Estimation and Inference with a (Nearly) Singular Jacobian*
[Preliminary and Incomplete]†

Sukjin Han
Department of Economics
University of Texas at Austin
sukjin.han@austin.utexas.edu

Adam McCloskey
Department of Economics
Brown University
adam_mccloskey@brown.edu

First Draft: February 15, 2014
This Draft: May 15, 2015

Abstract

This paper develops extremum estimation and inference results for nonlinear models with very general forms of potential identification failure when the source of this identification failure is known. We examine models that may have a deficient but nonzero rank Jacobian matrix in certain parts of the parameter space. We characterize weak identification in these models by examining sequences of parameters for which the parameter governing the potential identification failure drifts toward the point of identification failure as the sample size grows. This analysis leads to a "local to deficient-rank Jacobian" that does not necessarily have zero rank, allowing us to incorporate many models that have not been previously studied in the weak identification literature. In order to derive the local asymptotic theory for the estimators, the paper introduces a transformation of the parameter space as a key technical step. Asymptotic distributional results for extremum estimators are developed under a comprehensive class of identification strengths and uniformly valid inference procedures robust to identification strength are developed from these results. Importantly, the results allow one to conduct uniformly valid subvector inference. The paper focuses on four examples of models to illustrate the results: sample selection models, models of potential outcomes with endogenous treatment, threshold crossing models, and mixed proportional hazard models.

*The authors are grateful to Donald Andrews, Xiaohong Chen, Xu Cheng, Yanqin Fan, Bruce Hansen, Ivana Komunjer, Eric Renault, Edward Vytlacil, Tiemen Woutersen, and participants at the KAEA Summer Camp 2014 and the Texas Econometrics Camp 2015 for helpful comments. This paper is developed from an earlier work by Han (2009).

†Please do not cite or circulate.

Keywords: Weak identification, underidentification, singular Jacobian matrix, robust inference, uniform inference, sample selection models, bivariate probit models, models for treatment effects, mixed proportional hazard models.
JEL Classification Numbers: C12, C13.

1 Introduction

This paper develops estimation and inference results for nonlinear models with very general forms of potential identification failure when the source of this identification failure is known. A substantial portion of the recent econometrics literature has been devoted to estimation and inference that is robust to the strength of identification of the parameters in an underlying economic or statistical model. Earlier papers in this line of research focused upon the linear instrumental variables model, the behavior of standard estimators and inference procedures under weak identification of this model (e.g., Staiger and Stock, 1997), and the development of new inference procedures robust to the strength of identification in this model (e.g., Kleibergen, 2002 and Moreira, 2003). More recently, focus has shifted to nonlinear models, such as those defined through moment restrictions. In this more general setting, there have similarly been many attempts to characterize the behavior of standard estimators and inference procedures under weak identification (e.g., Stock and Wright, 2000) and to develop robust inference procedures (e.g., Kleibergen, 2005). In nonlinear models, identification failure and weak identification can be characterized via the rank of the Jacobian matrix. The earlier papers in this literature, such as Stock and Wright (2000) and Kleibergen (2005), focused upon special cases of identification failure and weak identification by characterizing how the Jacobian matrix of the underlying model could be (nearly) singular. Only very recently have researchers been able to develop inference procedures that are robust to completely general forms of (near) rank-deficiency in the Jacobian matrix. See Andrews and Mikusheva (2013) in the context of minimum distance estimation and Andrews and Guggenberger (2014) and Andrews and Mikusheva (2014) in the context of moment condition models.

Even though these latter papers develop inference procedures that are completely robust to general forms of the (nearly) singular Jacobian matrix, they are not designed explicitly for models in which the source of identification failure is known to the researcher. In contrast, the recent works of Andrews and Cheng (2012a; 2013; 2014a) develop inference procedures that indeed exploit such knowledge, when it exists, leading to potential gains in terms of the of tests power or volume of confidence sets, as well as the ability to directly conduct inference on a subvector of the parameters of the model. However, the models Andrews and Cheng (2012a;

2013; 2014a) focus upon lead to a specific form of identification failure that corresponds to a Jacobian matrix of rank zero. Yet there are many models that (i) may exhibit non-identification at certain parts of the parameter space for which the source of non-identification is known to the researcher and (ii) have *deficient but nonzero* rank Jacobian matrices at the point of non-identification. This paper characterizes the behavior of standard estimators when the model may be (nearly) unidentified and develops new identification-robust inference procedures for models with these two characteristics.

This paper analyzes the properties of extremum (e.g., generalized method of moments (GMM), maximum likelihood (ML), and minimum distance (MD)) estimators under a comprehensive class of identification strengths and develops inference procedures that are robust to the strength of identification. We characterize the identification strength of a given model in terms of how "close" the model's Jacobian matrix is to singularity. Let $0 = G^*(\theta)$ be a functional relationship between a model's parameters and $J^*(\theta) \equiv \partial G^*(\theta)/\partial\theta$ be its Jacobian matrix. Typically, θ is (locally) identified under the sufficient condition that $rank(J^*(\theta)) = d_\theta$ in a neighborhood of the true parameter (Rothenberg, 1971), where d_θ denotes the dimension of θ. When $rank(J^*(\theta_0)) = 0$ for some θ_0, the function G^* no longer reveals information about θ_0 and the parameter θ is completely unidentified at $\theta = \theta_0$ (under some regularity conditions). Andrews and Cheng (2012a, 2013, 2014a) analyze models of this form. In this paper, we are interested in a more general form of identification failure for which

$$0 \leq rank(J^*(\theta_0)) < d_\theta. \tag{1.1}$$

In cases for which the former inequality is strict, although the parameter θ is not identified at $\theta = \theta_0$, it is also not completely unidentified in the sense that its identification region is not equal to the entire parameter space in which it resides. Indeed, making use of an inverse function theorem due to Hadamard (1906a,b), we characterize the *nonidentification curves* within the parameter space for which parameter values along these curves are observationally equivalent.[1] This situation is sometimes referred to as "underidentification" (see e.g., Arellano et al., 2012).

In examining more general forms of identification failure, we characterize the known source of such failure in the same way as Andrews and Cheng (2012a, 2013, 2014a): when one component of the parameter vector θ is equal to a given value, the Jacobian matrix is singular. This general form of rank-deficient Jacobian at certain parameter values allows us to cover many interesting examples that are not nested in the settings of Andrews and Cheng (2012a, 2013, 2014a). These examples include sample selection models, models of potential outcomes with endogenous

[1] As a related work, see Qu and Tkachenko, 2012 for an example of this type of analysis in the context of macroeconomic DSGE models.

treatment, threshold crossing models with a dummy endogenous variable (e.g., bivariate probit models), mixed proportional hazard models, higher-order ARMA models, and nonlinear regression models. We focus on the first three of these examples to illustrate our approach throughout the paper.

In order to conduct a comprehensive analysis and develop uniformly valid inference procedures, we examine the asymptotic behavior of estimators and test statistics under various sequences of parameters indexing the data generating process (DGP). Under some of these parameter sequences, the component of θ governing the parameter's identification status converges to the value that induces identification failure as the sample size grows. Under such sequences, the Jacobian matrix becomes local-to-singular. This drifting sequence asymptotic device allows us to characterize weak identification in these models, yielding asymptotic distributions of estimators and test statistics that well approximate their finite-sample counterparts. Moreover, the formation of critical values (CVs) for our uniformly valid inference procedures relies on these asymptotic approximations.

Unlike the settings of Andrews and Cheng (2012a, 2013, 2014a), we do not necessarily know which elements of θ are weakly or strongly identified under weak-identification parameter sequences since the Jacobian matrix can be local-to-singular but nonzero rank. This is a key aspect of the problems we examine, making it necessary to develop new asymptotic theory and robust inference procedures. As a crucial step, we transform the parameter space using the nonidentification curves discussed above. One additional complication arises from the fact that, in some models, the nonidentification curves often depend upon the true DGP and must therefore be estimated. The estimation of the curve itself can lead to additional sampling variability in parameter estimates, leading to necessary adjustments in the asymptotic theory under various parameter sequences.

We develop uniformly valid tests and confidence intervals based on the Wald and QLR statistics that are robust to the identification strength of the underlying parameter. We do so by forming data-dependent CVs that adaptively adjust to the identification strength in the model. These CVs are versions of the Type I Robust CV of Andrews and Cheng (2012a) and the adjusted-Bonferroni CV of McCloskey (2012), adapted to the present context. Importantly, our tests and confidence intervals allow one to conduct inference on a subvector of parameters in potentially underidentified models for which such inference was previously unavailable.

The paper is organized as follows. In the next section, we introduce the general class of models under study and provide three examples of models in this class. Section 3 considers identification and the lack of identification, and based on the results derived there, Section 4 introduces the key technical step, i.e., the transformation of the parameter space. Section 5

defines criterion functions of the extremum estimators we examine and shows that a transformed criterion function satisfies a desirable property that is crucial in the subsequent asymptotic theory. Section 6 discusses the three examples in more details. The asymptotic theory under drifting parameter sequences is given in Sections 7–8, and inference in Section 9. Section 12 concludes.

2 Class of Models

Suppose that an economic model implies a relationship among the components of a finite-dimensional parameter $\tilde{\theta}$:

$$0 = \tilde{G}(\tilde{\theta}; \gamma^*) \equiv \tilde{G}^*(\tilde{\theta}) \in \mathbb{R}^{d_{\tilde{G}}} \tag{2.1}$$

when $\tilde{\theta} = \tilde{\theta}^*$, where $\tilde{\theta}^*$ is a component of the true parameter value $\gamma^* = (\tilde{\theta}^*, \tilde{\phi}^*)$ DGP. The function describing this relationship $\tilde{G}$ may depend on the true underlying parameter γ^* of $\gamma = (\tilde{\theta}, \tilde{\phi})$, and thus moment conditions may be involved in defining this relationship. The parameter $\tilde{\phi}$ captures the part of the distribution of the observed data that is not determined by $\tilde{\theta}$, which is typically infinite dimensional (Andrews and Cheng, 2012a). When $\tilde{G}$ does not depend on γ^*, the expression (2.1) specifies a known functional relationship among parameters in $\tilde{\theta}$. An important special case of (2.1) occurs when $\tilde{G}$ does not depend on γ^* and relates a structural parameter $\tilde{\theta}$ to a reduced-form parameter $\tilde{\zeta}$:

$$0 = \tilde{\zeta}^* - \tilde{g}(\tilde{\theta}) \in \mathbb{R}^{d_{\tilde{\zeta}}} \tag{2.2}$$

when $\tilde{\theta} = \tilde{\theta}^*$, where $\tilde{\zeta}^*$ is the true value of $\tilde{\zeta}$.

Oftentimes, econometric models imply a natural decomposition of $\tilde{\theta}$: $\tilde{\theta} = (\alpha, \delta, \tilde{\pi})$, where the parameter α determines the "identification status" of $\tilde{\pi}$. That is, when $\alpha \neq \bar{\alpha}$ for some $\bar{\alpha}$, $\tilde{\pi}$ is identified; when $\alpha = \bar{\alpha}$, $\tilde{\pi}$ is not identified; and when α is "close" to $\bar{\alpha}$ in a certain sense, then $\tilde{\pi}$ is weakly identified. The identification status of the parameter δ is not affected by the value of α, and within the context of (2.2), the same is trivially so for $\tilde{\zeta}$. For convenience, we us the normalization $\bar{\alpha} = 0$, which is without loss of generality.

We present four examples that satisfy the nonzero deficient rank of the Jacobian (1.1). The first two and the last examples fall into the framework of (2.1) and the third into (2.2):

Example 2.1 (Sample selection models)

$$
\begin{aligned}
&Y_i = X_i'\beta_1 + \varepsilon_i, \qquad D_i = \mathbf{1}[\delta + Z_{1i}'\alpha \geq \nu_i], \\
&(\varepsilon_i, \nu_i) \sim BVN\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon\sigma_\nu \\ \rho\sigma_\varepsilon\sigma_\nu & \sigma_\nu^2 \end{array}\right)\right),
\end{aligned}
$$

where $X_i \equiv (1, X_{1i}')'$ *is* $k \times 1$ *and* $Z_i \equiv (1, Z_{1i}')'$ *is* $l \times 1$*. Note that* Z_i *can include (components of)* X_i*. We observe* (D_iY_i, D_i, X_i, Z_i)*. Let* $W_i \equiv (Y_i, X_i, Z_i)$ *and* $\beta_2 \equiv -\rho\sigma_\varepsilon$*. Also normalize* σ_ν^2 *to* 1*. Then, we have*

$$
0 = \tilde{G}^*(\tilde{\theta}) = E_{\gamma^*}\varphi(W_i, \tilde{\theta}) \equiv E_*\varphi(W_i, \tilde{\theta}) \tag{2.3}
$$

when $\tilde{\theta} = \tilde{\theta}^*$*, where the moment function is*

$$
\varphi(w, \tilde{\theta}) = \left[\begin{array}{c} d\left[\begin{array}{c} x \\ \lambda(\delta + z'\alpha) \end{array}\right][y - x'\beta_1 - \beta_2\lambda(\delta + z'\alpha)] \\ \lambda(\delta + z'\alpha)\Phi^{-1}(-\delta - z'\alpha)\left[d - \Phi(\delta + z'\alpha)\right] z \end{array}\right],
$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ *is the inverse Mill's ratio.* □

Example 2.2 (Models of potential outcomes and endogenous treatment)

$$
\begin{aligned}
&\left.\begin{array}{l} Y_{1i} = X_i'\beta_1 + \varepsilon_{1i} \\ Y_{0i} = X_i'\beta_0 + \varepsilon_{0i} \end{array}\right., \qquad D_i = \mathbf{1}[\delta + Z_i'\alpha \geq \nu_i], \\
&Y_i = D_iY_{1i} + (1 - D_i)Y_{0i}, \\
&(\varepsilon_{1i}, \varepsilon_{0i}, \nu_i) \sim MVN(0_3, \Sigma_{10\nu}),
\end{aligned}
$$

and we observe (Y_i, D_i, X_i, Z_i)*. A Roy model (Heckman and Honore, 1990) is a special case of this model of regime switching. This model is slightly more general than the model in Example 2.1, but similar in the aspects that this paper focuses upon.* □

Example 2.3 (Threshold crossing models with a dummy endogenous variable)

$$
\left.\begin{array}{l} Y_i = \mathbf{1}[\beta_1 + \pi D_i - \varepsilon_i \geq 0] \\ D_i = \mathbf{1}[\delta + \alpha Z_i - \nu_i \geq 0] \end{array}\right., \quad \left(\begin{array}{c} \varepsilon_i \\ \nu_i \end{array}\right) \sim F_{\varepsilon_i\nu_i}(\varepsilon_i, v_i; \beta_3).
$$

where $Z_i \in \{0, 1\}$*. The model can be generalized by including common exogenous covariates* X_i *presented in both equations and allowing the instrument* Z_i *to take more than two values. We focus on this stylized version of the model in this paper only for simplicity. With*

$F_{\varepsilon\nu}(\varepsilon,\nu;\beta_3)=\Phi(\varepsilon,\nu;\beta_3)$, a bivariate normal distribution, the model becomes the usual bivariate probit model. *The model with $F_{\varepsilon\nu}(\varepsilon,\nu;\beta_3)=C(F_\varepsilon(\varepsilon),F_\nu(\nu);\beta_3)$, for $C(\cdot,\cdot;\beta_3)$ in a class of single parameter copulas, is considered in Han and Vytlacil (2015). Normalize F_v and F_ε to be uniform distributions for simplicity and let $\beta_2\equiv\beta_1+\pi$. Following Han and Vytlacil (2015), we assume that α and δ are identified from the D equation. The remaining informative fitted probabilities are*

$$\begin{aligned}
p_{11,0} &= C(\beta_2,\delta;\beta_3),\\
p_{11,1} &= C(\beta_2,\delta+\alpha;\beta_3),\\
p_{10,0} &= \beta_1 - C(\beta_1,\delta;\beta_3),\\
p_{10,1} &= \beta_1 - C(\beta_1,\delta+\alpha;\beta_3),\\
p_{01,0} &= \delta - C(\beta_2,\delta;\beta_3),\\
p_{01,1} &= \delta+\alpha - C(\beta_2,\delta+\alpha;\beta_3),
\end{aligned}$$

where $p_{yd,z}\equiv\Pr[Y=y,D=d|Z=z]$. Then, we have

$$0=\tilde{\zeta}^*-\tilde{g}(\tilde{\theta})=\begin{bmatrix} p_{11,0}\\ p_{11,1}\\ p_{10,0}\\ p_{10,1}\\ p_{01,0}\\ p_{01,1}\end{bmatrix}-\begin{bmatrix} C(\beta_2,\delta;\beta_3)\\ C(\beta_2,\delta+\alpha;\beta_3)\\ \beta_1-C(\beta_1,\delta;\beta_3)\\ \beta_1-C(\beta_1,\delta+\alpha;\beta_3)\\ \delta-C(\beta_2,\delta;\beta_3)\\ \delta+\alpha-C(\beta_2,\delta+\alpha;\beta_3)\end{bmatrix} \tag{2.4}$$

when $\tilde{\theta}=\tilde{\theta}^*$, where $\tilde{\zeta}^*$ and $\tilde{g}(\tilde{\theta})$ are defined implicitly. □

