E_{F_0} d_{\psi,i}(\pi) d_{\psi,i}(\pi)', \\ K(\pi;\gamma_0) &= -E_{F_0} h(X_i,\pi_0) d_{\psi,i}(\pi), \\ \Omega(\pi_1,\pi_2;\gamma_0) &= E_{F_0} U_i^2 d_{\psi,i}(\pi_1) d_{\psi,i}(\pi_2)', \\ J(\gamma_0) &= E_{\gamma_0} d_i(\pi_0) d_i(\pi_0)', \\ V(\gamma_0) &= E_{F_0} U_i^2 d_i(\pi_0) d_i(\pi_0)', \text{ where} \\ d_{\psi,i}(\pi) &= (h(X_i,\pi), Z_i')', \ d_i(\pi) = \left(h\left(X_i,\pi\right), Z_i, h_\pi\left(X_i,\pi\right)\right)', \end{aligned} \tag{5.6}$$

and $G(\pi;\gamma_0)$ is a Gaussian process with covariance kernel $\Omega(\pi_1,\pi_2;\gamma_0)$. □

6. Smooth Sample Average Criterion Function

This section provides relatively primitive sufficient conditions for most of the high-level assumptions given in Section 4 for the class of sample average criterion functions that are smooth in θ. This includes ML and LS estimators. Note that the high-level assumptions in Section 4 concern limit behavior under drifting sequences of true distributions. In contrast, the assumptions given here concern behavior under fixed true distributions and do not involve the sample size n.

In Assumptions S1-S3 below, the true distribution of $\{W_i : i \geq 1\}$ is F_{γ^*}. The conditions in Assumptions S1-S3 are assumed to hold for all $\gamma^* \in \Gamma$. Let C be a generic finite positive constant that does not necessarily take the same value when it appears in two different places. None of the constants that appear in Assumptions S1-S3 depend on $\gamma^* \in \Gamma$.

Assumption S1. $\{W_i : i \geq 1\}$ is a strictly stationary and strong mixing sequence with mixing coefficients $\alpha_m \leq Cm^{-A}$ for some $A > d_\theta q/(q - d_\theta)$ and some $q > d_\theta \geq 2$, or $\{W_i : i \geq 1\}$ is an i.i.d. sequence and the constant q (that appears in Assumption S3 below) equals $2 + \delta$ for some $\delta > 0$.

In Assumption S1, the decay rate of the strong mixing coefficients is used to obtain the stochastic equicontinuity of certain empirical processes using results in Hansen (1996b). The WLLN and CLT for strong mixing arrays also hold under this decay rate, see Andrews (1988) and de Jong (1997). In the i.i.d. case, the constant q is smaller than in the strong mixing case, which yields weaker moment restrictions in Assumption S3

below.

Assumption S2. (i) For some function $\rho(w,\theta) \in R$, $Q_n(\theta) = n^{-1}\sum_{i=1}^{n}\rho(W_i,\theta)$, where $\rho(w,\theta)$ is twice continuously differentiable in θ on an open set containing Θ^* $\forall w \in \mathcal{W}$.
(ii) $\rho(w,\theta)$ does not depend on π when $\beta = 0$ $\forall w \in \mathcal{W}$.
(iii) When $\beta^* = 0$, $E_{\gamma^*}\rho(W_i,\psi,\pi)$ is uniquely minimized by ψ^* $\forall \pi \in \Pi$.
(iv) When $\beta^* \neq 0$, $E_{\gamma^*}\rho(W_i,\theta)$ is uniquely minimized by θ^*.
(v) $\Psi(\pi)$ is compact $\forall \pi \in \Pi$, and Π and Θ are compact.
(vi) $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $d_H(\Psi(\pi_1), \Psi(\pi_2)) < \varepsilon$ $\forall \pi_1, \pi_2 \in \Pi$ with $\|\pi_1 - \pi_2\| < \delta$, where $d_H(\cdot)$ is the Hausdorff metric.

Examples of $\rho(w,\theta)$ functions that satisfy Assumption S2(ii) are functions of the form

$$\rho(w,\theta) = \rho^*(w, a(x,\beta)h(x,\pi),\zeta), \text{ where } a(x,0) = 0, \ \forall w \in \mathcal{W}, \tag{6.1}$$

x is a sub-vector of w, and $a(x,\beta)$ and $h(x,\pi)$ are known functions. In (6.1), $\rho(w,\theta)$ does not depend on π when $\beta = 0$ because $a(x,\beta) = 0$. Examples of $a(x,\beta)$ include (i) $a(x,\beta) = \beta$, (ii) $a(x,\beta) = \exp(\beta) - 1$, and (iii) $a(x,\beta) = x'\beta$. Example (i) covers the nonlinear regression example, where β is the coefficient of the nonlinear regressor. Example (ii) demonstrates that $a(x,\beta)$ can be nonlinear in β provided $a(x,\beta) = 0$ at $\beta = 0$. Example (iii) covers the weak IV example and the case in which β enters the model through a single index. The form in (6.1) does not require a regression model and it allows for complicated structural models by allowing different functional forms for $a(x,\beta)$, $h(x,\pi)$, and $\rho(w,\theta)$.

Returning now to the general $\rho(w,\theta)$ case, Assumption S2(vi) holds immediately in cases where $\Psi(\pi)$ does not depend on π. When $\Psi(\pi)$ depends on π, the boundary of $\Psi(\pi)$ is often a continuous linear function of π, as in the ARMA(1,1) example. In such cases, it is simple to verify Assumption S2(vi).

Let $\rho_\theta(w,\theta)$ and $\rho_{\theta\theta}(w,\theta)$ denote the first-order and second-order partial derivatives of $\rho(w,\theta)$ wrt θ, respectively. Let $\rho_\psi(w,\theta)$ and $\rho_{\psi\psi}(w,\theta)$ denote the first-order and second-order partial derivatives of $\rho(w,\theta)$ wrt ψ.

For $\beta \neq 0$, let

$$\begin{aligned} B^{-1}(\beta)\rho_\theta(w,\theta) &= \rho_\theta^\dagger(w,\theta) \text{ and} \\ B^{-1}(\beta)\rho_{\theta\theta}(w,\theta)B^{-1}(\beta) &= \rho_{\theta\theta}^\dagger(w,\theta) + \iota^{-1}(\beta)\varepsilon(w,\theta), \end{aligned} \tag{6.2}$$

where $\rho^\dagger_{\theta\theta}(w,\theta)$ is symmetric, $\rho^\dagger_\theta(w,\theta)$, $\rho^\dagger_{\theta\theta}(w,\theta)$, and $\varepsilon(w,\theta)$ satisfy Assumption S3 below, $\iota(\beta)=\beta$ when β is a scalar, and $\iota(\beta)=||\beta||$ when β is a vector. The re-scaling matrix $B^{-1}(\beta)$ in (6.2) is used to deal with the singularity issue that arises when $\beta=0$. In particular, the covariance matrix of $\rho_\theta(W_i,\theta)$ is singular when $\beta=0$ and close to singular when β is close to 0. In contrast, the re-scaled quantity $\rho^\dagger_\theta(W_i,\theta)$ has a covariance matrix that is not close to being singular even when β is close to 0. Similarly, $E_{\gamma^*}\rho_{\theta\theta}(W_i,\theta)$ is singular when $\beta=0$ and close to singular when β is close to 0. Re-scaling of $\rho_{\theta\theta}(W_i,\theta)$ yields a quantity $\rho^\dagger_{\theta\theta}(W_i,\theta)$ whose expectation is not close to singular even when β is close to 0 plus another term $\varepsilon(W_i,\theta)$ that is asymptotically negligible.

Below we illustrate the form of $\rho^\dagger_\theta(w,\theta)$, $\rho^\dagger_{\theta\theta}(w,\theta)$, and $\varepsilon(w,\theta)$ for $\rho(w,\theta)$ functions as in (6.1).

Define

$$V^\dagger(\theta_1,\theta_2;\gamma_0)=\sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho^\dagger_\theta(W_i,\theta_1),\rho^\dagger_\theta(W_{i+m},\theta_2)), \tag{6.3}$$

which does not depend on i because the observations are stationary under Assumption S1. Under Assumptions S1 and S3(iii) below, $V^\dagger(\theta_1,\theta_2;\gamma_0)$ exists by a standard strong mixing inequality.

