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: November 26, 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 general deficient rank Jacobian in certain parts of the parameter space, leading to an identified set that is a sub-manifold of the parameter space. We examine standard extremum estimators and Wald statistics under a comprehensive class of parameter sequences characterizing the strength of identification of the model parameters, ranging from non-identification to strong identification. Allowing for a general singular Jacobian as the limiting point of weak identification allows us to study estimation and inference in many models to which previous results in the weak identification literature do not apply. Using the asymptotic results, we propose two hypothesis testing methods that make use of a standard Wald statistic and data-dependent critical values, leading to tests with correct asymptotic size regardless of identification strength and good power properties. Importantly, this allows one to directly conduct uniform inference on low-dimensional functions of the model parameters, including one-dimensional subvectors. The paper focuses on three examples of models to illustrate the results: sample selection models, models of potential outcomes with endogenous treatment and threshold crossing models.

*The authors are grateful to Donald Andrews, Isaiah Andrews, Xiaohong Chen, Xu Cheng, Yanqin Fan, Bruce Hansen, Ivana Komunjer, Tassos Magdalinos, Peter Phillips, Eric Renault, Yixiao Sun, Edward Vytlacil and Tiemen Woutersen for helpful comments. This paper is developed from earlier work by Han (2009). The second author gratefully acknowledges support from the NSF under grant SES-1357607.

1 Introduction

Many models estimated by applied economists suffer the problem that, at some points in the parameter space, the model parameters lose point identification. It is often the case that at these points of identification failure, the identified set for each parameter is not characterized by the entire parameter space it lies in but rather a lower-dimensional manifold inside of this parameter space. Such a scenario is sometimes referred to as "under-identification", "partial identification" or simply "non-identification". The non-identification status of these models is not straightforwardly characterized in the sense that one cannot say that some parameters are "completely" unidentified while the others are identified. Instead, it can be characterized by a non-identification curve that describes the lower-dimensional manifold defining the identified set. Moreover, in practice the model parameters may be weakly identified in the sense that they are near the under-identified/partially-identified region of the parameter space relative to the number of observations and sampling variability present in the data.

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. We characterize identification failure in this paper as a lack of (global) first-order identification in that the Jacobian matrix of the model restrictions has deficient column rank at some points in the parameter space.[1] We examine models for which a vector of parameters governs the identification status of the model. The contributions of this paper are threefold. First, we characterize the non-identification curve for a general class of models at points of identification failure and transform these models to have straightforward identification status. Second, we derive the limit theory for standard extremum estimators (e.g., GMM, maximum likelihood and classical minimum distance) and Wald statistics for these models under a comprehensive class of identification strengths including non-identification, weak identification and strong identification. We find that the asymptotic distributions derived under certain sequences of data-generating processes (DGPs) indexed by the sample size provide much better approximations to the finite sample distributions of these objects than those derived under the standard limit theory that assumes strong identification. Third, we use the limit theory derived under weak identification DGP sequences to construct data-dependent critical values (CVs) for Wald statistics that yield (uniformly) correct asymptotic size and good power properties. Importantly, our robust inference procedures allow one to directly conduct hypothesis tests for low-dimensional functions of the model parameters, including one-dimesnional subvectors, that are uniformly valid regardless of identification strength.

[1]See Rothenberg (1971) for a discussion of local vs. global identification and Sargan (1983) for a discussion of first vs. higher-order (local) identification.

A substantial portion of the recent econometrics literature has been devoted to estimation and inference that is robust to the identification strength 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). Most papers in this literature, such as Stock and Wright (2000) and Kleibergen (2005), have focused upon special cases of identification failure and weak identification by explicitly specifying how the Jacobian matrix of the underlying model could become (nearly) singular. For example, Kleibergen (2005) focused on a zero rank Jacobian as the point of identification failure in moment condition models. In this case, the identified set becomes the entire parameter space at points of identification failure. The recent works of Andrews and Cheng (2012a, 2013, 2014) implicitly focus on models for which the Jacobian of the model restrictions has columns of zeros at points of identification failure. For these types of models, some parameters become "completely" unidentified (those corresponding to the zero columns) while others remain strongly identified. In this paper, we do not specify the form of singularity in the Jacobian at the point of identification failure. This complicates the analysis but allows us to cover many more economic models used in practice such as sample selection models, treatment effect models with endogenous treatment, mixed proportional hazards models and higher-order ARMA models. Indeed, this feature of a singular Jacobian without zero columns at points of identification failure is typical of, but not limited to, many nonlinear instrumental variables models.

Only very recently have researchers begun 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 classical minimum distance (CMD) estimation and Andrews and Guggenberger (2014) and Andrews and Mikusheva (2014) in the context of moment condition models. Andrews and Mikusheva (2013) provide methods to directly perform uniformly valid subvector inference while Andrews and Guggenberger (2014) and Andrews and Mikusheva (2014) do not.[2] Unlike these papers, but like Andrews and Cheng (2012a, 2013, 2014), we focus

[2] Andrews and Mikusheva (2014) provide a method of "concentrating out" strongly identified nuisance parameters for subvector inference when all potentially weakly identified parameters are included in the subvector. One may also "indirectly" perform subvector inference using the methods of either Andrews and Guggenberger (2014)

explicitly on models for which the source of identification failure (a finite-dimensional parameter) is known to the researcher. This enables us to directly conduct subvector inference in a large class of models that is not nested in the setup of Andrews and Mikusheva (2013). Also unlike these papers, but like Andrews and Cheng (2012a, 2013, 2014), we derive nonstandard limit theory for standard estimators and test statistics. This nonstandard limit theory sheds light on how (badly) the standard Gaussian and chi-squared distributional approximations can fail in practice. For example, one interesting feature of the models studied here is that the asymptotic size of standard Wald tests for the full parameter vector *is equal to one* no matter the nominal level of the test. This feature emerges from observing that the Wald statistic diverges to infinity under certain DGP sequences admissible under the null hypothesis.

Aside from those already mentioned, there are many papers in the literature that have studied various types of under-identification in various models. For example, Sargan (1983) studied regression models that are nonlinear in parameters and first-order locally under-identified. Phillips (1989) studied under-identified simultaneous equations models and spurious time series regressions. Arellano et al. (2012) proposed a way to test for under-identification in a GMM context. Qu and Tkachenko (2012) study under-identification in the context of dynamic stochastic general equilibrium models. Escanciano and Zhu (2013) studied under-identification in a class of semiparametric models.[3] Dovonon and Renault (2013) uncovered an interesting result that, when testing for common sources of conditional heteroskedasticity in a vector of time series, there is a loss of first-order identification under the null hypothesis while the model remains second-order identified. Although all of these papers study under-identification of various forms, none of them deal with the empirically relevant potential for near or local to under-identification, one of the main focuses of the present paper.

In order to derive our asymptotic results under a comprehensive class of identification strengths, we begin by examining a transformed extremum estimation problem that falls under the framework of Andrews and Cheng (2012a) (AC12 hereafter). More specifically, we "profile out" (i.e., minimize with respect to) a subvector of the parameters of interest and look at a concentrated objective function. The profiling yields a random vector-valued function that can be used to estimate the non-identification curve at points of identification failure. The concentrated objective function is a function of a subvector of the model parameters that we show satisfies the crucial assumption of AC12: at points of identification failure, the concentrated objective function does not depend upon the unidentified parameters.[4] Hence, we use the results

or Andrews and Mikusheva (2014) by using a projection or Bonferroni bound-based approach but these methods are known to often suffer from severe power loss.

[3] Both Qu and Tkachenko (2012) and Escanciano and Zhu (2013) use the phrase "conditional identification" to refer to "under-identification" as we use it here.

[4] This corresponds to Assumption A of AC12.

of AC12 to find the limit theory for a subvector of the model parameter estimates. The profiled/concentrated objective function approach we use is related to but distinct from approaches found in Sargan (1983) and Escanciano and Zhu (2013), who study different forms of (local) identification failure.

We subsequently derive the limit theory for the entire vector of model parameters by establishing convergence results for the remaining parameter estimators. These latter estimators are equal to a random function of the subvector estimators, where the random function comes from the profiling step. To obtain a full asymptotic characterization of the full vector parameter estimator, we rotate the estimator in different directions of the parameter space. The estimator converges at different rates in different directions of the parameter space when identification is not strong, with some directions leading to a standard parametric rate of convergence and others leading to slower rates. Under weak identification, some directions of the weakly identified part of the parameter are not consistently estimable, leading to inconsistency in the parameter estimator that is reflected in finite sample simulation results and our derived asymptotic approximations. The rotation technique we use in our asymptotic derivations has many antecedents in the literature. For example, Sargan (1983) and Phillips (1989) used similar rotations to derive limit theory for estimators under identification failure; Antoine and Renault (2009, 2012) used similar rotations to derive limit theory for estimators under "nearly-weak" identification;[5] Andrews and Cheng (2014) used similar rotations to find the asymptotic distributions of Wald statistics under weak and nearly-strong identification; and recently Phillips (2015) used similar rotations to find limit theory for regression estimators in the presence of near-multicollinearity in regressors.

Using the limit theory for the parameter estimators, we derive the asymptotic distributions of standard Wald statistics for general (possibly nonlinear) hypotheses under a comprehensive class of identification strengths. The nonstandard nature of these limit distributions implies that using standard quantiles from chi-squared distributions as CVs leads to asymptotic size-distortions. Finally, we provide two data-driven methods to construct CVs for standard Wald statistics that lead to tests with correct asymptotic size, regardless of identification strength. The first is a direct analog of the Type 1 Robust CVs of AC12. The second is a modified version of the adjusted-Bonferroni CVs of McCloskey (2012), where the modifications are designed to ease the computation of the CVs in the current setting of this paper. The former CV construction method is simpler to compute while the latter yields better finite-sample size and power properties.

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

[5] In this paper, we follow AC12 and describe such parameter sequences as "nearly-strong".

identification and the lack of identification, presenting the non-identification curve and the associated identified set. Based on this curve, Section 4 introduces a transformation of the parameter space and presents a result that is useful for calculating the limit distributions later derived in Sections 8 and 9. 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 three examples in more detail. The asymptotic theory for the parameter estimators under various strengths of identification is given in Sections 7–8 and for Wald Statistics in Section 9. We describe how to perform uniformly robust inference in Section 10. Section 11 contains further details for a threshold crossing model of a triangular system, including Monte Carlo simulations demonstrating how well the nonstandard limit distributions derived in Sections 7–9 approximate their finite-sample counterparts. Proofs of the main results of the paper are provided in Appendix A, while figures are collected at the end of the document.

Notationally, we let $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the smallest and largest eigenvalues of a generic matrix A and d_B denote the dimension of a generic vector B. All vectors in the paper are column vectors. However, to simplify notation, we occasionally abuse it by writing (c, d) instead of $(c', d')'$ and for a function $f(a)$ with $a = (c, d)$, we write $f(c, d)$ rather than $f(a)$.

2 Class of Models

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

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

when $\widetilde{\theta} = \widetilde{\theta}^*$. The function describing this relationship $\widetilde{g}$ may depend on the true underlying value $\gamma^* \equiv (\widetilde{\theta}^*, \widetilde{\phi}^*)$ of parameter $\gamma \equiv (\widetilde{\theta}, \widetilde{\phi})$, i.e., the true underlying DGP, and thus moment conditions may be involved in defining this relationship. The parameter $\widetilde{\phi}$ captures the part of the distribution of the observed data that is not determined by $\widetilde{\theta}$, which is typically infinite dimensional (AC12). An important special case of (2.1) occurs when $\widetilde{g}$ relates a "structural parameter" $\widetilde{\theta}$ to a reduced-form parameter δ and depends on γ^* only through the true value δ^* of δ:

$$0 = \delta^* - \widetilde{\mathfrak{g}}(\widetilde{\theta}) \in \mathbb{R}^{d_\delta} \tag{2.2}$$

when $\widetilde{\theta} = \widetilde{\theta}^*$.

Oftentimes, econometric models imply a natural decomposition of $\widetilde{\theta}$: $\widetilde{\theta} = (\beta, \zeta, \widetilde{\pi})$, where the parameter β determines the "identification status" of $\widetilde{\pi}$. That is, when $\beta \neq \bar{\beta}$ for some $\bar{\beta}$, $\widetilde{\pi}$

is identified; when $\beta = \bar{\beta}$, $\widetilde{\pi}$ is under-identified; and when β is "close" to $\bar{\beta}$ relative to sampling variability, then $\widetilde{\pi}$ is local-to-underidentified. The identification status of the parameter ζ is not affected by the value of β. For convenience and without loss of generality, we us the normalization $\bar{\beta} = 0$. In this paper, we characterize identification via the Jacobian of the model restrictions:

$$J^*(\widetilde{\theta}) \equiv \frac{\partial \widetilde{g}^*(\widetilde{\theta})}{\partial \widetilde{\theta}}. \tag{2.3}$$

When $\beta = 0$, $J^*(\widetilde{\theta})$ will have deficient rank. Although our results cover cases for which $J^*(\widetilde{\theta})$ has columns of zeros when $\beta = 0$, these cases are not of primary interest for this paper as they are already covered by the analysis of AC12. Rather, we focus on models for which the column rank of $J^*(\widetilde{\theta})$ lies strictly between $d_\beta + d_\zeta$ and $d_{\widetilde{\theta}}$ when $\beta = 0$ and this rank-deficiency is not the consequence of zero columns.[6]

We present three examples that have a deficient rank Jacobian (2.3) with nonzero columns when $\beta = 0$. The first two examples fall into the framework of (2.1) and the third into (2.2):

Example 2.1 (Sample selection models)**.**

$$\begin{aligned} &Y_i = X_i'\pi^1 + \varepsilon_i, \qquad D_i = \mathbf{1}[\zeta + Z_{1i}'\beta \geq \nu_i], \\ &(\varepsilon_i, \nu_i)' \sim F_{\varepsilon\nu}(\varepsilon, \nu; \pi), \end{aligned}$$

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.[7] We observe (D_iY_i, D_i, X_i, Z_i) and $F_{\varepsilon\nu}(\cdot, \cdot; \pi)$ is a parametric distribution of the unobservable variables (ε, ν) parameterized by scalar π. The mean and variance of each unobservable is normalized to be zero and one, respectively. Let $W_i \equiv (Y_i, X_i, Z_i)$. Then we have, when $\widetilde{\theta} = \widetilde{\theta}^*$,

$$0 = \widetilde{g}^*(\widetilde{\theta}) = E_{\gamma^*}\varphi(W_i, \widetilde{\theta}), \tag{2.4}$$

where $\widetilde{\theta} \equiv (\beta, \zeta, \pi^1, \pi)$ and the moment function is

$$\varphi(w, \widetilde{\theta}) = \begin{bmatrix} d \begin{bmatrix} x \\ \widetilde{\lambda}(\zeta + z_1'\beta; \pi) \end{bmatrix} \left[y - x'\pi^1 - \widetilde{\lambda}(\zeta + z_1'\beta; \pi) \right] \\ \widetilde{\lambda}(\zeta + z_1'\beta; \pi) F_\nu^{-1}(-\zeta - z_1'\beta) \left[d - F_\nu(\zeta + z_1'\beta) \right] z \end{bmatrix},$$

with $\widetilde{\lambda}(\cdot; \pi)$ being a known function. When $F_{\varepsilon\nu}(\varepsilon, \nu; \pi)$ is a bivariate standard normal distribution

[6] Recall that β and ζ are always identified so that $rank(J^*(\widetilde{\theta})) \geq d_\beta + d_\zeta$ for all $\widetilde{\theta}$.

[7] In this example and in Example 2.2, we can alternatively design the linear index for the D equation to be $X_i'\zeta + Z_{1i}'\beta$ so that β is the coefficient on the excluded instruments, where d_β is smaller than the current design. To keep the analysis simple, however, we maintain the current one.

with the correlation coefficient π, we have $\widetilde{\lambda}(\cdot;\pi) = \pi\lambda(\cdot)$ where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mill's ratio with the standard normal density and distribution functions $\phi(\cdot)$ and $\Phi(\cdot)$, and $F_\nu(\cdot) = \Phi(\cdot)$. $\square$

Example 2.2 (Models of potential outcomes with endogenous treatment)**.**

$$\begin{aligned} Y_{1i} &= X_i'\pi_1^1 + \varepsilon_{1i}, \\ Y_{0i} &= X_i'\pi_2^1 + \varepsilon_{0i}, \end{aligned} \qquad D_i = \mathbf{1}[\zeta + Z_{1i}'\beta \geq \nu_i],$$
$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i},$$
$$(\varepsilon_{1i}, \varepsilon_{0i}, \nu_i)' \sim F_{\varepsilon_1,\varepsilon_0,\nu}(\varepsilon_1, \varepsilon_0, \nu; \pi),$$

where $F_{\varepsilon_1,\varepsilon_0,\nu}(\cdot,\cdot,\cdot;\pi)$ is a parametric distribution of the unobserved variables $(\varepsilon_1, \varepsilon_0, \nu)$ parameterized by vector π. We observe (Y_i, D_i, X_i, Z_i). The Roy model (Heckman and Honore, 1990) is a special case of this model of regime switching. This model extends the model in Example 2.1, but is similar in the aspects that this paper focuses upon. $\square$

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

$$\begin{aligned} Y_i &= \mathbf{1}[\pi_1^1 + \pi_2 D_i - \varepsilon_i \geq 0] \\ D_i &= \mathbf{1}[\zeta + \beta Z_i - \nu_i \geq 0] \end{aligned} \;, \quad (\varepsilon_i, \nu_i)' \sim F_{\varepsilon\nu}(\varepsilon_i, v_i; \pi).$$

where $Z_i \in \{0, 1\}$. We observe (Y_i, D_i, Z_i). The model can be generalized by including common exogenous covariates X_i 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 $F_{\varepsilon\nu}(\varepsilon, \nu; \pi) = \Phi(\varepsilon, \nu; \pi)$, a bivariate standard normal distribution, the model becomes the usual bivariate probit model. *A more general model with $F_{\varepsilon\nu}(\varepsilon, \nu; \pi) = C(F_\varepsilon(\varepsilon), F_\nu(\nu); \pi)$, for $C(\cdot,\cdot;\pi)$ in a class of single parameter copulas, is considered in Han and Vytlacil (2015), whose generality we follow here. Normalize F_v and F_ε to be uniform distributions for simplicity and let $\pi_2^1 \equiv \pi_1^1 + \pi_2$. By Han and Vytlacil (2015), the non-redundant fitted probabilities are*

$$\begin{aligned} p_{11,0} &= C(\pi_2^1, \zeta; \pi), \\ p_{11,1} &= C(\pi_2^1, \zeta + \beta; \pi), \\ p_{10,0} &= \pi_1^1 - C(\pi_1^1, \zeta; \pi), \\ p_{10,1} &= \pi_1^1 - C(\pi_1^1, \zeta + \beta; \pi), \\ p_{01,0} &= \zeta - C(\pi_2^1, \zeta; \pi), \\ p_{01,1} &= \zeta + \beta - C(\pi_2^1, \zeta + \beta; \pi), \end{aligned}$$

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

$$0 = \delta^* - \widetilde{\mathfrak{g}}(\widetilde{\theta}) = \begin{bmatrix} p_{11,0} \\ p_{11,1} \\ p_{10,0} \\ p_{10,1} \\ p_{01,0} \\ p_{01,1} \end{bmatrix} - \begin{bmatrix} C(\pi_2^1, \zeta; \pi) \\ C(\pi_2^1, \zeta + \beta; \pi) \\ \pi_1^1 - C(\pi_1^1, \zeta; \pi) \\ \pi_1^1 - C(\pi_1^1, \zeta + \beta; \pi) \\ \zeta - C(\pi_2^1, \zeta; \pi) \\ \zeta + \beta - C(\pi_2^1, \zeta + \beta; \pi) \end{bmatrix}, \tag{2.5}$$

where δ^* *and* $\widetilde{\mathfrak{g}}(\widetilde{\theta})$ *are defined implicitly with* $\widetilde{\theta} \equiv (\beta, \zeta, \pi^1, \pi)$ *and* $\pi \equiv (\pi_1^1, \pi_2^1)$. $\square$

In Example 2.1, with $\lambda(\cdot)$ being the inverse Mill's ratio, the Jacobian matrix (2.3) satisfies

$$J^*(\widetilde{\theta}) = E_{\gamma^*} \begin{bmatrix} -\pi_2 D_i X_i \lambda_{1i} Z_i' & -D_i \lambda_i X_i & -D_i X_i X_i' \\ D_i Y_i \lambda_{1i} Z_i' - D_i X_i' \pi_1 \lambda_{1i} Z_i' - 2\pi_2 D_i \lambda_i \lambda_{1i} Z_i' & -D_i \lambda_i^2 & -D_i \lambda_i X_i' \\ L_i(\beta, \zeta) Z_i Z_i' & 0_{l \times 1} & 0_{l \times k} \end{bmatrix},$$

where $\lambda_i \equiv \lambda(\zeta + Z_{1i}'\beta)$, $\lambda_{1i} \equiv d\lambda(x)/dx|_{x=\zeta+Z_{1i}'\beta}$,

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

$\Phi_i \equiv \Phi(\zeta + Z_{1i}'\beta)$ and $\phi_i \equiv \phi(\zeta + Z_{1i}'\beta)$. Note that $d_\beta + d_\zeta < rank(J^*(\widetilde{\theta})) < d_{\widetilde{\theta}}$ when $\beta = 0$, since λ_i becomes a constant and $X_i = (1, X_{1i}')'$. This rank-deficient Jacobian with non-zero columns when $\beta = 0$ poses several challenges that make the existing asymptotic theory in the literature that considers a Jacobian with zero columns when $\beta = 0$ inapplicable here: (i) since none of the columns of $J^*(\widetilde{\theta})$ are equal to zero, it is not immediately clear which components of the $\widetilde{\pi}$ parameter are (un-)identified; (ii) key assumptions in the literature, such as Assumption A in AC12, do not hold; (iii) typically, $\widetilde{g}^*(\widetilde{\theta})$ or $J^*(\widetilde{\theta})$ is highly nonlinear in β. In what follows, we develop a framework to tackle these challenges and to obtain local asymptotic theory and uniform inference procedures.