Example 2.4 (Mixed proportional hazard models)

$$f(t|X,U;\alpha,\beta_1,\beta_2)=\lambda(t;\alpha,\beta_1)e^{\beta_2'X}e^U,$$

where $\lambda(t;\alpha,\beta_1)=\beta_1(t+\alpha)^{\beta_1-1}$ is the baseline hazard. The form of $\lambda(t;\alpha,\beta_1)$ is the translated Weibull distribution introduced by Ridder and Woutersen (2003), motivated by the well-known fact that with $\lambda(t;\alpha,\beta_1)$ being the Weibull distribution, the information matrix is singular (Hahn (1994)). Note that it is when $\alpha=0$ in the translated Weibull distribution. In this example, we have

$$0=\tilde{G}^*(\tilde{\theta})=E_*s(T,X,U;\alpha,\beta_1,\beta_2),$$

where $s(T, X, U; \alpha, \beta_1, \beta_2)$ *is the efficient score of* $f(t|X, U; \alpha, \beta_1, \beta_2)$. □

In Example 2.1, with $\tilde{\theta} = (\alpha, \delta, \beta_1, \beta_2)$, the Jacobian matrix satisfies (1.1):

$$J^*(\tilde{\theta}) = E_* \begin{bmatrix} -\beta_2 D_i X_i \lambda_{1i} Z_i' & -D_i X_i X_i' & -D_i \lambda_i X_i \\ D_i Y_i \lambda_{1i} Z_i' - D_i X_i' \beta_1 \lambda_{1i} Z_i' - 2\beta_2 D_i \lambda_i \lambda_{1i} Z_i' & -D_i \lambda_i X_i' & -D_i \lambda_i^2 \\ L_i(\alpha, \delta) Z_i Z_i' & 0_{l \times k} & 0_{l \times 1} \end{bmatrix},$$

where $\Phi_i \equiv \Phi(\delta + Z_{1i}'\alpha)$, $\phi_i \equiv \phi(\delta + Z_{1i}'\alpha)$, $\lambda_i \equiv \lambda(\delta + Z_{1i}'\alpha)$, $\lambda_{1i} \equiv d\lambda(x)/dx|_{x=\delta+Z_{1i}'\alpha}$, and

$$L_i(\alpha, \delta) \equiv \frac{\{\lambda_{1i}(D_i - \Phi_i) - \lambda_i \phi_i\}(1 - \Phi_i) + \lambda_i \phi_i (D_i - \Phi_i)}{(1 - \Phi_i)^2}.$$

Note that $rank(J^*(\tilde{\theta})) < d_{\tilde{\theta}}$ when $\alpha = 0$, since λ_i becomes a constant and $X_i = (1, X_{1i}')'$. Also, in Example 2.3, note that $0 < rank(J^*(\tilde{\theta})) < d_{\tilde{\theta}}$ when $\alpha = 0$.[2] This nonzero yet rank-deficient Jacobian when $\alpha = 0$ poses several challenges that make the existing asymptotic theory in the literature that considers a zero rank Jacobian when $\alpha = 0$ inapplicable here: (i) given (1.1), it is not known which components among $\tilde{\pi}$ are not identified; (ii) key assumptions in the literature, such as Assumption A in Andrews and Cheng (2012a), do not hold; (iii) typically, $G^*(\tilde{\theta})$ or $J^*(\tilde{\theta})$ is nonlinear in α. In what follows, we develop a framework to tackle these challenges and to obtain local asymptotic theory and inference procedures.

3 Identification and Lack of Identification

In this section, we formalize the class of problems we are interested in and introduce the main technical setup used in this paper. This section and the next are presented based on (2.1). The results with the special case (2.2) automatically follow without the asterisk sign in all the expressions below. Recall $\gamma = (\tilde{\theta}, \tilde{\phi})$ with $\tilde{\theta} = (\alpha, \delta, \tilde{\pi})$. Let Γ and $\tilde{\Theta}$ be the parameter spaces of γ and $\tilde{\theta}$, respectively. Let $\gamma_0 \equiv (\tilde{\theta}_0, \phi_0)$ and $\tilde{\theta}_0 \equiv (\alpha_0, \delta_0, \tilde{\pi}_0)$. Later we define a sequence of parameters that converge to these points. Let $\tilde{G}_0(\tilde{\theta}) \equiv \tilde{G}(\tilde{\theta}; \gamma_0)$.

Assumption 1 $\tilde{G}_0(\tilde{\theta}) = 0$ *holds at* $\tilde{\theta} = \tilde{\theta}_0$.

Assumption 2 *The function* $\tilde{G}_0 : \tilde{\Theta} \to \mathbb{R}^{d_{\tilde{G}}}$ *is continuously differentiable in* $\tilde{\theta}$ $\forall \gamma_0 \in \Gamma$.

Assumption 3 $\tilde{\theta}_0$ *is a regular point of the matrix* $\partial \tilde{G}_0(\tilde{\theta})/\partial \tilde{\pi}$.

[2] The explicit expression of $J^*(\tilde{\theta})$ can be found later in Section 6.

That is, there exists an open neighborhood of $\tilde{\theta}_0$ in which $\partial \tilde{G}_0(\tilde{\theta})/\partial \tilde{\pi}$ has contant rank. The next two assumptions define the role of α.

Assumption 4 *When* $\alpha_0 \neq 0$, *rank* $\left(\partial \tilde{G}_0(\tilde{\theta})/\partial \tilde{\pi}\right) = d_{\tilde{\pi}}$ *at* $\tilde{\theta} = \tilde{\theta}_0$ $\forall \gamma_0 = (\tilde{\theta}_0, \tilde{\phi}_0) \in \Gamma$.

Assumption 4 is closely related to the identification condition for $\tilde{\pi}$. For local identification of $\tilde{\pi}$, a local version of Assumption 4 is sufficient. For global identification, however, other regularity conditions are also needed; see, e.g., the global inverse function theorem in Han and Vytlacil (2015).

Assumption 5 *When* $\alpha_0 = 0$, *rank* $\left(\partial \tilde{G}_0(\tilde{\theta})/\partial \tilde{\pi}\right) = r < d_{\tilde{\pi}}$ *at* $\tilde{\theta} = \tilde{\theta}_0 = (0, \delta_0, \tilde{\pi}_0)$ $\forall \gamma_0 = (\tilde{\theta}_0, \tilde{\phi}_0) \in \Gamma$.

By allowing $r > 0$, Assumption 5 presents the key aspect of the problem of this paper: *a general form of deficient rank Jacobian.* This condition yields the lack of identification of $\tilde{\pi}$ in the sense of the lack of first-order identification by Sargan (1983). We follow this lack of identification concept throughout the paper. This concept is also used in relating the lack of identification with a criterion function in estimation; see, e.g., Theorem 5.2 below.

When $\alpha_0 = 0$, *not all* the parameters among $\tilde{\pi}$ are identified. Except in the special case of zero rank, in the case of a general deficient rank Jacobian, we typically do not know which parameters among $\tilde{\pi}$ are identified and which are not. Specifically, this happens when the Jacobian with general deficient rank does not have a simple form with zero columns. This motivates us to proceed as follows.

Assumption 6 *For* r*-dimensional subvectors* $\pi^1 \in \Pi^1$ *of* $\tilde{\pi} = (\pi^1, \pi^0)$ *and* G^1 *of* $\tilde{G} = (G^{1\prime}, G^{0\prime})'$, *it satisfies that*

$$rank(\partial G_0^1(\tilde{\theta})/\partial \pi^1) = r \tag{3.1}$$

at $\tilde{\theta} = \tilde{\theta}_0 = (0, \delta_0, \tilde{\pi}_0)$ $\forall \tilde{\gamma}_0 = (\tilde{\theta}_0, \phi_0) \in \tilde{\Gamma}$. *When* $d_\delta + d_{\pi^0} < r$, *there exist a reduced-form parameter* $\zeta^1 \in \mathcal{Z}^1$ *with* $d_{\zeta^1} \geq r - (d_\delta + d_{\pi^0})$ *such that*

$$0 = G_0^1(\tilde{\theta}) = \tilde{G}_0^1(\tilde{\theta}, \zeta_0^1). \tag{3.2}$$

Note that the existence of ζ^1 is not necessary. When it is, ζ_0^1 is an element of $\tilde{\phi}_0$ in $\gamma_0 = (\tilde{\theta}_0, \tilde{\phi}_0)$. Note that Examples 2.1–2.3 satisfy Assumption 6; see Section 6. This assumption trivially holds for cases described by (2.2), i.e., when $\tilde{G}^*(\tilde{\theta}) = \tilde{\zeta}^* - \tilde{g}(\tilde{\theta})$. In this case, ζ^1 in $\tilde{G}^1(\tilde{\theta}, \zeta^1) = \zeta^1 - g^1(\tilde{\theta})$ is a r-dimensional subvector of $\tilde{\zeta} = (\zeta^1, \zeta^0)$. Note that $r < d_{\tilde{\zeta}}$ since $r < d_{\tilde{\pi}}$. The following lemma is crucial to the main technical step of our paper. For a given

$\epsilon > 0$, define a *local parameter space* $\tilde{\Theta}_\epsilon \equiv \left\{\tilde{\theta} \in \tilde{\Theta} : \|\tilde{\theta} - \tilde{\theta}_0\| < \epsilon\right\}$ around the nonidentification region $\tilde{\theta}_0 \equiv (\alpha_0, \delta_0, \tilde{\pi}_0) = (0, \delta_0, \tilde{\pi}_0)$, and $\mathcal{Z}^1_\varepsilon \equiv \left\{\zeta^1 \in \mathcal{Z}^1 : \left\|\zeta^1 - \zeta^1_0\right\| < \varepsilon\right\}$. Also, define χ to be either $\chi \equiv (\delta, \pi^0)$ or $\chi \equiv (\delta, \zeta^1, \pi^0)$ in $\mathcal{X} \equiv \left\{\chi : \tilde{\theta} \in \tilde{\Theta}, \zeta^1 \in \mathcal{Z}^1\right\}$.

Lemma 3.1 *Under Assumptions 1–6, the following holds* $\forall \gamma_0 = (\tilde{\theta}_0, \tilde{\phi}_0) \in \Gamma$*: There exist an open neighborhood* $\mathcal{X}_\epsilon \equiv \left\{\chi : \tilde{\theta} \in \tilde{\Theta}_\epsilon, \zeta^1 \in \mathcal{Z}^1_\varepsilon\right\}$ *for some* $\epsilon > 0$ *and a continuously differentiable implicit function* $h^1 : \mathcal{X}_\epsilon \to \Pi^1$ *such that*

$$0 = \tilde{G}^1_0\left(0, h^1_0(\chi), \chi\right) \tag{3.3}$$

$\forall \chi \in \mathcal{X}_\epsilon$*, where* $h^1_0(\chi) \equiv h^1(\chi; \gamma_0)$*.*

Proof. Given Assumption 6, the result follows by the implicit function theorem. ■

Assumption 7 *The derivative* $\partial h^1_0(\chi)/\partial \chi$ *has full row rank* $\forall \chi \in \mathcal{X}_\epsilon$*.*

This assumption holds when $d_{\zeta^1} = r$ and $\partial \tilde{G}^1_0/\partial \zeta^1$ has full rank, which agrees with the fact that ζ^1 is assumed to be a reduced-form parameter.

When $\alpha_0 = 0$, by (3.1), there exists a matrix $\underset{(d_{\tilde{G}} - r) \times r}{M}$ such that $M \partial G^{1*}/\partial \tilde{\pi} = \partial G^{0*}/\partial \tilde{\pi}$. This implication is useful in proving Theorem 5.2 below. It is also useful for dealing with the singularity of the asymptotic variance of the estimator due to weak identification; see below. Our goal is to separate the parameters in $\tilde{\pi}$ that are not identified when $\alpha = 0$, presumably of dimension $d_{\tilde{\pi}} - r$, and eventually use a criterion function that does not depend on those parameters when $\alpha = 0$. The partition $\tilde{\pi} = (\pi^1, \pi^0)$ and the inverse function h^1 enable us to do so. More specifically, Lemma 3.1 shows "to what extent" $\tilde{\pi}$ is identified when $\alpha = 0$: the parameters of the model are identified up to π^0 when $\alpha = 0$.

4 Transformation

Define $\theta \equiv (\alpha, \zeta, \pi)$ in its parameter space Θ where (a) $\zeta = (\delta, \zeta^1)$ and $\pi = \pi^0$ if $r \neq 0$ and $\chi = (\delta, \zeta^1, \pi^0)$, (b) $\zeta = \delta$ and $\pi = \pi^0$ if $r \neq 0$ and $\chi = (\delta, \pi^0)$, and (c) $\zeta = \delta$ and $\pi = \tilde{\pi}$ if $r = 0$. Note that $d_\theta = d_{\tilde{\theta}}$. Define $\Theta_\epsilon \equiv \left\{\theta \in \Theta : \tilde{\theta} \in \tilde{\Theta}_\epsilon\right\}$. Also note that $\gamma = (\tilde{\theta}, \tilde{\phi})$ can be rearranged as $\gamma = (\theta, \phi)$, where ϕ is the part of the distribution of the observed data which is not determined by θ. Under Assumptions 1–5, we have the following result:

Theorem 4.1 (Transformation) *Under Assumptions 1–6,* $\forall \gamma^* \in \Gamma$*, there exists a function* $h^* : \tilde{\Theta}_\epsilon \to \Theta_\epsilon$ *such that*

$$\tilde{\theta} = h(\theta; \gamma^*) \equiv h^*(\theta). \tag{4.1}$$

Proof. When $r \neq 0$, the equation (4.1) is given by

$$(\alpha, \delta, \tilde{\pi}) = (\alpha, \delta, h^{1*}(\delta, \zeta^1, \pi^0), \pi^0) = (\alpha, \zeta, h^{1*}(\zeta, \pi), \pi) = h^*(\alpha, \zeta, \pi), \tag{4.2}$$

where $h^{1*}(\delta, \pi^0) \equiv h^1(\delta, \pi^0; \gamma^*)$ and the second equality holds by the simple change of notation in (a) above. The case (b) follows analogously. When $r = 0$, $(\alpha, \delta, \tilde{\pi}) = (\alpha, \zeta, \pi)$ by the change of notation in (c) above. ■

Theorem 4.1 defines a mapping between the transformed parameter $\theta = (\alpha, \zeta, \pi)$ and the original parameter $\tilde{\theta} = (\alpha, \delta, \tilde{\pi})$ within the local parameter space. Note that the transformation may depend on the true value γ^*. When h does not depend on γ^*, as in the case (2.2), this transformation is nothing but the reparametrization of $\tilde{\theta}$ into θ.

Define

$$G(\theta; \gamma) = \tilde{G}(h^*(\theta); \gamma), \tag{4.3}$$

$$g(\theta) = \tilde{g}(h(\theta)), \tag{4.4}$$

where each equation corresponds to the terms in (2.1) and (2.2), respectively. From the previous results, one can show that the Jacobian of $G(\theta; \gamma)$ when $\alpha = 0$ has a simpler expression with zero columns than the Jacobian of $\tilde{G}(\tilde{\theta}; \gamma)$; see Remark 5.3 for details. This simplification illustrates why we introduce the transformation.