Assumption S3. (i) $E_{\gamma^*}\varepsilon(W_i,\theta^*)=0$ and $||\beta^*||^{-1}||E_{\gamma^*}\varepsilon(W_i,\psi^*,\pi)||\leq C||\pi-\pi^*||$ $\forall\gamma^*\in\Gamma$ with $0<||\beta^*||<\delta$ for some $\delta>0$.
(ii) For all $\delta>0$ and some function $M(w):\mathcal{W}\to R_+$, $||\rho_{\psi\psi}(w,\theta_1)-\rho_{\psi\psi}(w,\theta_2)||+$ $||\rho^\dagger_{\theta\theta}(w,\theta_1)-\rho^\dagger_{\theta\theta}(w,\theta_2)||+||\varepsilon(w,\theta_1)-\varepsilon(w,\theta_2)||\leq M(w)\delta$, $\forall\theta_1,\theta_2\in\Theta$ with $||\theta_1-\theta_2||\leq\delta$, $\forall w\in\mathcal{W}$.
(iii) $E_{\gamma^*}\sup_{\theta\in\Theta}\{|\rho(W_i,\theta)|^q+||\rho^\dagger_\theta(W_i,\theta)||^q+||\rho_{\psi\psi}(W_i,\theta)||^q+||\rho^\dagger_{\theta\theta}(W_i,\theta)||^q+||\varepsilon(W_i,\theta)||^q+$ $M(W_i)^q\}\leq C$, where q is as in Assumption S1.
(iv) $\inf_{\pi\in\Pi}\lambda_{\min}(E_{\gamma^*}\rho_{\psi\psi}(W_i,\psi^*,\pi))>0$ when $\beta^*=0$ and $E_{\gamma^*}\rho^\dagger_{\theta\theta}(W_i,\theta^*)$ is positive definite $\forall\gamma^*\in\Gamma$.
(v) $V^\dagger(\theta_0,\theta_0;\gamma_0)$ is positive definite $\forall\gamma_0\in\Gamma$.

In Assumptions S1-S3, Assumptions S2(ii), S2(iii), S3(i), S3(iii), S3(iv) and S3(v) are particularly related to the weak identification problem. Assumption S2(ii) implies that the sample criterion function is flat in π when $\beta=0$, as in Assumption A. Assumption S2(iii) differs from a standard condition in the sense that the population criterion function is not uniquely minimized by the true value when $\beta^*=0$. The Lipschitz condition in Assumption S3(i) typically holds because the partial derivative of $E_{\gamma^*}\varepsilon(W_i,\psi^*,\pi)$ wrt π is approximately proportional to $||\beta^*||$ when $||\beta^*||$ is close to 0. Because parts of $B^{-1}(\beta)$

diverge as β converges to 0, the moment conditions for $\rho_\theta^\dagger(W_i,\theta)$ and $\rho_{\theta\theta}^\dagger(W_i,\theta)$ in Assumption S3(iii) are stronger than standard moment conditions on the first-order and second-order derivatives. These conditions hold in typical examples, see below, because the partial derivative of $\rho(w,\theta)$ wrt π is small when β is close to 0 under Assumption S2(ii). Hence, the rhs moments are uniformly bounded even after the scaling by $B^{-1}(\beta)$. In Assumptions S3(iv) and S3(v), $E_{\gamma^*}\rho_{\theta\theta}^\dagger(W_i,\theta^*)$ and $V^\dagger(\theta_0,\theta_0;\gamma_0)$ typically are positive definite due to the re-scaling in (6.2).

Let $S_\psi = [I_{d_\psi} : 0_{d_\psi \times d_\pi}]$ denote the $d_\psi \times d_\theta$ selector matrix that selects ψ out of θ.

Lemma 6.1. *Suppose Assumptions* B1 *and* B2 *hold. Assumptions* S1-S3 *imply that Assumptions* A, B3, C1-C4, C8, *and* D1-D3 *hold with*

$$Q(\theta;\gamma_0) = E_{\gamma_0}\rho(W_i,\theta),\ D_\psi Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_\psi(W_i,\theta),\ D_{\psi\psi}Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_{\psi\psi}(W_i,\theta),$$
$$m(W_i,\theta) = \rho_\psi(W_i,\theta),\ \Omega(\pi_1,\pi_2;\gamma_0) = S_\psi V^\dagger((\psi_0,\pi_1),(\psi_0,\pi_2);\gamma_0)S_\psi',$$
$$H(\pi;\gamma_0) = E_{\gamma_0}\rho_{\psi\psi}(W_i,\psi_0,\pi),\ DQ_n(\theta) = n^{-1}\sum_{i=1}^n \rho_\theta(W_i,\theta),$$
$$D^2Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_{\theta\theta}(W_i,\theta),\ J(\gamma_0) = E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0),\ \textit{and}\ V(\gamma_0) = V^\dagger(\theta_0,\theta_0;\gamma_0).$$

Next, we give some primitive sufficient conditions for Assumption C6 for the case where β is a scalar parameter. Let $\rho_\psi(w,\theta) = (\rho_\beta(w,\theta)', \rho_\zeta(w,\theta)')'$. When $\beta = 0$, $\rho_\zeta(w,\theta)$ does not depend on π by Assumption S2(ii) and is denoted by $\rho_\zeta(w,\psi)$. When $d_\beta = 1$ and $\beta_0 = 0$, define

$$\rho_\psi^*(W_i,\psi_0,\pi_1,\pi_2) = (\rho_\beta(W_i,\psi_0,\pi_1), \rho_\beta(W_i,\psi_0,\pi_2), \rho_\zeta(W_i,\psi_0)')'. \tag{6.4}$$

Assumption S4. (i) $d_\beta = 1$ (i.e., β is a scalar).
(ii) $\Omega_G(\pi_1,\pi_2;\gamma_0) = \sum_{m=-\infty}^\infty Cov_{\gamma_0}(\rho_\psi^*(W_i,\psi_0,\pi_1,\pi_2), \rho_\psi^*(W_{i+m},\psi_0,\pi_1,\pi_2))$ is positive definite, $\forall \pi_1,\pi_2 \in \Pi$ with $\pi_1 \neq \pi_2$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

Lemma 6.2. *Assumptions* S1-S4 *imply Assumption* C6.

Now, we illustrate the form of $\rho_\theta^\dagger(w,\theta)$, $\rho_{\theta\theta}^\dagger(w,\theta)$, and $\varepsilon(w,\theta)$ when $\rho(w,\theta)$ belongs to the class specified in (6.1) and show that Assumption S3(i) holds. For simplicity, we assume $a(x,\beta)$ and $h(x,\pi)$ are both scalars and no parameter ζ appears. Let $\rho'(\cdot)$

and $\rho''(\cdot)$ abbreviate the first- and second-order derivatives of $\rho^*(w, a(x,\beta)h(x,\pi))$ wrt $a(x,\beta)h(x,\pi)$. Let $a_\beta(x,\beta)$, $a_{\beta\beta}(x,\beta)$, $h_\pi(x,\pi)$, $h_{\pi\pi}(x,\pi)$ denote the first- and second-order partial derivatives of $a(x,\beta)$ and $h(x,\pi)$ wrt β and π. The first and second order partial derivatives of $\rho(w,\theta)$ wrt to β and π are

$$\begin{aligned}
\rho_\beta(w,\theta) &= \rho'(\cdot)a_\beta(x,\beta)h(x,\pi),\ \rho_\pi(w,\theta) = \rho'(\cdot)a(x,\beta)h_\pi(x,\pi),\\
\rho_{\beta\beta}(w,\theta) &= \rho''(\cdot)a_\beta(x,\beta)a_\beta(x,\beta)'h^2(x,\pi) + \rho'(\cdot)a_{\beta\beta}(x,\beta)h(x,\pi),\\
\rho_{\beta\pi}(w,\theta) &= \rho''(\cdot)a(x,\beta)h(x,\pi)a_\beta(x,\beta)h_\pi(x,\pi)' + \rho'(\cdot)a_\beta(x,\beta)h_\pi(x,\pi)',\ \text{and}\\
\rho_{\pi\pi}(w,\theta) &= \rho''(\cdot)a^2(x,\beta)h_\pi(x,\pi)h_\pi(x,\pi)' + \rho'(\cdot)a(x,\beta)h_{\pi\pi}(x,\pi). \qquad (6.5)
\end{aligned}$$

In this case, we define

$$\begin{aligned}
\rho^\dagger_\theta(w,\theta) &= \rho'(\cdot)a^\dagger(x,\theta),\ \rho^\dagger_{\theta\theta}(w,\theta) = \rho''(\cdot)a^\dagger(x,\theta)a^\dagger(x,\theta)',\ \text{where}\\
a^\dagger(x,\theta) &= (a_\beta(x,\beta)'h(x,\pi), \frac{a(x,\beta)}{\iota(\beta)}h_\pi(x,\pi)')' \text{ and}\\
\varepsilon(w,\theta) &= \rho'(\cdot)\begin{bmatrix} a_{\beta\beta}(x,\beta)h(x,\pi) & a_\beta(x,\beta)h_\pi(x,\pi)' \\ h_\pi(x,\pi)a_\beta(x,\beta)' & \frac{a(x,\beta)}{\iota(\beta)}h_{\pi\pi}(x,\pi)\end{bmatrix}. \qquad (6.6)
\end{aligned}$$

Note that $\beta^{-1}a(x,\beta)$ is continuous at $\beta = 0$ in the scalar β case. In particular, $\lim_{\beta\to 0}\beta^{-1}a(x,\beta) = a_\beta(x,0)$ by a mean-value expansion because $a(x,0) = 0$ and $a(x,\beta)$ is continuously differentiable in β. In the vector β case, $\lim_{\beta\to 0,\beta/||\beta||\to\omega_0}||\beta||^{-1}a(x,\beta) = a_\beta(x,0)\omega_0$.