3 Identification and Lack of Identification

In this section, through the discussions of identification and the lack of identification we formalize the class of problems we are interested in, and introduce the non-identification curve which may be of independent interest and is useful for subsequent analysis. Recall $\gamma \equiv (\widetilde{\theta}, \widetilde{\phi})$ with

$\widetilde{\theta} \equiv (\beta, \zeta, \widetilde{\pi})$. Let Γ and $\widetilde{\Theta}$ be the parameter spaces of γ and $\widetilde{\theta}$, respectively. Let $\gamma_0 \equiv (\widetilde{\theta}_0, \widetilde{\phi}_0)$ and $\widetilde{\theta}_0 \equiv (\beta_0, \zeta_0, \widetilde{\pi}_0)$. Later we define a sequence of parameters γ_n that converges to γ_0. Let $\widetilde{g}_0(\widetilde{\theta}) \equiv \widetilde{g}(\widetilde{\theta}; \gamma_0)$. We begin by assuming a mild regularity condition.

Assumption Reg1. *$\widetilde{g}_0 : \widetilde{\Theta} \to \mathbb{R}^{d_{\widetilde{g}}}$ is continuously differentiable in $\widetilde{\theta}$ $\forall \gamma_0 \in \Gamma$.*

The following assumption describes the lack of identification when $\beta = 0$.

Assumption ID1. *When $\beta = 0$, $rank\left(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}\right) = r < d_{\widetilde{\pi}}$ $\forall \widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$ $\forall \gamma_0 \in \Gamma$, where $\widetilde{\pi}$ is the smallest subvector of $\widetilde{\theta}$ such that $d_{\widetilde{\pi}} - rank(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}) = d_{\widetilde{\theta}} - rank(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\theta})$.*

By allowing $r > 0$, Assumption ID1 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 $\widetilde{\pi}$ in the sense of the lack of first-order identification detailed by Sargan (1983). This concept is also used in relating the lack of identification with a criterion function in estimation; see, e.g., Theorem FSTrans below.

When $\beta = 0$, the identified set for $\widetilde{\pi}$ is not necessarily equal to its entire parameter space. Rather, $\widetilde{\pi}$ is *under-identified.* Specifically, $\widetilde{\pi}$ is partially identified with an identified set that is a lower-dimensional manifold within the parameter space for $\widetilde{\pi}$. This identified set is characterized below by a non-identification curve. In the special case of zero rank $\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}$, this manifold shapes a linear subspace (or location shift of it). In general, however, this is not the case. This feature of the problem motivates us to proceed as follows.

Lemma ID1. *Assumption ID1 implies that there exist r-dimensional subvectors $\pi^1 \in \Pi^1$ of $\widetilde{\pi} = (\pi, \pi^1)$ and g_0^1 of $\widetilde{g}_0 = (g_0^{1\prime}, g_0^{0\prime})'$ such that when $\beta = 0$, $rank(\partial g_0^1(\widetilde{\theta})/\partial \pi^1) = r$ $\forall \widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$ $\forall \gamma_0 \in \Gamma$.*

The subvector π^1 of $\widetilde{\pi}$ in this lemma is not necessarily uniquely determined, but only its existence is necessary for the subsequent analysis. The next two assumptions are regularity conditions that are related to the global inversion of $\widetilde{g}_0$ at $\beta = 0$.

Assumption Reg2. *When $\beta = 0$, the function $\widetilde{g}_0(\beta, \zeta, \pi, \cdot)$ from Π^1 to $\mathbb{R}^{d_{\widetilde{g}}}$ is proper $\forall (0, \zeta, \pi)$ $\forall \gamma_0 \in \Gamma$.*

Under Assumption Reg1, a sufficient condition for Assumption Reg2 to hold is that Π^1 is bounded.

Assumption Reg3. *When $\beta = 0$, $\widetilde{g}_0(\beta, \zeta, \pi, \Pi^1)$ is simply connected $\forall (0, \zeta, \pi)$ $\forall \gamma_0 \in \Gamma$.*

For the remainder of this section, we suppose $r \neq 0$. Define $\Theta^0 \equiv \left\{(0, \zeta, \pi) : \widetilde{\theta} \in \widetilde{\Theta}\right\}$. The following lemma defines the non-identification curve as a solution to a subvector of the vector of equations (2.1) at $\beta = 0$.

Lemma ID2. *Suppose $r \neq 0$. Under Assumptions ID1 and Reg1–Reg3, there exists a unique solution $\pi^1 = h_0^1(\zeta, \pi) \equiv h^1(\zeta, \pi; \gamma_0)$ such that*

$$g_0^1\left(0, \zeta, \pi, h_0^1(\zeta, \pi)\right) = 0 \tag{3.1}$$

$\forall (0, \zeta, \pi) \in \Theta^0 \ \forall \gamma_0 \in \Gamma$.

When $\beta_0 = 0$, the non-identification curve $h_0^1(\zeta, \pi)$ defines the identified set for $\widetilde{\theta}$ by a curve (i.e., a lower dimensional manifold) in $\widetilde{\Theta}$ that depends on the true DGP, which can be denoted as $\widetilde{\Theta}_0(\gamma_0) \equiv \left\{(0, \zeta_0, \widetilde{\pi}) \in \widetilde{\Theta} : \pi^1 = h_0^1(\zeta_0, \pi)\right\}$.

4 Transformation

For a given π^1 such that $\widetilde{\pi} \equiv (\pi, \pi^1)$, define $\theta \equiv (\beta, \zeta, \pi)$ so that $\widetilde{\theta} \equiv (\beta, \zeta, \pi, \pi^1) \equiv (\theta, \pi^1) \in \widetilde{\Theta}$. Without loss of generality, $\widetilde{\Theta}$ can be written as

$$\widetilde{\Theta} = \{\widetilde{\theta} = (\theta, \pi^1) : \theta \in \Theta, \pi^1 \in \Pi^1(\theta)\},$$

where

$$\Theta = \{\theta : \text{ there is some } \pi^1 \text{ such that } (\theta, \pi^1) \in \widetilde{\Theta}\},$$
$$\Pi^1(\theta) = \{\pi^1 : (\theta, \pi^1) \in \widetilde{\Theta}\} \text{ for } \theta \in \Theta.$$

Finally, define

$$\Pi = \{\pi : \text{ there are some } \beta, \zeta, \pi^1 \text{ such that } (\beta, \zeta, \pi, \pi^1) \in \widetilde{\Theta}\}.$$

In order to transform the function $\widetilde{g}_0(\widetilde{\theta})$ to a function defined on Θ, we "extend" the non-identification curve $h_0^1(\zeta, \pi)$ to a function on Θ (which is a superset of Θ^0). We show that such a function can be defined as a concentrated true parameter value that solves a profiled optimization problem. We are ultimately interested in an estimator of $\widetilde{\theta}$ that minimizes a sample criterion function $\widetilde{Q}_n(\widetilde{\theta})$ but we first study the population criterion function $\widetilde{Q}_0(\widetilde{\theta}) \equiv \widetilde{Q}(\widetilde{\theta}; \gamma_0)$ to which $\widetilde{Q}_n(\widetilde{\theta})$ converges.[8]

Assumption CF1. *For some nonstochastic real-valued function $\widetilde{Q}_0(\widetilde{\theta})$ on $\widetilde{\Theta} \times \Gamma$, the solution $\widetilde{\theta}_0$ to $\widetilde{Q}_0(\widetilde{\theta}_0) = \inf_{\widetilde{\theta} \in \widetilde{\Theta}} \widetilde{Q}_0(\widetilde{\theta})$ exists and is unique $\forall \gamma_0 \in \Gamma$.*

The next assumption defines the class of criterion functions we consider in this paper.

[8]See Assumption FSCF2 for a formal expression.

Assumption CF2. $\widetilde{Q}_0(\widetilde{\theta})$ *can be written as*

$$\widetilde{Q}_0(\widetilde{\theta}) = \widetilde{\Psi}_0(\widetilde{g}_0(\widetilde{\theta}))$$

for some deterministic function $\widetilde{\Psi}_0(\cdot) \equiv \widetilde{\Psi}(\cdot;\gamma_0)$ *that is twice continuously differentiable such that* $\widetilde{\Psi}_{\widetilde{g}\widetilde{g},0} \equiv \frac{\partial^2 \widetilde{\Psi}_0}{\partial \widetilde{g} \partial \widetilde{g}'}$ *is positive definite* $\forall \gamma_0 \in \Gamma$.

Assumption CF2 is naturally satisfied when we construct GMM/CMD or maximum likelihood (ML) criterion functions, given (2.1) or (2.2). Note that models that generate likelihoods or minimum distance structures typically involve $\widetilde{g}_0(\widetilde{\theta}) = \delta_0 - \widetilde{\mathfrak{g}}(\widetilde{\theta})$ by (2.2). For a GMM/CMD criterion function, $\widetilde{\Psi}_0(\widetilde{g}_0(\widetilde{\theta})) = \left\| W_0 \left(\widetilde{g}_0(\widetilde{\theta}) \right) \right\|^2$ where W_0 is a weight matrix.[9] For a ML criterion function, $\widetilde{\Psi}_0(\widetilde{g}_0(\widetilde{\theta})) = -E_{\gamma_0} \ln f^\dagger \left(W_i, \delta_0 - \widetilde{\mathfrak{g}}(\widetilde{\theta}) \right)$ if the distribution of the data depends on $\widetilde{\theta}$ only through δ (Rothenberg, 1971). That is, there exists a function $f^\dagger(w;\delta)$ such that

$$f(w;\widetilde{\theta}) = f^\dagger(w;\widetilde{\mathfrak{g}}(\widetilde{\theta})) = f^\dagger(w;\delta). \tag{4.1}$$

The positive definiteness of $\widetilde{\Psi}_{\widetilde{g}\widetilde{g},0}$ can be ensured by the usual assumption that the weight matrix $W_0 = \widetilde{\Psi}_{\widetilde{g}\widetilde{g},0}$ is positive definite in the GMM/CMD case and by the fact that the information matrix with respect to the reduced-form parameter δ, $E_{\gamma_0} \frac{\partial^2 \ln f^\dagger}{\partial \delta \partial \delta'}$, is always non-singular in the ML case.

Remark 4.1. *Given the existence of* $f^\dagger(w;\delta)$ *in the ML framework, the setting of this paper can be characterized in terms of the information matrix. Let* $\mathcal{I}(\widetilde{\theta})$ *be the* $d_{\widetilde{\theta}} \times d_{\widetilde{\theta}}$ *information matrix*

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

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

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

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

Assumption CF3. *For any* $\theta \in \Theta$*, the solution* $\pi_0^1(\theta)$ *to* $\widetilde{Q}_0(\theta, \pi_0^1(\theta)) = \inf_{\pi^1 \in \Pi^1(\theta)} \widetilde{Q}_0(\theta, \pi^1)$ *exists and is unique and* $\pi_0^1(\theta) \in int(\Pi^1(\theta))$ $\forall \gamma_0 \in \Gamma$.

[9] Note that Assumption CF2 does not cover GMM with a continuously updating weight matrix $W_0(\widetilde{\theta}) \equiv W(\widetilde{\theta};\gamma_0)$.

This assumption holds if the population criterion function is well-behaved and the optimization parameter space is chosen to be "large enough".

Define

$$g_0(\theta) \equiv g(\theta; \gamma_0) \equiv \widetilde{g}(\theta, \pi_0^1(\theta); \gamma_0) \tag{4.2}$$

and

$$Q_0(\theta) \equiv \widetilde{Q}_0(\theta, \pi_0^1(\theta)) = \widetilde{\Psi}_0(g_0(\theta)).$$

Based on the analysis of the previous section, we show that the transformed model $0 = g_0(\theta)$ satisfies a useful property.

When $\beta = 0$, Lemma ID1 implies that there exists some $(d_{\widetilde{g}} - r) \times r$ matrix M, such that

$$\frac{\partial g_0^0(\widetilde{\theta})}{\partial \widetilde{\pi}} = M \frac{\partial g_0^1(\widetilde{\theta})}{\partial \widetilde{\pi}} \tag{4.3}$$

$\forall \widetilde{\theta} = (0, \zeta, \widetilde{\pi})$.

Assumption ID2. *The matrix M in* (4.3) *is only a function of (β, ζ) and not a function of $\widetilde{\pi}$, which is denoted as $M(\beta, \zeta; \gamma_0)$.*

This assumption holds for all the examples considered in this paper.

Our goal is now to transform the extremum estimation problem into one for which the criterion function does not depend upon the non-identified parameter when $\beta = 0$. The partition $\widetilde{\pi} = (\pi, \pi^1)$ and the function h_0^1 enable us to do so. More specifically, the following theorem shows "to what extent" $\widetilde{\pi}$ is identified when $\beta = 0$: conditional on knowing π, the parameters are identified. Note that (β, ζ) are always identified by assumption.

Theorem Trans. *Suppose Assumptions ID1–ID2, Reg1–Reg3 and CF1–CF3 hold. When $\beta_0 = 0$, $\pi_0^1(0, \zeta_0, \pi) = h_0^1(\zeta_0, \pi)$ and $g_0(0, \zeta_0, \pi) = 0$ $\forall \pi \in \Pi$ such that $(0, \zeta_0, \pi) \in \Theta$ $\forall \gamma_0 \in \Gamma$.*

Since h_0^1 can be obtained from inverting g_0^1 by Lemma ID2, the result of Theorem Trans can be useful for calculating the limit distributions of the estimators, which involve $\pi_0^1(\cdot)$, in Section 7. See Section 11.1 for an example of how this result can be used in practice.

Since $g_0(0, \zeta_0, \pi) = 0$ for all $\pi \in \Pi$ such that $(0, \zeta_0, \pi) \in \Theta$, it is not a function of π over the relevant parameter space. This result is reminiscent of Assumption GMM1(iii) (or GMM3(ii)) of Andrews and Cheng (2014) but for the transformed model defined by $g_0(\theta)$ rather than the original $\widetilde{g}_0(\widetilde{\theta})$, although the result is not restricted to the GMM setting. Note that in terms of

θ in the transformed model, the identified set when $\beta_0 = 0$ becomes a (location shifted) simple linear subspace of Θ, where π is entirely unidentified.

5 Criterion Functions

In this section, we define the original and transformed criterion functions and relevant estimators. We then establish a sample counterpart to Theorem Trans of the previous section.

We define the extremum estimator $\widehat{\widetilde{\theta}}_n$ as the minimizer of the criterion function $\widetilde{Q}_n(\widetilde{\theta})$ over the optimization parameter space $\widetilde{\Theta}$:

$$\widehat{\widetilde{\theta}}_n \in \widetilde{\Theta} \text{ and } \widetilde{Q}_n(\widehat{\widetilde{\theta}}_n) = \inf_{\widetilde{\theta} \in \widetilde{\Theta}} \widetilde{Q}_n(\widetilde{\theta}) + o(n^{-1}). \tag{5.1}$$

The function $\widetilde{Q}_n(\widetilde{\theta})$ depends on the observations $\{W_i : i \leq n\}$. Consider the following profiled extremum estimation, which defines the concentrated estimator $\widehat{\pi}_n^1(\theta)$:

$$\widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)) = \inf_{\pi^1 \in \Pi^1(\theta)} \widetilde{Q}_n(\theta, \pi^1) + o(n^{-1}). \tag{5.2}$$

This gives us the concentrated criterion function, which naturally defines the transformed criterion function analogous to the tranformation introduced in the previous section:

$$Q_n(\theta) \equiv \widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)). \tag{5.3}$$

Now we define the estimator $\widehat{\theta}_n$ of θ as

$$Q_n(\widehat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}). \tag{5.4}$$

We assume that $\widehat{\widetilde{\theta}}_n = (\widehat{\theta}_n, \widehat{\pi}_n^1)$ can be written as $(\widehat{\theta}_n, \widehat{\pi}_n^1(\widehat{\theta}_n))$, where $\widehat{\pi}_n^1(\cdot)$ and $\widehat{\theta}_n$ are defined in (5.3) and (5.4).

Define the sample counterpart of $g_0(\theta)$ as

$$\bar{g}(\theta) \equiv \bar{\widetilde{g}}(\theta, \widehat{\pi}_n^1(\theta))$$

where $\bar{\widetilde{g}}(\widetilde{\theta})$ is the sample counterpart of $\widetilde{g}_0(\widetilde{\theta})$. In the case of CMD and ML, $\bar{\widetilde{g}}(\widetilde{\theta}) = \widehat{\delta}_n - \widetilde{\mathfrak{g}}(\widetilde{\theta})$ analogous to (2.2). For GMM, $\bar{\widetilde{g}}(\widetilde{\theta}) = n^{-1} \sum_{i=1}^n \varphi(W_i, \widetilde{\theta})$. We list assumptions on the sample objects that are analogous to the assumptions on the population objects, Assumptions Reg1, ID1, CF2 and CF3.

Assumption FSReg1. *$\bar{\widetilde{g}} : \widetilde{\Theta} \to \mathbb{R}^{d_{\bar{g}}}$ is continuously differentiable in $\widetilde{\theta}$.*

Assumption FSID1. *When $\beta = 0$, rank $\left(\partial \bar{\widetilde{g}}(\widetilde{\theta})/\partial \widetilde{\pi}\right) = r < d_{\widetilde{\pi}}$ $\forall \widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$, where $\widetilde{\pi}$ is the smallest subvector of $\widetilde{\theta}$ such that $d_{\widetilde{\pi}} - rank(\partial \bar{\widetilde{g}}(\widetilde{\theta})/\partial \widetilde{\pi}) = d_{\widetilde{\theta}} - rank(\partial \bar{\widetilde{g}}(\widetilde{\theta})/\partial \widetilde{\theta})$.*

The proof of the following Corollary to Lemma ID1 is nearly identical to the proof for that lemma and is therefore omitted.

Corollary FSID. *Assumption FSID1 implies that there exist r-dimensional subvectors $\pi^1 \in \Pi^1$ of $\widetilde{\pi} = (\pi, \pi^1)$ and $\bar{g}^1$ of $\bar{\widetilde{g}} = (\bar{g}^{1\prime}, \bar{g}^{0\prime})'$ such that when $\beta = 0$, $rank(\partial \bar{g}^1(\widetilde{\theta})/\partial \pi^1) = r$ $\forall \widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$.*

Assumption FSCF1. *$\widetilde{Q}_n(\widetilde{\theta})$ can be written as*

$$\widetilde{Q}_n(\widetilde{\theta}) = \Psi_n(\bar{\widetilde{g}}_n(\widetilde{\theta}))$$

for some random function $\Psi_n(\cdot)$ that is twice continuously differentiable such that $\Psi_{n,\widetilde{g}\widetilde{g}} \equiv \frac{\partial^2 \Psi_n}{\partial \widetilde{g} \partial \widetilde{g}'}$ is positive definite.

Assumption FSReg2. *For any $\theta \in \Theta$, $\widehat{\pi}_n^1(\theta)$ in (5.2) satisfies $\partial \widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta))/\partial \pi^1 = 0$.*

For a GMM/CMD criterion function, $\Psi_n(\bar{\widetilde{g}}(\widetilde{\theta})) = \left\| W_n \bar{\widetilde{g}}(\widetilde{\theta}) \right\|^2$ where W_n is a (possibly random) weight matrix; for a ML criterion function, $\Psi_n(\bar{\widetilde{g}}(\widetilde{\theta})) = -\frac{1}{n}\sum_{i=1}^n \ln f^\dagger \left(W_i, \widehat{\delta}_n - \mathfrak{g}(\widetilde{\theta})\right)$. Defining $\bar{g}(\theta) \equiv \bar{\widetilde{g}}(\theta, \widehat{\pi}_n^1(\theta))$, note that $Q_n(\theta) = \Psi_n(\bar{\widetilde{g}}(\theta, \widehat{\pi}_n^1(\theta))) = \Psi_n(\bar{g}(\theta))$.

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

Theorem FSTrans. *Under Assumptions FSID1, FSReg1–FSReg2 and FSCF1, for all $\beta = 0$, $Q_n(\theta)$ does not depend upon π.*

This theorem is reminiscent of Assumption A in AC12. In sum, after transforming the problem, among the components of $\theta = (\beta, \zeta, \pi)$, β determines the identification status of θ, ζ is a parameter whose identification is not affected by the value of β, and π is a parameter which is not identified and does not appear in the criterion function when $\beta = 0$. 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 β is close to zero; (ii) it yields criterion functions that do not depend (in a generalized sense) on the unidentified parameters when $\beta = 0$.