5 Criterion Functions

We assume that the estimators of $\tilde{\theta}$ and θ minimize a criterion function. In order to define criterion functions, we define the sample counterparts $\bar{\tilde{G}}(\tilde{\theta})$ and $\bar{G}^1(\tilde{\theta})$ of $\tilde{G}^*(\tilde{\theta})$ and $\tilde{G}^{1*}(\tilde{\theta})$, whose "limits" are $\tilde{G}_0(\tilde{\theta})$ and $\tilde{G}_0^1(\tilde{\theta})$, respectively.

Assumption 8 *$\bar{\tilde{G}}(\tilde{\theta})$ is continuously differentiable and rank $\left(\partial \bar{\tilde{G}}(\tilde{\theta})/\partial \tilde{\pi}\right) = r < d_{\tilde{\pi}}$ a.s. at $\tilde{\theta} = \tilde{\theta}_0 = (0, \delta_0, \tilde{\pi}_0)$.*

This is a sample version of Assumptions 2 and 5. This condition holds for Examples 2.1 and 2.2 and trivially for Example 2.3 as $\bar{\tilde{G}}(\tilde{\theta}) = \hat{\tilde{\zeta}} - \tilde{g}(\tilde{\theta})$. The assumption on the differentiability of $\bar{\tilde{G}}(\tilde{\theta})$ is stronger than, e.g., Assumption GMM2 in Andrews and Cheng (2014a). Then by Assumption 6, $\bar{\tilde{G}}^1$ is implicitly defined in

$$0 = \bar{G}^1(0, \delta, \tilde{\pi}) = \bar{\tilde{G}}^1(0, \delta, \tilde{\pi}, \hat{\zeta}^1), \tag{5.1}$$

where the limit of $\hat{\zeta}^1$ is ζ_0^1. Given Assumption 8, by the sample analogue of (3.3) in Lemma 3.1, there exists a continuously differentiable function $\hat{h}^1 : \mathcal{X}_\epsilon \to \Pi^1$ such that

$$0 = \bar{\tilde{G}}^1(0, \hat{h}^1(\chi), \chi) \tag{5.2}$$

$\forall \chi \in \mathcal{X}_\epsilon$.[3] Then the sample analogue $\hat{h}(\theta)$ of the transformation $h^*(\theta)$ is defined accordingly. In the special case (2.2), $\bar{\tilde{G}}(\tilde{\theta}) = \hat{\tilde{\zeta}} - \tilde{g}(\tilde{\theta})$ and h does not depend on γ^* so that $\hat{h}(\theta) = h^*(\theta) = h(\theta)$ trivially. Lastly, the sample counterpart of $G^*(\theta) = G(\theta; \gamma^*)$ is defined as

$$\bar{G}(\theta) = \bar{\tilde{G}}(\hat{h}(\theta)). \tag{5.3}$$

Let $\tilde{Q}_n(\tilde{\theta})$ be the criterion function of the original parameter $\tilde{\theta}$. Given $\tilde{Q}_n(\tilde{\theta})$, we define the criterion function of the transformed parameter θ as

$$Q_n(\theta) \equiv \tilde{Q}_n(\hat{h}(\theta)). \tag{5.4}$$

Assumption 9 *$\tilde{Q}_n(\tilde{\theta})$ is a function of $\tilde{\theta}$ only through $\bar{\tilde{G}}(\tilde{\theta})$.*

By (5.4), Assumption 9 implies that the transformed criterion function $Q_n(\theta)$ is a function of θ only through $\bar{G}(\theta)$. We show that Assumption 9 is naturally satisfied when we construct GMM/MD or ML criterion function given (2.1) or (2.2). Note that models that generate likelihoods typically involve (2.2) as we see in the introduction.

(a) GMM/MD criterion functions: The original function $\tilde{Q}_n : \tilde{\Theta} \to \mathbb{R}$ is defined as

$$\tilde{Q}_n(\tilde{\theta}) = \left\| A_n \left(\bar{\tilde{G}}(\tilde{\theta}) \right) \right\|^2,$$

where A_n is a weight matrix, and the transformed function around the nonidentification region $Q_n : \Theta_\epsilon \to \mathbb{R}$ is defined as

$$Q_n(\theta) = \tilde{Q}_n(\hat{h}(\theta)) = \left\| A_n \left(\bar{G}(\theta) \right) \right\|^2. \tag{5.5}$$

In the special case (2.2) which typically arises in a MD framework, h no longer depends on the true parameter $\tilde{\gamma}^*$ and

$$Q_n(\theta) = \tilde{Q}_n(h(\theta)) = \left\| A_n(\hat{\tilde{\zeta}} - \bar{g}(\theta)) \right\|^2.$$

[3]Note that under Assumption 5, $0 = G_0^1(\tilde{\theta})$ at $\tilde{\theta} = \tilde{\theta}_0 = (0, \delta_0, \tilde{\pi}_0)$ can be seen as a moment condition where an r-dimensional parameter π^1 is exactly identified, and this is one way to motivate (5.1).

(b) ML criterion functions: Given (2.2), assume that the distribution of the data depends on $\tilde{\theta}$ only through $\tilde{\zeta}$ (Rothenberg (1971)), that is, there exists a function $f^{\dagger}(w;\tilde{\zeta})$ such that

$$f(w;\tilde{\theta}) = f^{\dagger}(w;\tilde{g}(\tilde{\theta})) = f^{\dagger}(w;\tilde{\zeta}). \tag{5.6}$$

Then, the original function $\tilde{Q}_n : \tilde{\Theta} \to \mathbb{R}$ is defined as

$$\tilde{Q}_n(\tilde{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \ln f^{\dagger}\left(W_i, \tilde{g}(\tilde{\theta})\right),$$

where $f^{\dagger}$ satisfies (5.6), and the transformed function $Q_n : \Theta_{\epsilon} \to \mathbb{R}$ is defined as

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln f^{\dagger}\left(W_i, g(\theta)\right). \tag{5.7}$$

Remark 5.1 *Given the existence of $f^{\dagger}(w;\tilde{\zeta})$ in the ML framework, the setting of this paper can be characterized in terms of the information matrix. Let $\mathcal{I}(\tilde{\theta})$ be the $d_{\tilde{\theta}} \times d_{\tilde{\theta}}$ information matrix*

$$\mathcal{I}(\tilde{\theta}) \equiv E\left[\frac{\partial \log f}{\partial \tilde{\theta}}\frac{\partial \log f}{\partial \tilde{\theta}'}\right].$$

Then, the general form of singularity of the Jacobian ($0 \leq rank(\partial\tilde{g}(\tilde{\theta}_0)/\partial\tilde{\theta}) < d_{\tilde{\theta}}$) can be characterized as the general form of singularity of the information matrix ($0 \leq rank(\mathcal{I}(\tilde{\theta}_0)) < d_{\tilde{\theta}}$), since

$$\frac{\partial \log f(w;\tilde{\theta})}{\partial \tilde{\theta}} = \frac{\partial \log f^{\dagger}(w;\tilde{g}(\tilde{\theta}))}{\partial \tilde{\zeta}'}\frac{\partial \tilde{g}(\tilde{\theta})}{\partial \tilde{\theta}}$$

and $\mathcal{I}^{\dagger}(\tilde{\zeta}) \equiv E\left(\partial \log f^{\dagger}/\partial\tilde{\zeta}\right)\left(\partial \log f^{\dagger}/\partial\tilde{\zeta}'\right)$ has full rank.

When $r > 0$, the original criterion function depends on $\tilde{\pi}$ when $\alpha = 0$. Only when $r = 0$ is $\tilde{Q}(0,\delta,\tilde{\pi})$ a constant function of $\tilde{\pi}$. Under the new set of parameters $\theta = (\alpha,\zeta,\pi)$, however, we show that the transformed criterion function does not depend on π when $\alpha = 0$:

Theorem 5.2 *Under Assumptions 1–9, $Q_n(\theta)$ is a constant function of π when $\alpha = 0$ $\forall\theta = (0,\zeta,\pi) \in \Theta_{\epsilon}$.*

Proof. We prove the case of $\chi = (\delta,\zeta^1,\pi^0)$. By Assumption 9, it suffices to consider

$$\bar{\tilde{G}}(\alpha,\delta,\hat{h}^1(\zeta,\pi),\pi) = \begin{bmatrix} \bar{G}^1(\alpha,\delta,\hat{h}^1(\zeta,\pi),\pi) \\ \bar{G}^0(\alpha,\delta,\hat{h}^1(\zeta,\pi),\pi) \end{bmatrix}.$$

When $\alpha = 0$,

$$0 = \bar{G}^1\left(0, \delta, \hat{h}^1(\zeta, \pi), \pi\right),$$

which is a constant function of π. Now we show that, when $\alpha = 0$, $\bar{G}^0\left(\alpha, \delta, \hat{h}^1(\zeta, \pi), \pi\right)$ is a constant function of π, or

$$\frac{\partial \bar{G}^0\left(0, \delta, \hat{h}^1(\zeta, \pi), \pi\right)}{\partial \pi} = 0.$$

By Assumption 8, for some $(d_{\tilde{G}} - r) \times r$ matrix M, $M\partial\bar{G}^1/\partial\tilde{\pi} = \partial\bar{G}^0/\partial\tilde{\pi}$ when $\alpha = 0$ around the neighborhood by Assumption 3, and therefore

$$\begin{aligned}
\frac{\partial \bar{G}^0\left(0, \delta, \hat{h}^1(\zeta, \pi), \pi\right)}{\partial \pi} &= \frac{\partial \bar{G}^0}{\partial \pi^1}\frac{\partial \hat{h}^1}{\partial \pi} + \frac{\partial \bar{G}^0}{\partial \pi} \\
&= -\frac{\partial \bar{G}^0}{\partial \pi^1}\left[\frac{\partial \bar{G}^1}{\partial \pi^1}\right]^{-1}\frac{\partial \bar{G}^1}{\partial \pi} + \frac{\partial \bar{G}^0}{\partial \pi} \\
&= -M\frac{\partial \bar{G}^1}{\partial \pi^1}\left[\frac{\partial \bar{G}^1}{\partial \pi^1}\right]^{-1}\frac{\partial \bar{G}^1}{\partial \pi} + M\frac{\partial \bar{G}^1}{\partial \pi} \\
&= 0,
\end{aligned}$$

where the second equality follows from differentiating (5.2). When h does not depend on γ^*, the same proof goes through after $\hat{h}^1$ is replaced by h^1. ■

Also note that by a similar proof

$$\begin{aligned}
\frac{\partial \bar{G}^0\left(0, \delta, \hat{h}^1(\zeta, \pi), \pi\right)}{\partial \zeta} &= \frac{\partial \bar{G}^0}{\partial \pi^1}\frac{\partial \hat{h}^1}{\partial \zeta} + \begin{bmatrix} \frac{\partial \bar{G}^0}{\partial \delta} \\ 0_r \end{bmatrix} \\
&= \frac{\partial \bar{G}^0}{\partial \pi^1}\left[\frac{\partial \bar{G}^1}{\partial \pi^1}\right]^{-1}\begin{bmatrix} -\frac{\partial \bar{G}^1}{\partial \delta} \\ I_r \end{bmatrix} + \begin{bmatrix} \frac{\partial \bar{G}^0}{\partial \delta} \\ 0_r \end{bmatrix} \\
&= M\frac{\partial \bar{G}^1}{\partial \pi^1}\left[\frac{\partial \bar{G}^1}{\partial \pi^1}\right]^{-1}\begin{bmatrix} -\frac{\partial \bar{G}^1}{\partial \delta} \\ I_r \end{bmatrix} + \begin{bmatrix} M\frac{\partial \bar{G}^1}{\partial \delta} \\ 0_r \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ M \end{bmatrix},
\end{aligned}$$

which implies that $Q_n(\theta)$ is a proper function of ζ even when $\alpha = 0$.

In sum, after the transformation, among the components of $\theta = (\alpha, \zeta, \pi)$, α determines the identification status of θ, ζ is the parameter whose identification is not affected by the value of α, and π is the parameter which is not identified and does not appear in the criterion function when $\alpha = 0$. This transformation facilitates our analysis in two ways: (i) it distinguishes the parameters that are strongly identified from the parameters that are weakly identified when α is close to zero; (ii) it yields criterion functions that do not depend on the weakly identified parameters when $\alpha = 0$. When h does not depend on γ^*, Theorem 5.2 provides the groundwork for asymptotic theory, just as Assumption A does in Andrews and Cheng (2012a). When h depends on γ^*, however, the fact that $Q_n(\theta)$ contains the sampling error of $\hat{h}$ complicates the asymptotic theory.

Here we formally define the estimators of $\tilde{\theta}$ and θ. First, for the transformed parameter θ, it is useful to define the estimator in two steps to facilitate the asymptotic theory. This is because π has different asymptotic behavior compared to $\psi = (\alpha, \zeta)$, which can be anticipated by the results of Theorem 5.2. Define a concentrated estimator $\hat{\psi}_n(\pi) \in \Psi(\pi) \equiv \{\psi : \theta \in \Theta \text{ for some } \pi \in \Pi\}$ of ψ for a given $\pi \in \Pi \equiv \{\pi : \theta \in \Theta\}$ as

$$Q_n^c(\pi) \equiv Q_n(\hat{\psi}_n(\pi), \pi) = \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o(n^{-1}), \tag{5.8}$$

and an estimator $\hat{\pi}_n \in \Pi$ of π as

$$Q_n^c(\hat{\pi}_n) = \inf_{\pi \in \Pi} Q_n^c(\pi) + o(n^{-1}). \tag{5.9}$$

Then $\hat{\theta}_n \equiv \left(\hat{\psi}_n, \hat{\pi}_n\right)$ where $\hat{\psi}_n \equiv \hat{\psi}_n(\hat{\pi}_n)$.

Assumption 10 *$\hat{\theta}_n$ also satisfies $Q_n(\hat{\theta}_n) = \inf_{\theta \in \Theta_\epsilon} Q_n(\theta) + o(n^{-1})$.*

First note that $\hat{\theta}_n$ satisfies

$$Q_n(\hat{\theta}_n) = \inf_{\pi \in \Pi} \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o(n^{-1}) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}).$$

Assumption 10 assumes that $Q_n(\theta)$ reaches its infimum within the nonidentification region Θ_ϵ. This setting is relevant in the weak and semi-strong identification cases defined below. Lastly, we show that $\hat{h}(\hat{\theta}_n)$ can be seen as the estimator of the original parameter $\tilde{\theta}$ around the nonidentification region. Let $\hat{\tilde{\theta}}_n \equiv \hat{h}(\hat{\theta}_n)$.

Lemma 5.1 *Under Assumptions 1–8 and 10,* $\hat{\tilde{\theta}}_n$ *satisfies*

$$\tilde{Q}_n(\hat{\tilde{\theta}}_n) = \inf_{\tilde{\theta}\in\hat{h}(\Theta_\epsilon)} \tilde{Q}_n(\tilde{\theta}) + o(n^{-1}). \tag{5.10}$$

Proof. The result follows by

$$\begin{aligned}
\inf_{\tilde{\theta}\in\hat{h}(\Theta_\epsilon)} \tilde{Q}_n(\tilde{\theta}) &= \inf\left\{\tilde{Q}_n(\tilde{\theta}) : \tilde{\theta}\in\hat{h}(\Theta_\epsilon)\right\} \\
&= \inf\left\{\tilde{Q}_n(\hat{h}(\theta)) : \theta\in\Theta_\epsilon\right\} \\
&= \inf\left\{Q_n(\theta) : \theta\in\Theta_\epsilon\right\} \\
&= \inf_{\theta\in\Theta_\epsilon} Q_n(\theta) \\
&= Q_n(\hat{\theta}_n) + o(n^{-1}) \\
&= \tilde{Q}_n(\hat{h}(\hat{\theta}_n)) + o(n^{-1})
\end{aligned}$$

■

Note that we define the estimator $\hat{\tilde{\theta}}$ of the original parameter $\tilde{\theta}$ on a random set $\hat{h}(\Theta_\epsilon)$. When h is known, then $\hat{h} = h$, so $\hat{h}(\Theta_\epsilon) = \tilde{\Theta}_\epsilon$. Henceforth, for a function $f(x)$, $f_{x_1}(x)$ denotes $\partial f(x)/\partial x_1$, where x_1 is a subvector of x.