When $\varepsilon(w,\theta)$ takes the form in (6.6), Assumption S3* below implies Assumption S3(i). In Assumption S3*(i), X_i is a sub-vector of W_i that takes the place of x in w.

Assumption S3*. (i) X_i is a vector of weakly exogenous variables such that $E_{\gamma^*}(\rho'(W_i, a(X_i,\beta^*)h(X_i,\pi^*))|X_i) = 0$ a.s. $\forall\gamma^*\in\Gamma$.
(ii) $E_{\gamma^*}\sup_{\pi\in\Pi}|\rho''(W_i, a(X_i,\beta^*)h(X_i,\pi))|\cdot||h_\pi(X_i,\pi)||\cdot(|h(X_i,\pi)| + ||h_\pi(X_i,\pi)|| + ||h_{\pi\pi}(X_i,\pi)||)\cdot\sup_{||\beta||<\delta}||a_\beta(X_i,\beta)||\cdot(||a_\beta(X_i,\beta)|| + ||a_{\beta\beta}(X_i,\beta)||) \le C$ for some $C<\infty$ and $\delta > 0$ $\forall\gamma^*\in\Gamma$.

Several of the derivatives in Assumption S3*(ii) are constants in many examples, which makes the moment condition in Assumption S3*(ii) less restrictive than it may appear. For example, when $a(X_i,\beta) = \beta$, $a_\beta(X_i,\beta) = 1$ and $a_{\beta\beta}(X_i,\beta) = 0$.

Lemma 6.3. *Suppose $\rho(w,\theta)$ belongs to the class in* (6.1)*, where $a(x,\beta)\in R$ and*

$h(x,\pi) \in R$ are twice differentiable wrt β and π, respectively, and no parameter ζ appears. Then, $\varepsilon(w,\theta)$ takes the form in (6.6) *and Assumption* S3(i) *is implied by Assumption* S3*.

Comment. The case where $\rho(w,\theta)$ belongs to the class in (6.1) and the parameter ζ appears is analyzed in Appendix A [IT IS NOT THERE YET].

7. Wald Confidence Sets and Tests

In this section, we consider a CS for a function $r(\theta)$ of θ by inverting a Wald or t test of the hypotheses $H_0 : r(\theta) = v$ for $v \in r(\Theta)$. We also consider Wald and t tests of H_0. We determine the asymptotic size of standard Wald and t CS's. We introduce robust Wald and t CS's whose asymptotic size is guaranteed to equal their nominal size.

7.1. Wald and t Statistics

The Wald and t statistics are defined as follows. Let

$$\Sigma(\gamma_0) = J^{-1}(\gamma_0)' V(\gamma_0) J^{-1}(\gamma_0) \text{ and } \widehat{\Sigma}_n = \widehat{J}_n^{-1} \widehat{V}_n \widehat{J}_n^{-1}, \tag{7.1}$$

where $\widehat{J}_n$ and $\widehat{V}_n$ are estimators of $J(\gamma_0)$ and $V(\gamma_0)$ that do not depend on the nuisance parameter ϕ. The Wald statistic takes the form

$$W_n(v) = n(r(\widehat{\theta}_n) - v)'(r_\theta(\widehat{\theta}_n) B^{-1}(\widehat{\beta}_n) \widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n) r_\theta(\widehat{\theta}_n)')^{-1}(r(\widehat{\theta}_n) - v), \tag{7.2}$$

where $r_\theta(\theta) = (\partial/\partial\theta') r(\theta) \in R^{d_r \times d_\theta}$.

When $d_r = 1$, the t statistic takes the form

$$T_n(v) = \frac{n^{1/2}(r(\widehat{\theta}_n) - v)}{(r_\theta(\widehat{\theta}_n) B^{-1}(\widehat{\beta}_n) \widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n) r_\theta(\widehat{\theta}_n)')^{1/2}}. \tag{7.3}$$

Although these definitions of the Wald and t statistics involve $B^{-1}(\widehat{\beta}_n)$, they are the same as the standard definitions used in practice. By Theorem 5.2(a), when $\beta_0 \neq 0$, $B^{-1}(\beta_0)\Sigma(\gamma_0)B^{-1}(\beta_0)$ is the asymptotic covariance matrix of $\widehat{\theta}_n$. In the Wald and t statistics, the asymptotic covariance is replaced by the estimator $B^{-1}(\widehat{\beta}_n)\widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n)$.

The same form of the Wald and t statistics is used under all sequences of true parameters $\gamma_n \in \Gamma(\gamma_0)$.

In the results below, we consider the behavior of the Wald and t statistics when the null hypothesis holds. Thus, under a sequence $\{\gamma_n\}$, we consider the sequence of null hypotheses $H_0 : r(\theta) = v_n$, where v_n equals $r(\theta_n)$ and $\gamma_n = (\theta_n, \phi_n)$. We employ the following notational simplification:

$$T_n = T_n(v_n) \text{ and } W_n = W_n(v_n), \text{ where } v_n = r(\theta_n). \tag{7.4}$$

7.2. Rotation

To obtain the asymptotic distribution of the Wald statistic we consider a rotation of $r(\widehat{\theta}_n)$ and $r_\theta(\widehat{\theta}_n)$ by a matrix $A(\widehat{\theta}_n)$. The rotation is designed to separate the effects of the randomness in $\widehat{\psi}_n$ and $\widehat{\pi}_n$, which have different rates of convergence for some sequences $\{\gamma_n\}$. We partition $r_\theta(\theta)$ conformably with $\theta = (\psi, \pi)$:

$$r_\theta(\theta) = [r_\psi(\theta) : r_\pi(\theta)]. \tag{7.5}$$

Suppose $rank(r_\pi(\theta)) = d_\pi^*$ $(\leq \min(d_r, d_\pi))$ $\forall \theta \in \Theta_\delta$ for some $\delta > 0$. (Assumption R1(iii) below). For $\theta \in \Theta_\delta$, let $A(\theta) = [A_1(\theta)' : A_2(\theta)']' \in O(d_r)$, where the rows of $A_1(\theta) \in R^{(d_r - d_\pi^*) \times d_r}$ span the null space of $r_\pi(\theta)'$, the rows of $A_2(\theta) \in R^{d_\pi^* \times d_r}$ span the column space of $r_\pi(\theta)$, and $O(d_r)$ stands for the orthogonal group of degree d_r over the real space. Hence,

$$A(\theta) r_\pi(\theta) = \begin{bmatrix} A_1(\theta) r_\pi(\theta) \\ A_2(\theta) r_\pi(\theta) \end{bmatrix} = \begin{bmatrix} 0_{(d_r - d_\pi^*) \times d_\pi} \\ r_\pi^*(\theta) \end{bmatrix}, \tag{7.6}$$

where $r_\pi^*(\theta) \in R^{d_\pi^* \times d_\pi}$ has full row rank d_π^*. For simplicity, hereafter we write the 0 matrix as 0 when there is no confusion about its dimension.

With the $A(\theta)$ rotation, the derivative matrix $r_\theta(\theta)$ becomes

$$r_\theta^A(\theta) = A(\theta) r_\theta(\theta) = \begin{bmatrix} r_\psi^*(\theta) & 0 \\ r_\psi^0(\theta) & r_\pi^*(\theta) \end{bmatrix}, \tag{7.7}$$

where the $(d_r - d_\pi^*) \times d_\psi$ matrix $r_\psi^*(\theta)$ has full row rank $d_r - d_\pi^*$. When $d_\pi^* = d_r$, $A_1(\theta)$ and $[r_\psi^*(\theta) : 0]$ disappear. When $d_\pi^* = 0$, $A_2(\theta)$ and $[r_\psi^0(\theta) : r_\pi^*(\theta)]$ disappear.

The effect of randomness in $\widehat{\pi}_n$ is concentrated in the full rank matrix $r^*_\pi(\widehat{\theta}_n)$ because the upper right corner of $r^A_\theta(\widehat{\theta}_n)$ is 0. The effect of randomness in $\widehat{\psi}_n$ is incorporated in both $r^*_\psi(\widehat{\theta}_n)$ and $r^0_\psi(\widehat{\theta}_n)$.

Using the rotation by $A(\widehat{\theta}_n)$, the Wald statistic in (7.2) can be written as

$$W_n = n(r(\widehat{\theta}_n) - v)'A(\widehat{\theta}_n)'(r^A_\theta(\widehat{\theta}_n)B^{-1}(\widehat{\beta}_n)\widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n)r^A_\theta(\widehat{\theta}_n)')^{-1}A(\widehat{\theta}_n)(r(\widehat{\theta}_n) - v), \quad (7.8)$$

where the first $d_r - d^*_\pi$ rows of $A(\widehat{\theta}_n)r(\widehat{\theta}_n)$ only depend on the randomness in $\widehat{\psi}_n$, not $\widehat{\pi}_n$, asymptotically by the choice of $A(\widehat{\theta}_n)$.