6 Examples

6.1 Sample selection models and models of potential outcomes

Since Examples 2.1 and 2.2 share similar features, we focus our attention on Example 2.1. Let $\widetilde{\pi} = (\pi, \pi^1)$ in that model. Given $\widetilde{g}_0(\widetilde{\theta})$ defined by (2.4) with a bivariate standard normal distribution, the part of the Jacobian relevant to our discussions in Sections 3–5 is

$$\frac{\partial \widetilde{g}_0(\widetilde{\theta})}{\partial \widetilde{\pi}} = -E_{\gamma_0} \begin{bmatrix} D_i \lambda_i X_i & D_i X_i X_i' \\ D_i \lambda_i^2 & D_i \lambda_i X_i' \\ 0_{l \times 1} & 0_{l \times k} \end{bmatrix}.$$

When $\beta = 0$, we have $rank(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}) = d_{\widetilde{\pi}} - 1 = r = k$ since

$$\frac{\partial \widetilde{g}_0(0, \zeta, \widetilde{\pi})}{\partial \widetilde{\pi}} = -E_{\gamma_0} \begin{bmatrix} \lambda(\zeta) D_i X_i & D_i X_i X_i' \\ \lambda^2(\zeta) D_i & \lambda(\zeta) D_i X_i' \\ 0_{l \times 1} & 0_{l \times k} \end{bmatrix},$$

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

$$g_0^1(\widetilde{\theta}) = E_{\gamma_0} \left[D_i X_i Y_i - D_i X_i X_i' \pi^1 - \pi D_i \lambda_i X_i \right]$$

when $\beta = 0$, note that $0 = g_0^1(0, \zeta, \widetilde{\pi})$ is equivalent to

$$\begin{aligned} 0_{k \times 1} &= E_{\gamma_0} \left[D_i X_i Y_i - D_i X_i X_i' \pi^1 - \pi \lambda(\zeta) D_i X_i \right] \\ &\equiv Q_{0,DXY} - Q_{0,DXX} \pi_1 - \lambda(\zeta) \pi Q_{0,DX}, \end{aligned}$$

where $Q_{0,DXY}$, $Q_{0,DXX}$, and $Q_{0,DX}$ are implicitly defined. Observe that the result of Lemma ID1 holds. Also when $\beta = 0$, $M \partial g_0^1(0, \zeta, \widetilde{\pi})/\partial \widetilde{\pi} = \partial g_0^0(0, \zeta, \widetilde{\pi})/\partial \widetilde{\pi}$ with M being a $(l+1) \times k$ zero matrix but with the $(1,1)$ element replaced $\lambda(\zeta)$. In this example, the function h_0^1 has a closed form solution:

$$h_0^1(\zeta, \pi) = Q_{0,DXX}^{-1} \left(Q_{0,DXY} - \lambda(\zeta) \pi Q_{0,DX} \right).$$

6.2 Threshold crossing models with a dummy endogenous variable

We now continue to discuss Example 2.3. Let $\widetilde{\pi} = (\pi, \pi_1^1, \pi_2^1)$. Given $\widetilde{g}_0(\widetilde{\theta}) = \delta_0 - \widetilde{\mathfrak{g}}(\widetilde{\theta})$ from the expression in (2.5), the relevant Jacobian is

$$\frac{\partial \widetilde{g}_0(\widetilde{\theta})}{\partial \widetilde{\pi}} = -\frac{\partial \widetilde{\mathfrak{g}}(\widetilde{\theta})}{\partial \widetilde{\pi}} = -\begin{bmatrix} C_3\left(\pi_2^1, \zeta; \pi\right) & 0 & C_1\left(\pi_2^1, \zeta; \pi\right) \\ C_3\left(\pi_2^1, \zeta+\beta; \pi\right) & 0 & C_1\left(\pi_2^1, \zeta+\beta; \pi\right) \\ -C_3\left(\pi_1^1, \zeta; \pi\right) & 1 - C_1\left(\pi_1^1, \zeta; \pi\right) & 0 \\ -C_3\left(\pi_1^1, \zeta+\beta; \pi\right) & 1 - C_1\left(\pi_1^1, \zeta+\beta; \pi\right) & 0 \\ -C_3\left(\pi_2^1, \zeta; \pi\right) & 0 & -C_1\left(\pi_2^1, \zeta; \pi\right) \\ -C_3\left(\pi_2^1, \zeta+\beta; \pi\right) & 0 & -C_1\left(\pi_2^1, \zeta+\beta; \pi\right) \end{bmatrix},$$

where $C_1(\cdot, \cdot; \pi)$ and $C_3(\cdot, \cdot; \pi)$ denote the derivatives of $C(\cdot, \cdot; \pi)$ with respect to the first argument and π, respectively. When $\beta = 0$, we have $rank(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}) = d_{\widetilde{\pi}} - 1 = r = 2$, since there are only two linearly independent rows in $\partial \widetilde{\mathfrak{g}}(0, \zeta, \pi)/\partial \widetilde{\pi}$. When $\beta = 0$, note that $0 = \delta_0^1 - g^1(0, \zeta, \widetilde{\pi})$ is equivalent to

$$0 = \begin{bmatrix} p_{11,0} \\ p_{10,0} \\ p_{01,0} \end{bmatrix} - \begin{bmatrix} C(\pi_2^1, \zeta; \pi) \\ \pi_1^1 - C(\pi_1^1, \zeta; \pi) \\ \zeta - C(\pi_2^1, \zeta; \pi) \end{bmatrix}.$$

Observe that the result of Lemma ID1 holds, as $rank\left(\partial g_0^1(0, \zeta, \widetilde{\pi})/\partial \pi^1\right) = rank\left(\partial g^1(0, \zeta, \widetilde{\pi})/\partial \pi^1\right) = r$. Also when $\beta = 0$, $M \partial g_0^1(0, \zeta, \widetilde{\pi})/\partial \widetilde{\pi} = \partial g_0^0(0, \zeta, \widetilde{\pi})/\partial \widetilde{\pi}$, where $M = I_3$. In this example, the function h_0^1 may or may not have a closed form solution, depending upon the copula used. See Section 11.1 for an example.

7 Concentrated Estimation

We proceed to derive the limit theory for $\widehat{\widehat{\theta}}_n$ under a comprehensive class of identification strengths in two steps: (i) using the results of Section 5, we derive the joint limit theory for $\widehat{\theta}_n$ and $\widehat{\pi}_n^1(\cdot)$ and (ii) we use the results of (i) to find the limit theory for the parameter of interest. This section is devoted to step (i).

We formally characterize a local-to-deficient rank Jacobian by modeling the β parameter as local-to-zero. This allows us to 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.

The parameter space Γ for γ is of the form

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

where $\widetilde{\Theta}^*$ is a compact subset of $\mathbb{R}^{d_{\widetilde{\theta}}}$ and $\Phi^*(\widetilde{\theta}) \subset \Phi^* \; \forall \widetilde{\theta} \in \widetilde{\Theta}$ for some compact metric space Φ^* with a metric that induces weak convergence of the bivariate distributions of the data (W_i, W_{i+m}) for all $i, m \geq 1$. Define sets of sequences of parameters $\{\gamma_n\}$ as follows:

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

where $\gamma_0 \equiv \left(\beta_0, \zeta_0, \widetilde{\pi}_0, \widetilde{\phi}_0\right)$ and $\gamma_n \equiv \left(\beta_n, \zeta_n, \widetilde{\pi}_n, \widetilde{\phi}_n\right)$, and $\mathbb{R}_\infty \equiv \mathbb{R} \cup \{\pm\infty\}$. When $\|b\| < \infty$, $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ are weak or non-identification sequences, otherwise, when $\|b\| = \infty$, they characterize semi-strong identification. Sequences $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ characterize semi-strong identification when $\beta_n \to 0$, otherwise, when $\lim_{n\to\infty} \beta_n \neq 0$, they are strong identification sequences.

We first establish the limit theory for the first step estimator $\widehat{\pi}_n^1(\theta)$, as a function of $\theta \in \Theta$. The following assumption establishes the population criterion function discussed in Section 4 as the limit (in a uniform sense) of the sample criterion function.

Assumption FSCF2. *The function $\widetilde{Q}_0(\widetilde{\theta})$ is such that*

$$\sup_{\widetilde{\theta} \in \widetilde{\Theta}} |\widetilde{Q}_n(\widetilde{\theta}) - \widetilde{Q}_0(\widetilde{\theta})| \xrightarrow{p} 0$$

under $\{\gamma_n\} \in \Gamma(\gamma_0) \; \forall \gamma_0 \in \Gamma$.

Assumption ID3. *For $\pi_0^1 : \Theta \to \Pi^1(\Theta) \subset \mathbb{R}^{d_{\pi^1}}$ and every neighborhood $\Pi_0^1(\theta)$ of $\pi_0^1(\theta)$ at any given $\theta \in \Theta$,*

$$\inf_{\theta \in \Theta} \left(\inf_{\pi^1 \in \Pi^1(\theta) \setminus \Pi_0^1(\theta)} \widetilde{Q}_0(\theta, \pi^1) - \widetilde{Q}_0(\theta, \pi_0^1(\theta)) \right) > 0$$

$\forall\gamma_0 \in \Gamma$.

In conjunction with Assumption CF3, Assumption ID3 states that *conditional on* θ, the population criterion function locally identifies the parameter $\pi_0^1(\theta)$. For any $\{\gamma_n\} \in \Gamma(\gamma_0)$ with $\gamma_0 \in \Gamma$ and any $\theta \in \Theta$, define $\pi_n^1(\theta)$ implicitly as

$$\widetilde{Q}(\theta, \pi_n^1(\theta); \gamma_n) = \inf_{\pi^1 \in \Pi^1(\theta)} \widetilde{Q}(\theta, \pi^1; \gamma_n).$$

The next assumption is a weak continuity condition on $\widetilde{Q}(\theta, \pi^1; \cdot)$.

Assumption Reg4. *Under any* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *with* $\gamma_0 \in \Gamma$, $\sup_{\theta\in\Theta} \|\pi_n^1(\theta) - \pi_0^1(\theta)\| \to 0$.

Let $X_n(\theta) = o_{p\theta}(1)$ mean that $\sup_{\theta\in\Theta} \|X_n(\theta)\| = o_p(1)$. Similarly, $X_n(\pi) = o_{p\pi}(1)$ means that $\sup_{\pi\in\Pi} \|X_n(\pi)\| = o_p(1)$ and analogously for $O_{p\theta}(1)$ and $O_{p\pi}(1)$. The next assumptions supposes the sample criterion function has a partial quadratic expansion in the parameter π^1.

Assumption FSReg3. *Under any* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *with* $\gamma_0 \in \Gamma$*, the following statements hold:*
(i) The sample criterion function $\widetilde{Q}_n(\theta, \pi^1)$ *has a quadratic expansion in* π^1 *around* $\pi_n^1(\theta)$ *for given* θ*:*

$$\begin{aligned}\widetilde{Q}_n(\theta, \pi^1) &= \widetilde{Q}_n(\theta, \pi_n^1(\theta)) + D_{\pi^1}\widetilde{Q}_n(\theta, \pi_n^1(\theta))'(\pi^1 - \pi_n^1(\theta)) \\ &\quad + \frac{1}{2}(\pi^1 - \pi_n^1(\theta))' D_{\pi^1\pi^1}\widetilde{Q}_n(\theta, \pi_n^1(\theta))(\pi^1 - \pi_n^1(\theta)) + R_n(\theta, \pi^1),\end{aligned}$$

where $D_{\pi^1}\widetilde{Q}_n(\theta, \pi_n^1(\theta)) \in \mathbb{R}^{d_{\pi^1}}$ *is a first partial-derivative vector (with respect to* π^1*, evaluated at* $\pi_n^1(\theta)$*) and* $D_{\pi^1\pi^1}\widetilde{Q}_n(\theta, \pi_n^1(\theta)) \in \mathbb{R}^{d_{\pi^1}\times d_{\pi^1}}$ *is a second partial-derivative matrix (with respect to* π^1*, evaluated at* $\pi_n^1(\theta)$*) that is symmetric and may be stochastic or nonstochastic.*

(ii) The remainder term $R_n(\theta, \pi^1)$ *satisfies*

$$\sup_{\pi^1\in\Pi^1(\theta):\|\pi^1 - \pi_n^1(\theta)\| \le \epsilon_n} \frac{|nR_n(\theta, \pi^1)|}{(1 + \|\sqrt{n}(\pi^1 - \pi_n^1(\theta))\|)^2} = o_{p\theta}(1)$$

for all constants $\epsilon_n \to 0$.

It will be convenient to let $\psi \equiv (\beta, \zeta)$ denote the subvector of well-identified parameters. Let

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

The following is a joint convergence assumption on the (partial) generalized stochastic derivatives of the (concentrated) criterion function. Let $\psi_{0,n} \equiv (0, \zeta_n)$.

Assumption FSReg4. *(i) Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$,

$$\sqrt{n} \begin{pmatrix} D_{\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot)) \\ D_\psi Q_n(\psi_{0,n}, \cdot) - E_{\gamma_n} D_\psi Q_n(\psi_{0,n}, \cdot) \end{pmatrix} \Rightarrow \begin{pmatrix} \widetilde{G}_0(\cdot) \\ G_0(\cdot) \end{pmatrix},$$

where $D_\psi Q_n(\psi_{0,n}, \cdot)$ *is defined in Assumption C1(i)-(ii) of AC12 and* $\widetilde{G}_0(\cdot) \equiv \widetilde{G}(\cdot; \gamma_0)$ *and* $G_0(\cdot) \equiv G(\cdot; \gamma_0)$ *are mean zero Gaussian processes indexed by* $\theta \in \Theta$ *and* $\pi \in \Pi$ *with bounded continuous sample paths and some covariance kernels* $\widetilde{\Omega}_0(\theta_1, \theta_2) \equiv \widetilde{\Omega}(\theta_1, \theta_2; \gamma_0)$ *and* $\Omega_0(\pi_1, \pi_2) \equiv \Omega(\pi_1, \pi_2; \gamma_0)$*, respectively, for* $\theta_1, \theta_2 \in \Theta$ *and* $\pi_1, \pi_2 \in \Pi$.

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n} \begin{pmatrix} D_{\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot)) \\ B^{-1}(\beta_n) D Q_n(\theta_n) \end{pmatrix} \Rightarrow \begin{pmatrix} \widetilde{G}_0(\cdot) \\ \mathcal{N}(0_{d_\theta}, V(\gamma_0)) \end{pmatrix},$$

where $DQ_n(\theta_n)$ *is defined in Assumption D1 of AC12 and* $\widetilde{G}_0(\cdot)$ *is defined in part (i) of this assumption.*

Assumptions C2(i) and C3 of AC12 ensure the marginal convergence of the empirical process $D_\psi Q_n(\psi_{0,n}, \cdot) - E_{\gamma_n} D_\psi Q_n(\psi_{0,n}, \cdot)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. Assumption D3(i) of AC12 ensures the marginal convergence of the random variable $DQ_n(\theta_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The next assumption is a standard regularity condition on the second (stochastic) partial derivative matrix of the criterion function with respect to π^1.

Assumption FSReg5. *(i) Under any* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *with* $\gamma_0 \in \Gamma$, $\sup_{\theta \in \Theta} \|D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) - \widetilde{H}_{0,\pi^1\pi^1}(\theta)\| \xrightarrow{p} 0$ *for some nonstochastic symmetric* $d_{\pi^1} \times d_{\pi^1}$ *matrix-valued function* $\widetilde{H}_{0,\pi^1\pi^1}(\theta) \equiv \widetilde{H}_{\pi^1\pi^1}(\theta; \gamma_0)$ *on* $\Theta \times \Gamma$ *that is continuous on* Θ $\forall \gamma_0 \in \Gamma$.

(ii) $\lambda_{\min}(\widetilde{H}_{0,\pi^1\pi^1}(\theta)) > 0$ *and* $\lambda_{\max}(\widetilde{H}_{0,\pi^1\pi^1}(\theta)) < \infty$ $\forall \theta \in \Theta$, $\forall \gamma_0 \in \Gamma$.

We are now ready to state a result that extends Theorems 3.1(a) and 3.2(a) of AC12, but applied to the concentrated estimator $\widehat{\theta}_n$ and the random function $\widehat{\pi}_n^1(\cdot)$.

Theorem Conc. *(i) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(i), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4 and Assumptions B1-B3 and C1-C6 of AC12, applied to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Under parameter sequences* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\psi}_n - \psi_n) \\ \widehat{\pi}_n \\ \sqrt{n}(\widehat{\pi}_n^1(\cdot) - \pi_n^1(\cdot)) \end{pmatrix} \Rightarrow \begin{pmatrix} \tau_{0,b}(\pi^*_{0,b}) \\ \pi^*_{0,b} \\ -\widetilde{H}^{-1}_{0,\pi^1\pi^1}(\cdot) \widetilde{G}_0(\cdot) \end{pmatrix},$$

where

$$\pi^*_{0,b} \equiv \pi^*(\gamma_0, b) \equiv \arg\min_{\pi\in\Pi} -\frac{1}{2}(G_0(\pi) + K_0(\pi)b)' H_0^{-1}(\pi)(G_0(\pi) + K_0(\pi)b),$$
$$\tau_{0,b}(\pi) \equiv \tau(\pi; \gamma_0, b) \equiv -H_0^{-1}(\pi; \gamma_0)(G_0(\pi) + K_0(\pi)b) - (b, 0_{d_\zeta})$$

with $(b, 0_{d_\zeta}) \in \mathbb{R}^{d_\psi}$, $\pi^*_{0,b}$ being a random vector and $\{\tau_{0,b}(\pi) : \pi \in \Pi\}$ being a Gaussian process. The underlying functions $H_0(\pi) \equiv H(\pi; \gamma_0)$ and $K_0(\pi) \equiv K(\pi; \gamma_0)$ are defined in Assumptions C4(i) and C5(ii) of AC12, respectively.

(ii) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(ii), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4 and Assumptions B1-B3, C1-C5, C7-C8 and D1-D3 of AC12, applied to the θ and $Q_n(\theta)$ of this paper, hold. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n}\begin{pmatrix} B(\beta_n)(\widehat{\theta}_n - \theta_n) \\ \widehat{\pi}^1_n(\cdot) - \pi^1_n(\cdot) \end{pmatrix} \Rightarrow \begin{pmatrix} \mathcal{N}(0_{d_\theta}, J^{-1}(\gamma_0)V(\gamma_0)J^{-1}(\gamma_0) \\ -\widetilde{H}^{-1}_{0,\pi^1\pi^1}(\cdot)\widetilde{G}_0(\cdot) \end{pmatrix},$$

where $J(\gamma_0)$ and $V(\gamma_0)$ are defined in Assumptions D2 and D3 of AC12.

This theorem is instrumental to deriving the limit theory for the ultimate parameter estimates of interest: $\widehat{\widehat{\theta}}_n = (\widehat{\theta}_n, \widehat{\pi}^1_n(\widehat{\theta}_n))$. This is the goal of the following section.

8 Limit Theory for Original Parameter Estimates

We now proceed to find the limit theory for the original parameter estimator of interest $\widehat{\widehat{\theta}}_n$ under a comprehensive class of identification strengths using the results of the previous section.

Let $\pi^1_{n,\theta}(\theta^*) = \partial\pi^1_n(\theta)/\partial\theta'|_{\theta=\theta^*}$. Partition $\pi^1_{n,\theta}(\theta^*)$ conformably with $\theta' = (\psi, \pi)$: $\pi^1_{n,\theta}(\theta^*) = [\pi^1_{n,\psi}(\theta^*) : \pi^1_{n,\pi}(\theta^*)]$. Suppose $rank(\pi^1_{n,\pi}(\theta)) = d^*_{\pi,n}$ for all $\theta \in \Theta_\epsilon \equiv \{\theta \in \Theta : \|\beta\| < \epsilon\}$. For $\theta \in \Theta_\epsilon$, let $A_n(\theta) \equiv [A_{1,n}(\theta)' : A_{2,n}(\theta)']'$ be an orthogonal $d_{\pi^1} \times d_{\pi^1}$ matrix such that $A_{1,n}(\theta)$ is a $(d_{\pi^1} - d^*_{\pi,n}) \times d_{\pi^1}$ matrix whose rows span the null space of $\pi^1_{n,\pi}(\theta)'$ and $A_{2,n}(\theta)$ is a $d^*_{\pi,n} \times d_{\pi^1}$ matrix whose rows span the column space of $\pi^1_{n,\pi}(\theta)$. The matrix $A_{1,n}(\theta)$ essentially rotates $\pi^1_n(\theta)$ "off" the direction of π while the matrix $A_{2,n}(\theta)$ rotates $\pi^1_n(\theta)$ in the direction of π. The estimate $\widehat{\pi}^1_n = \widehat{\pi}^1_n(\widehat{\theta}_n)$ has very different limiting behavior after being rotated by either of these two matrices, with one "direction" converging at the $\sqrt{n}$-rate and the other being inconsistent. Similar asymptotic behavior can be found in the related contexts of, e.g., Phillips (1989) and Antoine and Renault (2009, 2012), where parameters of interest are functions of quantities with different convergence rates. Indeed, the rotation approach used in the limit theory here has antecedents in many distinct but related contexts including Sargan (1983), Phillips (1989), Antoine and Renault (2009, 2012), Andrews and Cheng (2014) and Phillips (2015). Let $\pi^1_{0,\theta}(\theta^*) = \partial\pi^1_0(\theta)/\partial\theta'|_{\theta=\theta^*}$ with the analogous partition $\pi^1_{0,\theta}(\theta^*) = [\pi^1_{0,\psi}(\theta^*) : \pi^1_{0,\pi}(\theta^*)]$.

The following assumptions impose regularity conditions on the mapping $\pi^1_n : \Theta \to \Pi^1(\Theta)$.

Assumption Reg5. *For all $n \geq 1$, the following statements hold:*

(i) $\pi^1_n(\theta)$ is continuously differentiable on Θ.

(ii) $rank(\pi^1_{n,\pi}(\theta)) = d^*_{\pi,n}$ for some constant $d^*_{\pi,n} \leq \min(d_{\pi^1}, d_\pi)$ and $d^*_{\pi,n} - rank(\pi^1_{0,\pi}(\theta)) \to 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ $\forall\theta \in \Theta_\epsilon$ for some $\epsilon > 0$.