Remark 5.3 *Since the derivative* h_θ *has full rank, the rank deficiency of* $\tilde{G}_{\tilde{\theta}}$ *implies that* $G_\theta(\alpha,\zeta,\beta;\gamma_0)$ *also has deficient rank when* $\alpha = 0$*. In fact, note that with* $h_0^1(\zeta,\pi) \equiv h^1(\zeta,\pi;\gamma_0)$*,*

$$G_\pi(\alpha,\zeta,\pi;\gamma_0) = \frac{\partial\tilde{G}(\alpha,\delta,h_0^1(\zeta,\pi),\pi;\gamma_0)}{\partial\pi} = \begin{bmatrix} \frac{\partial G^1}{\partial\pi_1}\frac{\partial h_0^1}{\partial\pi} + \frac{\partial G^1}{\partial\pi} \\ \frac{\partial G^0}{\partial\pi_1}\frac{\partial h_0^1}{\partial\pi} + \frac{\partial G^0}{\partial\pi} \end{bmatrix}$$

becomes a zero matrix when $\alpha = 0$*. This is because, when* $\alpha = 0$*,* $\frac{\partial G^1}{\partial\pi_1}\frac{\partial h_0^1}{\partial\pi} + \frac{\partial G^1}{\partial\pi} = 0$ *and hence* $\frac{\partial G^0}{\partial\pi_1}\frac{\partial h_0^1}{\partial\pi} + \frac{\partial G^0}{\partial\pi} = M\left(\frac{\partial G^1}{\partial\pi_1}\frac{\partial h_0^1}{\partial\pi} + \frac{\partial G^1}{\partial\pi}\right) = 0$ *since* $M\partial G^{1*}/\partial\tilde{\pi} = \partial G^{0*}/\partial\tilde{\pi}$*. The transformation we employ helps establish a Jacobian matrix* G_θ *that has a simpler form of deficient rank than* $\tilde{G}_{\tilde{\theta}}$*, which is useful to the following step.*

Remark 5.4 *The rank deficiency of* $G_\theta(0,\zeta,\beta;\gamma_0)$ *leads to a singular asymptotic variance matrix of the estimators. Therefore, we treat* α *separately in the derivation of asymptotic theory below. Given the result above, consider the following element-wise mean value expansion around*

$\alpha = 0$*:*

$$[G_\pi(\alpha, \zeta, \pi; \gamma_0)]_{jk} = \left[\frac{\partial \tilde{G}(\alpha, \delta, h_0^1(\zeta, \pi), \pi; \gamma_0)}{\partial \pi}\right]_{jk}$$
$$= \frac{\partial}{\partial \alpha'}\left[\frac{\partial \tilde{G}(\alpha^\dagger, \delta, h_0^1(\zeta, \pi), \pi; \gamma_0)}{\partial \pi}\right]_{jk} \alpha \tag{5.11}$$

for $d_\psi \leq j \leq d_{\tilde{\zeta}}$ *and* $d_\psi \leq k \leq d_\theta$*, or*

$$vec\left(G_\pi(\alpha, \zeta, \pi; \gamma_0)\right) = \frac{\partial}{\partial \alpha'} vec\left(G_\pi(\alpha^\dagger, \zeta, \pi; \gamma_0)\right) \alpha$$

or

$$G_\pi(\theta; \gamma_0) = D_{\alpha'} G_\pi(\alpha^\dagger, \zeta, \pi; \gamma_0) \circ \alpha, \tag{5.12}$$

where $D_{\alpha'}[\cdot]$ *is an element-wise matrix derivative and* $\circ$ *is multiplication according to* (5.11)*, and* $\alpha^\dagger$ *lies between* 0 *and* α *which is not necessarily identical across elements. Then by* (5.12)*, one can rewrite the Jacobian as*

$$G_\theta(\theta; \gamma_0) = \left[G_\psi(\theta; \gamma_0) : D_{\alpha'} G_\pi(\theta^\dagger; \gamma_0) \circ \alpha\right]$$

where $\theta^\dagger \equiv (\alpha^\dagger, \zeta, \pi)$*.*

6 Examples

6.1 Sample selection models and models of potential outcomes

We continue to discuss Examples 2.1 and 2.2. Since the two examples share similar features of sample selection, we focus our attention on Example 2.1. Let $\tilde{\pi} = \tilde{\beta} = (\beta_1, \beta_2)$ and $\pi^1 = \beta_1$. Given $\tilde{G}^*(\tilde{\theta}) = E_* \varphi(W_i, \tilde{\theta})$ in (2.3), the Jacobian relevant to our discussions in Sections 3–5 is

$$\tilde{G}^*_{\tilde{\pi}} = -E_* \begin{bmatrix} D_i X_i X_i' & D_i \lambda_i X_i \\ D_i \lambda_i X_i' & D_i \lambda_i^2 \\ 0_{l \times k} & 0_{l \times 1} \end{bmatrix}.$$

When $\alpha = 0$, we have $rank(\tilde{G}^*_{\tilde{\pi}}) = d_{\tilde{\pi}} - 1 = r = k$ since

$$\tilde{G}^*_{\tilde{\pi}}(0, \delta, \tilde{\pi}) = -E_{\tilde{\gamma}^*} \begin{bmatrix} D_i X_i X_i' & \lambda(\delta) D_i X_i \\ \lambda(\delta) D_i X_i' & \lambda^2(\delta) D_i \\ 0_{l \times k} & 0_{l \times 1} \end{bmatrix},$$

and the $(k+1)$-th row is a scalar multiple of the first row since $X_i = (1, X_{1i}')'$. Given

$$G^{*1}(\tilde{\theta}) = E_* \left[D_i X_i Y_i - D_i X_i X_i' \beta_1 - \beta_2 D_i \lambda_i X_i \right],$$

when $\alpha = 0$, note that $0 = G^{*1}(0, \delta, \tilde{\pi})$ is equivalent to

$$\begin{aligned} 0_{k \times 1} &= E_* \left[D_i X_i Y_i - D_i X_i X_i' \beta_1 - \beta_2 \lambda(\delta) D_i X_i \right] \\ &\equiv Q^*_{DXY} - Q^*_{DXX} \beta_1 - \lambda(\delta) \beta_2 Q^*_{DX}. \end{aligned}$$

Observe that Assumptions 6 and 7 hold with $\zeta^{1*} = Q^*_{DXY}$. Also, $rank\left(G^{*1}_{\pi^1}(0, \delta, \tilde{\pi})\right) = r$, and when $\alpha = 0$, $M G^{*1}_{\tilde{\pi}} = G^{*0}_{\tilde{\pi}}$ with M being a $(l+1) \times k$ zero matrix with the (1×1) element being $\lambda(\delta)$. In this example, the function h^{1*} has a closed form solution:

$$h^{1*}(\zeta, \beta_2) = h^{1*}(\delta, \zeta^1, \beta_2) = Q^{*-1}_{DXX} \left(\zeta^1 - \lambda(\delta) \beta_2 Q^*_{DX} \right).$$

6.2 Threshold crossing models with dummy endogenous variable

We now continue to discuss Example 2.3. Let $\tilde{\pi} = \tilde{\beta} = (\beta_1, \beta_2, \beta_3)$ and $\pi^1 = (\beta_1, \beta_2)$. Given $\tilde{G}^*(\tilde{\theta}) = \tilde{\zeta}^* - \tilde{g}(\tilde{\theta})$ from the expression in (2.4), the relevant Jacobian is

$$\tilde{G}^*_{\tilde{\pi}} = -\tilde{g}_{\tilde{\pi}} = -\begin{bmatrix} 0 & C_1(\beta_2, \delta; \beta_3) & C_3(\beta_2, \delta; \beta_3) \\ 0 & C_1(\beta_2, \delta+\alpha; \beta_3) & C_3(\beta_2, \delta+\alpha; \beta_3) \\ 1 - C_1(\beta_1, \delta; \beta_3) & 0 & -C_3(\beta_1, \delta; \beta_3) \\ 1 - C_1(\beta_1, \delta+\alpha; \beta_3) & 0 & -C_3(\beta_1, \delta+\alpha; \beta_3) \\ 0 & -C_1(\beta_2, \delta; \beta_3) & -C_3(\beta_2, \delta; \beta_3) \\ 0 & -C_1(\beta_2, \delta+\alpha; \beta_3) & -C_3(\beta_2, \delta+\alpha; \beta_3) \end{bmatrix}.$$

When $\alpha = 0$, we have $rank(\tilde{G}^*_{\tilde{\pi}}) = d_{\tilde{\pi}} - 1 = r = 2$, since there are only two linearly independent rows in the following Jacobian matrix:

$$\tilde{g}_{\tilde{\pi}} = \begin{bmatrix} 0 & C_1(\beta_2, \delta; \beta_3) & C_3(\beta_2, \delta; \beta_3) \\ 0 & C_1(\beta_2, \delta; \beta_3) & C_3(\beta_2, \delta; \beta_3) \\ 1 - C_1(\beta_1, \delta; \beta_3) & 0 & -C_3(\beta_1, \delta; \beta_3) \\ 1 - C_1(\beta_1, \delta; \beta_3) & 0 & -C_3(\beta_1, \delta; \beta_3) \\ 0 & -C_1(\beta_2, \delta; \beta_3) & -C_3(\beta_2, \delta; \beta_3) \\ 0 & -C_1(\beta_2, \delta; \beta_3) & -C_3(\beta_2, \delta; \beta_3) \end{bmatrix}.$$

The deficient rank of the Jacobian can be easily seen by the fact that when $\alpha = 0$, the fitted probabilities are reduced to

$$\begin{aligned} p_{11,0} &= p_{11,1} = C(\beta_2, \delta; \beta_3), \\ p_{10,0} &= p_{10,1} = \beta_1 - C(\beta_1, \delta; \beta_3), \\ p_{01,0} &= p_{01,1} = \delta - C(\beta_2, \delta; \beta_3), \end{aligned}$$

where the number of equations is less than the number of unknowns. Observe that Assumptions 6 and 7 trivially hold with $0 = \zeta^{1*} - g^1(0, \delta, \tilde{\pi})$, i.e.,

$$0 = \begin{bmatrix} p_{11,0} \\ p_{10,0} \\ p_{01,0} \end{bmatrix} - \begin{bmatrix} C(\beta_2, \delta; \beta_3) \\ \beta_1 - C(\beta_1, \delta; \beta_3) \\ \delta - C(\beta_2, \delta; \beta_3) \end{bmatrix},$$

and therefore $rank\left(g^1_{\pi^1}(0, \delta, \tilde{\pi})\right) = r$. Also when $\alpha = 0$, $Mg^1_{\tilde{\pi}} = \partial g^0_{\tilde{\pi}}$ where $M = I_3$. In this example, the function h^{1*} does not have a closed form solution unless we introduce a copula of a simple form.

6.3 Mixed proportional hazard models

Lastly, we discuss Example 2.4. Let $\tilde{\pi} = (\beta_1, \beta_2)$. When $\alpha = 0$, we have $rank(\tilde{G}^*_{\tilde{\pi}}) = 1 = r$ since

$$\tilde{G}^*_{\tilde{\pi}}(0, \tilde{\pi}) = E_*(1 - SE_*[e^U|S]) \begin{bmatrix} \frac{\beta_2'}{\beta_1^2} X & -\frac{1}{\beta_1} X' \\ 0_{d_{\beta_2} \times 1} & 0_{d_{\beta_2} \times d_{\beta_2}} \end{bmatrix},$$

where

$$S = \Lambda(T, \beta_1^*) e^{\beta_2^{*\prime} X}$$

with $\Lambda(t, \beta_1) = \int_0^t \lambda(s, \beta_1) ds$. Also when $\alpha = 0$, $MG^{*1}_{\tilde{\pi}} = G^{*0}_{\tilde{\pi}}$ with M being a $d_{\beta_2} \times 1$ zero matrix. Note that $G^{1*}(0, \tilde{\pi})$ is linear in β_2. In this example, $d_{\pi^0} > d_{\pi^1}$, and we can verify that Assumptions 6 and 7 hold without the existence of ζ^1.

7 Drifting Sequences of Distributions

We formally characterize a local-to-deficient rank Jacobian by modeling the α parameter as local-to-zero. In doing so, we may fully characterize different strengths of identification, namely,

strong, semi-strong, and weak. Ultimately, we derive asymptotic theory under parameters with different strengths of identification in order to conduct uniformly valid inference robust to identification strength.

Define sets of sequences of parameters $\{\gamma_n\}$ as follows:

$$\begin{aligned}
\Gamma(\gamma_0) &\equiv \{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \to \gamma_0 \in \Gamma\}, \\
\Gamma(\gamma_0, 0, a) &\equiv \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : \alpha_0 = 0 \text{ and } n^{1/2}\alpha_n \to a \in \mathbb{R}_\infty^{d_\alpha}\right\}, \\
\Gamma(\gamma_0, \infty, \omega_0) &\equiv \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : n^{1/2}\|\alpha_n\| \to \infty \text{ and } \frac{\alpha_n}{\|\alpha_n\|} \to \omega_0 \in \mathbb{R}^{d_\alpha}\right\},
\end{aligned}$$

where $\gamma_0 \equiv (\alpha_0, \zeta_0, \pi_0, \phi_0)$ and $\gamma_n \equiv (\alpha_n, \zeta_n, \pi_n, \phi_n)$, and $\mathbb{R}_\infty \equiv \mathbb{R} \cup \{\pm\infty\}$. When $\|a\| < \infty$, $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ are weak identification sequences, otherwise, when $\|a\| = \infty$, they characterize semi-strong identification. Sequences $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ characterize semi-strong identification when $\alpha_n \to 0$, otherwise, when $\lim_{n\to\infty} \alpha_n \neq 0$, they are strong identification sequences.

8 Asymptotic Theory

The asymptotic theory is based on

$$Q_n(\theta) = (\tilde{Q}_n \circ \hat{h})(\theta).$$

We proceed with generic sample and population criterion functions $Q_n(\theta)$ and $Q(\theta)$. They can be GMM, MD, or ML criterion functions. To start the analysis, we make use of Theorem 3.1 of Andrews and Cheng (2012a) to obtain the limit theory for the estimators of θ under weak identification and subsequently obtain the weak identification limit theory for the estimators of the original structural parameter of interest $\tilde{\theta}$.