Define a $d_r \times d_\theta$ matrix

$$r^*_\theta(\theta) = \begin{bmatrix} r^*_\psi(\theta) & 0 \\ 0 & r^*_\pi(\theta) \end{bmatrix}. \quad (7.9)$$

Because $\widehat{\psi}_n$ converges faster than $\widehat{\pi}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, as shown in Theorems 5.1 and 5.2, the effect of randomness in $\widehat{\pi}_n$ is of an order of magnitude larger than that in $\widehat{\psi}_n$. As a result, the limit of $r^0_\psi(\widehat{\theta}_n)$ does not show up in the asymptotic distributions of the Wald and t statistics. On the other hand, the limit of $r^*_\psi(\widehat{\theta}_n)$ does appear in the asymptotic distribution because it is the effect of randomness in $\widehat{\psi}_n$ separated from that in $\widehat{\pi}_n$. Hence, the matrix $r^*_\theta(\theta)$, rather than $r^A_\theta(\theta)$, appears in the asymptotic distribution below.

When $r_\pi(\theta)$ has full row rank, i.e., $d^*_\pi = d_r$, for $\theta \in \Theta_\delta$, we have $A(\theta) = I_{d_r}$, $r^A_\theta(\theta) = r_\theta(\theta)$, and $r^*_\theta(\theta) = [0 : r_\pi(\theta)]$. In this case, rotation is not needed to concentrate the randomness in $\widehat{\pi}_n$. Also, when $d_r = 1$, we have $A(\theta) = 1$, so no rotation is employed.

Define

$$\eta_n(\theta) = \begin{cases} n^{1/2}A_1(\theta)(r(\psi_n, \pi) - r(\psi_n, \pi_n)) & \text{if } d^*_\pi < d_r \\ 0 & \text{if } d^*_\pi = d_r. \end{cases} \quad (7.10)$$

7.3. Function of Interest

The function of interest, $r(\theta)$, satisfies the following assumptions.

Assumption R1. (i) $r(\theta)$ is continuously differentiable on Θ.
(ii) $r_\theta(\theta)$ is full row rank d_r $\forall \theta \in \Theta$.
(iii) $rank(r_\pi(\theta)) = d^*_\pi$ for some constant $d^*_\pi \leq \min(d_r, d_\pi)$ $\forall \theta \in \Theta_\delta = \{\theta \in \Theta : ||\beta|| < \delta\}$ for some $\delta > 0$.

Assumption R2. $\eta_n(\widehat{\theta}_n) \to_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ $\forall b \in R^{d_\beta}_{[\pm\infty]}$.

Three different sufficient conditions for the high-level Assumption R2 are given by Assumptions R2*(i)-(iii) below. Any one of them is sufficient for Assumption R2 (under the conditions in Lemma 7.1 below).

Assumption R2*. (i) $d^*_\pi = d_r$.
(ii) $d_r = 1$.
(iii) The column space of $r_\pi(\theta)$ is the same $\forall \theta \in \Theta_\delta$ for some $\delta > 0$.

Assumption R2*(i) requires that the restrictions only involve π. Assumption R2*(ii) requires that only one restriction appears. Assumption R2*(iii) is satisfied when $r_\pi(\theta) = a(\theta)R_\pi$, where $a(\theta) : \Theta_\delta \to R$, $a(\theta) \neq 0$, and $R_\pi \in R^{d_r \times d_\pi}$. A special case is when $r_\pi(\theta)$ is constant due to the restrictions being linear.

Assumption $\mathrm{R_L}$. $r(\theta) = R\theta$, where $R \in R^{d_r \times d_\theta}$ has full row rank d_r.

Assumption $\mathrm{R_L}$ is a sufficient condition for Assumptions R1 and R2.

Lemma 7.1. *Assumptions* R2*(i) *and* R2*(ii) *each* (*separately*) *implies Assumption* R2. *Assumption* R2*(iii) *combined with Assumptions* A *and* B3(i)-(ii) *implies Assumption* R2.

Lemma 7.2. *Assumption* $\mathrm{R_L}$ *implies Assumptions* R1 *and* R2.

7.4. Variance Matrix Estimators

The estimators of the components of the asymptotic variance matrix are assumed to satisfy the following assumptions. Two forms are given for Assumption V1 that follows. The first applies when β is a scalar and the second applies when β is a vector. The reason for the difference is that the normalizing matrix $B(\beta)$ is different in these two cases.

When β is a scalar, let $J(\theta; \gamma_0)$ and $V(\theta; \gamma_0)$ for $\theta \in \Theta$ be some non-stochastic $d_\theta \times d_\theta$ matrix-valued functions such that $J(\theta_0; \gamma_0) = J(\gamma_0)$ and $V(\theta_0; \gamma_0) = V(\gamma_0)$, where $J(\gamma_0)$ and $V(\gamma_0)$ are as in Assumptions D2 and D3. Let

$$\Sigma(\theta; \gamma_0) = J^{-1}(\theta; \gamma_0)V(\theta; \gamma_0)J^{-1}(\theta; \gamma_0) \text{ and } \Sigma(\pi; \gamma_0) = \Sigma(\psi_0, \pi; \gamma_0). \tag{7.11}$$

Let $\Sigma_{\beta\beta}(\pi; \gamma_0)$ denote the upper left (1,1) element of $\Sigma(\pi; \gamma_0)$.

Assumption V1 below applies when β is a scalar.

Assumption V1 (scalar $\boldsymbol{\beta}$). (i) $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n)$ for some (stochastic) $d_\theta \times d_\theta$ matrix-valued functions $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ on Θ that satisfy $\sup_{\theta\in\Theta} ||\widehat{J}_n(\theta) - J(\theta;\gamma_0)|| \to_p 0$ and $\sup_{\theta\in\Theta} ||\widehat{V}_n(\theta) - V(\theta;\gamma_0)|| \to_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$.
(ii) $J(\theta;\gamma_0)$ and $V(\theta;\gamma_0)$ are continuous in θ on Θ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.
(iii) $0 < \inf_{\pi\in\Pi} \Sigma_{\beta\beta}(\pi;\gamma_0)) \leq \sup_{\pi\in\Pi} \Sigma_{\beta\beta}(\pi;\gamma_0)) < \infty$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.

When β is a vector, i.e., $d_\beta > 1$, we reparameterize β as $(||\beta||, \omega)$, where $\omega = \beta/||\beta||$ if $\beta \neq 0$ and by definition $\omega = 1_{d_\beta}/||1_{d_\beta}||$ with $1_{d_\beta} = (1, ..., 1) \in R^{d_\beta}$ if $\beta = 0$. Correspondingly, θ is reparameterized as $\theta^+ = (||\beta||, \omega, \zeta, \pi)$. Let $\Theta^+ = \{\theta^+ : \theta^+ = (||\beta||, \beta/||\beta||, \zeta, \pi), \theta \in \Theta\}$. Let $\widehat{\theta}_n^+$ and θ_0^+ be the counterparts of $\widehat{\theta}_n$ and θ_0 after reparameterization.

When β is a vector, let $J(\theta^+;\gamma_0)$ and $V(\theta^+;\gamma_0)$ denote some non-stochastic $d_\theta \times d_\theta$ matrix-valued functions such that $J(\theta_0^+;\gamma_0) = J(\gamma_0)$ and $V(\theta_0^+;\gamma_0) = V(\gamma_0)$. Let

$$\begin{aligned} \Sigma(\theta^+;\gamma_0) &= J^{-1}(\theta^+;\gamma_0)V(\theta^+;\gamma_0)J^{-1}(\theta^+;\gamma_0) \text{ and} \\ \Sigma(\pi,\omega;\gamma_0) &= \Sigma(||\beta_0||, \omega, \zeta_0, \pi;\gamma_0). \end{aligned} \tag{7.12}$$

Let $\Sigma_{\beta\beta}(\pi,\omega;\gamma_0)$ denote the upper left $d_\beta \times d_\beta$ sub-matrix of $\Sigma(\pi,\omega;\gamma_0)$.

Assumption V1 below applies when β is a vector.

Assumption V1 (vector $\boldsymbol{\beta}$). (i) $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n^+)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n^+)$ for some (stochastic) $d_\theta \times d_\theta$ matrix-valued functions $\widehat{J}_n(\theta^+)$ and $\widehat{V}_n(\theta^+)$ on Θ^+ that satisfy $\sup_{\theta^+\in\Theta^+} ||\widehat{J}_n(\theta^+) - J(\theta^+;\gamma_0)|| \to_p 0$ and $\sup_{\theta^+\in\Theta^+} ||\widehat{V}_n(\theta^+) - V(\theta^+;\gamma_0)|| \to_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$.[14]
(ii) $J(\theta^+;\gamma_0)$ and $V(\theta^+;\gamma_0)$ are continuous in θ^+ on Θ^+ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.
(iii) $0 < \inf_{\pi\in\Pi} \lambda_{\min}(\Sigma_{\beta\beta}(\pi,\omega;\gamma_0)) \leq \sup_{\pi\in\Pi} \lambda_{\min}(\Sigma_{\beta\beta}(\pi,\omega;\gamma_0)) < \infty$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.
(iv) $P(\tau_\beta(\pi^*(\gamma_0, b), \gamma_0, b) = 0) = 0$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$ and $\forall b \in R^{d_\beta}$.