Analogous assumptions can be found in, e.g., Assumptions R1 and R2 of Andrews and Cheng (2014). The major difference with these assumptions is the condition in Assumption Reg5(ii) on the limit of $d^*_{\pi,n}$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. This essentially restricts the class of sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ under study to those that yield a limit for $d^*_{\pi,n}$. This assumption is not restrictive since for any given sequence $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, one could find a subsequence $\{\gamma_{\omega_n}\}$ along which d^*_{π,ω_n} converges since d^*_{π,ω_n} is limited to a finite number of values.

Define

$$\eta_n(\theta) \equiv \begin{cases} \sqrt{n}A_{1,n}(\theta)\{\pi^1_n(\psi_n, \pi) - \pi^1_n(\psi_n, \pi_n)\}, & \text{if } d^*_{\pi,n} < d_{\pi^1} \\ 0, & \text{if } d^*_{\pi,n} = d_{\pi^1}. \end{cases}$$

Assumption Reg6. *Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\eta_n(\widehat{\theta}_n) \xrightarrow{p} 0$.*

Denote the Gaussian random vector to which $\sqrt{n}B(\beta_n)(\widehat{\theta}_n - \theta_n)$ converges in Theorem Conc (ii) as $Z_\theta = (Z'_\psi, Z'_\pi)' \sim \mathcal{N}(0_{d_\theta}, J^{-1}(\gamma_0)V(\gamma_0)J^{-1}(\gamma_0))$, partitioned conformably with ψ and π. We are now ready to state the main result of this section. In what follows, $A_{1,0}(\cdot)$ and $A_{2,0}(\cdot)$ are defined analogously to $A_{1,n}(\cdot)$ and $A_{2,n}(\cdot)$, replacing $\pi^1_{n,\theta}(\theta)$ with $\pi^1_{0,\theta}(\theta)$ in the definition.

Theorem Est. *(i) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(i), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4–Reg6 and B1-B3 and C1-C6 of AC12, applied to the θ and $Q_n(\theta)$ of this paper, hold. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\widehat{\psi}_n - \psi_n) \\ \widehat{\pi}_n \\ \sqrt{n}A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}^1_n - \pi^1_n) \\ A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}^1_n - \pi^1_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{0,b}(\pi^*_{0,b}) \\ \pi^*_{0,b} \\ A_{1,0}(\psi_0, \pi^*_{0,b})\{\pi^1_{0,\psi}(\psi_0, \pi^*_{0,b})\tau_{0,b}(\pi^*_{0,b}) - \widetilde{H}^{-1}_{0,\pi^1\pi^1}(\psi_0, \pi^*_{0,b})\widetilde{G}_0(\psi_0, \pi^*_{0,b})\} \\ A_{2,0}(\psi_0, \pi^*_{0,b})\{\pi^1_0(\psi_0, \pi^*_{0,b}) - \pi^1_0(\psi_0, \pi_0)\} \end{pmatrix}.$$

(ii) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(ii), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4–Reg6 and Assumptions B1-B3, C1-C5, C7-C8 and D1-D3 of AC12, applied to

the θ and $Q_n(\theta)$ of this paper, hold. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n}\begin{pmatrix} B(\beta_n)(\widehat{\theta}_n - \theta_n) \\ A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) \\ \iota(\beta_n)A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_\theta \\ A_{1,0}(\theta_0)[\pi^1_{0,\psi}(\theta_0)Z_\psi - \widetilde{H}^{-1}_{0,\pi^1\pi^1}(\theta_0)\widetilde{G}_0(\theta_0)] \\ A_{2,0}(\theta_0)\pi^1_{0,\pi}(\theta_0)Z_\pi \end{pmatrix},$$

if $\beta_0 = 0$ and

$$\sqrt{n}(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n) \xrightarrow{d} \begin{pmatrix} B^{-1}(\beta_0)Z_\theta \\ \pi^1_{0,\theta}(\theta_0)B^{-1}(\beta_0)Z_\theta - \widetilde{H}^{-1}_{0,\pi^1\pi^1}(\theta_0)\widetilde{G}_0(\theta_0) \end{pmatrix}$$

if $\beta_0 \neq 0$.

Due to the rotation by $A_{1,n}(\widehat{\theta}_n)$ and $A_{2,n}(\widehat{\theta}_n)$, Theorem Est does not directly express the limiting distribution of $\widehat{\pi}_n^1$. However this is can be obtained as a corollary.

Let $A_n(\theta) = [A_n^1(\theta) : A_n^2(\theta)]$, which forms a conformable partition so that $A_n^1(\theta)$ is a $d_{\pi^1} \times (d_{\pi^1} - d^*_{\pi,n})$ matrix and $A_n^2(\theta)$ is a $d_{\pi^1} \times d^*_{\pi,n}$ matrix. Define $A_0(\theta) = [A_0^1(\theta) : A_0^2(\theta)]$ analogously. Let

$$\widetilde{B}(\beta) \equiv \begin{pmatrix} I_{d_\psi} & 0_{d_\psi \times d_{\widetilde{\pi}}} \\ 0_{d_{\widetilde{\pi}} \times d_{d_\psi}} & \iota(\beta)I_{d_{\widetilde{\pi}}} \end{pmatrix},$$

$$\bar{B}(\beta) \equiv \begin{pmatrix} \iota(\beta)I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & I_{d_\pi} \end{pmatrix} = \iota(\beta)B^{-1}(\beta),$$

and $\bar{H}^{-1}_{0,\pi^1\pi^1}(\theta_0) \equiv \iota(\beta_0)\widetilde{H}^{-1}_{0,\pi^1\pi^1}(\theta_0)$.

Corollary Est. *(i) Under the assumptions of Theorem Est(i) and $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\widehat{\psi}_n - \psi_n) \\ \widehat{\pi}_n \\ \widehat{\pi}_n^1 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{0,b}(\pi^*_{0,b}) \\ \pi^*_{0,b} \\ \pi^{1*}_{0,b} \end{pmatrix},$$

where

$$\pi^{1*}_{0,b} \equiv \pi_0^1 + A_0^2(\psi_0, \pi^*_{0,b})A_{2,0}(\psi_0, \pi^*_{0,b})(\pi_0^1(\psi_0, \pi^*_{0,b}) - \pi_0^1(\psi_0, \pi_0)).$$

(ii) Under the assumptions of Theorem Est(ii) and $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n}\widetilde{B}(\beta_n)(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n) \xrightarrow{d} \begin{pmatrix} Z_\theta \\ \pi^1_{0,\theta}(\theta_0)\bar{B}(\beta_0)Z_\theta - \bar{H}^{-1}_{0,\pi^1\pi^1}(\theta_0)\widetilde{G}_0(\theta_0) \end{pmatrix} \equiv Z_{\widetilde{\theta}}.$$

Remark 8.1. *Though $\pi_{0,b}^{1*}$ is the limiting random vector for $\widehat{\pi}_n^1$ in Corollary Est, including the asymptotic counterpart to the $O_p(n^{-1/2})$ term $A_n^1(\widehat{\theta}_n)A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1)$ provides a better approximation to the finite sample distribution of $\widehat{\pi}_n^1$. That is, the distributional approximation for $\widehat{\pi}_n^1$ by*

$$\pi_{0,b}^{1a} = \pi_{0,b}^{1*} + n^{-1/2}A_0^1(\psi_0, \pi_{0,b}^*)A_{1,0}(\psi_0, \pi_{0,b}^*)\{\pi_{0,\psi}^1(\psi_0, \pi_{0,b}^*)\tau_{0,b}(\pi_{0,b}^*) - \widetilde{H}_{0,\pi^1\pi^1}^{-1}(\psi_0, \pi_{0,b}^*)\widetilde{G}_0(\psi_0, \pi_{0,b}^*)\}$$

serves better in finite samples than that by $\pi_{0,b}^{1*}$, even though they are asymptotically equivalent. See the Monte Carlo results for the threshold-crossing model example in Section 11.1.

9 Wald Statistics

We are interested in testing general nonlinear hypotheses of the form

$$H_0 : \widetilde{r}(\widetilde{\theta}) = v \in \mathbb{R}^{d_{\widetilde{r}}}$$

using the standard Wald statistic. The usual Wald statistic for H_0 based upon $\widehat{\widetilde{\theta}}_n$ can be written as

$$\widetilde{W}_n(v) \equiv n(\widetilde{r}(\widehat{\widetilde{\theta}}_n) - v)'(\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)\widetilde{B}^{-1}(\widehat{\beta}_n)\widehat{\widetilde{\Sigma}}_n\widetilde{B}^{-1}(\widehat{\beta}_n)\widetilde{r}_{\widetilde{\theta}}^{-1}(\widehat{\widetilde{\theta}}_n)')(r(\widehat{\widetilde{\theta}}_n) - v),$$

where $\widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}) \equiv \partial\widetilde{r}(\widetilde{\theta})/\partial\widetilde{\theta}' \equiv [\widetilde{r}_{\psi}(\widetilde{\theta}) : \widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta})] \in \mathbb{R}^{d_{\widetilde{r}} \times d_{\widetilde{\theta}}}$ and the covariance matrix estimator of $\widehat{\widetilde{\theta}}_n$ is specified as

$$\widehat{\widetilde{\Sigma}}_n \equiv \begin{pmatrix} \widehat{\Sigma}_n & \widehat{\Sigma}_n^{12} \\ \widehat{\Sigma}_n^{12\prime} & \widehat{\Sigma}_n^{22} \end{pmatrix}$$

with $\widehat{\Sigma}_n \equiv \widehat{J}_n^{-1}\widehat{V}_n\widehat{J}_n^{-1}$ being as defined in AC12 to be an estimator of $\Sigma(\gamma_0) = E_{\gamma_0}[Z_\theta Z_\theta']$, $\widehat{\Sigma}_n^{12}$ being an estimator of

$$\Sigma^{12}(\gamma_0) = \Sigma(\gamma_0)\bar{B}(\beta_0)\pi_{0,\theta}^1(\theta_0)' - E_{\gamma_0}[Z_\theta\widetilde{G}(\theta_0)']\bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)$$

and $\widehat{\Sigma}_n^{22}$ being an estimator of

$$\begin{aligned}\Sigma^{22}(\gamma_0) = {} & \pi_{0,\theta}^1(\theta_0)\bar{B}(\beta_0)\Sigma(\gamma_0)\bar{B}(\beta_0)\pi_{0,\theta}^1(\theta_0)' - \pi_{0,\theta}^1(\theta_0)\bar{B}(\beta_0)E_{\gamma_0}[Z_\theta\widetilde{G}_0(\theta_0)']\bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0) \\ & - \bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)E_{\gamma_0}[\widetilde{G}_0(\theta_0)Z_\theta']\bar{B}(\beta_0)\pi_{0,\theta}^1(\theta_0)' + \bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)\widetilde{\Omega}_0(\theta_0,\theta_0)\bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)\end{aligned}$$

under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The following assumptions are similar to R1 and R2 of AC14, applied to $\widetilde{r}(\widetilde{\theta})$.

Assumption Res1. *(i) $\widetilde{r}(\widetilde{\theta})$ is continuously differentiable on $\widetilde{\Theta}$.*
(ii) rank$(\widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta})) = d^*_{\widetilde{\pi}}$ for some constant $d^*_{\widetilde{\pi}} \leq \min\{d_{\widetilde{r}}, d_{\widetilde{\pi}}\}$ for all $\widetilde{\theta} \in \widetilde{\Theta}_\epsilon \equiv \{\widetilde{\theta} \in \widetilde{\Theta} : \|\beta\| < \epsilon\}$ for some $\epsilon > 0$.

In a similar spirit to the rotations used to characterize the limit theory for $\widehat{\pi}^1_n = \widehat{\pi}^1_n(\widehat{\theta}_n)$, we rotate the restrictions being tested when evaluated at $\widehat{\widetilde{\theta}}_n$, $\widetilde{r}(\widehat{\widetilde{\theta}}_n)$. Let $\widetilde{A}(\widetilde{\theta}) = [\widetilde{A}_1(\widetilde{\theta})' : \widetilde{A}_2(\widetilde{\theta})']'$ be an orthogonal $d_{\widetilde{r}} \times d_{\widetilde{r}}$ matrix such that $\widetilde{A}_1(\widetilde{\theta})$ is a $(d_{\widetilde{r}} - d^*_{\widetilde{\pi}}) \times d_{\widetilde{r}}$ matrix whose rows span the null space of $\widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta})'$ and $\widetilde{A}_2(\widetilde{\theta})$ is a $d^*_{\widetilde{\pi}} \times d_{\widetilde{r}}$ matrix whose rows span the column space of $\widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta})$ with $\widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}) \equiv [\widetilde{r}_\psi(\widetilde{\theta}) : \widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta})]$. Let

$$\widetilde{\eta}_n(\widetilde{\theta}) \equiv \begin{cases} n^{1/2}\widetilde{A}_1(\widetilde{\theta})\left\{\widetilde{r}(\psi_n, \widetilde{\pi}) - \widetilde{r}(\psi_n, \widetilde{\pi}_n)\right\}, & \text{if } d^*_{\widetilde{\pi}} < d_{\widetilde{r}} \\ 0, & \text{if } d^*_{\widetilde{\pi}} = d_{\widetilde{r}}. \end{cases}$$

Assumption Res2. *Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\widetilde{\eta}_n(\widehat{\widetilde{\theta}}_n) \xrightarrow{p} 0$.*

When β is scalar, let $\Sigma^{12}_0(\widetilde{\theta}) \equiv \Sigma^{12}(\widetilde{\theta}; \gamma_0)$ and $\Sigma^{22}_0(\widetilde{\theta}) \equiv \Sigma^{22}(\widetilde{\theta}; \gamma_0)$ be some nonstochastic $d_\theta \times d_{\pi^1}$ and $d_{\pi^1} \times d_{\pi^1}$ matrix-valued functions for $\widetilde{\theta} \in \widetilde{\Theta}$. Let

$$\begin{aligned}\widetilde{\Sigma}_0(\theta) \equiv \widetilde{\Sigma}(\theta; \gamma_0) &= \begin{pmatrix} \Sigma_0(\theta) & \Sigma_0(\theta)\begin{pmatrix} 0_{d_\psi \times d_{\pi^1}} \\ \pi^1_{0,\pi}(\theta) \end{pmatrix} \\ (0_{d_{\pi^1} \times d_\psi} : \pi^1_{0,\pi}(\theta)')\Sigma_0(\theta) & (0_{d_{\pi^1} \times d_\psi} : \pi^1_{0,\pi}(\theta)')\Sigma_0(\theta)\begin{pmatrix} 0_{d_\psi \times d_{\pi^1}} \\ \pi^1_{0,\pi}(\theta) \end{pmatrix} \end{pmatrix} \\ &\equiv \begin{pmatrix} \Sigma_0(\theta) & \Sigma^{12}_0(\theta) \\ \Sigma^{12}_0(\theta)' & \Sigma^{22}_0(\theta) \end{pmatrix}, \\ \widetilde{\Sigma}_0(\pi) \equiv \widetilde{\Sigma}(\pi; \gamma_0) &= \widetilde{\Sigma}(0, \zeta_0, \widetilde{\pi}; \gamma_0),\end{aligned}$$

where $\Sigma_0(\theta) \equiv \Sigma(\theta; \gamma_0)$ is defined in (4.4) of AC12. In addition to Assumption V1 of AC12, the following assumption describes the limiting behavior of $\widehat{\widetilde{\Sigma}}_n$ under weak identification sequences. For this assumption, $\Sigma_0(\pi) \equiv \Sigma(\pi; \gamma_0)$ is defined in (4.4) of AC12.

Assumption Var1. *(Scalar β) (i) $\widehat{\Sigma}^{12}_n = \widehat{\Sigma}^{12}_n(\widehat{\theta}_n)$ and $\widehat{\Sigma}^{22}_n = \widehat{\Sigma}^{22}_n(\widehat{\theta}_n)$ for some (stochastic) $d_\theta \times d_{\pi^1}$ and $d_{\pi^1} \times d_{\pi^1}$ matrix-valued functions $\widehat{\Sigma}^{12}_n(\cdot)$ and $\widehat{\Sigma}^{22}_n(\cdot)$ on $\widetilde{\Theta}$ that satisfy $\sup_{\theta \in \Theta} \|\widehat{\Sigma}^{12}_n(\theta) - \Sigma^{12}_0(\theta)\| \xrightarrow{p} 0$ and $\sup_{\theta \in \Theta} \|\widehat{\Sigma}^{22}_n(\theta) - \Sigma^{22}_0(\theta)\| \xrightarrow{p} 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $|b| < \infty$.*

(ii) $\Sigma^{12}_0(\theta)$ and $\Sigma^{22}_0(\theta)$ are continuous on $\theta \in \Theta$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(iii) $\lambda_{\min}(\Sigma_0(\pi)) > 0$ and $\lambda_{\max}(\widetilde{\Sigma}_0(\pi)) < \infty$ $\forall \pi \in \Pi$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

We deal with the scalar β case here in the main text. See Appendix B for details on the vector β case.

When estimating $\Sigma^{12}(\gamma_0)$ and $\Sigma^{22}(\gamma_0)$ by $\widehat{\Sigma}_n^{12}$ and $\widehat{\Sigma}_n^{22}$ in practice, the β_0's appearing in $\bar{B}(\beta_0)$ and $\bar{H}^{-1}_{0,\pi^1\pi^1}(\theta_0)$ would be replaced by the estimator $\widehat{\beta}_n$. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $|b| < \infty$, along with Assumption Var1 and the results of Corollary Conc(i), this implies

$$\begin{aligned}
\widehat{\Sigma}_n^{12}(\widehat{\widetilde{\theta}}_n) &\xrightarrow{d} \Sigma_0(\theta^*_{0,b})\bar{B}(0)\pi^1_{0,\theta}(\theta^*_{0,b})' = \Sigma_0(\theta^*_{0,b})[0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})]' \\
\widehat{\Sigma}_n^{22}(\widehat{\widetilde{\theta}}_n) &\xrightarrow{d} \pi^1_{0,\theta}(\theta^*_{0,b})\bar{B}(0)\Sigma_0(\theta^*_{0,b})\bar{B}(0)\pi^1_{0,\theta}(\theta^*_{0,b})' \\
&= [0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})]\Sigma_0(\theta^*_{0,b})[0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})]'.
\end{aligned}$$

since $\bar{H}^{-1}_{0,\pi^1\pi^1}(\theta_0) = \iota(0)H^{-1}_{0,\pi^1\pi^1}(\theta_0) = 0$. This implies the limiting matrix $\Sigma_0(\theta^*_{0,b})$ has reduced rank equal to the rank of $\Sigma_0(\theta^*_{0,b})$, which is equal to d_θ by Assumption Var1(iii). This problem also arises under semi-strong identification sequences. Indeed, note that by Corollary Est(ii), under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ with $\beta_0 = 0$,

$$\sqrt{n}\widetilde{B}(\beta_n)(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n) \xrightarrow{d} \begin{pmatrix} I_{d_\theta} \\ \pi^1_{0,\theta}(\theta_0)\bar{B}(0) \end{pmatrix} Z_\theta \tag{9.1}$$

Standard variance matrix estimators $\widehat{\widetilde{\Sigma}}_n$ are consistent for the variance matrix of this Gaussian limiting random vector under semi-strong identification sequences. Hence, $\widehat{\widetilde{\Sigma}}_n$ is singular in the limit under these sequences as well.

Depending upon the restrictions being tested, these singularities in the limit of $\widehat{\widetilde{\Sigma}}_n$ under weak and semi-strong identification sequences can cause the standard Wald statistic to diverge. This can perhaps be most easily seen for the standard full vector test restriction $\widetilde{r}(\widetilde{\theta}) = \widetilde{\theta}$ under semi-strong identification sequences: $\widetilde{W}_n(\widetilde{\theta}_n) = n(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n)'\widetilde{B}(\widehat{\beta}_n)\widehat{\widetilde{\Sigma}}_n^{-1}\widetilde{B}(\widehat{\beta}_n)(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n)$. Under semi-strong identification and H_0, $\sqrt{n}\widetilde{B}(\widehat{\beta_n})(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n)$ weakly converges to the right hand side of (9.1) and with $\Sigma(\gamma_0) = E_{\gamma_0}[Z_\theta Z_\theta']$,

$$\widehat{\widetilde{\Sigma}}_n \xrightarrow{p} \begin{pmatrix} \Sigma(\gamma_0) & \Sigma(\gamma_0)\bar{B}(\beta_0)\pi^1_{0,\theta}(\theta_0)' \\ \pi^1_{0,\theta}(\theta_0)\bar{B}(\beta_0)\Sigma(\gamma_0) & \pi^1_{0,\theta}(\theta_0)\bar{B}(\beta_0)\Sigma(\gamma_0)\bar{B}(\beta_0)\pi^1_{0,\theta}(\theta_0)' \end{pmatrix}.$$

Consider inverting $\widehat{\widetilde{\Sigma}}_n$. The lower right hand block of $\widehat{\widetilde{\Sigma}}_n^{-1}$ is

$$(\widehat{\widetilde{\Sigma}}_{n,22} - \widehat{\widetilde{\Sigma}}'_{n,12}\widehat{\widetilde{\Sigma}}^{-1}_{n,11}\widehat{\widetilde{\Sigma}}_{n,12})^{-1}$$

but under semi-strong identification,

$$\widehat{\widehat{\Sigma}}_{n,22} - \widehat{\widehat{\Sigma}}'_{n,12}\widehat{\widehat{\Sigma}}^{-1}_{n,11}\widehat{\widehat{\Sigma}}_{n,12} \xrightarrow{p} \pi^1_{0,\theta}(\theta_0)\bar{B}(0)\Sigma(\gamma_0)\bar{B}(0)\pi^1_{0,\theta}(\theta_0)'$$
$$- \pi^1_{0,\theta}(\theta_0)\bar{B}(0)\Sigma(\gamma_0)\Sigma^{-1}(\gamma_0)\Sigma(\gamma_0)\bar{B}(0)\pi^1_{0,\theta}(\theta_0)' = 0.$$

Therefore, the Wald statistic diverges *under the null*.