Proposition 8.1 *Suppose Assumptions 1–8 and Assumptions B1-B3 and C1-C6 of Andrews and Cheng (2012a), adapted to the θ and $Q_n(\theta)$ of this paper, hold. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ with $\|a\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\hat{\psi}_n - \psi_n) \\ \hat{\pi}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau(\pi^*(\gamma_0, a); \gamma_0, a) \\ \pi^*(\gamma_0, a) \end{pmatrix},$$

where

$$\pi^*(\gamma_0, a) \equiv \arg\min_{\pi\in\Pi} -\frac{1}{2}(\tilde{\mathcal{G}}(\pi; \gamma_0) + K(\pi; \gamma_0)a)' H^{-1}(\pi; \gamma_0)(\tilde{\mathcal{G}}(\pi; \gamma_0) + K(\pi; \gamma_0)a),$$

$$\tau(\pi;\gamma_0,a) \equiv -H^{-1}(\pi;\gamma_0)(\tilde{\mathcal{G}}(\pi;\gamma_0)+K(\pi;\gamma_0)a)-(a,0_{d_\zeta})$$

with $(a,0_{d_\zeta}) \in \mathbb{R}^{d_\psi}$, $\pi^*(\gamma_0,a)$ *being a random vector and* $\{\tau(\pi;\gamma_0,a):\pi\in\Pi\}$ *being a Gaussian process. The underlying functions* $H(\pi;\gamma_0)$ *and* $K(\pi;\gamma_0)$ *and Gaussian process* $\tilde{\mathcal{G}}(\pi,\gamma_0)$ *are defined in Assumptions C4(i), C5(ii) and C3 of Andrews and Cheng (2012a), respectively.*

Proof: By Theorem 5.2, Assumptions 1–9 imply Assumption A of Andrews and Cheng (2012a). The remaining conditions of Theorem 3.1 of Andrews and Cheng (2012a) are satisfied by direct assumption. ■

Since the parameter of interest is given by $\tilde{\theta} = h^*(\theta)$, Proposition 8.1 is not directly useful for obtaining distributional approximations for the estimator $\hat{\tilde{\theta}}_n$ of the parameter of interest. However, Lemma 5.1 provides us with sufficient conditions under which $\hat{\tilde{\theta}}_n = \hat{h}(\hat{\theta}_n)$. In conjunction with Proposition 8.1, we can use this representation to obtain the weak identification limit theory for $\hat{\tilde{\theta}}_n$. Recall from Section 5 that in the special case (2.2), the function h^* does not depend upon γ^*, and can therefore be written simply as h and does not need to be estimated. The asymptotic analysis for the distribution of $\hat{\tilde{\theta}}_n$ is simpler in this special case. Moreover, the analysis for the case in which h^* depends upon γ^* builds upon the results obtained for this $h^* = h$ case. We begin this section by analyzing the limiting distribution of $\hat{\tilde{\theta}}_n$ in the former case and move to the latter case in a subsequent subsection. Using the transformed parameter space, redefine the space $\mathcal{X}_\epsilon \equiv \{(\zeta,\pi):\theta=(\alpha,\zeta,\pi)\in\Theta_\varepsilon\}$ for some $\epsilon > 0$.

8.1 Known Transformation

In this subsection, we assume that $r \neq 0$ and $\hat{\tilde{\theta}}_n = (\hat{\alpha}_n,\hat{\zeta}_n,\hat{\pi}^1_n,\hat{\pi}_n) = (\hat{\alpha}_n,\hat{\zeta}_n,h^1(\hat{\zeta}_n,\hat{\pi}_n),\hat{\pi}_n) \equiv h(\hat{\theta}_n)$, where $h^1 : \mathcal{X}_\varepsilon \to \Pi^1$ (and therefore $h(\cdot)$) is a known, nonrandom function. If we can establish the asymptotic distribution of $\hat{\pi}^1_n = h^1(\hat{\zeta}_n,\hat{\pi}_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0,0,a)$, then the joint convergence in distribution of $\hat{\tilde{\theta}}_n$ follows from the joint convergence of Proposition 8.1.

We wish to obtain the asymptotic distribution of $\hat{\pi}^1_n$ that is a function of parameters with different rates of convergence. To do so, we make use of a "rotation" technique that has been used in various similar contexts (e.g., Sargan, 1983; Phillips, 1989; Antoine and Renault, 2009, 2012; Andrews and Cheng, 2014a). By defining a certain rotation of the parameters $\hat{\pi}^1_n$, we are able to characterize which "directions" of $\hat{\pi}^1_n$ converge at different rates and to different limiting processes. This is conducive to subsequent analysis on inference. To proceed, let $h^1_\chi(\chi) \equiv [h^1_\zeta(\chi) : h^1_\pi(\chi)]$ according to a conformable partition. For $\chi \in \mathcal{X}_\epsilon$, let $A(\chi) \equiv [A_1(\chi)' : A_2(\chi)']'$ be an orthogonal $d_{\pi^1} \times d_{\pi^1}$ matrix such that $A_1(\chi)$ is a $(d_{\pi^1} - d^*_\pi) \times d_{\pi^1}$ matrix whose rows span the null space of $h^1_\pi(\chi)'$ and $A_2(\chi)$ is a $d^*_\pi \times d_{\pi^1}$ matrix whose rows span the column space of

$h^1_\pi(\chi)$, where $d^*_\pi \equiv rank(h^1_\pi(\chi))$ for $\chi \in \mathcal{X}_\epsilon$. Define

$$\eta_n(\chi) \equiv \begin{cases} n^{1/2} A_1(\chi) \left\{ h^1(\zeta_n, \pi) - h^1(\zeta_n, \pi_n) \right\} & \text{if } d^*_\pi < d_{\pi^1} \\ 0 & \text{if } d^*_\pi = d_{\pi^1}. \end{cases}$$

We impose the following analog of Assumption R2 of Andrews and Cheng (2014a):

Assumption 11 *$\eta_n(\hat{\chi}_n) \xrightarrow{p} 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ for all $a \in \mathbb{R}^{d_\alpha}$.*

Now we can state the following Corollary to Theorem 5.1 of Andrews and Cheng (2014a) which establishes the asymptotic distribution of $\hat{\hat{\theta}}_n$ under weak identification sequences:

Proposition 8.2 *Suppose Assumptions 1–11 and Assumptions B1–B3 and C1–C6 of Andrews and Cheng (2012a), adapted to the θ and $Q_n(\theta)$ of this paper, hold. Then under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ with $\|a\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha}_n - \alpha_n) \\ \sqrt{n}(\hat{\delta}_n - \delta_n) \\ \sqrt{n} A_1(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \\ A_2(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \\ \hat{\pi}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau^\alpha(\pi^*_{0,a}; \gamma_0, a) \\ \tau^\delta(\pi^*_{0,a}; \gamma_0, a) \\ A_1(\zeta_0, \pi^*_{0,a}) h^1_\zeta(\zeta_0, \pi^*_{0,a}) \tau^\zeta(\pi^*_{0,a}; \gamma_0, a) \\ A_2(\zeta_0, \pi^*_{0,a}) \left\{ h^1(\zeta_0, \pi^*_{0,a}) - h^1(\zeta_0, \pi_0) \right\} \\ \pi^*_{0,a} \end{pmatrix},$$

where $\pi^*_{0,a}$ abbreviates $\pi^*(\gamma_0, a)$ and τ^α, τ^δ and τ^ζ denote the components of τ corresponding to α, δ and ζ.

Proof: First, note that Lemma 3.1 provides that h^1 is continuously differentiable and Assumption 7 provides that $h^1_\chi(\chi)$ is full row rank d_{π^1} for all $\chi \in \mathcal{X}_\epsilon$. This implies that Assumption R1(i)-(ii) of Andrews and Cheng (2014a) holds, where $r(\theta) = h^1(\chi)$ here. Next, note that by differentiating (3.3) in Lemma 3.1 we obtain

$$h^1_\pi = -\left(\frac{\partial G^1}{\partial \pi^1} \right)^{-1} \frac{\partial G^1}{\partial \pi}$$

for $\chi \in \mathcal{X}_\epsilon$ and the analog of Assumption R1(iii) of Andrews and Cheng (2014a) holds in our context as well. Hence, using similar arguments to those in the proof of Theorem 5.1 of Andrews and Cheng (2014a), note that

$$\begin{aligned} h^1(\hat{\chi}_n) - h^1(\chi_n) &= h^1(\hat{\zeta}_n, \hat{\pi}_n) - h^1(\zeta_n, \hat{\pi}_n) + h^1(\zeta_n, \hat{\pi}_n) - h^1(\zeta_n, \pi_n) \\ &= h^1_\zeta(\hat{\chi}_n)(\hat{\zeta}_n - \zeta_n) + (h^1(\zeta_n, \hat{\pi}_n) - h^1(\zeta_n, \pi_n)) + o_p(n^{-1/2}) \end{aligned}$$

by Proposition 8.1 and the continuous differentiability of h^1. Thus,

$$\begin{aligned}
\begin{pmatrix} \sqrt{n}A_1(\hat{\chi}_n)\left\{h^1(\hat{\chi}_n) - h^1(\chi_n)\right\} \\ A_2(\hat{\chi}_n)\left\{h^1(\hat{\chi}_n) - h^1(\chi_n)\right\} \end{pmatrix} &= \begin{pmatrix} \sqrt{n}A_1(\hat{\chi}_n)h^1_\zeta(\hat{\chi}_n)(\hat{\zeta}_n - \zeta_n) \\ A_2(\hat{\chi}_n)\left\{h^1(\zeta_n, \hat{\pi}_n) - h^1(\zeta_n, \pi_n)\right\} \end{pmatrix} \\
&\quad + \begin{pmatrix} \sqrt{n}A_1(\hat{\chi}_n)\left\{h^1(\zeta^1_n, \hat{\pi}_n) - h^1(\zeta^1_n, \pi_n)\right\} \\ A_2(\hat{\chi}_n)h^1_\zeta(\hat{\chi}_n)(\hat{\zeta}_n - \zeta_n) \end{pmatrix} + o_p(1) \\
&= \begin{pmatrix} A_1(\hat{\chi}_n)h^1_\zeta(\hat{\chi}_n)\sqrt{n}(\hat{\zeta}_n - \zeta_n) \\ A_2(\hat{\chi}_n)\left\{h^1(\zeta_n, \hat{\pi}_n) - h^1(\zeta_n, \pi_n)\right\} \end{pmatrix} + o_p(1) \\
&\xrightarrow{d} \begin{pmatrix} A_1(\zeta_0, \pi^*(\gamma_0, a))h^1_\zeta(\zeta_0, \pi^*(\gamma_0, a))\tau^\zeta(\pi^*(\gamma_0, a); \gamma_0, a) \\ A_2(\zeta_0, \pi^*(\gamma_0, a))\left\{h^1(\zeta_0, \pi^*(\gamma_0, a)) - h^1(\zeta_0, \pi_0)\right\} \end{pmatrix},
\end{aligned}$$

where the second equality follows from Assumption 11 and Proposition 8.1 and the weak convergence follows from Proposition 8.1, the continuous differentiability of h^1 and the CMT (see 11.13 Andrews and Cheng, 2014b). Lemma 5.1 then implies

$$\begin{pmatrix} \sqrt{n}A_1(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \\ A_2(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} A_1(\zeta_0, \pi^*(\gamma_0, a))h^1_\zeta(\zeta_0, \pi^*(\gamma_0, a))\tau^\zeta(\pi^*(\gamma_0, a); \gamma_0, a) \\ A_2(\zeta_0, \pi^*(\gamma_0, a))\left\{h^1(\zeta_0, \pi^*(\gamma_0, a)) - h^1(\zeta_0, \pi_0)\right\} \end{pmatrix}$$

and the marginal convergence results for $\hat{\alpha}_n$, $\hat{\delta}_n$ and $\hat{\pi}_n$ follow directly from Proposition 8.1. Finally, note that $\hat{\alpha}_n$, $\hat{\delta}_n$, $\hat{\pi}_n$, $A_1(\hat{\chi}_n)$, $A_2(\hat{\chi}_n)$ and $\hat{\pi}^1_n$ are all continuous functions of $\hat{\psi}(\hat{\pi}_n)$ and $\hat{\pi}_n$ in Proposition 8.1 so that, with the CMT, joint convergence also follows from that proposition. ■

Due to the rotation by $A_1(\hat{\chi}_n)$ and $A_2(\hat{\chi}_n)$, Proposition 8.2 does not directly express the limiting distribution of $\hat{\pi}^1_n$. However, this is easily obtained as a corollary to the proposition. Let $A(\chi)^{-1} = [A^1(\chi) : A^2(\chi)]$, which forms a conformable partition such that $A^1(\chi)$ is a $d_{\pi^1} \times (d_{\pi^1} - d^*_\pi)$ matrix and $A^2(\chi)$ is a $d_{\pi^1} \times d^*_\pi$ matrix.

Corollary 8.1 *Under the assumptions of Proposition 8.2 and* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ *with* $\|a\| < \infty$,

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha}_n - \alpha_n) \\ \sqrt{n}(\hat{\delta}_n - \delta_n) \\ \hat{\pi}^1_n \\ \hat{\pi}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau^\alpha(\pi^*_{0,a}; \gamma_0, a) \\ \tau^\delta(\pi^*_{0,a}; \gamma_0, a) \\ \pi^{1*}_{0,a} \\ \pi^*_{0,a} \end{pmatrix},$$

where

$$\pi^{1*}_{0,a} = \pi^1_0 + A^2(\zeta_0, \pi^*_{0,a})A_2(\zeta_0, \pi^*_{0,a})(h^1(\zeta_0, \pi^*_{0,a}) - h^1(\zeta_0, \pi_0)).$$

Proof: Note that by Proposition 8.2,

$$\begin{aligned}(\hat{\pi}_n^1 - \pi_n^1) &= A^{-1}(\hat{\chi}_n)A(\hat{\chi}_n)(\hat{\pi}_n^1 - \pi_n^1) \\ &= A^1(\hat{\chi}_n)A_1(\hat{\chi}_n)(\hat{\pi}_n^1 - \pi_n^1) + A^2(\hat{\chi}_n)A_2(\hat{\chi}_n)(\hat{\pi}_n^1 - \pi_n^1) \\ &\xrightarrow{d} A^2(\zeta_0, \pi_{0,a}^*)A_2(\zeta_0, \pi_{0,a}^*)(h^1(\zeta_0, \pi_{0,a}^*) - h^1(\zeta_0, \pi_0))\end{aligned}$$

since $A_1(\hat{\chi}_n)(\hat{\pi}_n^1 - \pi_n^1) = O_p(n^{-1/2})$ and $A^1(\hat{\chi}_n) = O_p(1)$. The joint convergence follows immediately from Proposition 8.2. ■

Remark 8.2 *Though $\pi_{0,a}^{1*}$ is the limiting random variable for $\hat{\pi}_n^1$ in Corollary 8.1, including the asymptotic counterpart to the $O_p(n^{-1/2})$ term $A^1(\hat{\chi}_n)A_1(\hat{\chi}_n)(\hat{\pi}_n^1 - \pi_n^1)$ should provide a better approximation to the finite sample distribution of $\hat{\pi}_n^1$. That is, the distributional approximation for $\hat{\pi}_n^1$ by*

$$\pi_{0,a}^{1*} + n^{-1/2}A^1(\zeta_0, \pi_{0,a}^*)A_1(\zeta_0, \pi_{0,a}^*)h_\zeta^1(\zeta_0, \pi_{0,a}^*)\tau^\zeta(\pi_{0,a}^*; \gamma_0, a)$$

should serve better in small samples than that by $\pi_{0,a}^{1*}$.

8.2 Estimated Transformation

In this subsection, we again assume that $r \neq 0$, but now we allow for the (possibly implicit) function $h^1(\cdot)$ describing π^1 in terms of ζ and π to depend upon any parameters γ^* in the true underlying DGP, viz., $\pi^1 = h^1(\zeta, \pi; \gamma^*) \equiv h^{1*}(\zeta, \pi)$. Using the sample analog $\hat{h}_n^1$ of h^{1*}, $\hat{\hat{\theta}}_n = (\hat{\alpha}_n, \hat{\delta}_n, \hat{\pi}_n^1, \hat{\pi}_n) = (\hat{\alpha}_n, \hat{\delta}_n, \hat{h}_n^1(\hat{\zeta}_n, \hat{\pi}_n), \hat{\pi}_n)$, under the conditions of Lemma 5.1.

Notationally, let $h_n^1(\zeta, \pi) \equiv h^1(\zeta, \pi; \gamma_n)$ under drifting sequences of parameters $\{\gamma_n\}$. Also, define

$$Z_n(\pi) = -n^{1/2}(D_{\psi\psi}Q_n(\psi_{0,n}, \pi))^{-1}D_\psi Q_n(\psi_{0,n}, \pi),$$

where $\psi_{0,n} = (0, \zeta_n)$. Note that Assumption C4 of Andrews and Cheng (2012a) and Lemma 9.1(a) of Andrews and Cheng (2012b) imply that $Z_n(\cdot) \Rightarrow \tau(\cdot; \gamma_0, a) + (a, 0_{d_\zeta})$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ with $\|a\| < \infty$. With this empirical process convergence result in mind, we impose the following high level assumption that a joint empirical process convergence result holds for $(\hat{h}_n^1(\cdot), Z_n(\cdot))$:

Assumption 12 *Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ with $\|a\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\hat{h}_n^1(\cdot) - h_n^1(\cdot)) \\ Z_n(\cdot) \end{pmatrix} \Rightarrow \begin{pmatrix} \mathcal{G}(\cdot; \gamma_0) \\ \tau(\cdot; \gamma_0, a) + (a, 0_{d_\zeta}) \end{pmatrix},$$

where $\mathcal{G}(\cdot;\gamma_0)$ is a mean zero Gaussian process indexed by $(\zeta,\pi)=\chi\in\mathcal{X}_\epsilon$ with bounded continuous sample paths and some covariance kernel $\Omega(\chi_1,\chi_2;\gamma_0)$ for $\chi_1,\chi_2\in\mathcal{X}_\epsilon$.