The following assumption applies with both scalar and vector β.

Assumption V2. Under $\Gamma(0, \infty, \omega_0)$, $\widehat{J}_n \to_p J(\gamma_0)$ and $\widehat{V}_n \to_p V(\gamma_0)$.

[14] The functions $J(\theta^+;\gamma_0)$ and $V(\theta^+;\gamma_0)$ do not depend on ω_0, only γ_0.

Example 1. (cont.). In this example, we estimate $J(\gamma_0)$ and $V(\gamma_0)$ by $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n)$, respectively, where

$$\begin{aligned}
\widehat{J}_n(\theta) &= n^{-1}\sum_{i=1}^{n} d_i(\pi)\, d_i(\pi)' \text{ and} \\
\widehat{V}_n(\theta) &= n^{-1}\sum_{i=1}^{n} U_i^2(\theta) d_i(\pi)\, d_i(\pi)' = n^{-1}\sum_{i=1}^{n} U_i^2 d_i(\pi)\, d_i(\pi)' \\
&\quad + 2n^{-1}\sum_{i=1}^{n} U_i \left(\beta_n h(X_i,\pi_n) - \beta h(X_i,\pi) + (\zeta_n - \zeta)' Z_i\right) d_i(\pi)\, d_i(\pi)' \\
&\quad + n^{-1}\sum_{i=1}^{n} \left(\beta_n h(X_i,\pi_n) - \beta h(X_i,\pi) + (\zeta_n - \zeta)' Z_i\right)^2 d_i(\pi)\, d_i(\pi)'. \qquad (7.13)
\end{aligned}$$

Assumption V1(i) (scalar β) holds with

$$\begin{aligned}
&J(\theta;\gamma_0) = E_{\gamma_0} d_i(\pi)\, d_i(\pi)' \text{ and } V(\theta;\gamma_0) = E_{\gamma_0} U_i^2 d_i(\pi)\, d_i(\pi)' \\
&+ E_{\gamma_0}\left(\beta_0 h(X_i,\pi_0) - \beta h(X_i,\pi) + (\zeta_0 - \zeta)' Z_i\right)^2 d_i(\pi) d_i(\pi)', \qquad (7.14)
\end{aligned}$$

by Lemma 10.7 in Appendix B using the conditions in (4.3). Assumption V1(ii) holds by the continuity of $h(x,\pi)$ and $h_\pi(x,\pi)$ in π and the moment conditions in (4.3).

The quantity $\Sigma(\pi;\gamma_0)$ in (7.11) takes the form

$$\Sigma(\pi;\gamma_0) = \left(E_{\gamma_0} d_i(\pi)\, d_i(\pi)'\right)^{-1} E_{\gamma_0} U_i^2 d_i(\pi)\, d_i(\pi)' \left(E_{\gamma_0} d_i(\pi)\, d_i(\pi)'\right)^{-1}. \qquad (7.15)$$

Given this, Assumption V1(iii) holds by the nonsingularity conditions in (4.3).

Assumptions V1(i) and V1(ii) hold not only under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, but also under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ in this example. This and $\widehat{\theta}_n \to_p \theta_0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, which holds by Lemma 5.3, imply that Assumption V2 holds. □

7.5. Asymptotic Distribution of the Wald Statistic

The asymptotic null distribution of the Wald statistic under H_0 depends on the following quantities. The limit distribution of $\widehat{\omega}_n(\pi) = \widehat{\beta}_n(\pi)/||\widehat{\beta}_n(\pi)||$ under $\Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$ is given by

$$\omega^*(\pi;\gamma_0,b) = \frac{\tau_\beta(\pi;\gamma_0,b)}{||\tau_\beta(\pi;\gamma_0,b)||} \text{ for } \pi \in \Pi, \qquad (7.16)$$

where $\tau_\beta(\pi;\gamma_0,b)$ is defined in (5.5). Let $\overline{B}(\pi;\gamma_0,b)$ be a $d_r \times d_r$ matrix-valued function of $\tau_\beta(\pi;\gamma_0,b)$ defined as

$$\overline{B}(\pi;\gamma_0,b) = \begin{bmatrix} I_{(d_r-d_\pi^*)} & 0 \\ 0 & \iota(\tau_\beta(\pi;\gamma_0,b))I_{d_\pi^*} \end{bmatrix} \tag{7.17}$$

where $\iota(\cdot)$ is defined in (4.22).

Let

$$\begin{aligned} r_\theta^*(\pi) = r_\theta^*(\psi_0,\pi),\ r_\psi^*(\pi) = r_\psi^*(\psi_0,\pi) \text{ and} \\ \overline{\Sigma}(\pi;\gamma_0,b) = \begin{cases} \Sigma(\pi;\gamma_0) & \text{if } \beta \text{ is a scalar} \\ \Sigma(\pi,\omega^*(\pi;\gamma_0,b);\gamma_0) & \text{if } \beta \text{ is a vector,} \end{cases} \end{aligned} \tag{7.18}$$

where $\Sigma(\pi;\gamma_0)$ and $\Sigma(\pi,\omega;\gamma_0)$ are defined in (7.11) and (7.12), respectively.

Define a stochastic process $\{\lambda(\pi;\gamma_0,b) : \pi \in \Pi\}$ by

$$\begin{aligned} &\lambda(\pi;\gamma_0,b) \\ &= \tau^A(\pi;\gamma_0,b)'\overline{B}(\pi;\gamma_0,b)(r_\theta^*(\pi)\overline{\Sigma}(\pi;\gamma_0,b)r_\theta^*(\pi)')^{-1}\overline{B}(\pi;\gamma_0,b)\tau^A(\pi;\gamma_0,b), \text{ where} \\ &\tau^A(\pi;\gamma_0,b) = \begin{pmatrix} r_\psi^*(\pi)\tau(\pi;\gamma_0,b) \\ A_2(\psi_0,\pi)(r(\psi_0,\pi) - r(\psi_0,\pi_0)) \end{pmatrix} \in R^{d_r}. \end{aligned} \tag{7.19}$$

Under Assumption $\mathrm{R_L}$, $r_\theta(\theta) = R$ does not depend on θ, and, hence, $A(\theta)$ and $r_\theta^*(\theta)$ do not depend on θ. Define $R^* = r_\theta^*(\theta)$ under Assumption $\mathrm{R_L}$. Specifically,

$$R^A = AR = \begin{bmatrix} R_\psi^* & 0 \\ R_\psi^0 & R_\pi^* \end{bmatrix} \text{ and } R^* = \begin{bmatrix} R_\psi^* & 0 \\ 0 & R_\pi^* \end{bmatrix}, \tag{7.20}$$

where $R_\psi^* \in R^{(d_r-d_\pi^*)\times d_\psi}$ and $R_\pi^* \in R^{d_\pi^* \times d_\pi}$.

Define a stochastic process $\{\lambda_L(\pi;\gamma_0,b) : \pi \in \Pi\}$ by

$$\begin{aligned} &\lambda_L(\pi;\gamma_0,b) \\ &= \overline{\tau}(\pi;\gamma_0,b)'R^{*\prime}\overline{B}(\pi;\gamma_0,b)(R^*\overline{\Sigma}(\pi;\gamma_0,b)R^{*\prime})^{-1}\overline{B}(\pi;\gamma_0,b)R^*\overline{\tau}(\pi;\gamma_0,b), \text{ where} \\ &\overline{\tau}(\pi;\gamma_0,b) = (\tau(\pi;\gamma_0,b)', (\pi-\pi_0)')' \in R^{d_\theta}. \end{aligned} \tag{7.21}$$

Under the linear restriction of Assumption $\mathrm{R_L}$, $\lambda_L(\pi;\gamma_0,b) = \lambda(\pi;\gamma_0,b)$ and the asymptotic distribution of the Wald statistic can be simplified by replacing the stochastic

process $\{\lambda(\pi;\gamma_0,b):\pi\in\Pi\}$ with $\{\lambda_L(\pi;\gamma_0,b):\pi\in\Pi\}$ in the asymptotic results below.

The following theorem establishes the asymptotic null distribution of the Wald statistic for nonlinear restrictions that satisfy Assumption R2. (The null holds by the definition $W_n = W_n(v_n)$ in (7.4).)

Theorem 7.1. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1-R2, *and* V1-V2 *hold.*
(a) *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ *with* $b\in R^{d_\beta}$, $W_n\to_d\lambda(\pi^*(\gamma_0,b);\gamma_0,b)$.
(b) *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,\infty,\omega_0)$, $W_n\to_d\chi^2_{d_r}$.

A special case of Theorem 7.1 is the following result for linear restrictions.