Because the Wald statistic diverges under the null and weak or semi-strong identification for some null hypotheses, standard Wald tests that make use of $\chi^2_{d_{\widetilde{r}}}$ CVs exhibit size distortion of the most extreme kind: their asymptotic size is equal to one. In order to construct Wald tests with uniform asymptotic size control, we limit our attention to null hypotheses that do no cause Wald statistic divergence under the null. Standard Wald tests still exhibit size distortions in these cases since the Wald statistics are not asymptotically $\chi^2_{d_{\widetilde{r}}}$-distributed under weak identification sequences (see Theorem Wald(i) below). We know the form of singularity in the asymptotic variance matrix under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ that causes the Wald statistic to diverge for certain restrictions, e.g., $rank(\widetilde{\Sigma}(\gamma_0)) = rank(\Sigma(\gamma_0)) = d_\theta$. This enables us to write down verifiable sufficient rank conditions on the restrictions that ensure the Wald statistic does not diverge.

Assumption Res3. *(i)* $d_{\widetilde{r}} \le d_\theta$
(ii) $rank(\widetilde{A}_1(\widetilde{\theta})\widetilde{r}_\psi(\widetilde{\theta})) = d_{\widetilde{r}} - d^*_{\widetilde{\pi}}$ *and* $rank(\widetilde{A}_2(\widetilde{\theta})(\widetilde{r}_\pi(\widetilde{\theta}) + \widetilde{r}_{\pi^1}(\widetilde{\theta})\partial\pi^1_0(\theta)/\partial\pi')) = d^*_{\widetilde{\pi}}$ $\forall\widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$, $\forall\gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

One-dimensional restrictions trivially satisfy this assumption. Subvector restrictions satisfy this assumption so long as the subvector is not too large (it must have dimension smaller than d_θ) and does not contain more than d_π entries of $\widetilde{\pi}$. It is interesting to note that while Andrews and Guggenberger (2014) and Andrews and Mikusheva (2014) cannot generally directly conduct one-dimensional inference that is uniformly valid, we can by properly constructing CVs (see the following section). Conversely, we cannot directly conduct full vector inference that is uniformly valid while this is precisely what the methods Andrews and Guggenberger (2014) and Andrews and Mikusheva (2014) are for (in moment condition models).[10]

Along with Assumption V2 of AC12 applied to $\widehat{\Sigma}_n$, the following assumption presumes that $\widehat{\widehat{\Sigma}}_n$ is consistent under (semi-)strong identification.

Assumption Var2. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\widehat{\Sigma}^{12}_n \xrightarrow{p} \Sigma^{12}(\gamma_0)$ *and* $\widehat{\Sigma}^{22}_n \xrightarrow{p} \Sigma^{22}(\gamma_0)$.

Before stating the main result of this section, we must introduce some notation. First, let

[10] Andrews and Mikusheva (2014) cannot handle moment conditions for which the asymptotic variance matrix of the moments is singular. This occurs for the ML estimators of this paper.

$\widetilde{\theta}^*_{0,b} = (\psi_0, \pi^*_{0,b}, \pi^{1*}_{0,b})$. Second, let

$$q^{\widetilde{A}}_{0,b}(\widetilde{\theta}) \equiv q^{\widetilde{A}}(\widetilde{\theta};\gamma_0,b) = \left(\begin{array}{c} \widetilde{A}_1(\widetilde{\theta})\widetilde{r}_\psi(\widetilde{\theta})\tau_{0,b}(\pi) \\ \iota(b+\tau^\beta_{0,b}(\pi))\widetilde{A}_2(\widetilde{\theta})(\widetilde{r}(\widetilde{\theta})-\widetilde{r}(\widetilde{\theta}_0)) \end{array} \right),$$

where $\tau^\beta_{0,b}(\pi)$ denotes the subvector of the first d_β elements of $\tau_{0,b}(\pi)$. Third, let

$$\widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}) = \left(\begin{array}{cc} \widetilde{A}_1(\widetilde{\theta})\widetilde{r}_\psi(\widetilde{\theta}) & 0 \\ 0 & \widetilde{A}_2(\widetilde{\theta})\widetilde{r}_{\widetilde{\pi}}(\widetilde{\theta}) \end{array} \right).$$

Finally, let

$$\bar{\bar{\widetilde{\Sigma}}}_{0,b}(\widetilde{\pi}) \equiv \bar{\bar{\widetilde{\Sigma}}}(\widetilde{\pi};\gamma_0,b) = \begin{cases} \widetilde{\Sigma}_0(\widetilde{\pi}), & \text{if } \beta \text{ is scalar} \\ \widetilde{\Sigma}_0(\widetilde{\pi},\omega^*_{0,b}(\widetilde{\pi})), & \text{if } \beta \text{ is a vector} \end{cases}$$

where for $\pi \in \Pi$,

$$\omega^*_{0,b}(\widetilde{\pi}) \equiv \omega^*(\pi;\gamma_0,b) \equiv \frac{\tau^\beta_{0,b}(\pi)}{\|\tau^\beta_{0,b}(\pi)\|}.$$

Under a sequence $\{\gamma_n\}$, we consider the sequence of null hypotheses $H_0 : \widetilde{r}(\widetilde{\theta}) = v_n$, where $v_n = \widetilde{r}(\widetilde{\theta}_n)$.

Theorem Wald. *(i) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(i), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4–Reg6, Res1–Res3, Var1 and Assumptions B1-B3, C1-C6 and V1 of AC12, applied to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$,

$$\widetilde{W}_n(v_n) \xrightarrow{d} \widetilde{W}(b,\gamma_0) \equiv q^{\widetilde{A}}_{0,b}(\widetilde{\theta}^*_{0,b})'(\widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}^*_{0,b})\bar{\bar{\widetilde{\Sigma}}}_{0,b}(\widetilde{\pi}^*_{0,b})\widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}^*_{0,b})')^{-1}q^{\widetilde{A}}_{0,b}(\widetilde{\theta}^*_{0,b}).$$

(ii) Suppose Assumptions FSID1, FSReg1–FSReg3, FSReg4(ii), FSReg5, FSCF1–FSCF2, CF3, ID3, Reg4–Reg6, Res1–Res3, Var2 and Assumptions B1-B3, C1-C5, C7-C8, D1-D3 and V2 of AC12, applied to the θ *and* $Q_n(\theta)$ *of this paper, hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\widetilde{W}_n(v_n) \xrightarrow{d} \chi^2_{d_{\widetilde{r}}}.$$

10 Robust Wald Inference

As can be seen, e.g., in Figures 5–8, for some $b \in \mathbb{R}^{d_\beta}_\infty$, the limit distribution of $\widetilde{W}(b,\gamma_0)$ given in Theorem Wald(i) provides a good approximation to the finite-sample distribution of $\widetilde{W}_n(v)$. This limit distribution depends upon the unknown nuisance parameters b and γ_0. Letting $c_{1-\alpha}(b,\gamma_0)$ denote the $1-\alpha$ quantile of this distribution, a standard approach to CV construction for a test of size α would be to evaluate $c_{1-\alpha}(\cdot)$ at a consistent estimate of (b, γ_0). However, the nuisance parameter b and some elements in γ_0 are not consistently estimable under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$, lending such an approach to size distortions. This feature of the problem leads us to consider more sophisticated CV construction methods that lead to correct asymptotic size for the test. We will restrict our focus to testing problems for which the distribution function of $\widetilde{W}(b, \gamma_0)$ in Theorem Wald(i) only depends upon γ_0 through the parameters ζ_0 and π_0 and an additional consistently-estimable finite-dimensional parameter δ_0. This is the case in all of the examples we have encountered. Without loss of generality, we will assume δ is a component of $\widetilde{\phi}$ so we can write $\widetilde{\phi} = (\delta, \phi)$.[11]

Assumption FD. *The distribution function of $\widetilde{W}(b, \gamma_0)$ depends upon γ_0 only through ζ_0, π_0, and some $\delta_0 \in \mathbb{R}^{d_\delta}_{\infty}$ such that under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ or $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ there is an estimator $\widehat{\delta}_n$ with $\widehat{\delta}_n \xrightarrow{p} \delta_0$.*

We will "plug-in" consistent estimators for ζ_0 and δ_0, $\widehat{\zeta}_n$ and $\widehat{\delta}_n$, when constructing the CVs. The first construction is more computationally straightforward while the second leads to tests with better finite-sample properties.

10.1 Identification Category Selection CVs

The first type of CV we consider is the direct analog of AC12's (plug-in and null-imposed) Type I Robust CV. Define $t_n \equiv (n\widehat{\beta}_n' \widehat{\Sigma}^{-1}_{\beta\beta,n} \widehat{\beta}_n / d_\beta)^{1/2}$, where $\widehat{\Sigma}_{\beta\beta,n}$ is equal to the upper left $d_\beta \times d_\beta$ block of $\widehat{\Sigma}_n$ and suppose $\{\kappa_n\}$ is a sequence of constants such that $\kappa_n \to \infty$ and $\kappa_n / n^{1/2} \to 0$ (Assumption K of AC12). Then the ICS CV for a test of size α is defined as follows:

$$c^{ICS}_{1-\alpha,n} \equiv \begin{cases} \chi^2_{d_{\widetilde{r}}}(1-\alpha)^{-1} & \text{if } t_n > \kappa_n, \\ c^{LF}_{1-\alpha,n} & \text{if } t_n \leq \kappa_n \end{cases}$$

where $\chi^2_{d_{\widetilde{r}}}(1-\alpha)^{-1}$ is the $(1-\alpha)$ quantile of a $\chi^2_{d_{\widetilde{r}}}$-distributed random variable and $c^{LF}_{1-\alpha,n} \equiv \sup_{\lambda \in \widehat{\Lambda}_n \cap \Lambda(v)} c_{1-\alpha}(\lambda)$ with $\widehat{\Lambda}_n \equiv \{\lambda = (b, \gamma) \in \Lambda : \gamma = (\beta, \widehat{\zeta}_n, \pi, \widehat{\delta}_n, \phi)\}$, $\Lambda(v) \equiv \{\lambda = (b, \gamma_0) \in \Lambda : \widetilde{r}(\widetilde{\theta}_0) = v\}$, and $\Lambda \equiv \{\lambda = (b, \gamma_0) \in \mathbb{R}^{d_\beta}_{\infty} \times \Gamma :$ for some $\{\gamma_n\} \in \Gamma(\gamma_0), n^{1/2}\beta_n \to b\}$. That is, we both impose H_0 and "plug-in" consistent estimators $\widehat{\zeta}_n$ and $\widehat{\delta}_n$ of ζ_0 and δ_0 in the construction of the CV. This leads to tests with smaller CVs and hence better power (see, e.g., AC12 for a discussion).[12] Under the assumptions of Theorem Wald, Assumption FD and the following

[11] It is possible to relax this restriction and modify the CVs accordingly. However, we have not found an example where this is necessary.

[12] As in AC12, one may also choose not to impose H_0 in the CV construction since it is misspecified under the alternative. Then, simply replace $\widehat{\Lambda}_n \cap \Lambda(v)$ with $\widehat{\Lambda}_n$ in the expression for $c^{LF}_{1-\alpha}$. Also, any consistent estimators of the components of γ_0 may be analogously "plugged-in".

assumption, we can establish the correct asymptotic size of tests using the Wald statistic and ICS CVs.

Assumption DF1. *The distribution function of* $\widetilde{W}(b,\gamma_0)$ *is continuous at* $\chi^2_{d_{\widetilde{r}}}(1-\alpha)^{-1}$ *and* $\sup_{\lambda\in\Lambda_0\cap\Lambda(v)} c_{1-\alpha}(\lambda)$*, where* $\Lambda_0 \equiv \{\lambda=(b,\gamma)\in\Lambda : \gamma=(\beta,\zeta_0,\pi,\delta_0,\phi)\}$.

Proposition ICS. *Under the assumptions of Theorem Wald, Assumption K of AC12 and Assumptions FD and DF1,* $\limsup_{n\to\infty}\sup_{\gamma\in\Gamma:\widetilde{r}(\widetilde{\theta})=v} P_\gamma(\widetilde{W}_n(v) > c^{ICS}_{1-\alpha,n}) = \alpha$.

10.2 Adjusted-Bonferroni CVs

The second type of CV we consider is a modification of the adjusted-Bonferroni CV of McCloskey (2012). The basic idea here is to use the data to narrow down the set of localization parameters b and parameters π from the entire space $\mathcal{P}(\widehat{\zeta}_n,\widehat{\delta}_n) \equiv \{(b,\pi)\in\mathbb{R}^{d_\beta+d_\pi}_\infty$: for some $\gamma_0\in\Gamma$ with $\zeta_0=\widehat{\zeta}_n$ and $\delta_0=\widehat{\delta}_n$, $\pi=\pi_0$ and for some $\{\gamma_n\}\in\Gamma(\gamma_0), n^{1/2}\beta_n\to b\}$, as in the construction of least-favorable CVs, to a data-dependent set and subsequently maximize $c_{1-\alpha}(b,\gamma)$ over b and π in this restricted set. Roughly speaking, this allows the CV to randomly adapt to the data to determine how "guarded" we should be against potential weak identification and which part of the parameter space Π is relevant to the finite-sample testing problem.

Let $\widehat{b}_n = n^{1/2}\widehat{\beta}_n$. Using the results of Theorem Est, we can determine the joint asymptotic distribution of $(\widehat{b}_n,\widehat{\pi}_n)$ under sequences $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ with $\|b\|<\infty$, and consequently construct an asymptotically valid confidence set for (b,π_0). In the context of this paper, the adjusted-Bonferroni CV of McCloskey (2012) uses such a confidence set for (b,π_0) as the data-dependent set from which to form the data-adaptive CV. Though this may be feasible in principle, the formation of such a confidence set would be quite computationally burdensome in our context since the quantiles of the limit random vector $(\tau^{\beta}_{0,b}(\pi^*_{0,b}),\pi^*_{0,b})$ depend upon the underlying parameters (b,π_0) themselves.[13] As a modification, we instead here propose the use of the set $\widehat{I}^a_n(\widehat{b}_n,\widehat{\pi}_n)=\{(b,\pi)\in\mathcal{P}(\widehat{\zeta}_n,\widehat{\delta}_n):[(\widehat{b}_n-b)',(\widehat{\pi}_n-\pi)']\widehat{\widehat{\Sigma}}_n^{-1}[(\widehat{b}_n-b)',(\widehat{\pi}_n-\pi)']'\le\chi^2_{d_\beta+d_\pi}(1-a)^{-1}\}$, where

$$\widehat{\widehat{\Sigma}}_n \equiv \begin{pmatrix} \widehat{\Sigma}_{\beta\beta,n} & n^{-1/2}\|\widehat{\beta}_n\|^{-1}\widehat{\Sigma}_{\beta\pi,n} \\ n^{-1/2}\|\widehat{\beta}_n\|^{-1}\widehat{\Sigma}'_{\beta\pi,n} & n^{-1}\|\widehat{\beta}_n\|^{-2}\widehat{\Sigma}_{\pi\pi,n} \end{pmatrix}$$

with $\widehat{\Sigma}_{\beta\pi,n}$ denoting the upper right $d_\beta\times d_\pi$ block of $\widehat{\Sigma}_n$ and $\widehat{\Sigma}_{\pi\pi,n}$ denoting the lower right $d_\pi\times d_\pi$ block of $\widehat{\Sigma}_n$. This is set is akin to an a-level Wald confidence set for (b,π_0). Though this confidence set does not have asymptotically correct coverage under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ with $\|b\|<\infty$ sequences, it attains nearly correct coverage as $\|b\|\to\infty$. Similarly to the ICS CV in

[13] A similar complication arises in e.g., the formation of an asymptotically valid confidence set for the localization parameter in a local-to-unit root autoregressive model.

the previous subsection, one may both impose H_0 and "plug-in" the values of $\widehat{\zeta}_n$ and $\widehat{\delta}_n$ since they are consistent estimators.

Let $\widetilde{\Lambda}_n^a(b,\gamma_0)=\{\lambda=(\widetilde{b},\gamma)\in\widehat{\Lambda}_n:(\widetilde{b},\pi)\in\widehat{I}_n^a(b+\tau_{0,b}^{\beta}(\pi_{0,b}^*),\pi_{0,b}^*)\}$ and $\widehat{\Lambda}_n^a=\{\lambda=(b,\gamma)\in\widehat{\Lambda}_n:(b,\pi)\in\widehat{I}_n^a(\widehat{b}_n,\widehat{\pi}_n)\}$. For a size-$\alpha$ test, the construction of the CV proceeds in two steps:

1. Compute the smallest value $\eta=\eta(\widehat{\zeta}_n,\widehat{\delta}_n,\widehat{\Sigma}_n(\cdot))$ such that

$$P\bigg(\widetilde{W}(b,\gamma_0)\geq\sup_{\lambda\in\widetilde{\Lambda}_n^a(b,\gamma_0)\cap\Lambda(v)}c_{1-\alpha}(\lambda)+\eta\bigg)\leq\alpha$$

   for all $(b,\gamma_0)\in\widehat{\Lambda}_n\cap\Lambda(v)$.

2. Construct the quantity $c_{1-\alpha,n}^{AB}=\sup_{\lambda\in\widehat{\Lambda}_n^a\cap\Lambda(v)}c_{1-\alpha}(\lambda)+\eta(\widehat{\zeta}_n,\widehat{\delta}_n,\widehat{\Sigma}_n(\cdot))$. This is the adjusted-Bonferroni CV.

The computations in Step 1 can be achieved by simulating from the joint distribution of $\widetilde{W}(b,\gamma_0)$, $\tau_{0,b}^{\beta}(\pi_{0,b}^*)$ and $\pi_{0,b}^*$ over a grid of (b,γ_0) values in $\widehat{\Lambda}_n\cap\Lambda(v)$. See Algorithm Bonf-Adj in McCloskey (2012) for additional details on the computation of this CV. Under the assumptions of Theorem Wald, Assumption FD and the following assumption, we can establish the correct asymptotic size of tests using the Wald statistic and adjusted-Bonferroni CVs.

Let $\Lambda_{\gamma_0}^a(b,\gamma_0)=\{\lambda=(\widetilde{b},\gamma)\in\Lambda_{\gamma_0}:(\widetilde{b},\pi)\in I_0^a(b+\tau_{0,b}^{\beta}(\pi_{0,b}^*),\pi_{0,b}^*)\}$, where $\Lambda_{\gamma_0}\equiv\{\lambda=(b,\gamma)\in\Lambda:\gamma=(\beta,\zeta_0,\pi,\delta_0,\phi)\}$ and

$$\begin{aligned}&I_0^a(b+\tau_{0,b}^{\beta}(\pi_{0,b}^*),\pi_{0,b}^*)=\{(b,\pi)\in\mathcal{P}(\zeta_0,\delta_0):\\&[(\tau_{0,b}^{\beta}(\pi_{0,b}^*)',(\pi_{0,b}^*-\pi)']\bar{\Sigma}_0^{-1}(b+\tau_{0,b}^{\beta}(\pi_{0,b}^*),\pi_{0,b}^*)[(\tau_{0,b}^{\beta}(\pi_{0,b}^*)',(\pi_{0,b}^*-\pi)']'\leq\chi_{d_\beta+d_\pi}^2(1-a)^{-1}\}\end{aligned}$$

with

$$\bar{\Sigma}_0(b+\tau_{0,b}^{\beta}(\pi_{0,b}^*),\pi_{0,b}^*)\equiv\begin{pmatrix}\Sigma_{\beta\beta,0}(\theta_{0,b}^*) & \|b+\tau_{0,b}^{\beta}(\pi_{0,b}^*)\|^{-1}\Sigma_{\beta\pi,0}(\theta_{0,b}^*)\\ \|b+\tau_{0,b}^{\beta}(\pi_{0,b}^*)\|^{-1}\Sigma_{\beta\pi,0}(\theta_{0,b}^*)' & \|b+\tau_{0,b}^{\beta}(\pi_{0,b}^*)\|^{-2}\Sigma_{\pi\pi,0}(\theta_{0,b}^*)\end{pmatrix}$$

and $\Sigma_{\beta\beta,0}(\theta_{0,b}^*)$ denoting the upper left $d_\beta\times d_\beta$ block of $\Sigma_0(\theta_{0,b}^*)$, $\Sigma_{\beta\pi,0}(\theta_{0,b}^*)$ denoting the upper right $d_\beta\times d_\pi$ block of $\Sigma_0(\theta_{0,b}^*)$, and $\Sigma_{\pi\pi,0}(\theta_{0,b}^*)$ denoting the lower right $d_\pi\times d_\pi$ block of $\Sigma_0(\theta_{0,b}^*)$.

Assumption DF2. *There exists some* $(b^*,\gamma_0^*)\in\Lambda$ *such that*

(i) $P(\widetilde{W}(b^*,\gamma_0^*)\geq\sup_{\lambda\in\Lambda_{\gamma_0^*}^a(b^*,\gamma_0^*)\cap\Lambda(v)}c_{1-\alpha}(\lambda)+\eta(\zeta_0^*,\delta_0^*,\Sigma(\cdot;\gamma_0^*)))=\alpha$,

(ii) $P(\widetilde{W}(b^*,\gamma_0^*)=\sup_{\lambda\in\Lambda_{\gamma_0^*}^a(b^*,\gamma_0^*)\cap\Lambda(v)}c_{1-\alpha}(\lambda)+\eta(\zeta_0^*,\delta_0^*,\Sigma(\cdot;\gamma_0^*)))=0$.