Due to the marginal convergence results for $Z_n(\cdot)$ mentioned above, verification of Assumption 12 only requires verification that an empirical process CLT holds for $\hat{h}_n^1(\cdot)$ and that the weak convergence of $\sqrt{n}(\hat{h}_n^1(\cdot)-h_n^1(\cdot))$ and $Z_n(\cdot)$ occurs jointly.

For an illustration of how to verify this assumption, we return to Example 2.1, letting the dimension of x_i be equal to one for notational simplicity. Here, we have

$$\hat{h}_n^1(\zeta,\pi)=(n^{-1}\sum d_ix_i^2)^{-1}[\zeta^1-\lambda(\delta)\pi(n^{-1}\sum d_ix_i)].$$

Suppose a CLT for triangular arrays holds:

$$\frac{1}{\sqrt{n}}\begin{pmatrix}\sum(d_ix_i^2-E_{\gamma_n}[d_ix_i^2])\\ \sum(d_ix_i-E_{\gamma_n}[d_ix_i])\end{pmatrix}\xrightarrow{d}\begin{pmatrix}Z_1\\ Z_2\end{pmatrix}\overset{d}{\sim}\mathcal{N}(0,\Sigma_{\gamma_0})$$

under $\{\gamma_n\}\in\Gamma(\gamma_0,0,a)$. Then,

$$\begin{aligned}
&\sqrt{n}(\hat{h}_n^1(\zeta,\pi)-h_n^1(\zeta,\pi))\\
&=\zeta^1\sqrt{n}\left(\frac{1}{n^{-1}\sum d_ix_i^2}-\frac{1}{E_{\gamma_n}[d_ix_i^2]}\right)\\
&\quad-\lambda(\delta)\pi\left(\frac{n^{-1/2}\sum d_ix_i}{n^{-1}\sum d_ix_i^2}-\frac{\sqrt{n}E_{\gamma_n}[d_ix_i]}{E_{\gamma_n}[d_ix_i^2]}\right)\\
&=\zeta^1\sqrt{n}\left(\frac{E_{\gamma_n}[d_ix_i^2]-n^{-1}\sum d_ix_i^2}{E_{\gamma_n}[d_ix_i^2](n^{-1}\sum d_ix_i^2)}\right)\\
&\quad-\lambda(\delta)\pi\left[\frac{n^{-1/2}\sum(d_ix_i-E_{\gamma_n}[d_ix_i])}{n^{-1}\sum d_ix_i^2}+\frac{\sqrt{n}E_{\gamma_n}[d_ix_i]}{n^{-1}\sum d_ix_i^2}-\frac{\sqrt{n}E_{\gamma_n}[d_ix_i]}{E_{\gamma_n}[d_ix_i^2]}\right]\\
&=(\zeta^1-\lambda(\delta)\pi E_{\gamma_n}[d_ix_i])\sqrt{n}\left(\frac{E_{\gamma_n}[d_ix_i^2]-n^{-1}\sum d_ix_i^2}{E_{\gamma_n}[d_ix_i^2](n^{-1}\sum d_ix_i^2)}\right)\\
&\quad-\left(\frac{\lambda(\delta)\pi}{n^{-1}\sum d_ix_i^2}\right)n^{-1/2}\sum(d_ix_i-E_{\gamma_n}[d_ix_i])\\
&\Rightarrow(\lambda(\delta)\pi E_{\gamma_0}[d_ix_i]-\zeta^1)\frac{Z_1}{E_{\gamma_0}[d_ix_i^2]^2}-\left(\frac{\lambda(\delta)\pi}{E_{\gamma_0}[x_i^2]}\right)Z_2\equiv\mathcal{G}(\zeta,\pi;\gamma_0).
\end{aligned}$$

The joint convergence of $\hat{h}_n^1(\cdot)$ and $Z_n(\cdot)$ occurs here by a similar argument to that made on p. 25 of Andrews and Cheng (2012b).

Finally, we need to adapt Assumption 11 to the present context under which the true h_n^1

function depends upon the parameter γ_n. To do so, let $h^1_{\chi,n}(\chi) = \partial h^1_n(\chi)/\partial\chi' \equiv [h^1_{\zeta,n}(\chi) : h^1_{\pi,n}(\chi)]$. For $\chi \in \mathcal{X}_\epsilon$, let $A_n(\chi) = [A_{1,n}(\chi)' : A_{2,n}(\chi)']'$ be an orthogonal $d_{\pi^1} \times d_{\pi^1}$ matrix such that $A_{1,n}(\chi)$ is a $(d_{\pi^1} - d^*_\pi) \times d_{\pi^1}$ matrix whose rows span the null space of $h^1_{\pi,n}(\chi)'$ and $A_{2,n}(\chi)$ is a $d^*_\pi \times d_{\pi^1}$ matrix whose rows span the column space of $h^1_{\pi,n}(\chi)$, where $d^*_\pi = rank(h^1_{\pi,n}(\chi))$ for $\chi \in \mathcal{X}_\epsilon$. [Will it be possible to have the following condition in terms of $h^1(\cdot,\cdot;\gamma_0)$?] Define

$$\bar{\eta}_n(\chi) \equiv \begin{cases} n^{1/2} A_{1,n}(\chi) \left\{ h^1_n(\zeta_n, \pi) - h^1_n(\zeta_n, \pi_n) \right\} & \text{if } d^*_\pi < d_{\pi^1} \\ 0 & \text{if } d^*_\pi = d_{\pi^1}. \end{cases}$$

Assumption 10′ $\bar{\eta}_n(\hat{\chi}_n) \xrightarrow{p} 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ *for all* $a \in \mathbb{R}^{d_\alpha}$.

Given Assumptions 10′ and 12, we may now state the distributional convergence result when using an estimated $\hat{h}^1_n$.

Theorem 8.3 *Suppose Assumptions 1–10, 10′, 12 and Assumptions B1-B3 and C1-C6 of Andrews and Cheng (2012a), adapted to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Then under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ *with* $\|a\| < \infty$,

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha}_n - \alpha_n) \\ \sqrt{n}(\hat{\delta}_n - \delta_n) \\ \sqrt{n} A_{1,n}(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \\ A_{2,n}(\hat{\chi}_n)(\hat{\pi}^1_n - \pi^1_n) \\ \hat{\pi}_n \end{pmatrix}$$

$$\xrightarrow{d} \begin{pmatrix} \tau^\alpha(\pi^*_{0,a}; \gamma_0, a) \\ \tau^\delta(\pi^*_{0,a}; \gamma_0, a) \\ A_{1,0}(\zeta_0, \pi^*_{0,a}) \left\{ \mathcal{G}_0(\zeta_0, \pi^*_{0,a}) + h^1_{\zeta,0}(\zeta_0, \pi^*_{0,a}) \tau^\zeta(\pi^*_{0,a}; \gamma_0, a) \right\} \\ A_{2,0}(\zeta_0, \pi^*_{0,a}) \left\{ h^1_0(\zeta_0, \pi^*_{0,a}) - h^1_0(\zeta_0, \pi_0) \right\} \\ \pi^*_{0,a} \end{pmatrix},$$

where $A_{1,0}(\cdot) \equiv A_1(\cdot;\gamma_0)$, $A_{2,0}(\cdot) \equiv A_2(\cdot;\gamma_0)$, $\mathcal{G}_0(\cdot) \equiv \mathcal{G}(\cdot;\gamma_0)$, $h^1_0(\cdot) \equiv h^1(\cdot;\gamma_0)$ *and* $h^1_{\zeta,0}(\cdot) \equiv h^1_\zeta(\cdot;\gamma_0)$ *as abbreviations.*

Proof: Similarly to the proof of Proposition 8.2, Lemma 3.1 and Theorem 4.1 imply that Assumption R1 of Andrews and Cheng (2014a) holds for $r(\theta) = h^1_n(\chi)$ in this context. [Or, again, a condition in terms of $h^1(\cdot,\cdot;\gamma_0)$?] Now, note that Proposition 8.1, and the CMT imply

$$\sqrt{n}(\hat{h}^1_n(\hat{\zeta}_n, \hat{\pi}_n) - h^1_n(\hat{\zeta}_n, \hat{\pi}_n)) \xrightarrow{d} \mathcal{G}(\zeta_0, \pi^*(\gamma_0, a); \gamma_0)$$

under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ since $\hat{\zeta}_n = \zeta_n + O_p(n^{-1/2}) \xrightarrow{p} \zeta_0$ and $\hat{\pi}_n$ is a continuous function of $Z_n(\cdot)$ and $D_{\psi\psi}Q_n(\psi_{0,n}, \cdot)$, which converge jointly with each other and with $\hat{h}_n^1(\cdot)$ by Assumption C4 of Andrews and Cheng (2012a) and Assumption 12. Hence,

$$\begin{aligned}
&\sqrt{n}A_{1,n}(\hat{\chi}_n)\left\{\hat{h}_n^1(\hat{\chi}_n) - h_n^1(\chi_n)\right\} \\
&= A_{1,n}(\hat{\zeta}_n, \hat{\pi}_n)\sqrt{n}\left\{\hat{h}_n^1(\hat{\zeta}_n, \hat{\pi}_n) - h_n^1(\hat{\zeta}_n, \hat{\pi}_n)\right\} + \sqrt{n}A_{1,n}(\hat{\zeta}_n, \hat{\pi}_n)\left\{h_n^1(\hat{\zeta}_n, \hat{\pi}_n) - h_n^1(\zeta_n, \pi_n)\right\} \\
&\xrightarrow{d} A_1(\zeta_0, \pi^*(\gamma_0, a); \gamma_0)\left\{\mathcal{G}(\zeta_0, \pi^*(\gamma_0, a); \gamma_0) + h_\zeta^1(\zeta_0, \pi^*(\gamma_0, a); \gamma_0)\tau^\zeta(\pi^*(\gamma_0, a); \gamma_0, a)\right\}
\end{aligned}$$

by Proposition 8.1, the CMT and (a slightly modified version of) Proposition 8.2. On the other hand,

$$\begin{aligned}
&A_{2,n}(\hat{\chi}_n)\left\{\hat{h}_n^1(\hat{\chi}_n) - h_n^1(\chi_n)\right\} \\
&= A_{2,n}(\hat{\zeta}_n, \hat{\pi}_n)\left\{\hat{h}_n^1(\hat{\zeta}_n, \hat{\pi}_n) - h_n^1(\hat{\zeta}_n, \hat{\pi}_n)\right\} + A_{2,n}(\hat{\zeta}_n, \hat{\pi}_n)\left\{h_n^1(\hat{\zeta}_n, \hat{\pi}_n) - h_n^1(\zeta_n, \pi_n)\right\} \\
&\xrightarrow{d} A_2(\zeta_0, \pi^*(\gamma_0, a); \gamma_0)\left\{h^1(\zeta_0, \pi^*(\gamma_0, a); \gamma_0) - h^1(\zeta_0, \pi_0; \gamma_0)\right\}
\end{aligned}$$

by Proposition 8.1, the CMT, (a slightly modified version of) Proposition 8.2 and the fact that $\hat{h}_n^1(\hat{\zeta}_n, \hat{\pi}_n) - h_n^1(\hat{\zeta}_n, \hat{\pi}_n) = O_p(n^{-1/2})$. Similarly to the proof of Proposition 8.2, the marginal convergence results follow directly from Proposition 8.1. Also, $\hat{\alpha}_n$, $\hat{\delta}_n$, $\hat{\pi}_n$, $A_{1,n}(\hat{\chi}_n)$, $A_{2,n}(\hat{\chi}_n)$ and $\hat{\pi}_n^1$ are all continuous functions of $\hat{\psi}(\hat{\pi}_n)$, $\hat{\pi}_n$ and $\hat{h}_n^1$. In turn, $\hat{\psi}(\hat{\pi}_n)$ and $\hat{\pi}_n$ are continuous functions of $Z_n(\cdot)$, $D_{\psi\psi}Q_n(\psi_{0,n}, \cdot)$ and $\hat{h}_n^1(\cdot)$ (see Andrews and Cheng, 2012b) so that joint convergence of all random variables follows from Proposition 8.1, the CMT, Assumption 12 and Assumption C4 of Andrews and Cheng (2012a). ■

In analogy with the previous subsection, we may directly express the limiting joint distributional behavior of $\hat{\tilde{\theta}}_n$ in the following corollary. Let $A_n(\chi)^{-1} = [A_n^1(\chi) : A_n^2(\chi)]$, which forms a conformable partition such that $A_n^1(\chi)$ is a $d_{\pi^1} \times (d_{\pi^1} - d_\pi^*)$ matrix and $A_n^2(\chi)$ is a $d_{\pi^1} \times d_\pi^*$ matrix. Similarly, let $A_0(\chi)^{-1} = [A_0^1(\chi) : A_0^2(\chi)]$. The proof of the following corollary is very similar to that of Corollary 8.1 and therefore omitted.

Corollary 8.4 *Under the assumptions of Theorem 8.3 and* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ *with* $\|a\| < \infty$,

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha}_n - \alpha_n) \\ \sqrt{n}(\hat{\delta}_n - \delta_n) \\ \hat{\pi}_n^1 \\ \hat{\pi}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau^\alpha(\pi_{0,a}^*; \gamma_0, a) \\ \tau^\delta(\pi_{0,a}^*; \gamma_0, a) \\ \pi_{0,a}^{1*} \\ \pi_{0,a}^* \end{pmatrix},$$

where

$$\pi_{0,a}^{1*} = \pi_0^1 + A_0^2(\zeta_0, \pi_{0,a}^*)A_{2,0}(\zeta_0, \pi_{0,a}^*)(h_0^1(\zeta_0, \pi_{0,a}^*) - h_0^1(\zeta_0, \pi_0)).$$

Remark 8.5 *A similar remark to Remark 8.2 applies here, in the estimated h case. That is,*

$$\pi_{0,a}^{1*} + n^{-1/2}A^1(\zeta_0, \pi_{0,a}^*)A_1(\zeta_0, \pi_{0,a}^*)\{\mathcal{G}_0(\zeta_0, \pi_{0,a}^*) + h_\zeta^1(\zeta_0, \pi_{0,a}^*)\tau^\zeta(\pi_{0,a}^*; \gamma_0, a)\}$$

should serve as a better finite sample approximation to the distribution of $\hat{\pi}_n^1$ than $\pi_{0,a}^{1*}$ alone. This is especially true in light of the estimation error involved with $\hat{h}_n^1$.

9 Inference

In this section, we develop robust inference procedures for the original parameter $\tilde{\theta}$. In order to do so, we first determine the asymptotic properties of test statistics.

9.1 Test Statistics and Asymptotic Distributions

We are interested in general nonlinear hypotheses of the form

$$H_0 : \tilde{r}(\tilde{\theta}) = v \in \tilde{r}(\tilde{\Theta}).$$

9.1.1 Wald Statistics

We begin this section by defining the Wald statistics we will be analyzing. In order to define the test statistics, we make the following two additional assumptions regarding the behavior of the extremum objective function under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ (strong and semi-strong identification sequences). They are identical to Assumptions D2 and D3 of Andrews and Cheng (2012a). [The introduction of B needs to be done with care. See Remark 5.2.]First, let

$$\tilde{B}(\alpha) \equiv \begin{pmatrix} I_{d_\alpha + d_\delta} & 0_{(d_\alpha+d_\delta)\times d_{\tilde{\pi}}} \\ 0_{d_{\tilde{\pi}}\times d_{(d_\alpha+d_\delta)}} & \iota(\alpha)I_{d_{\tilde{\pi}}} \end{pmatrix}, \quad \iota(\alpha) \equiv \begin{cases} \alpha, & \text{if } \alpha\text{is a scalar,} \\ \|\alpha\|, & \text{if } \alpha\text{is a vector.} \end{cases}$$

Second, define $D\tilde{Q}_n(\tilde{\theta}) \in \mathbb{R}^{d_{\tilde{\theta}}}$ as a stochastic generalized first derivative vector of $\tilde{Q}_n(\tilde{\theta})$ and $D^2\tilde{Q}_n(\tilde{\theta}) \in \mathbb{R}^{d_{\tilde{\theta}}\times d_{\tilde{\theta}}}$ as a generalized second derivative matrix of $\tilde{Q}_n(\tilde{\theta})$ that is symmetric and may be stochastic or nonstochastic.