Corollary 7.1. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, $\mathrm{R_L}$, *and* V1-V2 *hold.*
(a) *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ *with* $b\in R^{d_\beta}$, $W_n\to_d\lambda_L(\pi^*(\gamma_0,b);\gamma_0,b)$.
(b) *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,\infty,\omega_0)$, $W_n\to_d\chi^2_{d_r}$.

Specific forms of the stochastic process $\lambda(\pi;\gamma_0,b)$ are provided in the following examples. In Examples 1-4, $r(\theta)$ is linear in θ and Corollary 7.1 applies. In Example 5, $r(\theta)$ is nonlinear in θ and Assumption R2 is verified.

Example 1. *When* $r(\theta)=\psi$, $R=R^*=[I_{d_\psi}:0]$, *and* $\lambda_L(\pi;\gamma_0,b)=\tau(\pi;\gamma_0,b)'\overline{\Sigma}^{-1}_{\psi\psi}(\pi;\gamma_0,b)\tau(\pi;\gamma_0,b)$, *where* $\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)$ *is the upper left* $d_\psi\times d_\psi$ *block of* $\overline{\Sigma}(\pi;\gamma_0,b)$.

Example 2. *When* $r(\theta)=\pi$, $R=R^*=[0:I_{d_\pi}]$, *and* $\lambda_L(\pi;\gamma_0,b)=||\tau_\beta(\pi;\gamma_0,b)||^2(\pi-\pi_0)'\overline{\Sigma}^{-1}_{\pi\pi}(\pi;\gamma_0,b)(\pi-\pi_0)$, *where* $\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)$ *is the lower right* $d_\pi\times d_\pi$ *block of* $\overline{\Sigma}(\pi;\gamma_0,b)$.

Example 3. *When* $d_\psi=d_\pi$ *and* $r(\theta)=\psi+\pi$, $R=[I_{d_\psi}:I_{d_\pi}]$, $R^*=[0_{d_\psi}:I_{d_\pi}]$, *and* $\lambda_L(\pi;\gamma_0,b)=||\tau_\beta(\pi;\gamma_0,b)||^2(\pi-\pi_0)'\overline{\Sigma}^{-1}_{\pi\pi}(\pi;\gamma_0,b)(\pi-\pi_0)$. *Note that* $\lambda_L(\pi;\gamma_0,b)$ *is the same in this example as in Example* 2. *This occurs because* $d^*_\pi=d_r$ *so that the randomness in* $\widehat{\psi}_n$ *is completely dominated by that in* $\widehat{\pi}_n$. *Although* R *is different in Examples* 2 *and* 3, R^* *is the same in both examples.*

Example 4. *When* $r(\theta)=\theta$, $R=R^*=I_{d_\theta}$, *and* $\lambda_L(\pi;\gamma_0,b)=\overline{\tau}(\pi;\gamma_0,b)'\overline{B}(\pi;\gamma_0,b)\overline{\Sigma}^{-1}(\pi;\gamma_0,b)\overline{B}(\pi;\gamma_0,b)\overline{\tau}(\pi;\gamma_0,b)$.

Example 5. *When* $\theta = (\beta, \pi)'$, $r(\theta) = (\beta, \pi^2)'$, *and* β *and* π *are scalars, we have*

$$r_\theta(\theta) = r^*_\theta(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & 2\pi \end{bmatrix}, \text{ and } A(\theta) = I_2. \tag{7.22}$$

Assumption R2*(iii) *holds because* $A_2(\theta)$ *does not depend on* θ. *This implies that Assumption* R2 *holds. The stochastic process* $\{\tau^A(\pi; \gamma_0, b) : \pi \in \Pi\}$ *can be simplified to* $\tau^A(\pi; \gamma_0, b) = (\tau(\pi; \gamma_0, b), \pi - \pi_0^2)$.

Next we show that Assumption R2 is not superfluous. In certain cases, the Wald statistic diverges to infinity in probability under H_0.

Theorem 7.2. *Suppose Assumptions* A, B1-B3, C1-C8, R1, *and* V1 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $W_n \to_p \infty$ *if* $||\eta_n(\widehat{\theta}_n)|| \to_p \infty$.

Comment. This theorem provides a high-level condition under which the Wald statistic diverges to infinity in probability. The Wald statistic, which uses $r_\theta(\widehat{\theta}_n)$ in the covariance matrix estimation, is designed for the standard case in which $\widehat{\theta}_n$ converges to θ_n at rate $n^{-1/2}$. When $\widehat{\pi}_n$ is inconsistent or converges to π_n slower than $n^{-1/2}$, the estimator of the covariance matrix does not necessarily provide a proper normalization for the Wald statistic to have a non-degenerate limit.

Example 6. *When* $\theta = (\beta, \pi)'$, $r(\theta) = ((\beta+1)\pi, \pi^2)'$, *and* β *and* π *are both scalars, we have*

$$r_\theta(\theta) = \begin{bmatrix} \pi & \beta+1 \\ 0 & 2\pi \end{bmatrix}, \; A_1(\theta) = \frac{1}{||(-2\pi, \beta+1)||}(-2\pi, \beta+1), \text{ and}$$
$$\eta_n(\theta) = -\frac{n^{1/2}}{||(-2\pi, \beta+1)||}[-2\pi(\beta_n+1)(\pi - \pi_n) + (\beta+1)(\pi^2 - \pi_n^2)].$$

Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $b \in R^{d_\beta}$, $\eta_n(\widehat{\theta}_n) = ||(-2\widehat{\pi}_n, 1)||^{-1} n^{1/2}(\widehat{\pi}_n - \pi_0)^2 + o_p(n^{1/2})$ $\to_p \infty$. *The divergence to infinity in probability holds because* $\widehat{\pi}_n \to_d \pi^*(\gamma_0, b)$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *by Theorem* 5.1(a) *under Assumptions* A, B1-B3, *and* C1-C6 *and* $P(\pi^*(\gamma_0, b) = \pi_0) = 0$ *provided* Π *contains more than one element.*

7.6. Asymptotic Distribution of the t Statistic

Next, we provide the asymptotic distribution of the t statistic under H_0. Define

$$T_\psi(\pi;\gamma_0,b) = \frac{r_\psi(\pi)\tau(\pi;\gamma_0,b)}{(r_\psi(\pi)\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)r_\psi(\pi)')^{1/2}}, \tag{7.23}$$

where $\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)$ is the upper left $d_\psi \times d_\psi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$, $r_\psi(\pi) = r_\psi(\psi_0,\pi)$, and $\tau_\beta(\pi;\gamma_0,b)$ is defined in (5.5).

Define

$$T_\pi(\pi;\gamma_0,b) = \frac{\iota(\tau_\beta(\pi;\gamma_0,b))(r(\psi_0,\pi) - r(\psi_0,\pi_0))}{(r_\pi(\pi)\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)r_\pi(\pi)')^{1/2}}, \tag{7.24}$$

where $\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)$ is the lower right $d_\pi \times d_\pi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$ and $r_\pi(\pi) = r_\pi(\psi_0,\pi)$.

The following theorem provides the asymptotic null distribution of the t statistic for a scalar restriction. (The null holds by the definition $T_n = T_n(v_n)$ in (7.4).)

Theorem 7.3. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1, *and* V1-V2 *hold and* $d_r = 1$.
(a) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ *with* $b \in R^{d_\beta}$ *and* $d_\pi^* = 0$, $T_n \to_d T_\psi(\pi^*(\gamma_0,b);\gamma_0,b)$.
(b) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ *with* $b \in R^{d_\beta}$ *and* $d_\pi^* = 1$, $T_n \to_d T_\pi(\pi^*(\gamma_0,b);\gamma_0,b)$.
(c) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$, $T_n \to_d N(0,1)$.

Comments. 1. When $d_\pi^* = 0$, the scalar restriction is only on ψ by Assumption R1(iii). When $d_\pi^* = 1$, the restriction can be on π as well as ψ. However, the randomness in $\widehat{\psi}_n$ is dominated by that in $\widehat{\pi}_n$ under the condition in Theorem 7.3(b) because $\widehat{\psi}_n$ is consistent but $\widehat{\pi}_n$ is not. In consequence, the asymptotic distribution in Theorem 7.3(b) appears as if the restriction is only on π.

2. The numerators in $T_\psi(\cdot;\gamma_0,b)$ and $T_\pi(\cdot;\gamma_0,b)$ are special cases of $\overline{B}(\cdot;\gamma_0,b)\tau^A(\cdot;\gamma_0,b)$ used in Theorem 7.1 and the denominators of $T_\psi(\cdot;\gamma_0,b)$ and $T_\pi(\cdot;\gamma_0,b)$ correspond to special cases of $(r_\theta^*(\cdot)\Sigma(\cdot)r_\theta^*(\cdot)')^{1/2}$. When $d_\pi^* = 0$, $A(\theta) = A_1(\theta) = 1$ and $A_2(\theta)$ disappears, so that $\overline{B}(\cdot;\gamma_0,b) = 1$ and $\tau^A(\cdot;\gamma_0,b) = r_\psi(\cdot)\tau(\cdot;\gamma_0,b)$. When $d_\pi^* = 1$, $A(\theta) = A_2(\theta) = 1$ and $A_1(\theta)$ disappears, so that $\overline{B}(\cdot;\gamma_0,b) = \iota(\tau_\beta(\cdot;\gamma_0,b))$ and $\tau^A(\cdot;\gamma_0,b) = r(\psi_0,\cdot) - r(\psi_0,\pi_0)$.