Proposition AB. *Under the assumptions of Theorem Wald and Assumptions FD and DF2,* $\limsup_{n\to\infty}\sup_{\gamma\in\Gamma:\widetilde{r}(\widetilde{\theta})=v}P_\gamma(\widetilde{W}_n(v)>c_{1-\alpha,n}^{AB})=\alpha$.

11 Threshold-Crossing Model Example

To illustrate our approach in this section, we examine a particular version of the threshold-crossing model (Example 2.3) that uses the Ali-Makhail-Haq copula, defined for $\pi \in (-1, 1)$ by

$$C(u_1, u_2; \pi) = \frac{u_1 u_2}{1 - \pi(1 - u_1)(1 - u_2)}.$$

The data is given by the vector $W_i \equiv (Y_i, D_i, Z_i)$ for $i = 1, \ldots, n$. We also suppose the instrument $Z_i \in \{0, 1\}$ is independent of (ε_i, ν_i) with $P_{\gamma_0}(Z_i = z) \equiv \phi_{z,0}$.

The maximum likelihood estimator $\widehat{\widetilde{\theta}}_n$ minimizes the following criterion function in $\widetilde{\theta} \equiv (\beta, \zeta, \pi, \pi_1^1, \pi_2^1)$ over the parameter space $\widetilde{\Theta} \equiv \{\widetilde{\theta} \in [-0.98, 0.98] \times [0.01, 0.99] \times [-0.99, 0.99] \times [0.01, 0.99] \times [0.01, 0.99] : 0.01 \leq \beta + \zeta \leq 0.99\}$:

$$\widetilde{Q}_n(\widetilde{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \log p_{yd,z}(\widetilde{\theta}),$$

where $\mathbf{1}_{ydz}(W_i) \equiv \mathbf{1}\{W_i = (y, d, z)\}$ and

$$\begin{aligned}
p_{11,0}(\widetilde{\theta}) &\equiv C(\pi_2^1, \zeta; \pi), \\
p_{11,1}(\widetilde{\theta}) &\equiv C(\pi_2^1, \zeta + \beta; \pi), \\
p_{10,0}(\widetilde{\theta}) &\equiv \pi_1^1 - C(\pi_1^1, \zeta; \pi), \\
p_{10,1}(\widetilde{\theta}) &\equiv \pi_1^1 - C(\pi_1^1, \zeta + \beta; \pi), \\
p_{01,0}(\widetilde{\theta}) &\equiv \zeta - C(\pi_2^1, \zeta; \pi), \\
p_{01,1}(\widetilde{\theta}) &\equiv \zeta + \beta - C(\pi_2^1, \zeta + \beta; \pi), \\
p_{00,0}(\widetilde{\theta}) &\equiv 1 - p_{11,0}(\widetilde{\theta}) - p_{10,0}(\widetilde{\theta}) - p_{01,0}(\widetilde{\theta}), \\
p_{00,1}(\widetilde{\theta}) &\equiv 1 - p_{11,1}(\widetilde{\theta}) - p_{10,1}(\widetilde{\theta}) - p_{01,1}(\widetilde{\theta}).
\end{aligned} \tag{11.1}$$

After concentrating out $\pi^1 = (\pi_1^1, \pi_2^1)$, the concentrated objective function is

$$Q_n(\theta) = \widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \log p_{yd,z}(\theta, \widehat{\pi}_n^1(\theta)),$$

where $\theta \equiv (\beta, \zeta, \pi)$ and $\widehat{\pi}_n^1(\theta)$ solves

$$\widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)) = \inf_{\pi^1 \in [0.01, 0.99] \times [0.01, 0.99]} \widetilde{Q}_n(\widetilde{\theta}).$$

The population objective function evaluated at γ_0 is

$$\widetilde{Q}_0(\widetilde{\theta}) = -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}\log p_{yd,z}(\widetilde{\theta}).$$

The concentrated population objective function evaluated at γ_0 is

$$Q_0(\theta) = \widetilde{Q}_0(\theta, \pi_0^1(\theta)) = -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}\log p_{yd,z}(\theta, \pi_0^1(\theta)),$$

where $\pi_0^1(\theta)$ solves

$$\widetilde{Q}_0(\theta, \pi_0^1(\theta)) = \inf_{\pi^1 \in [0.01,0.99]\times[0.01,0.99]} \widetilde{Q}_0(\widetilde{\theta}).$$

11.1 Asymptotic Distributional Approximations for the Estimators

In this subsection, we describe the quantities composing the asymptotic distributions of the estimators in the Threshold Crossing example under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ found in Theorems Conc and Est and Remark 8.1. The derivations used to obtain these quantities are given in Appendix C.

The first deterministic function is

$$\widetilde{H}_{0,\pi^1\pi^1}(\theta) \equiv D_{\pi^1\pi^1}\widetilde{Q}_0(\theta, \pi_0^1(\theta)) = -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}D_{\pi^1\pi^1}\log p_{yd,z}(\theta, \pi_0^1(\theta)).$$

We define the Gaussian processes jointly as follows. In what follows, we use the notation $D_x f(x,y,z) = \partial f(x,y,z)/\partial x$ and $D_{xy} f(x,y,z) = \partial f(x,y,z)/\partial x \partial y$ for a generic function $f(x,y,z)$ of (x,y,z). Let $\widetilde{Z}$ be an eight-dimensional random vector with entries indexed by $y,d,z \in \{0,1\}$ such that

$$\widetilde{Z} = \begin{pmatrix} \widetilde{Z}_{000} \\ \vdots \\ \widetilde{Z}_{111} \end{pmatrix} \overset{d}{\sim} \mathcal{N}(0, \widetilde{\mathcal{V}}_0)$$

with $\widetilde{\mathcal{V}}_0$ defined such that

$$\begin{aligned} Var_{\gamma_0}(\widetilde{Z}_{ydz}) &= p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}(1 - p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}) \\ Cov_{\gamma_0}(\widetilde{Z}_{ydz}, \widetilde{Z}_{y'd'z'}) &= -p_{yd,z}(\widetilde{\theta}_0)p_{y'd',z'}(\widetilde{\theta}_0)\phi_{z,0}\phi_{z',0}. \end{aligned}$$

for $(y,d,z) \neq (y',d',z')$. The first Gaussian process is defined as

$$\widetilde{G}_0(\theta) = -\sum_{y,d,z=0,1} \widetilde{Z}_{ydz} D_{\pi^1} \log p_{yd,z}(\theta, \pi_0^1(\theta)).$$

Letting

$$\begin{aligned}\bar{G}_0(\theta) \equiv \bar{G}(\theta;\gamma_0) = &\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0} D_{\psi\pi^1} \log p_{yd,z}(\theta, \pi_0^1(\theta)) \widetilde{H}^{-1}_{0,\pi^1\pi^1}(\theta)\widetilde{G}_0(\theta) \\ &- \sum_{y,d,z=0,1} \widetilde{Z}_{ydz} D_{\psi} \log p_{yd,z}(\theta, \pi_0^1(\theta)),\end{aligned}$$

the second Gaussian process is defined by $G_0(\pi) = \bar{G}(0,\zeta_0,\pi;\gamma_0)$. For the other two deterministic functions, let

$$\begin{aligned}\bar{H}_0(\theta) &\equiv \bar{H}(\theta;\gamma_0) = -\sum_{y,d,z} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0} D_{\psi\psi} \log p_{yd,z}(\theta, \pi_0^1(\theta)) \\ \bar{K}_0(\theta) &\equiv \bar{K}(\theta;\gamma_0) = \frac{\partial D_\psi Q(\theta;\gamma_0)}{\partial \beta_0'} = -\sum_{y,d,z=0,1} D_\psi \log p_{yd,z}(\theta, \pi_0^1(\theta))[D_{\beta_0} p_{yd,z}(\widetilde{\theta}_0)]'\phi_{z,0}.\end{aligned}$$

The functions are then defined as $H_0(\pi) = \bar{H}_0(0,\zeta_0,\pi)$ and $K_0(\pi) = \bar{K}_0(0,\zeta_0,\pi)$.

Finally, noting that $\pi_0^1(\theta)$ is always evaluated at $\theta = (0,\zeta_0,\pi)$ in the construction of the above quantities, we use the result that $\pi_0^1(0,\zeta_0,\pi) = h_0^1(\zeta_0,\pi)$ from Theorem Trans so that we can solve explicitly for $\pi_0^1(0,\zeta_0,\pi)$ in this model using (3.1). This is useful for simulating from the distributions of Corollary Est, Remark 8.1 and Theorem Wald(i). More specifically, upon setting $\beta = 0$, we invert the fitted probabilities in (11.1) to find

$$\pi_0^1(0,\zeta_0,\pi) = \begin{pmatrix} -b_0(\zeta_0,\pi) + \sqrt{b_0(\zeta_0,\pi)^2 - 4a(\zeta_0,\pi)c_0(\zeta_0,\pi)}/2a(\zeta_0,\pi) \\ \delta_{1,0}[1-\pi(1-\zeta_0)]/[\zeta_0 - \delta_{1,0}\pi(1-\zeta_0)] \end{pmatrix},$$

where $a(\zeta_0,\pi) = \pi(1-\zeta_0)$, $b_0(\zeta_0,\pi) = (1-\zeta_0)[\pi(1+\delta_{2,0})-1]$, $c_0(\zeta_0,\pi) = \delta_{2,0}[\pi(1-\zeta_0)-1]$, $\delta_{1,0} = P_{\gamma_0}(Y_i = 1, D_i = 1|Z_i = 0)$ and $\delta_{2,0} = P_{\gamma_0}(Y_i = 1, D_i = 0|Z_i = 0)$.[14]

We conclude this subsection with a brief simulation study illustrating how well the weak identification asymptotic distributions for the parameter estimators approximate their finite sample counterparts. Figures 1–4 provide the simulated finite-sample density functions of the estimators of the threshold-crossing model parameters in red and their approximations that

[14] As may be gleaned from the formula, the expression for $\pi_{1,0}^1(0,\zeta_0,\pi)$ comes from solving a quadratic equation. This equation has two solutions, one of which is always negative and one of which is always positive. Given that π_1^1 must be strictly positive, $\pi_{1,0}^1(0,\zeta_0,\pi)$ is equal to the positive solution.

arise from simulating the distributions in Theorem Conc/Est (to approximate the distributions of $\widehat{\beta}$, $\widehat{\zeta}$ and $\widehat{\pi}$) and Remark 8.1 (to approximate the distributions of $\widehat{\pi}_1^1$ and $\widehat{\pi}_2^1$) in blue. For the finite-sample distributions, we examine the parameter values $\beta \in 0, 0.1, 0.2, 0.4$, $\zeta = 0.2$, $\pi = 0.4$, $\pi_1^1 = 0.6$ and $\pi_2^1 = 0.4$. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ the asymptotic distributional approximations use the corresponding parameter values with $b = \sqrt{n}\beta$, $\zeta_0 = \zeta$, $\pi_0 = \pi$ and $\pi_0^1 = \pi^1$. Figures 1–4 show that (i) the distributions of the parameter estimators can be highly non-Gaussian under weak/non-identification; (ii) as β grows larger, the distributions become approximately Gaussian; and (iii) the new asymptotic distributional approximations perform well overall, especially in contrast with usual Gaussian approximations.

11.2 Asymptotic Distributional Approximations for Wald Statistics

In this subsection, we describe how to approximate the asymptotic distributions of Wald statistics in the threshold-crossing model. For the threshold-crossing model estimated by maximum likelihood, the estimator of the asymptotic covariance matrix of the parameter estimators is equal to an estimator of the inverse information matrix so that for a generic null hypothesis the Wald statistic takes the form

$$\widetilde{W}_n(v) = n(\widetilde{r}(\widehat{\widetilde{\theta}}_n) - v)'(\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)\mathcal{I}^{-1}(\widehat{\widetilde{\theta}}_n, \widehat{\phi}_{1,n})\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)')^{-1}(\widetilde{r}(\widehat{\widetilde{\theta}}_n) - v),$$

where

$$\mathcal{I}(\widehat{\widetilde{\theta}}_n, \widehat{\phi}_{1,n}) = \sum_{y,d,z=0,1} \frac{\widehat{\phi}_{z,n}}{p_{yd,z}(\widehat{\widetilde{\theta}}_n)} D_{\widetilde{\theta}} p_{yd,z}(\widehat{\widetilde{\theta}}_n) D_{\widetilde{\theta}} p_{yd,z}(\widehat{\widetilde{\theta}}_n)'$$

with $\widehat{\phi}_{1,n} = n^{-1}\sum_{i=1}^n Z_i$ and $\widehat{\phi}_{0,n} = 1 - \widehat{\phi}_{1,n}$. In the notation of Section 9, $\widehat{\widetilde{\Sigma}}_n = \widetilde{B}(\widehat{\beta}_n)\mathcal{I}^{-1}(\widehat{\widetilde{\theta}}_n, \widehat{\phi}_{1,n})\widetilde{B}(\widehat{\beta}_n)$. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $|b| < \infty$, using the CMT and Corollary Est, $\mathcal{I}(\widehat{\widetilde{\theta}}_n, \widehat{\phi}_{1,n}) \xrightarrow{d} \mathcal{I}(\widetilde{\theta}^*_{0,b}, \phi_{1,0})$ and $\mathcal{I}(\widetilde{\theta}^*_{0,b}, \phi_{1,0})$ is a singular matrix and is therefore not invertible in the limit. However, we may use an asymptotically equivalent approximation to $\mathcal{I}(\widetilde{\theta}^*_{0,b}, \phi_{1,0})$ when constructing an approximation to $\widetilde{W}(b, \gamma_0)$ in Theorem Wald that is easy to implement. Let $\widetilde{\theta}^a_{0,b} = (\psi^a_{0,b}, \pi^*_{0,b}, \pi^{1a}_{0,b})$, where $\psi^a_{0,b} = \psi_0 + \tau_{0,b}(\pi^*_{0,b})/\sqrt{n}$ and $\pi^{1a}_{0,b}$ is given by the display in Remark 8.1. We use $\widetilde{\theta}^a_{0,b}$ to construct the asymptotically equivalent approximation to $\widetilde{W}(b, \gamma_0)$:

$$\widetilde{W}^a(b, \gamma_0) = n(\widetilde{r}(\widetilde{\theta}^a_{0,b}) - v)'(\widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}^a_{0,b})\mathcal{I}^{-1}(\widetilde{\theta}^a_{0,b}, \phi_{1,0})\widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}^a_{0,b})')^{-1}(\widetilde{r}(\widetilde{\theta}^a_{0,b}) - v).$$

The random vector $\widetilde{\theta}^a_{0,b}$ is asymptotically equivalent to $\widetilde{\theta}^*_{0,b}$ so that $\widetilde{W}^a(b, \gamma_0)$ is asymptotically equivalent to $\widetilde{W}(b, \gamma_0)$. Analogous to the case in Remark 8.1, this $\widetilde{W}^a(b, \gamma_0)$ serves as a better approximation in finite samples than $\widetilde{W}(b, \gamma_0)$. Moreover, since for any finite n, $P(\psi^a_{0,b} = \psi_0) =$

0, $\mathcal{I}(\widetilde{\theta}^a_{0,b}, \phi_{1,0})$ is invertible with probability 1.[15]

Similarly to the previous subsection, we provide a brief simulation study to illustrate how well the random variables $\widetilde{W}^a(b, \gamma_0)$ approximate their finite-sample counterparts. Figures 5–8 provide the simulated finite sample density functions of $\widetilde{W}_n(v)$ for one-dimensional null hypotheses on the separate elements of the parameter vector $\widetilde{\theta}$. This type of null hypothesis is a special case of those satisfying Assumptions Res1–Res2 in Section 9 and we therefore describe the limiting Wald statistic under weak identification in Theorem Wald(i). We emphasize the one-dimensional subvector testing case here, since it is often of primary interest in applied work and, to the best of our knowledge, no other studies in the literature have developed weak identification asymptotic results for test statistics of this form. As in the previous subsection, the finite-sample density functions for the Wald statistics are given in red and the densities of $\widetilde{W}^a(b, \gamma_0)$ are given in blue. In addition, the solid black line graphs the density function of a χ^2_1 distribution for comparison. We look at identical true parameter values as in the previous subsection. Figures 5–8 show similar features to the corresponding figures for the estimators (Figures 1–4): (i) the distributions of the Wald statistics can depart significantly from the usual asymptotic χ^2_1 approximations in the presence of weak/non-identification; (ii) as β grows larger, the distributions become approximately χ^2_1; and (iii) the new asymptotic distributional approximation perform very well, especially compared to the usual χ^2_1 approximation when β is small. One interesting additional feature to note is that, although the distributions of the parameter estimates when $\beta = 0.2$ in Figure 3 appear highly non-Gaussian (especially for π and π^1_2), the corresponding distributions in Figure 7 look well-approximated by the χ^2_1 distribution. This is perhaps due to the self-normalizing nature of Wald statistics.

12 Appendix A: Proofs of Main Results

Proof of Lemma ID1: By Assumption ID1, the following arguments hold for all $\widetilde{\theta} = (0, \zeta, \widetilde{\pi}) \in \widetilde{\Theta}$ and $\gamma_0 \in \Gamma$. Since $rank(\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}) = r$, there are r linearly independent columns in $\partial \widetilde{g}_0(\widetilde{\theta})/\partial \widetilde{\pi}$. These columns form the matrix $\partial \widetilde{g}_0(\widetilde{\theta})/\partial \pi^1$ and the rank of this matrix is r. Thus, $\partial \widetilde{g}_0(\widetilde{\theta})/\partial \pi^1$ has r linearly independent rows. These rows form the matrix $\partial g^1_0(\widetilde{\theta})/\partial \pi^1$ which has rank equal to r. ■

Proof of Lemma ID2: Under the imposed assumptions, by Hadamard's global inverse function theorem, $g^1_0(0, \zeta, \pi, \pi^1)$ as a function of π^1, is a homeomorphism at a given $(0, \zeta, \pi)$.

[15] The invertibility issue for $\mathcal{I}(\widehat{\widetilde{\theta}}_n, \widehat{\phi}_{1,n})$ arises when $\widehat{\beta}_n$ is exactly equal to zero, as is the limiting case under weak and semi-strong identification. However, the probability that $\widehat{\beta}_n = 0$ is equal to zero in finite samples, which is reflected in the refined asymptotic approximation.

Therefore,

$$\pi^1 = \left(g_0^1\right)^{-1}(0,\zeta,\pi,0) \equiv h_0^1(\zeta,\pi) \tag{12.1}$$

$\forall(0,\zeta,\pi) \in \Theta^0$. ■

Proof of Theorem Trans: When $\beta = 0$, (4.3) and Assumption ID2 imply that

$$g_0^0(\widetilde{\theta}) = M(\beta,\zeta;\gamma_0)g_0^1(\widetilde{\theta}) + C(\beta,\zeta;\gamma_0)$$

$\forall\widetilde{\theta} = (0,\zeta,\widetilde{\pi})$, where $C(\beta,\zeta;\gamma_0)$ is $(d_G - r)$-vector, which also does not depend on $\widetilde{\pi}$. When $\widetilde{\theta} = \widetilde{\theta}_0 = (\beta_0,\zeta_0,\widetilde{\pi}_0) = (0,\zeta_0,\widetilde{\pi}_0)$, both $g_0^0(\widetilde{\theta})$ and $g_0^1(\widetilde{\theta})$ are zero vectors by (2.1), and hence $C(0,\zeta_0;\gamma_0) = 0$. Therefore

$$\begin{aligned}
\widetilde{g}_0(0,\zeta_0,\pi,h_0^1(0,\zeta_0,\pi)) &= \begin{bmatrix} g_0^1(0,\zeta_0,\pi,h_0^1(0,\zeta_0,\pi)) \\ g_0^0(0,\zeta_0,\pi,h_0^1(0,\zeta_0,\pi)) \end{bmatrix} \\
&= \begin{bmatrix} I \\ M_0 \end{bmatrix} g_0^1(0,\zeta_0,\pi,h_0^1(0,\zeta_0,\pi)) \\
&= 0
\end{aligned} \tag{12.2}$$

where $M_0 \equiv M(0,\zeta_0;\gamma_0)$ and the last equality holds by Lemma ID2.