Assumption 13 *Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\tilde{J}_n = \tilde{B}^{-1}(\alpha_n)D^2\tilde{Q}_n(\tilde{\theta}_n)\tilde{B}^{-1}(\alpha_n) \xrightarrow{p} \tilde{J}(\gamma_0) \in \mathbb{R}^{d_{\tilde{\theta}}\times d_{\tilde{\theta}}}$, where $\tilde{J}(\gamma_0)$ is nonsingular and symmetric.*

Assumption 14 *(i) Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $n^{1/2}\tilde{B}^{-1}(\alpha_n)D\tilde{Q}_n(\tilde{\theta}_n) \xrightarrow{d} \tilde{\mathcal{G}}(\gamma_0) \sim \mathcal{N}(0_{d_\theta}, \tilde{V}(\gamma_0))$ *for some symmetric* $d_{\tilde{\theta}} \times d_{\tilde{\theta}}$ *matrix* $\tilde{V}(\gamma_0)$.
(ii) $\tilde{V}(\gamma_0)$ *is positive definite for all* $\gamma_0 \in \Gamma$.

With these assumptions in hand, let

$$\hat{\tilde{\Sigma}}_n \equiv \hat{\tilde{J}}_n^{-1}\hat{\tilde{V}}_n\hat{\tilde{J}}_n^{-1},$$

where $\hat{\tilde{J}}_n$ and $\hat{\tilde{V}}_n$ are estimators of $\tilde{J}(\gamma_0)$ and $\tilde{V}(\gamma_0)$. The Wald statistic for H_0 based upon $\hat{\tilde{\theta}}_n$ can be written as

$$W_n(v) \equiv n(\tilde{r}(\hat{\tilde{\theta}}_n) - v)'(\tilde{r}_{\tilde{\theta}}(\hat{\tilde{\theta}}_n)\tilde{B}^{-1}(\hat{\alpha}_n)\hat{\tilde{\Sigma}}_n\tilde{B}^{-1}(\hat{\alpha}_n)\tilde{r}_{\tilde{\theta}}(\hat{\tilde{\theta}}_n)')^{-1}(r(\hat{\tilde{\theta}}_n) - v),$$

where $\tilde{r}_{\tilde{\theta}}(\tilde{\theta}) \in \mathbb{R}^{d_{\tilde{r}} \times d_{\tilde{\theta}}}$. To obtain the asymptotic distribution of $W_n(v)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$, in addition to Assumptions 13–14 and the assumptions imposed in Proposition 8.2 or Theorem 8.3, we also impose the following assumptions.

Assumption 15 *(i)* $\tilde{r}(\tilde{\theta})$ *is continuously differentiable on* $\tilde{\Theta}$.
(ii) $\tilde{r}_{\tilde{\theta}}(\tilde{\theta})$ *has full row rank* d_r *for all* $\tilde{\theta} \in \tilde{\Theta}$.
(iii) $rank(\tilde{r}_{\tilde{\pi}}(\tilde{\theta})) = d_{\tilde{\pi}}^*$ *for some constant* $d_{\tilde{\pi}}^* \leq \min\{d_r, d_{\tilde{\pi}}\}$ *for all* $\tilde{\theta} \in \tilde{\Theta}_\epsilon \equiv \{\tilde{\theta} \in \tilde{\Theta} : \|\alpha\| < \epsilon\}$ *for some* $\epsilon > 0$.

This assumption is identical to Assumption R1 of Andrews and Cheng (2014a), applied to $\tilde{r}$. For the next assumption, let $\tilde{A}(\tilde{\theta}) = [\tilde{A}_1(\tilde{\theta})' : \tilde{A}_2(\tilde{\theta})']'$ be an orthogonal $d_{\tilde{r}} \times d_{\tilde{r}}$ matrix such that $\tilde{A}_1(\tilde{\theta})$ is a $(d_{\tilde{r}} - d_{\tilde{\pi}}^*) \times d_{\tilde{r}}$ matrix whose rows span the null space of $\tilde{r}_{\tilde{\pi}}(\tilde{\theta})'$ and $\tilde{A}_2(\tilde{\theta})$ is a $d_{\tilde{\pi}}^* \times d_{\tilde{r}}$ matrix whose rows span the column space of $\tilde{r}_{\tilde{\pi}}(\tilde{\theta})$. Let

$$\tilde{\eta}_n(\tilde{\theta}) = \begin{cases} n^{1/2}\tilde{A}_1(\tilde{\theta})\left\{\tilde{r}(\alpha_n, \delta_n, \tilde{\pi}) - \tilde{r}(\alpha_n, \delta_n, \tilde{\pi}_n)\right\}, & \text{if } d_{\tilde{\pi}}^* < d_{\tilde{r}} \\ 0, & \text{otherwise.} \end{cases}$$

Assumption 16 *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$, $\tilde{\eta}_n(\hat{\tilde{\theta}}_n) \xrightarrow{p} 0$ *for all* $a \in \mathbb{R}^{d_\alpha}$.

Next we impose an assumption on the variance matrix estimator $\hat{\tilde{\Sigma}}_n$ that is directly analogous to Assumption V1 of Andrews and Cheng (2014a). As in that, paper, we must treat the cases when α is scalar and when α is a vector separately. When α is scalar, $\hat{\tilde{J}}_n = \hat{\tilde{J}}_n(\hat{\tilde{\theta}}_n)$ and $\hat{\tilde{V}}_n = \hat{\tilde{V}}_n(\hat{\tilde{\theta}}_n)$, where $\sup_{\tilde{\theta} \in \tilde{\Theta}} \|\hat{\tilde{J}}_n(\tilde{\theta}) - \tilde{J}(\tilde{\theta}; \gamma_0)\|, \sup_{\tilde{\theta} \in \tilde{\Theta}} \|\hat{\tilde{V}}_n(\tilde{\theta}) - \tilde{V}(\tilde{\theta}; \gamma_0)\| \xrightarrow{p} 0$ for some nonstochastic $d_{\tilde{\theta}} \times d_{\tilde{\theta}}$ matrix-valued functions $\tilde{J}(\tilde{\theta}; \gamma_0)$ and $\tilde{V}(\tilde{\theta}; \gamma_0)$ such that $\tilde{J}(\tilde{\theta}_0; \gamma_0) = \tilde{J}(\gamma_0)$

and $\tilde{V}(\tilde{\theta}_0;\gamma_0) = \tilde{V}(\gamma_0)$. Finally, the limiting variance matrix under weak identification sequences in this case will equal

$$\tilde{\Sigma}(\tilde{\pi};\gamma_0) = \tilde{\Sigma}(\alpha_0,\delta_0,\tilde{\pi};\gamma_0) \text{ where } \tilde{\Sigma}(\tilde{\theta};\gamma_0) = \tilde{J}^{-1}(\tilde{\theta};\gamma_0)\tilde{V}(\tilde{\theta};\gamma_0)\tilde{J}^{-1}(\tilde{\theta};\gamma_0),$$

evaluated at $\tilde{\pi} = (\pi_{0,a}^{1*}, \pi_{0,a}^{*})$.

For the vector α case, reparameterize α as $(\|\alpha\|,\omega)$, where $\omega = \alpha/\|\alpha\|$ if $\alpha \neq 0$ and define $\omega = 1_{d_\alpha}/\|1_{d_\alpha}\|$ if $\alpha = 0$. Correspondingly, reparameterize $\tilde{\theta}$ as $\tilde{\theta}^+ = (\|\alpha\|,\omega,\delta,\tilde{\pi})$. Let $\hat{\tilde{\theta}}_n^+$ and $\tilde{\theta}_0^+$ be the correspondingly reparameterized versions of $\hat{\tilde{\theta}}_n$ and $\tilde{\theta}_0$. In the vector α case, $\hat{\tilde{J}}_n = \hat{\tilde{J}}_n(\hat{\tilde{\theta}}_n^+)$ and $\hat{\tilde{V}}_n = \hat{\tilde{V}}_n(\hat{\tilde{\theta}}_n^+)$, where $\sup_{\tilde{\theta}^+\in\tilde{\Theta}^+} \|\hat{\tilde{J}}_n(\tilde{\theta}^+) - \tilde{J}(\tilde{\theta}^+;\gamma_0)\|, \sup_{\tilde{\theta}^+\in\tilde{\Theta}^+} \|\hat{\tilde{V}}_n(\tilde{\theta}^+) - \tilde{V}(\tilde{\theta}^+;\gamma_0)\| \xrightarrow{p} 0$ with $\tilde{\Theta}^+ \equiv \{\tilde{\theta}^+ : \tilde{\theta} \in \tilde{\Theta}\}$ for some nonstochastic $d_{\tilde{\theta}} \times d_{\tilde{\theta}}$ matrix-valued functions $\tilde{J}(\tilde{\theta}^+;\gamma_0)$ and $\tilde{V}(\tilde{\theta}^+;\gamma_0)$ such that $\tilde{J}(\tilde{\theta}_0^+;\gamma_0) = \tilde{J}(\gamma_0)$ and $\tilde{V}(\tilde{\theta}_0;^+\gamma_0) = \tilde{V}(\gamma_0)$. For $\pi \in \Pi$, let

$$\omega^*(\pi;\gamma_0,a) = \frac{\tau^\alpha(\pi;\gamma_0,a)}{\|\tau^\alpha(\pi;\gamma_0,a)\|},$$

where $\tau^\alpha(\pi;\gamma_0,a)$ denotes the first d_α entries of $\tau(\pi;\gamma_0,a)$ (defined in Proposition 8.1). The limiting variance matrix under weak identification sequences in this case will equal

$$\tilde{\Sigma}(\tilde{\pi},\omega;\gamma_0) = \tilde{\Sigma}(\|\alpha_0\|,\omega,\delta_0,\tilde{\pi};\gamma_0) \text{ where } \tilde{\Sigma}(\tilde{\theta}^+;\gamma_0) = \tilde{J}^{-1}(\tilde{\theta}^+;\gamma_0)\tilde{V}(\tilde{\theta}^+;\gamma_0)\tilde{J}^{-1}(\tilde{\theta}^+;\gamma_0),$$

evaluated at $\tilde{\pi} = (\pi_{0,a}^{*}, \pi_{0,a}^{1*})$ and $\omega = \omega^*(\pi_{0,a}^{*};\gamma_0,a)$.

Assumption 17 *Assumption V1 of Andrews and Cheng (2014a) holds for $\hat{\tilde{\Sigma}}_n$ and its associated quantities.*

We are now ready to state the result for the asymptotic distribution of the Wald statistic under weak identification sequences but we must first define the quantities that appear in the limiting random variable. First, let $\tilde{\theta}_{0,a}^{*} = (\alpha_0,\delta_0,\pi_{0,a}^{*},\pi_{0,a}^{1*})$, where $\pi_{0,a}^{1*}$ is defined according to either Corollary 8.1 or 8.4, depending on the context. Second, let

$$q^{\tilde{A}}(\tilde{\theta};\gamma_0,a) = \begin{pmatrix} \tilde{A}_1(\tilde{\theta})\tilde{r}_{(\alpha,\delta)}(\tilde{\theta})[\tau^\alpha(\pi;\gamma_0,a) : \tau^\delta(\pi;\gamma_0,a)] \\ \iota(a + \tau^\alpha(\pi;\gamma_0,a))\tilde{A}_2(\tilde{\theta})(\tilde{r}(\tilde{\theta}) - \tilde{r}(\tilde{\theta}_0)) \end{pmatrix}.$$

Third, let

$$\tilde{r}_{\tilde{\theta}}^{\tilde{A}}(\tilde{\theta}) = \begin{pmatrix} \tilde{A}_1(\tilde{\theta})\tilde{r}_{(\alpha,\delta)}(\tilde{\theta}) & 0 \\ 0 & \tilde{A}_2(\tilde{\theta})\tilde{r}_{\tilde{\pi}}(\tilde{\theta}) \end{pmatrix}.$$

Finally, let

$$\bar{\tilde{\Sigma}}(\tilde{\pi};\gamma_0,a)=\begin{cases}\tilde{\Sigma}(\tilde{\pi};\gamma_0) & \text{if } \alpha\text{is scalar}\\ \tilde{\Sigma}(\tilde{\pi},\omega^*(\pi;\gamma_0,a);\gamma_0) & \text{if } \alpha\text{is a vector.}\end{cases}$$

Under a sequence $\{\gamma_n\}$[a weak identification sequence?], we consider the sequence of null hypotheses $H_0:\tilde{r}(\tilde{\theta})=v_n$, where $v_n=\tilde{r}(\tilde{\theta}_n)$.

Proposition 9.1 *Suppose Assumptions 1–10, 13–17 and Assumptions B1-B3 and C1-C6 of Andrews and Cheng (2012a), adapted to the θ and $Q_n(\theta)$ of this paper, hold.*

(i) In the case of known h, if Assumption 11 also holds, under $\{\gamma_n\}\in\Gamma(\gamma_0,0,a)$ with $\|a\|<\infty$,

$$W_n(v_n)\xrightarrow{d}q^{\tilde{A}}(\tilde{\theta}^*_{0,a};\gamma_0,a)'(\tilde{r}^{\tilde{A}}_{\tilde{\theta}}(\tilde{\theta}^*_{0,a})\bar{\tilde{\Sigma}}(\pi^*_{0,a},\pi^{1*}_{0,a};\gamma_0,a)\tilde{r}^{\tilde{A}}_{\tilde{\theta}}(\tilde{\theta}^*_{0,a})')^{-1}q^{\tilde{A}}(\tilde{\theta}^*_{0,a};\gamma_0,a),$$

where $\pi^{1*}_{0,a}$ is defined according to Corollary 8.1.

(ii) In the case of estimated h, if Assumptions 10′ and 12 also hold, under $\{\gamma_n\}\in\Gamma(\gamma_0,0,a)$ with $\|a\|<\infty$,

$$W_n(v_n)\xrightarrow{d}q^{\tilde{A}}(\tilde{\theta}^*_{0,a};\gamma_0,a)'(\tilde{r}^{\tilde{A}}_{\tilde{\theta}}(\tilde{\theta}^*_{0,a})\bar{\tilde{\Sigma}}(\pi^*_{0,a},\pi^{1*}_{0,a};\gamma_0,a)\tilde{r}^{\tilde{A}}_{\tilde{\theta}}(\tilde{\theta}^*_{0,a})')^{-1}q^{\tilde{A}}(\tilde{\theta}^*_{0,a};\gamma_0,a),$$

where $\pi^{1*}_{0,a}$ is defined according to Corollary 8.4.