7.7. Asymptotic Size of Standard Wald Confidence Sets

First, we consider the asymptotic size of a standard CS for $r(\theta) \in R^{d_r}$ obtained by inverting a Wald statistic, i.e.,

$$CS_{W,n} = \{v : W_n(v) \leq \chi^2_{d_r}(1-\alpha)\}, \tag{7.25}$$

where the Wald statistic $W_n(v)$ is as in (7.2), $\chi^2_{d_r}(1-\alpha)$ is the $1-\alpha$ quantile of a chi-square distribution with d_r degree of freedom, and $1-\alpha$ is the nominal size of the CS.

The asymptotic size of the CS above is established by verifying the high-level conditions in Andrews, Cheng, and Guggenberger (2009), hereafter ACG. In particular, assumptions in ACG require the asymptotic distribution of W_n, which abbreviates $W_n(r(\theta_n))$, under drifting sequences of true parameters. Such asymptotic distributions are given in Theorems 7.1 and 7.2.

Define

$$\begin{aligned} h &= (b, \gamma_0), \\ H &= \{h = (b, \gamma_0) : b \in R^{d_\beta}, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0\}, \\ W(h) &= \lambda(\pi^*(\gamma_0, b); \gamma_0, b) \text{ for } b \in R^{d_\beta}, \text{ and} \\ T(h) &= \begin{cases} T_\psi(\pi^*(\gamma_0, b); \gamma_0, b) & \text{if } d^*_\pi = 0 \\ T_\pi(\pi^*(\gamma_0, b); \gamma_0, b) & \text{if } d^*_\pi = 1 \end{cases} \end{aligned} \tag{7.26}$$

for $b \in R^{d_\beta}$. As defined, $W(h)$ is the asymptotic distribution of W_n under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $b \in R^{d_\beta}$ determined in Theorem 7.1(a). Also, $T(h)$ is the asymptotic distribution of T_n under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $b \in R^{d_\beta}$ given in Theorem 7.3(a) or 7.3(b) depending on the rank of $r_\pi(\theta)$, which is denoted by d^*_π. Only one of the cases applies for any particular parameter of interest $r(\theta)$ and it is known which applies.

As in (3.8), $AsySz$ denotes the asymptotic size of a CS of nominal level $1-\alpha$. The asymptotic size results use the following distribution function (df) continuity assumption, which typically is not restrictive.

Assumption V3. (i) The df of $W(h)$ is continuous at $\chi^2_{d_r}(1-\alpha)$ and $\sup_{h \in H} C_{W,1-\alpha}(h)$ $\forall h \in H$.
(ii) The df of $T(h)$ is continuous at $z(\alpha/2)$, $z(\alpha)$, $z(1-\alpha)$, $z(1-\alpha/2)$, $\sup_{h \in H} C_{t,1-\alpha}(h)$, $\sup_{h \in H} C_{-t,1-\alpha}(h)$, and $\sup_{h \in H} C_{|t|,1-\alpha}(h)$ $\forall h \in H$.

Theorem 7.4. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1-R2, *and* V1-V3 *hold. Then, the standard nominal* $1-\alpha$ *Wald CS has*

$$AsySz = \min\{\inf_{h\in H} \Pr(W(h) \leq \chi^2_{d_r}(1-\alpha)),\ 1-\alpha\}.$$

Comment. Under Assumption $\mathrm{R_L}$ (i.e., linearity of $r(\theta)$), Theorem 7.4 holds with $W(h)$ replaced by the equivalent, but simpler, quantity $W_L(h) = \lambda_L(\pi^*(\gamma_0, b); \gamma_0, b)$ for $h = (b, \gamma_0)$. This holds by Corollary 7.1(a).

Theorem 7.2 shows that the Wald statistic W_n diverges to infinity in some circumstances, e.g., see Example 6 in Section 7.5 above. In such cases, the standard Wald CS has an asymptotic size equal to 0.

Corollary 7.2. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R_1, *and* V1 *hold. If* $||\eta_n(\widehat{\theta}_n)|| \to_p \infty$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *for some* $\gamma_0 \in \Gamma$ *and* $b \in R^{d_\beta}$, *the standard nominal* $1-\alpha$ *Wald CS has* $AsySz = 0$.

7.8. Asymptotic Size of Standard t Confidence Intervals

Next, we establish the asymptotic size of a standard confidence interval (CI) obtained by inverting a t statistic. The usual symmetric two-sided t CI takes the form

$$CI_{t,n} = \{v : |T_n(v)| \leq z(1-\alpha/2)\}, \tag{7.27}$$

where the t statistic $T_n(v)$ is as in (7.3), $z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of a standard normal distribution, and $1-\alpha$ is the nominal size of the CI. Standard upper and lower one-sided t CI's are obtained by replacing $|T_n(v)|$ with $T_n(v)$ and $-T_n(v)$, respectively, and using $z(1-\alpha)$ and $z(\alpha)$ as the critical values, respectively.

Theorem 7.5. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1, *and* V1-V3 *hold and* $d_r = 1$. *The standard nominal* $1-\alpha$ *symmetric two-sided, upper one-sided, and lower one-sided t CI's have* $AsySz = \min\{\inf_{h\in H} \Pr(|T(h)| \leq z(1-\alpha/2)),\ 1-\alpha\}$, $\min\{\inf_{h\in H} \Pr(T(h) \leq z(1-\alpha)), 1-\alpha\}$, *and* $\min\{\inf_{h\in H} \Pr(T(h) \geq z(\alpha)), 1-\alpha\}$, *respectively.*

7.9. Robust Wald and t Confidence Sets

Now we construct a robust CS for $r(\theta)$ that has correct asymptotic size. First we consider the general case where $r(\theta)$ is a vector, i.e., $d_r \geq 1$, and may be nonlinear in θ. The robust CS for $r(\theta)$ is obtained by inverting a Wald statistic as in (7.25), but with a critical value that is different from $\chi^2_{d_r}(1-\alpha)$.

The new critical value takes into account the fact that the Wald statistic W_n has a non-standard asymptotic distribution under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$. As a result, a larger critical value may be required under weak identification, i.e., $b \in R^{d_\beta}$, than under semi-strong or strong identification, i.e., $||b|| = \infty$. One way to deal with this problem is to construct a "least-favorable" critical value that is large enough for all identification situations. But, the corresponding least-favorable CS typically is overly large and not as informative as desirable when the model is strongly identified.

The robust CS improves upon the least-favorable CS by using a model-selection procedure to choose the critical value. The idea is to use the data to determine whether b is finite. If b is deemed to be finite, i.e., π is deemed to be weakly identified (or unidentified), the least-favorable critical value is applied. Otherwise, the standard critical value is used. This model-selection critical value is analogous to the generalized moment selection method used in Andrews and Soares (2010).

The model-selection procedure chooses between $\mathcal{M}_0 : b \in R^{d_\beta}$ and $\mathcal{M}_1 : ||b|| = \infty$. The statistic used for model selection is

$$A_n = \left(n \widehat{\beta}_n' \widehat{\Sigma}^{-1}_{\beta\beta,n} \widehat{\beta}_n \right)^{1/2}, \tag{7.28}$$

where $\widehat{\Sigma}_{\beta\beta,n}$ is the upper left $d_\beta \times d_\beta$ block of $\widehat{\Sigma}_n$ and $\widehat{\Sigma}_n$ is an estimator of the covariance matrix defined in (7.1). We use A_n to assess the degree of identification.

Let $\{\kappa_n : n \geq 1\}$ be a sequence of constants that diverges to infinity as $n \to \infty$. We call κ_n the tuning parameter. One selects $\mathcal{M}_0$ if $A_n \leq \kappa_n$ and one selects $\mathcal{M}_1$ otherwise. Under $\mathcal{M}_0$, A_n is $O_p(1)$. Hence, one consistently selects $\mathcal{M}_0$ provided the tuning parameter κ_n diverges to infinity. Suitable choices of κ_n include $(\ln n)^{1/2}$ and $(2 \ln \ln n)^{1/2}$, which are analogous to BIC and Hannan-Quinn information criteria, respectively.

Let $C_{W,1-\alpha}(h)$ denote the $1-\alpha$ quantile of $W(h)$ for $h \in H$.

Using the model-selection procedure described above, the robust CS with nominal level $1-\alpha$ is obtained by inverting the Wald statistic with critical value $\widehat{C}_{W,n}(1-\alpha)$,

where

$$\widehat{C}_{W,n}(1-\alpha) = \begin{cases} \max\{\sup_{h\in H} C_{W,1-\alpha}(h),\ \chi^2_{d_r}(1-\alpha)\} & \text{if } A_n \le \kappa_n, \\ \chi^2_{d_r}(1-\alpha), & \text{if } A_n > \kappa_n. \end{cases} \tag{7.29}$$

Assumption K. (i) $\kappa_n \to \infty$ and (ii) $\kappa_n / n^{1/2} \to 0$.