Now, by Assumption CF2, $\widetilde{Q}_0(0,\zeta_0,\pi,h_0^1(\zeta_0,\pi)) = \Psi_0(0) = \widetilde{Q}_0(\widetilde{\theta}_0)$. Since $\widetilde{Q}_0(\widetilde{\theta}_0)$ is the smallest value that can be achieved by Assumption CF1, Assumption CF3a implies

$$\pi_0^1(0,\zeta_0,\pi) = h_0^1(\zeta_0,\pi). \tag{12.3}$$

This establishes the theorem's first claim. Combining this result with (12.2) establishes the theorem's second claim. ■

Proof of Theorem FSTrans: Note that

$$\frac{dQ_n(\theta)}{d\pi} = \left(\frac{d\bar{g}(\theta)}{d\pi}\right)' \frac{d\Psi_n(\bar{g}(\theta))}{dG}. \tag{12.4}$$

We want to show that, for all $\theta = (0,\zeta,\pi) \in \Theta$,

$$\frac{d\bar{g}(\theta)}{d\pi} = 0.$$

By Assumptions FSCF1 and FSReg2, for a given $\theta = (0,\zeta,\pi)$, $\widehat{\pi}_n^1(\theta)$ satisfies the first order condition:

$$0 = \frac{\partial \Psi_n\left(\bar{\tilde{g}}(\theta, \widehat{\pi}_n^1(\theta))\right)}{\partial \widetilde{g}} \frac{\partial \bar{\tilde{g}}(\theta, \widehat{\pi}_n^1(\theta))}{\partial \pi^1}$$

$$= \frac{\partial \Psi_n\left(\bar{\tilde{g}}(\theta, \widehat{\pi}_n^1(\theta))\right)}{\partial \widetilde{g}} \begin{bmatrix} I_r \\ M_n \end{bmatrix} \frac{\partial \bar{g}^1(\theta, \widehat{\pi}_n^1(\theta))}{\partial \pi^1}, \tag{12.5}$$

where the second equality follows from Assumption FSID1 so that

$$\frac{\partial \bar{g}^0((\theta, \widehat{\pi}_n^1(\theta))}{\partial \pi^1} = M_n \frac{\partial \bar{g}^1((\theta, \widehat{\pi}_n^1(\theta))}{\partial \pi^1}.$$

Since $\partial \bar{g}^1/\partial \pi^1$ is invertible, post-multiplying by its inverse on both sides of (12.5) yields

$$0 = \frac{\partial \Psi_n\left(\bar{\tilde{g}}(\theta, \widehat{\pi}_n^1(\theta))\right)}{\partial \widetilde{g}} \begin{bmatrix} I_r \\ M_n \end{bmatrix}, \tag{12.6}$$

which, by differentiating with respect to π then yields (suppressing the argument $(\theta, \widehat{\pi}_n^1(\theta))$ in all relevant parts)

$$0 = \left[\frac{\partial \bar{\tilde{g}}}{\partial \pi} + \frac{\partial \bar{\tilde{g}}}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right]' \frac{\partial^2 \Psi_n}{\partial \widetilde{g} \partial \widetilde{g}'} \begin{bmatrix} I_r \\ M_n \end{bmatrix}$$

$$= \left[\frac{\partial \bar{g}^1}{\partial \pi} + \frac{\partial \bar{g}^1}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right]' \begin{bmatrix} I_r \\ M_n \end{bmatrix}' \Psi_{n,\widetilde{g}\widetilde{g}} \begin{bmatrix} I_r \\ M_n \end{bmatrix}$$

$$= \left[\frac{\partial \bar{g}^1}{\partial \pi} + \frac{\partial \bar{g}^1}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right]' \Psi^\dagger_{n,\widetilde{g}\widetilde{g}}, \tag{12.7}$$

where the second equality holds by Assumption FSID1 and $\Psi^\dagger_{n,\widetilde{g}\widetilde{g}}$ is implicitly defined. Since $\widetilde{\Psi}_{n,\widetilde{g}\widetilde{g}}$ is positive definite by Assumption FSCF1, it can be decomposed as $\Psi_{n,\widetilde{g}\widetilde{g}} = R_n' R_n$ where R_n is a matrix of full column rank. Therefore,

$$\Psi^\dagger_{n,\widetilde{g}\widetilde{g}} = \begin{bmatrix} I_r \\ M_n \end{bmatrix}' R_n' R_n \begin{bmatrix} I_r \\ M_n \end{bmatrix}$$

and $R_n \begin{bmatrix} I_r \\ M_n \end{bmatrix}$ has rank of r and hence full column rank, which implies that $\Psi^\dagger_{n,\widetilde{g}\widetilde{g}}$ is positive definite and thus invertible. Then (12.7) gives the first desired result:

$$0 = \left[\frac{\partial \bar{g}^1}{\partial \pi} + \frac{\partial \bar{g}^1}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right] = \frac{d\bar{g}^1(\theta, \widehat{\pi}_n^1(\theta))}{d\pi}.$$

We also have

$$\begin{aligned}\frac{d\bar{g}^0(\theta, \widehat{\pi}_n^1(\theta))}{d\pi} &= \left[\frac{\partial \bar{g}^0}{\partial \pi} + \frac{\partial \bar{g}^0}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right] \\ &= M_n\left[\frac{\partial \bar{g}^1}{\partial \pi} + \frac{\partial \bar{g}^1}{\partial \pi^1}\frac{\partial \widehat{\pi}_n^1(\theta)}{\partial \pi}\right] \\ &= 0. \quad \blacksquare\end{aligned}$$

Lemma Conc1. *Suppose Assumptions CF3, FSCF2, ID4 and Reg4 hold. Then,* $\sup_{\theta\in\Theta}\|\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)\| \xrightarrow{p} 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *for any* $\gamma_0 \in \Gamma$.

Proof: Begin by noting the following:

$$\begin{aligned}0 &\le \inf_{\theta\in\Theta}[\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta))] \\ &\le \sup_{\theta\in\Theta}[\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta))] \\ &\le \sup_{\theta\in\Theta}[\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta))] + \sup_{\theta\in\Theta}[\widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta))] \\ &\le \sup_{\theta\in\Theta}[\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta))] + \sup_{\theta\in\Theta}[\widetilde{Q}_n(\theta, \pi_0^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta))] + o(n^{-1}) \\ &\le 2\sup_{\pi^1\in\Pi^1(\theta),\theta\in\Theta}|\widetilde{Q}_n(\theta, \pi^1) - \widetilde{Q}_0(\theta, \pi^1)| + o(n^{-1}) = o_p(1),\end{aligned}$$

where the first inequality holds by Assumption CF3, the fourth inequality holds by (5.2) and the equality holds by Assumption FSCF2. Now, Assumption ID3 implies that for any given $\theta \in \Theta$ and any neighborhood $\Pi_0^1(\theta)$ of $\pi_0^1(\theta)$, there is some $\varepsilon > 0$ such that $\inf_{\pi^1\in\Pi^1(\theta)\setminus\Pi_0^1(\theta)}\widetilde{Q}_0(\theta, \pi^1) - \widetilde{Q}_0(\theta, \pi_0^1(\theta)) \ge \varepsilon$. Thus,

$$\begin{aligned}&P(\widehat{\pi}_n^1(\theta) \in \Pi^1(\theta) \setminus \Pi_0^1(\theta) \text{ for some } \theta \in \Theta) \\ &\le P(\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta)) \ge \varepsilon \text{ for some } \theta \in \Theta) \to 0\end{aligned}$$

since $\sup_{\theta\in\Theta}[\widetilde{Q}_0(\theta, \widehat{\pi}_n^1(\theta)) - \widetilde{Q}_0(\theta, \pi_0^1(\theta))] \xrightarrow{p} 0$. The statement of the lemma then follows from Assumption Reg4. $\blacksquare$

Lemma Conc2. *Suppose Assumptions CF3, FSCF2, ID3, Reg4 and FSReg3–FSReg5 hold. Then,* $\sqrt{n}(\widehat{\pi}_n^1(\cdot) - \pi_n^1(\cdot)) \Rightarrow -\widetilde{H}_{0,\pi^1\pi^1}^{-1}(\cdot)\widetilde{G}_0(\cdot)$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *for any* $\gamma_0 \in \Gamma$.

Proof: Let $\kappa_n(\theta) = [D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{1/2}\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta))$. Since $\pi_n^1(\theta) \in \Pi^1(\theta)$, we have

$$\begin{aligned}
o_{p\theta}(1) &\geq n(\widetilde{Q}_n(\theta,\widehat{\pi}_n^1(\theta)) - \widetilde{Q}_n(\theta,\pi_n^1(\theta))) \\
&= \sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))'[D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1/2}\kappa_n(\theta) + \frac{1}{2}\|\kappa_n(\theta)\|^2 + nR_n(\theta,\widehat{\pi}_n^1(\theta)) \\
&= O_{p\theta}(\|\kappa_n(\theta)\|) + \frac{1}{2}\|\kappa_n(\theta)\|^2 + (1 + \|[D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1/2}\kappa_n(\theta)\|)^2 o_{p\theta}(1) \\
&= O_{p\theta}(\|\kappa_n(\theta)\|) + \frac{1}{2}\|\kappa_n(\theta)\|^2 + o_{p\theta}(\|\kappa_n(\theta)\|) + o_{p\theta}(\|\kappa_n(\theta)\|^2) + o_{p\theta}(1),
\end{aligned}$$

where the first equality follows from Assumption FSReg3(i), the second equality follows from Assumptions FSReg3(ii), FSReg4 and FSReg5 and Lemma Conc1 and the third equality follows from Assumption FSReg5. Rearranging the previous display provides

$$2\|\kappa_n(\theta)\|O_{p\theta}(1) + \|\kappa_n(\theta)\|^2 + \|\kappa_n(\theta)\|o_{p\theta}(1) + \|\kappa_n(\theta)\|^2 o_{p\theta}(1) + o_{p\theta}(1) \leq o_{p\theta}(1)$$

or

$$2\|\kappa_n(\theta)\|O_{p\theta}(1) + \|\kappa_n(\theta)\|^2 \leq o_{p\theta}(1),$$

which in turn implies

$$\|\kappa_n(\theta)\|^2 \leq 2\|\kappa_n(\theta)\|O_{p\theta}(1) + o_{p\theta}(1).$$

Let $O_{p\theta}(1) = \xi_{n,\theta}$. Then

$$\xi_{n,\theta}^2 + \|\kappa_n(\theta)\|^2 - 2\|\kappa_n(\theta)\|\xi_{n,\theta} \leq \xi_{n,\theta}^2 + o_{p\theta}(1),$$

so that $\|\kappa_n(\theta)\| = O_{p\theta}(1)$ by taking square roots. Then $\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) = O_{p\theta}(1)$ under any $\{\gamma_n\} \in \Gamma(\gamma_0)$ by Assumption FSReg5.

Applying the quadratic approximation of Assumption FSReg3(i) with $\pi^1 = \widehat{\pi}_n^1(\theta)$,

$$\begin{aligned}
& n(\widetilde{Q}_n(\theta,\widehat{\pi}_n^1(\theta)) - \widetilde{Q}_n(\theta,\pi_n^1(\theta))) \\
&= nD_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))'[D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1}D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) \\
&\quad + \frac{1}{2}n(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta))'D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) + o_{p\theta}(1) \\
&= \frac{1}{2}\left(\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) + [D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1}\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))\right)' D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta)) \\
&\quad \times \left(\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) + [D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1}\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))\right) \qquad (12.8)
\end{aligned}$$

$$- \frac{1}{2} n D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))' [D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) + o_{p\theta}(1)$$

where the $o_{p\theta}(1)$ term follows from Assumption FSReg3(ii) and the fact that $\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) = O_{p\theta}(1)$. Similarly applying the quadratic approximation of Assumption FSReg3(i) with $\pi^1 = \pi_n^{1\dagger}(\theta) \equiv \pi_n^1(\theta) - [D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))$,

$$n(\widetilde{Q}_n(\theta, \pi_n^{1\dagger}(\theta)) - \widetilde{Q}_n(\theta, \pi_n^1(\theta))) = -\frac{1}{2} n D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))' [D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) + o_{p\theta}(1), \tag{12.9}$$

where the $o_{p\theta}(1)$ term follows from Assumption FSReg3(ii) and the facts that $\pi_n^{1\dagger}(\theta) - \pi_n^1(\theta) = o_{p\theta}(1)$ and $\pi_n^{1\dagger}(\theta) \in \Pi^1(\theta)$ with probability approaching 1. The first of these facts follows from Assumptions FSReg4 and FSReg5. The second follows from the first and Assumption CF3.

Since $\pi_n^{1\dagger}(\theta) \in \Pi^1(\theta)$ with probability approaching 1, $\widetilde{Q}_n(\theta, \widehat{\pi}_n^1(\theta)) \leq \widetilde{Q}_n(\theta, \pi_n^{1\dagger}(\theta)) + o_{p\theta}(n^{-1})$. Along with (12.8) and (12.9), this implies

$$\begin{aligned} &\frac{1}{2} \left(\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) + [D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} \sqrt{n} D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) \right)' D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) \\ &\quad \times \left(\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) + [D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} \sqrt{n} D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) \right) \leq o_{p\theta}(1). \end{aligned}$$

Finally, Assumption FSReg5 then implies

$$\sqrt{n}(\widehat{\pi}_n^1(\theta) - \pi_n^1(\theta)) = -[D_{\pi^1\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta))]^{-1} \sqrt{n} [D_{\pi^1} \widetilde{Q}_n(\theta, \pi_n^1(\theta)) + o_{p\theta}(1) \Rightarrow -\widetilde{H}_{0,\pi^1\pi^1}^{-1}(\theta) \widetilde{G}_0(\theta),$$

where the weak convergence is a direct result of Assumptions FSReg4 and FSReg5. ■

Proof of Theorem Conc: First note that Theorem FSTrans implies that Assumption A of AC12 holds for the concentrated criterion function $Q_n(\theta)$.

(i) Theorem 3.1(a) of AC12 implies the marginal convergence of $(\sqrt{n}(\widehat{\psi}_n - \psi_n), \widehat{\pi}_n)$. Lemma Conc2 provides the marginal convergence of $\sqrt{n}(\widehat{\pi}_n^1(\cdot) - \pi_n^1(\cdot))$. Hence, it suffices to show that these quantities converge jointly. The proof of Theorem 3.1 of AC12 shows that $\sqrt{n}(\widehat{\psi}_n(\cdot) - \psi_n)$ and $\widehat{\pi}_n$ are continuous functions of $\sqrt{n} D_\psi Q_n(\psi_{0,n}, \cdot) + o_{p\pi}(1)$ and $D_{\psi\psi} Q_n(\psi_{0,n}, \cdot) + o_{p\pi}(1)$, where $\widehat{\psi}_n = \widehat{\psi}_n(\widehat{\pi}_n)$. Similarly, the proof of Lemma Conc2 shows that $\sqrt{n}(\widehat{\pi}_n^1(\cdot) - \pi_n^1(\cdot))$ is a continuous function of $\sqrt{n} D_{\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot)) + o_{p\theta}(1)$ and $D_{\pi^1\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot)) + o_{p\theta}(1)$. Since $D_{\psi\psi} Q_n(\psi_{0,n}, \cdot)$ and $D_{\pi^1\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot))$ converge to nonrandom limits $H_0(\cdot)$ and $\widetilde{H}_{0,\pi^1\pi^1}(\cdot)$ by Assumptions C4(i) of AC12 and FSReg5(i) of this paper, it suffices that $\sqrt{n} D_\psi Q_n(\psi_{0,n}, \cdot)$ and $\sqrt{n} D_{\pi^1} \widetilde{Q}_n(\cdot, \pi_n^1(\cdot))$ converge jointly. This is given by Assumption FSReg4(i).

(ii) Similarly to the proof of part (i), it suffices to show that $\sqrt{n} B(\beta_n)(\widehat{\theta}_n - \theta_n)$ and $\sqrt{n}(\widehat{\pi}_n(\cdot) -$

$\pi_n(\cdot)$) converge jointly since their marginal convergence has been shown in Theorem 3.2(a) of AC12 and Lemma Conc2. The proof of Theorem 3.2 of AC12 shows that $\sqrt{n}B(\beta_n)(\widehat{\theta}_n - \theta_n)$ is a continuous function of $\sqrt{n}B^{-1}(\beta_n)DQ_n(\theta) + o_p(1)$ and $B^{-1}(\beta_n)D^2Q_n(\theta_n)B^{-1}(\beta_n) + o_p(1)$, the latter of which converges to the nonrandom limit $J(\gamma_0)$ by Assumption D2 of AC12. Thus, in analogy with the proof of part (i), it suffices that $\sqrt{n}B^{-1}DQ_n(\theta_n)$ and $\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\cdot, \pi_n^1(\cdot))$ converge jointly. This is given by Assumption FSReg4(ii). ■

Proof of Theorem Est: (i) Begin by decomposing $\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n) = \widehat{\pi}_n^1 - \pi_n^1$ as follows:

$$\begin{aligned}
\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n) &= \{\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)\} + \{\pi_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)\} \\
&= \{\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)\} + \{[\pi_n^1(\widehat{\psi}_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \widehat{\pi}_n)] + [\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)]\} \\
&= \{\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)\} + \{\pi_{n,\psi}(\widehat{\theta}_n)(\widehat{\psi}_n - \psi_n) + [\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)]\} + o_p(n^{-1/2}),
\end{aligned}$$

where the final equality uses a mean value expansion (with respect to ψ) that holds by Theorem Conc(i) and Assumption Reg5(i). Using this decomposition, we have

$$\begin{aligned}
&\begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \\ A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \end{pmatrix} \\
&= \begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)[\pi_{n,\psi}^1(\widehat{\theta}_n)(\widehat{\psi}_n - \psi_n) + (\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n))] \\ A_{2,n}(\widehat{\theta}_n)(\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)) \end{pmatrix} \\
&\quad + \begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)(\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)) \\ A_{2,n}(\widehat{\theta}_n)[\pi_{n,\psi}^1(\widehat{\theta}_n)(\widehat{\psi}_n - \psi_n) + (\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n))] \end{pmatrix} + o_p(1) \\
&= \begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)[\pi_{n,\psi}^1(\widehat{\theta}_n)(\widehat{\psi}_n - \psi_n) + (\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n))] \\ A_{2,n}(\widehat{\theta}_n)(\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)) \end{pmatrix} + o_p(1) \\
&\xrightarrow{d} \begin{pmatrix} A_{1,0}(\psi_0, \pi_{0,b}^*)\{\pi_{0,\psi}^1(\psi_0, \pi_{0,b}^*)\tau_{0,b}(\pi_{0,b}^*) - \widetilde{H}_{0,\pi^1\pi^1}^{-1}(\psi_0, \pi_{0,b}^*)\widetilde{G}_0(\psi_0, \pi_{0,b}^*)\} \\ A_{2,0}(\psi_0, \pi_{0,b}^*)\{\pi_0^1(\psi_0, \pi_{0,b}^*) - \pi_0^1(\psi_0, \pi_0)\} \end{pmatrix}
\end{aligned}$$

under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, where the second equality follows from Assumptions Reg5(i) and Reg6, Theorem Conc(i) and the uniform CMT and the weak convergence follows from Assumption Reg5, Theorem Conc(i) and the uniform CMT. Assumption Reg5(ii) is used to ensure $A_n(\theta) \rightarrow A_0(\theta)$ for all $\theta \in \Theta_\epsilon$ (see e.g., Andrews, 1987). The marginal convergence $(\sqrt{n}(\widehat{\psi}_n - \psi_n), \widehat{\pi}_n) \xrightarrow{d} (\tau_{0,b}(\pi_{0,b}^*), \pi_{0,b}^*)$ immediately follows from Theorem Conc(i) and, noting that $\widehat{\psi}_n$, $\widehat{\pi}_n$, $A_{1,n}(\widehat{\theta}_n)$, $A_{1,n}(\widehat{\theta}_n)$ and $\widehat{\pi}_n^1$ are all continuous functions of $\widehat{\psi}_n$, $\widehat{\pi}_n$ and $\widehat{\pi}_n^1(\cdot)$, the uniform CMT and Theorem Conc(i) yield the joint convergence stated in the theorem.

(ii) For the $\beta_0 = 0$ case, the same decomposition of $\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)$ as that used in the proof of part (i) and similar reasoning imply

$$\begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \\ \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \end{pmatrix} = \begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)[\pi_{n,\psi}^1(\widehat{\theta}_n)(\widehat{\psi}_n - \psi_n) + (\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n))] \\ \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)(\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)) \end{pmatrix} + o_p(1).$$

A mean-value expansion, Assumption Reg5(i) and the consistency of $\widehat{\theta}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ given by Theorem Conc(ii) provide that

$$\begin{aligned} \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)(\pi_n^1(\psi_n, \widehat{\pi}_n) - \pi_n^1(\psi_n, \pi_n)) &= \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)[(\pi_{n,\pi}^1(\psi_n, \widehat{\pi}_n) + o_p(1))(\widehat{\pi}_n - \pi_n)] \\ &= \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)\pi_{n,\pi}^1(\psi_n, \widehat{\pi}_n)(\widehat{\pi}_n - \pi_n) + o_p(1), \end{aligned}$$

where the second equality follows from Assumption Reg5(i) and Theorem Conc(ii). Putting these results together, we have

$$\begin{pmatrix} \sqrt{n}A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \\ \sqrt{n}\iota(\beta_n)A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} A_{1,0}(\theta_0)[\pi_{0,\psi}^1(\theta_0)Z_\psi - \widetilde{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)\widetilde{G}_0(\theta_0)] \\ A_{2,0}(\theta_0)\pi_{0,\pi}^1(\theta_0)Z_\pi \end{pmatrix}$$

by Assumption Reg5, Theorem Conc(ii) and the uniform CMT. Analogous reasoning to that used in the proof of part (i) then yields the stated result for $\beta_0 = 0$. Finally, for the $\beta_0 \neq 0$ case, note that

$$\begin{aligned} \widehat{\pi}_n^1 - \pi_n^1 &= \widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n) \\ &= [\pi_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)] + [\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)] \\ &= \pi_{\theta,n}^1(\widehat{\theta}_n)[\widehat{\theta}_n - \theta_n] + [\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)] + o_p(n^{-1/2}) \end{aligned}$$

where the third equality follows from a mean value expansion which holds by Assumption Reg5(i) and Theorem Conc(ii). By Theorem Conc(ii), Assumption Reg5(i) and the CMT, this implies

$$\sqrt{n}(\widehat{\pi}_n^1 - \pi_n^1) \xrightarrow{d} \pi_{0,\theta}^1(\theta_0)B^{-1}(\beta_0)Z_\theta - \widetilde{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)\widetilde{G}_0(\theta_0)$$

jointly with $\sqrt{n}(\widehat{\theta}_n - \theta_n) \xrightarrow{d} B^{-1}(\beta_0)Z_\theta$. ■

Proof of Corollary Est: (i) By Theorem Est(i),

$$\begin{aligned} (\widehat{\pi}_n^1 - \pi_n^1) &= A_n^{-1}(\widehat{\theta}_n)A_n(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) \\ &= A_n^1(\widehat{\theta}_n)A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) + A_n^2(\widehat{\theta}_n)A_{2,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) \\ &\xrightarrow{d} A_0^2(\psi_0, \pi_{0,b}^*)A_{2,0}(\psi_0, \pi_{0,b}^*)(\pi_0^1(\psi_0, \pi_{0,b}^*) - \pi_0^1(\psi_0, \pi_0)), \end{aligned}$$

since $A_{1,n}(\widehat{\theta}_n)(\widehat{\pi}_n^1 - \pi_n^1) = O_p(n^{-1/2})$ and $A_n^1(\widehat{\theta}_n) = O_p(1)$. The joint convergence follows immediately from Theorem Est(i).