Proof: We provide the proof for part (ii) only since the proof for part (i) is nearly identical. Begin by noting that under H_0 we can express the Wald statistic as

$$W_n(v_n)=q^{\tilde{A}}_n(\hat{\tilde{\theta}}_n)'(\tilde{r}^{\tilde{A}}_{\tilde{\theta},n}(\hat{\tilde{\theta}}_n)\hat{\tilde{\Sigma}}_n\tilde{r}^{\tilde{A}}_{\tilde{\theta},n}(\hat{\tilde{\theta}}_n)')^{-1}q^{\tilde{A}}_n(\hat{\tilde{\theta}}_n),$$

where

$$q^{\tilde{A}}_n(\hat{\tilde{\theta}}_n)=n^{1/2}B^*(\hat{\alpha}_n)\tilde{A}(\hat{\tilde{\theta}}_n)(\tilde{r}(\hat{\tilde{\theta}}_n)-\tilde{r}(\tilde{\theta}_n))\text{ and }r^{\tilde{A}}_{\tilde{\theta},n}(\hat{\tilde{\theta}}_n)=B^*(\hat{\alpha}_n)\tilde{A}(\hat{\tilde{\theta}}_n)\tilde{r}_{\tilde{\theta}}(\hat{\tilde{\theta}}_n)\tilde{B}^{-1}(\hat{\alpha}_n)$$

with

$$B^*(\hat{\alpha}_n)=\begin{pmatrix}I_{d_{\tilde{r}}-d^*_{\tilde{\pi}}} & 0\\ 0 & \iota(\hat{\alpha}_n)I_{d^*_\pi}\end{pmatrix}.$$

Note that

$$r^{\tilde{A}}_{\tilde{\theta},n}(\hat{\tilde{\theta}}_n)=B^*(\hat{\alpha}_n)\begin{pmatrix}\tilde{A}_1(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n) & 0\\ \tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n) & \tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{\tilde{\pi}}(\hat{\tilde{\theta}}_n)\end{pmatrix}\tilde{B}^{-1}(\hat{\alpha}_n)$$

$$= \begin{pmatrix} \tilde{A}_1(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n) & 0 \\ \iota(\hat{\alpha}_n)\tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n) & \tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{\tilde{\pi}}(\hat{\tilde{\theta}}_n) \end{pmatrix} \xrightarrow{d} \tilde{r}_{\tilde{\theta}}^{\tilde{A}}(\tilde{\theta}_{0,a}^*), \tag{9.1}$$

where the convergence follows from Corollary 8.4, Assumption 14(i) and the CMT since $\iota(\hat{\alpha}_n) = o_p(1)$ by Corollary 8.4 and $\tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n) = O_p(1)$ by Assumption 14(i). Turning to the $q_n^{\tilde{A}}(\hat{\tilde{\theta}}_n)$ term, note that

$$\begin{aligned}\tilde{r}(\hat{\tilde{\theta}}_n) - \tilde{r}(\tilde{\theta}_n) &= \{\tilde{r}(\hat{\alpha}_n, \hat{\delta}_n, \hat{\tilde{\pi}}_n) - \tilde{r}(\alpha_n, \delta_n, \hat{\tilde{\pi}}_n)\} + \{\tilde{r}(\alpha_n, \delta_n, \hat{\tilde{\pi}}_n) - \tilde{r}(\alpha_n, \delta_n, \tilde{\pi}_n)\} \\ &= \tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n)[(\hat{\alpha}_n, \hat{\delta}_n) - (\alpha_n, \delta_n)] + \{\tilde{r}(\alpha_n, \delta_n, \hat{\tilde{\pi}}_n) - \tilde{r}(\alpha_n, \delta_n, \tilde{\pi}_n)\} + o_p(n^{-1/2}),\end{aligned}$$

where the second equality follows from a mean-value expansion, the fact that $(\hat{\alpha}_n, \hat{\delta}_n) - (\alpha_n, \delta_n) = O_p(n^{-1/2})$ by Corollary 8.4 and Assumption 14(i). Hence,

$$q_n^{\tilde{A}}(\hat{\tilde{\theta}}_n) = \begin{pmatrix} n^{1/2}\tilde{A}_1(\hat{\tilde{\theta}}_n)(\tilde{r}(\hat{\tilde{\theta}}_n) - \tilde{r}(\tilde{\theta}_n)) \\ n^{1/2}\iota(\hat{\alpha}_n)\tilde{A}_2(\hat{\tilde{\theta}}_n)(\tilde{r}(\hat{\tilde{\theta}}_n) - \tilde{r}(\tilde{\theta}_n)) \end{pmatrix} = q_{1,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) + q_{2,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) + o_p(1),$$

where

$$\begin{aligned} q_{1,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) &= \begin{pmatrix} n^{1/2}\tilde{A}_1(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n)((\hat{\alpha}_n, \hat{\delta}_n) - (\alpha_n, \delta_n)) \\ n^{1/2}\iota(\hat{\alpha}_n)\tilde{A}_2(\hat{\tilde{\theta}}_n)(\tilde{r}(\alpha_n, \delta_n, \hat{\tilde{\pi}}_n) - \tilde{r}(\alpha_n, \delta_n, \tilde{\pi}_n)) \end{pmatrix} \\ q_{2,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) &= \begin{pmatrix} \tilde{\eta}_n(\hat{\tilde{\theta}}_n) \\ n^{1/2}\iota(\hat{\alpha}_n)\tilde{A}_2(\hat{\tilde{\theta}}_n)\tilde{r}_{(\alpha,\delta)}(\hat{\tilde{\theta}}_n)((\hat{\alpha}_n, \hat{\delta}_n) - (\alpha_n, \delta_n)) \end{pmatrix}. \end{aligned}$$

Note that Assumption 15, the fact that $(\hat{\alpha}_n, \hat{\delta}_n) - (\alpha_n, \delta_n) = O_p(n^{-1/2})$ and $\iota(\hat{\alpha_n}) = o_p(1)$ by Corollary 8.4 and Assumption 14(i) imply that $q_{2,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) = o_p(1)$. Hence,

$$q_n^{\tilde{A}}(\hat{\tilde{\theta}}_n) = q_{1,n}^{\tilde{A}}(\hat{\tilde{\theta}}_n) + o_p(1) \xrightarrow{d} q^{\tilde{A}}(\tilde{\theta}_{0,a}^*; \gamma_0, a) \tag{9.2}$$

by Corollary 8.4, Assumption 14(i) and the CMT. Now, for the case of scalar α,

$$\hat{\tilde{\Sigma}}_n = \hat{\tilde{J}}(\hat{\tilde{\theta}}_n)^{-1}\hat{\tilde{V}}(\hat{\tilde{\theta}}_n)\hat{\tilde{J}}(\hat{\tilde{\theta}}_n)^{-1} \xrightarrow{d} \tilde{\Sigma}(\pi_{0,a}^*, \pi_{0,a}^{1*}; \gamma_0) \tag{9.3}$$

by Assumption 16, Corollary 8.4 and the CMT. The analogous argument holds for the vector α case. Finally, the convergence of (9.1), (9.2) and (9.3) occurs jointly by Corollary 8.4 and the CMT, providing the result of the theorem. ■

9.1.2 QLR Statistics

...

9.2 Confidence Sets

...

9.3 Robust Inference

Let $T_n(v)$ denote a generic test statistic for the null hypothesis for a test of level $\tilde{\alpha}$. For some $a \in \mathbb{R}_\infty^{d_\alpha}$, the limit distribution of $T_n(v_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$, where $v_n = r(h_n(\theta_n))$, provides a good approximation to the finite-sample distribution of $T_n(v)$. Let $\lambda \equiv (a, \gamma_0) \in \Lambda \equiv \{(a, \gamma_0) :$ for some $\{\gamma_n\} \in \Gamma(\gamma_0), n^{1/2}\alpha_n \to a \in \mathbb{R}_\infty^{d_\alpha}\}$ and let $T(\lambda)$ denote the asymptotic distribution of $T_n(v_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$. Let $c_{T,1-\tilde{\alpha}}(\lambda)$ denote the $1-\tilde{\alpha}$ quantile of this distribution. Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $T_n(v_n)$ is assumed to have a standard asymptotic distribution. Let $c_{T,1-\tilde{\alpha}}(\infty)$ denote the $1-\tilde{\alpha}$ quantile of this distribution.

We construct critical values that (uniformly) control the asymptotic size of a test based upon test statistic $T_n(v)$. The first construction is more computationally straightforward while the second leads to tests with more desirable finite-sample properties (in terms of size and power).

9.3.1 ICS Critical Values

The first type of CV is the direct analog of Andrews and Cheng's (2012a) Type I Robust CV. Define $t_n \equiv (n\hat{\alpha}_n' \hat{\Sigma}_{\alpha\alpha,n}^{-1} \hat{\alpha}_n / d_\alpha)^{1/2}$, where $\hat{\Sigma}_{\alpha\alpha,n}$ is a consistent estimator of the asymptotic covariance matrix of $n^{1/2}\hat{\alpha}_n$ under $\{\gamma_n\}$ with $\alpha_0 \neq 0$ (strong identification sequences) and suppose $\{\kappa_n\}$ is a sequence of constants such that $\kappa_n \to \infty$ and $\kappa_n / n^{1/2} \to 0$. Then the ICS CV is defined as follows:

$$\tilde{c}_{T,1-\tilde{\alpha},n} \equiv \begin{cases} c_{T,1-\tilde{\alpha}}^{LF} & \text{if } t_n \leq \kappa_n \\ c_{T,1-\tilde{\alpha}}(\infty) & \text{if } t_n > \kappa_n, \end{cases}$$

where $c_{T,1-\tilde{\alpha}}^{LF} \equiv \sup_{\lambda \in \hat{\Lambda}_n \cap \Lambda(v)} c_{T,1-\tilde{\alpha}}(\lambda)$ with $\hat{\Lambda}_n \equiv \{\lambda = (a, \gamma_0) \in \Lambda : \gamma_0 = (\alpha_0, \hat{\zeta}_n, \pi_0, \phi_0)\}$ and $\Lambda(v) \equiv \{\lambda = (a, \gamma_0) : r(h(\theta_0)) = v\}$. That is, we both impose H_0 and "plug-in" a consistent estimator $\hat{\zeta}_n$ of ζ_0 in the construction of the CV. This leads to tests with smaller CVs and hence better power (see, e.g., Andrews and Cheng, 2012a for a discussion).

9.3.2 Adjusted-Bonferroni Critical Values

The second type of CV is the adjusted-Bonferroni CV of McCloskey (2012). Let $\hat{a}_n = n^{1/2}\hat{\alpha}_n$. Using the asymptotic distributional results of Proposition 8.1, one can form asymptotically valid confidence sets for the vector (a, ζ_0, π_0) using the sample counterpart $(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$. Similarly to the ICS CV in the previous subsection, one may both impose H_0 and "plug-in" for the value of $\hat{\zeta}_n$ since it is consistent for ζ_0. Denote a $(1-b)$-level, estimate-based confidence set for (a, ζ_0, π_0) as $I_b(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n)$. We assume this confidence set has the following properties: (i) $P_{\gamma_n}((a, \zeta_0, \pi_0) \in I_b(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n)) \to 1 - b$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ (asymptotically correct coverage), (ii) $\tilde{I}_b(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n) \equiv \{\lambda = (a, \gamma) \in \Lambda : (a, \zeta, \pi) \in I_b(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n)\} \subset \hat{\Lambda}_n \cap \Lambda(v)$ with probability one (plug-in and null-imposed) and (iii) as a correspondence, $I_b : \mathbb{R}^{d_\alpha}_\infty \times \mathbb{R}^{d_\zeta} \times \mathbb{R}^{d_\pi} \rightrightarrows \mathbb{R}^{d_\alpha}_\infty \times \mathbb{R}^{d_\zeta} \times \mathbb{R}^{d_\pi}$ is continuous. For a given $b \in [0, 1]$, the adjusted-Bonferroni CV is defined as follows. First, compute the largest value $\bar{\alpha} \in [0, \tilde{\alpha}]$ such that the following inequality approximately holds:

$$\sup_{\lambda \in \hat{\Lambda}_n \cap \Lambda(v)} P(T(\lambda) \geq \sup_{\ell \in \tilde{I}_b(\tilde{a}, \zeta_0, \pi^*(a;\gamma_0))} c_{T,1-\bar{\alpha}}(\ell)) \leq \tilde{\alpha},$$

where $\hat{a}_n \xrightarrow{d} a + \tau^\alpha(\pi^*(\gamma_0, a); \gamma_0, a) \equiv \tilde{a}$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, a)$ by Proposition 8.1. This computation can be achieved by simulating from the joint distributions of $(T(\lambda), \tilde{a}, \pi^*(a; \gamma_0))$. The adjusted-Bonferroni CV is then defined as $\sup_{\ell \in \tilde{I}_b(\hat{a}_n, \hat{\zeta}_n, \hat{\pi}_n)} c_{T,1-\bar{\alpha}}(\ell)$. See Algorithm Bonf-Adj in McCloskey (2012) for details on the computation of this CV.

10 Simulations

...

11 Application

...

12 Conclusions

...

References

Andrews, D. W. K., Cheng, X., 2012a. Estimation and inference with weak, semi-strong, and strong identificationith weak, semi-strong, and strong identification. Econometrica 80, 2153–2211. 1, 1, 2, 2, 5, 8, 8.1, 8, 8.2, 8.2, 8.3, 8.2, 9.1.1, 9.1, 9.3.1

Andrews, D. W. K., Cheng, X., 2012b. Supplement to 'estimation and inference with weak, semi-strong and strong identification'. Econometrica Supplementary Material. 8.2, 8.2, 8.2

Andrews, D. W. K., Cheng, X., 2013. Maximum likelihood estimation and uniform inference with sporadic identification failure. Journal of Econometrics 173, 36–56. 1, 1

Andrews, D. W. K., Cheng, X., 2014a. GMM estimation and uniform subvector inference with possible identification failure. Econometric Theory 30, 287–333. 1, 1, 5, 8.1, 8.1, 8.1, 8.2, 9.1.1, 9.1.1, 17

Andrews, D. W. K., Cheng, X., 2014b. Supplementary material on 'GMM estimation and uniform subvector inference with possible identification failure', econometric Theory Supplement. 8.1

Andrews, D. W. K., Guggenberger, P., 2014. Identification- and singularity-robust inference for moment condition models, Cowles Foundation Discussion Paper No. 1978. 1

Andrews, I., Mikusheva, A., 2013. A geometric approach to weakly identified econometric models, Unpublished Manuscript, Department of Economics, Massachusetts Institute of Technology. 1

Andrews, I., Mikusheva, A., 2014. Conditional inference with a functional nuisance parameter, Unpublished Manuscript, Department of Economics, Massachusetts Institute of Technology. 1

Antoine, B., Renault, E., 2009. Efficient GMM with nearly-weak instruments. Econometrics Journal 12, 135–171. 8.1

Antoine, B., Renault, E., 2012. Efficient minimum distance estimation with multiple rates of convergence. Journal of Econometrics 170, 350–367. 8.1

Arellano, M., Hansen, L. P., Sentana, E., 2012. Underidentification? Journal of Econometrics 170, 256–290. 1

Hadamard, J., 1906a. Sur les transformations planes. Comptes Rendus des Seances de l'Academie des Sciences, Paris 74, 142. 1

Hadamard, J., 1906b. Sur les transformations ponctuelles. Bulletin de la Societe Mathematique de France 34, 71–84. 1

Hahn, J., 1994. The efficiency bound of the mixed proportional hazard model. Review of Economic Studies 61, 607–629. 2.4

Han, S., 2009. Identification and inference in a bivariate probit model with weak instruments. Unpublished Manuscript, Department of Economics, Yale University. ∗

Han, S., Vytlacil, E., 2015. Identification in a generalization of bivariate probit models with dummy endogenous regressors, working Paper, University of Texas at Austin and New York University. 2.3, 3

Heckman, J. J., Honore, B. E., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149. 2.2

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental varaibles regression. Econometrica 70, 1781–1803. 1

Kleibergen, F., 2005. Testing parameters in GMM without assuming that they are identified. Econometrica 73, 1103–1123. 1

McCloskey, A., 2012. Bonferroni-based size-correction for nonstandard testing problems, Working Paper, Department of Economics, Brown University. 1, 9.3.2

Moreira, M. J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71, 1027–1048. 1

Phillips, P. C. B., 1989. Partially identified econometric models. Econometric Theory 5, 181–240. 8.1

Qu, Z., Tkachenko, D., 2012. Identification and frequency domain quasi-maximum likelihood estimation of linearized dynamic stochastic general equilibrium models. Quantitative Economics 3, 95–132. 1

Ridder, G., Woutersen, T. M., 2003. The singularity of the information matrix of the mixed proportional hazard model. Econometrica 71, 1579–1589. 2.4

Rothenberg, T. J., 1971. Identification in parametric models. Econometrica 39, 577–591. 1, 5

Sargan, J. D., 1983. Identification and lack of identification. Econometrica 51, 1605–1633. 3, 8.1

Staiger, D., Stock, J. H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557–586. 1

Stock, J. H., Wright, J. H., 2000. GMM with weak identification. Econometrica 68, 1055–1096. 1