Theorem 7.6. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1-R2, V1-V3, *and* K *hold. Then, the nominal* $1-\alpha$ *robust Wald CS has* $AsySz = 1-\alpha$.

Comment. When $r(\theta)$ is linear in θ, as in Assumption $\mathrm{R_L}$, $C_{W,1-\alpha}(h)$ reduces to the $1-\alpha$ quantile of $W_L(h)$ for $h \in H$.

When $d_r = 1$, a robust CI can be constructed for $r(\theta)$ by inverting the t statistic and applying the model-selection procedure above to choose the critical value.

Let $C_{|t|,1-\alpha}(h)$, $C_{t,1-\alpha}(h)$, and $C_{-t,1-\alpha}(h)$ denote the $1-\alpha$ quantile of $|t(h)|$, $t(h)$, and $-t(h)$ for $h \in H$. The critical values of the symmetric two-sided, upper one-sided, and lower one-sided robust t CI's take the form in (7.29) with $C_{W,1-\alpha}(h)$ replaced by $C_{|t|,1-\alpha}(h)$, $C_{t,1-\alpha}(h)$, and $C_{-t,1-\alpha}(h)$, respectively, and with $\chi^2_{d_r}(1-\alpha)$ replaced by $z(1-\alpha/2)$, $z(1-a)$, and $z(\alpha)$, respectively.

Theorem 7.7. *Suppose Assumptions* A, B1-B3, C1-C8, D1-D3, R1, V1-V3, *and* K *hold and* $d_r = 1$. *Then, the nominal* $1-\alpha$ *symmetric two-sided, upper one-sided, and lower one-sided robust* t *CI's all have* $AsySz = 1-\alpha$.

8. Examples

TO BE ADDED.

REFERENCES

Anderson, T. W. and H. Rubin (1949): "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics,* 21, 570-582.

Andrews, D. W. K. (1988): "Laws of Large Numbers for Dependent Non-identically Distributed Random Variables," *Econometric Theory*, 4, 458-467.

——— (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241-257.

——— (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.

——— (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: North-Holland.

——— (1997): "Estimation When a Parameter Is on a Boundary: Part II," unpublished manuscript, Cowles Foundation, Yale University.

——— (1999): "Estimation When a Parameter Is on a Boundary," *Econometrica*, 67, 1341-1383.

——— (2000): "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica,* 68, 399–405.

——— (2001): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683-734.

——— (2002): "Generalized Method of Moments Estimation When a Parameter Is on a Boundary," *Journal of Business and Economic Statistics*, 20, 530-544.

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2009): "Generic Results for Establishing the Asymptotic Size of Confidence Intervals and Tests," unpublished manuscript, Cowles Foundation, Yale University.

Andrews, D. W. K. and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica,* 77, 721-762.

——— (2010): "Asymptotic Size and a Problem with Subsampling and with the m Out of n Bootstrap," *Econometric Theory,* 26, forthcoming.

Andrews, D. W. K., X. Liu, and W. Ploberger (1998): "Tests for White Noise Against Alternatives with Both Seasonal and Non-seasonal Serial Correlation," *Biometrika,* 85, 727–740.

Andrews, D. W. K. and C. J. McDermott (1995): "Nonlinear Econometric Models with Deterministically Trending Variables," *Review of Economic Studies,* 62, 343–360.

Andrews, D. W. K. and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative," *Econometrica,* 62, 1383–1414.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica,* 78, 119-157.

Andrews, D. W. K. and J. H. Stock (2007): "Inference with Weak Instruments," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society,* Vol. III, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge, UK: Cambridge University Press.

Arcones, M. A. and B. Yu (1994): "Central Limit Theorems for Empirical and U-processes of Stationary Mixing Sequences," *Journal of Theoretical Probability*, 7, 47–71.

Bai, J. (1997): "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics,* 79, 551-563.

Chan, K. S. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *Annals of Statistics,* 21, 520-533.

Chan, K. S. and R. S. Tsay (1998): "Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model," *Biometrika,* 85, 413-426.

Cheng, X. (2008): "Robust Confidence Intervals in Nonlinear Regression under Weak Identification," unpublished working paper, Department of Economics, Yale University.

Chernoff, H. (1954): "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 54, 573-578.

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

Davidson, J. (1994): *Stochastic Limit Theory.* Oxford: Oxford University Press.

Davies, R. B. (1977): "Hypothesis Testing When a Nuisance Parameter Is Present Only Under the Alternative," *Biometrika,* 64, 247-254.

——— (1987): "Hypothesis Testing When a Nuisance Parameter Is Present Only Under the Alternatives," *Biometrika,* 74, 33-43.

de Jong, R. M. (1997): "Central Limit Theorems for Dependent Heterogeneous Random Variables," *Econometric Theory*, 13, 353-367.

Dong, Y. (2009): "Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration," unpublished manuscript, Department of Economics, California State University at Fullerton.

Doukhan, P., P. Massart, and E. Rio (1995): "Invariance Principles for Absolutely Regular Empirical Processes," *Annals of the Institute of Henri Poincaré*, 31, 393–427.

Drton, M. (2009): "Likelihood Ratio Tests and Singularities," *Annals of Statistics,* 37, 979-1012.

Dufour, J.-M. (1997): "Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica,* 65, 1365-1387.

Elliott, G. and U. K. Müller (2007): "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions," *Journal of Econometrics,* 141, 1196-1218.

——— (2008): "Pre and Post Break Parameter Inference," unpublished manuscript, Department of Economics, Princeton University.

Gleser, L. J. and J. T. Hwang (1987): "The Nonexistence of $100(1-\alpha)$ Confidence Sets of Finite Expected Diameter in Errors in Variables and Related Models," *Annals of Statistics,* 15, 1351-1362.

Han, S. (2009): "Identification and Inference in a Binary Probit Model with Weak Instruments," unpublished manuscript, Department of Economics, Yale University.

Hansen, B. E. (1996a): "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica,* 64, 413-430.

——— (1996b): "Stochastic Equicontinuity for Unbounded Dependent Heterogeneous Arrays," *Econometric Theory*, 12, 347-359.

——— (2000): "Sample Splitting and Threshold Estimation," *Econometrica,* 68, 575-603.

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

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

——— (2005): "Testing Parameters in GMM without Assuming That They Are Identified," *Econometrica,* 73, 1103-1123.

Liu, X. and Y. Shao (2003): "Asymptotics for Likelihood Ratio Tests Under Loss of Identifiability," *Annals of Statistics,* 31, 807-832.

Ma, J. and C. R. Nelson (2006): "Valid Inference for a Class of Models Where Standard Inference Performs Poorly; Including Nonlinear Regression, ARMA, GARCH, and Unobserved Components," unpublished manuscript, Department of Economics, U. of Washington.

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

Nelson, C. R. and R. Startz (1990): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator," *Econometrica,* 58, 967-976.

——— (2007): "The Zero-information-limit Condition and Spurious Inference in Weakly Identified Models," *Journal of Econometrics,* 138, 47-62.

Nelson, F. and L. Olson (1978): "Specification and Estimation of a Simultaneous-Equation Model with Limited Dependent Variables," *International Economic Review,* 19, 695-709.

Newey, W. K. (1987): "Efficient Estimation of Limited Dependent Variable Models with Endogenous Regressors," *Journal of Econometrics,* 36, 231-250.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics,* Vol. IV, Ch. 36, ed. by R. F. Engle and D. McFadden. Amsterdam: North-Holland.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057.

Park, J. Y. and P. C. B. Phillips (1999): "Asymptotics for Nonlinear Transformations of Integrated Time Series," *Econometric Theory,* 15, 269-298.

——— (2001): "Nonlinear Regressions with Integrated Time Series," *Econometrica,* 69, 117-161.

Phillips, P. C. B. (1989): "Partially Identified Econometric Models," *Econometric Theory,* 5, 181-240.

Picard, D. (1985): "Testing and Estimating Change-point in Time Series," *Advances in Applied Probability,* 17, 841-867.

Pollard, D. (1984): *Convergence of Stochastic Processes*. New York: Springer–Verlag.

——— (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295-313.

——— (1990): *Empirical Processes: Theory and Applications*. CBMS Conference Series in Statistics and Probability, Vol. 2. Hayward, CA: Institute of Mathematical Statistics.

Rivers, D. and Q. H. Vuong (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics,* 39, 347-366.

Shi, X. and P. C. B. Phillips (2009): "Nonlinear Cointegrating Regression Under Weak Identification," unpublished manuscript, Cowles Foundation, Yale University.

Smith, R. J. and R. W. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica,* 54, 679-685.

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

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Instruments," *Econometrica,* 68, 1055-1096.

Tripathi, G. T. (1999): "A Matrix Extension of the Cauchy-Schwarz Inequality," *Economics Letters,* 63, 1-3.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes.* New York: Springer.