(ii) When $\beta_0 \neq 0$, the stated result follows directly from Theorem Est(ii). For the $\beta_0 = 0$ case, note that

$$\begin{aligned}
\widehat{\pi}_n^1 - \pi_n^1 &= [\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)] + [\pi_n^1(\widehat{\theta}_n) - \pi_n^1(\theta_n)] \\
&= [\widehat{\pi}_n^1(\widehat{\theta}_n) - \pi_n^1(\widehat{\theta}_n)] + \pi_{\theta,n}^1(\widehat{\theta}_n)[\widehat{\theta}_n - \theta_n] + o_p(\iota(\beta_n)^{-1}n^{-1/2}) \\
&= \pi_{\pi,n}^1(\widehat{\theta}_n)[\widehat{\pi}_n - \pi_n] + o_p(\iota(\beta_n)^{-1}n^{-1/2})
\end{aligned}$$

where the second equality follows from a mean value expansion which holds by Assumption Reg5(ii) and Theorem Conc(ii) and the third equality follows from Theorem 1(ii). Again applying Theorem Conc(ii), with Assumption Reg5(ii) and the CMT, this implies

$$n^{1/2}\iota(\beta_n)[\widehat{\pi}_n^1 - \pi_n^1] \xrightarrow{d} \pi_{0,\pi}^1(\theta_0)Z_\pi$$

jointly with $\sqrt{n}B(\beta_n)(\widehat{\theta}_n - \theta_n) \xrightarrow{d} Z_\theta$. To obtain the stated result, then note that when $\beta_0 = 0$, $\bar{H}_{0,\pi^1\pi^1}(\theta_0) = 0$ and

$$\pi_{0,\theta}^1(\theta_0)\bar{B}(\beta_0)Z_\theta - \bar{H}_{0,\pi^1\pi^1}^{-1}(\theta_0)\widetilde{G}_0(\theta_0) = \pi_{0,\theta}^1(\theta_0)\begin{pmatrix} 0_{d_\psi \times d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & I_{d_\pi} \end{pmatrix}\begin{pmatrix} Z_\psi \\ Z_\pi \end{pmatrix} = \pi_{0,\pi}^1(\theta_0)Z_\pi. \blacksquare$$

Proof of Theorem Wald: Under H_0 we can express the Wald statistic as

$$\widetilde{W}_n(v_n) = q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n)'(\widetilde{r}_{\widetilde{\theta},n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n)\widehat{\widetilde{\Sigma}}_n\widetilde{r}_{\widetilde{\theta},n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n)')^{-1}q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n),$$

where

$$q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = n^{1/2}B^*(\widehat{\beta}_n)\widetilde{A}(\widehat{\widetilde{\theta}}_n)(\widetilde{r}(\widehat{\widetilde{\theta}}_n) - \widetilde{r}(\widetilde{\theta}_n)) \text{ and } r_{\widetilde{\theta},n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = B^*(\widehat{\beta}_n)\widetilde{A}(\widehat{\widetilde{\theta}}_n)\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)\widetilde{B}^{-1}(\widehat{\beta}_n)$$

with

$$B^*(\widehat{\beta}_n) = \begin{pmatrix} I_{d_{\widetilde{r}} - d_{\widetilde{\pi}}^*} & 0 \\ 0 & \iota(\widehat{\beta}_n)I_{d_{\widetilde{\pi}}^*} \end{pmatrix}.$$

(i) Note that

$$r_{\widetilde{\theta},n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = B^*(\widehat{\beta}_n)\begin{pmatrix} \widetilde{A}_1(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n) & 0 \\ \widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n) & \widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_{\widetilde{\pi}}(\widehat{\widetilde{\theta}}_n) \end{pmatrix}\widetilde{B}^{-1}(\widehat{\beta}_n)$$

$$= \begin{pmatrix} \widetilde{A}_1(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n) & 0 \\ \iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n) & \widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_{\widetilde{\pi}}(\widehat{\widetilde{\theta}}_n) \end{pmatrix} \xrightarrow{d} \widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\widetilde{\theta}_{0,b}^*), \tag{12.10}$$

where the convergence follows from Corollary Est(i), Assumption Res1(i) and the CMT since $\iota(\widehat{\beta}_n) = o_p(1)$ and $\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n) = O_p(1)$. Turning to the $q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n)$ term, note that

$$\begin{aligned} \widetilde{r}(\widehat{\widetilde{\theta}}_n) - \widetilde{r}(\widetilde{\theta}_n) &= \{\widetilde{r}(\widehat{\psi}_n, \widehat{\widetilde{\pi}}_n) - \widetilde{r}(\psi_n, \widehat{\widetilde{\pi}}_n)\} + \{\widetilde{r}(\psi_n, \widehat{\widetilde{\pi}}_n) - \widetilde{r}(\psi_n, \widetilde{\pi}_n)\} \\ &= \widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n)(\widehat{\psi}_n - \psi_n) + \{\widetilde{r}(\psi_n, \widehat{\widetilde{\pi}}_n) - \widetilde{r}(\psi_n, \widetilde{\pi}_n)\} + o_p(n^{-1/2}), \end{aligned}$$

where the second equality follows from a mean-value expansion, the fact that $\widehat{\psi}_n - \psi_n = O_p(n^{-1/2})$ by Corollary Est(i) and Assumption Res1(i). Hence,

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

where

$$\begin{aligned} q_{1,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) &= \begin{pmatrix} n^{1/2}\widetilde{A}_1(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n)(\widehat{\psi}_n - \psi_n) \\ n^{1/2}\iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)(\widetilde{r}(\psi_n, \widehat{\widetilde{\pi}}_n) - \widetilde{r}(\psi_n, \widetilde{\pi}_n)) \end{pmatrix} \\ q_{2,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) &= \begin{pmatrix} \widetilde{\eta}_n(\widehat{\widetilde{\theta}}_n) \\ n^{1/2}\iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_\psi(\widehat{\widetilde{\theta}}_n)(\widehat{\psi}_n - \psi_n) \end{pmatrix}. \end{aligned}$$

Note that Assumption Res2, the fact that $\widehat{\psi}_n - \psi_n = O_p(n^{-1/2})$ and $\iota(\widehat{\beta}_n) = o_p(1)$ by Corollary Est(i) and Assumption Res1(i) imply that $q_{2,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = o_p(1)$. Hence,

$$q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = q_{1,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) + o_p(1) \xrightarrow{d} q_{0,b}^{\widetilde{A}}(\widetilde{\theta}_{0,b}^*) \tag{12.11}$$

by Corollary Est(i), Assumption Res1(i) and the CMT. Now, for the case of scalar β,

$$\widehat{\widetilde{\Sigma}}_n = \begin{pmatrix} \widehat{J}^{-1}(\widehat{\theta}_n)\widehat{V}(\widehat{\theta}_n)\widehat{J}^{-1}(\widehat{\theta}_n) & \widehat{\Sigma}_n^{12}(\widehat{\widetilde{\theta}}_n) \\ \widehat{\Sigma}_n^{12}(\widehat{\widetilde{\theta}}_n)' & \widehat{\Sigma}_n^{22}(\widehat{\widetilde{\theta}}_n) \end{pmatrix} \xrightarrow{d} \widetilde{\Sigma}_0(\pi_{0,b}^*) \tag{12.12}$$

by Assumption V1 of AC12, Assumption Var1, Corollary Est(i) and the CMT. The analogous argument holds for the vector β case. Finally, the convergence of (12.10) and (12.12) occurs jointly by Corollary Est(i) and the CMT so that

$$\widetilde{r}^{\widetilde{A}}_{\widehat{\widetilde{\theta}},n}(\widehat{\widetilde{\theta}}_n)\widehat{\widetilde{\Sigma}}_n\widetilde{r}^{\widetilde{A}}_{\widehat{\widetilde{\theta}},n}(\widehat{\widetilde{\theta}}_n)' \xrightarrow{d} \widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}^*_{0,b})\widetilde{\Sigma}_0(\pi^*_{0,b})\widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}^*_{0,b})'$$
$$= [\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b}) : \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})]\widetilde{\Sigma}_0(\pi^*_{0,b})[\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b}) : \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})]', \tag{12.13}$$

where

$$\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}) \equiv \begin{pmatrix} \widetilde{A}_1(\widetilde{\theta})\widetilde{r}_{\psi}(\widetilde{\theta}) & 0 \\ 0 & \widetilde{A}_2(\widetilde{\theta})\widetilde{r}_{\pi}(\widetilde{\theta}) \end{pmatrix},$$
$$\widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}) \equiv \begin{pmatrix} 0 \\ \widetilde{A}_2(\widetilde{\theta})\widetilde{r}_{\pi^1}(\widetilde{\theta}) \end{pmatrix}.$$

By Assumption Var1(i), this in turn equals

$$[\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b}) : \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})]\times$$
$$\begin{pmatrix} \Sigma_0(\pi^*_{0,b}) & \Sigma_0(\pi^*_{0,b})\begin{pmatrix} 0_{d_\psi\times d_{\pi^1}} \\ \pi^1_{0,\pi}(\theta^*_{0,b}) \end{pmatrix} \\ (0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})')\Sigma_0(\pi^*_{0,b}) & (0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})')\Sigma_0(\pi^*_{0,b})\begin{pmatrix} 0_{d_\psi\times d_{\pi^1}} \\ \pi^1_{0,\pi}(\theta^*_{0,b}) \end{pmatrix} \end{pmatrix}$$
$$\times\begin{pmatrix} \widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b})' \\ \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})' \end{pmatrix} \tag{12.14}$$
$$= \left(\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b}) + \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})(0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})')\right)\Sigma_0(\pi^*_{0,b})\left(\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b}) + \widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})(0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta)')\right)'.$$

Now note that

$$\widetilde{r}^{\widetilde{A}}_{\theta}(\widetilde{\theta}^*_{0,b})+\widetilde{r}^{\widetilde{A}}_{\pi^1}(\widetilde{\theta}^*_{0,b})(0_{d_{\pi^1}\times d_\psi} : \pi^1_{0,\pi}(\theta^*_{0,b})') = \begin{pmatrix} \widetilde{A}_1(\widetilde{\theta}^*_{0,b})\widetilde{r}_{\psi}(\widetilde{\theta}^*_{0,b}) & 0 \\ 0 & \widetilde{A}_2(\widetilde{\theta}^*_{0,b})[\widetilde{r}_{\pi}(\widetilde{\theta}^*_{0,b}) + \widetilde{r}_{\pi^1}(\widetilde{\theta}^*_{0,b})\pi^1_{0,\pi}(\theta^*_{0,b})'] \end{pmatrix}$$

so that by Assumption Var1(iii), (12.14) has full rank when Assumption Res3 holds. Finally, the convergence of (12.11) and (12.13) is joint by Corollary Est(i) and the CMT, yielding the statement of the theorem.

(ii) In the case that $\beta_0 = 0$, very similar arguments to those providing the convergence in (12.10) but instead using Corollary Est(ii) provide that $r^{\widetilde{A}}_{\widetilde{\theta},n}(\widehat{\widetilde{\theta}}_n) \xrightarrow{p} \widetilde{r}^{\widetilde{A}}_{\widetilde{\theta}}(\widetilde{\theta}_0)$. In addition, very similar arguments to those leading up to (12.11), but instead using Corollary Est(ii), provide that $q_n^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = q_{1,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) + o_p(1)$. In this case,

$$\begin{aligned} n^{1/2}\iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)(\widetilde{r}(\psi_n,\widehat{\widetilde{\pi}}_n) - \widetilde{r}(\psi_n,\widetilde{\pi}_n)) &= n^{1/2}\iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)(\widetilde{r}_{\widetilde{\pi}}(\widehat{\widetilde{\theta}}_n) + o_p(1))(\widehat{\widetilde{\pi}}_n - \widetilde{\pi}_n) \\ &= n^{1/2}\iota(\widehat{\beta}_n)\widetilde{A}_2(\widehat{\widetilde{\theta}}_n)\widetilde{r}_{\widetilde{\pi}}(\widehat{\widetilde{\theta}}_n)(\widehat{\widetilde{\pi}}_n - \widetilde{\pi}_n) + o_p(1) \end{aligned}$$

where the first equality follows from a mean value expansion, Assumption Res1(i) and Corollary Est(ii) and the second equality follows from Corollary Est(ii). Hence,

$$q_{1,n}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n) = \widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\widehat{\widetilde{\theta}}_n)n^{1/2}\widetilde{B}(\beta_n)(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n) \xrightarrow{d} \widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)Z_{\widetilde{\theta}}$$

by Corollary Est(ii) and Assumption Res1(i). By Assumption Var2 and Assumption V2 of AC12, $\widehat{\widetilde{\Sigma}}_n \xrightarrow{p} E_{\gamma_0}[Z_{\widetilde{\theta}}Z_{\widetilde{\theta}}'] \equiv \widetilde{\Sigma}(\gamma_0)$. Very similar arguments to those used in part (i) imply that $\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)\widetilde{\Sigma}(\gamma_0)\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)'$ is invertible. Putting the parts together, we have

$$\widetilde{W}_n(v_n) \xrightarrow{d} Z_{\widetilde{\theta}}'\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)'\left(\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)\widetilde{\Sigma}(\gamma_0)\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)'\right)^{-1}\widetilde{r}_{\widetilde{\theta}}^{\widetilde{A}}(\theta_0)Z_{\widetilde{\theta}} \sim \chi^2_{d_{\widetilde{r}}}$$

by the CMT. Finally, for the $\beta_0 \neq 0$ case, note that

$$\sqrt{n}(\widetilde{r}(\widehat{\widetilde{\theta}}_n) - \widetilde{r}(\widetilde{\theta}_n)) = (\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n) + o_p(1))\sqrt{n}(\widehat{\widetilde{\theta}}_n - \widetilde{\theta}_n) \xrightarrow{d} \widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}_0)\widetilde{B}^{-1}(\beta_0)Z_{\widetilde{\theta}}$$

by a mean-value expansion, Theorem Est(ii) and Assumption Res1(i) and

$$\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)\widetilde{B}^{-1}(\widehat{\beta}_n)\widehat{\widetilde{\Sigma}}_n\widetilde{B}^{-1}(\widehat{\beta}_n)\widetilde{r}_{\widetilde{\theta}}(\widehat{\widetilde{\theta}}_n)' \xrightarrow{p} \widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}_0)\widetilde{B}^{-1}(\beta_0)\widetilde{\Sigma}(\gamma_0)\widetilde{B}^{-1}(\beta_0)\widetilde{r}_{\widetilde{\theta}}(\widetilde{\theta}_0)',$$

by the CMT, Theorem Est(ii), Assumption Res1(i), Assumption Var2 and Assumption V2 of AC12. The stated result then follows. ■

Proof of Proposition ICS: The proof is nearly identical to the proof of Theorem 5.1(b)(iv) of AC12, using Theorem Wald in the place of Theorems 4.2 and 4.3 of AC12. ■

Proof of Proposition AB: The proof of this proposition verifies that the assumptions of Theorem Bonf-Adj of McCloskey (2012) hold, with some modifications. First, Assumption PS of McCloskey (2012) holds with $\gamma_1 = (\beta, \pi)$, $\gamma_2 = (\zeta, \delta)$, no γ_3 and $\gamma_4 = \phi$. For the definition of $\{\gamma_{n,h}\}$, $\gamma_{n,h,1} = (\beta_{n,h}, n^{-1/2}\pi_{n,h})$ and $\gamma_{n,h,2} = (\zeta_{n,h}, \delta_{n,h})$. Note that $h_{1,1} = b$, where $h_{1,1}$ denotes the first d_β elements of h_1. In the notation of McCloskey (2012), sequences $\{\gamma_{n,h}\}$ with $\|h_{1,1}\| < \infty$ ($\|h_{1,1}\| = \infty$) correspond to weak (semi-strong or strong) identification sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ ($\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$) in the notation of this paper.

Second, for Assumption DS of McCloskey (2012), $T_n(\theta_n) = \widetilde{W}_n(v_n)$ $\widehat{h}_{n,1} = (\widehat{b}_n, \widehat{\pi}_n)$ and

$\widehat{h}_{n,2} = (\widehat{\zeta}_n, \widehat{\delta}_n)$. Theorem Wald provides the marginal weak convergence of $T_{\omega_n}(\theta_{\omega_n})$ for all sequences $\{\gamma_{\omega_n,h}\}$, where in the notation of McCloskey (2012), $W_h = \widetilde{W}(b, \gamma_0)$ when $\|h_{1,1}\| < \infty$ and W_h is distributed $\chi^2_{d_{\widetilde{r}}}$ when $\|h_{1,1}\| = \infty$. Corollary Est and Assumption FD provide the marginal weak convergence of $\widehat{h}_{\omega_n} = (\widehat{h}_{\omega_n,1}, \widehat{h}_{\omega_n,2})$ for all sequences $\{\gamma_{\omega_n,h}\}$, where in the notation of McCloskey (2012), $\widetilde{h}_1 = (b + \tau^{\beta}_{0,b}(\pi^*_{0,b}), \pi^*_{0,b})$ when $\|h_{1,1}\| < \infty$, $\widetilde{h}_1 = (b + Z_\beta, \pi_0)$ when $\|h_{1,1}\| = \infty$ and $h_2 = (\zeta_0, \delta_0)$. Joint convergence of $(T_{\omega_n}(\theta_{\omega_n}), \widehat{h}_{\omega_n})$ follows from Corollary Est similarly to the joint convergence statements made in Theorem Wald.

Third, for Assumption MLLD of McCloskey (2012), we are in what McCloskey (2012) refers to as "the usual case" for which $u = 1$, $\widetilde{W}^{(1)}_h = \widetilde{W}(b, \gamma_0)$ and $\bar{H}^{(1),c} = \emptyset$ since $P(|\widetilde{W}(b, \gamma_0)| < \infty) = 1$ under the assumptions of Theorem Wald. Since we are in the usual case, there is no need to define the auxiliary sequence of parameters $\{\zeta_n\}$ (it can be any arbitrary sequence in $\mathbb{R}^r$ for arbitrary $r > 0$) and $P = \mathbb{R}^r_\infty$ for any $r > 0$. Since $W_h = \widetilde{W}(b, \gamma_0) = \widetilde{W}^{(1)}_h$ when $\|h_{1,1}\| < \infty$ and $W_h = \widetilde{W}^{(1)}_h$ is distributed $\chi^2_{d_{\widetilde{r}}}$ when $\|h_{1,1}\| = \infty$, the only item left to verify is that $\widetilde{W}(b, \gamma_0)$ is completely characterized by $h^{(1)} = h = (b, \pi_0, \zeta_0, \delta_0)$. This holds by Assumption FD.

Fourth, for Assumption Cont-Adj of McCloskey (2012), $\bar{H}^{(1)} = H$. This assumption holds for any $\underline{\delta}^{(1)} > 0$ and $\bar{\delta}^{(1)} \leq \alpha$ since $\widetilde{W}(b, \gamma_0)$ is an absolutely continuous random variable with quantiles that are continuous in b and $\widetilde{W}(b, \gamma_0) \overset{d}{\sim} \chi^2_{d_{\widetilde{r}}}$ for any b such that $\|b\| = \infty$. Fifth, Assumption Sel holds trivially since we are in the "usual case".

Sixth, Assumption CS of McCloskey (2012) can be modified and applied to $\widehat{I}^a_n(\cdot)$ and its limit counterpart $I^a_0(\cdot)$ so that: (i)

$$\sup_{(b,\pi_0)\in\{(\widetilde{b},\widetilde{\pi})\in\mathbb{R}^{d_\beta+d_\pi}_\infty:(\widetilde{b},\widetilde{\gamma})\in\Lambda\}} d_H(\widehat{I}^a_n(b, \pi_0), I^a_0(b, \pi_0)) \xrightarrow{p} 0$$

under any $\{\gamma_n\} \in \Gamma(\gamma_0)$, where $d_H(A, B)$ denotes the Hausdorff distance between the two sets A and B; (ii) $I^a_0(\cdot)$ is a continuous and compact-valued correspondence; (iii) $P_{\gamma_n}(\widehat{I}^a_n(\widehat{b}_n, \widehat{\pi}_n) \subset \bar{H}^{(1)}_1(\widehat{h}_{n,2})) = 1$ for all $n \geq 1$ and $\{\gamma_n\} \in \Gamma(\gamma_0)$ and $P(I^a_0(b + \tau^{\beta}_{0,b}(\pi^*_{0,b}), \pi^*_{0,b}) \subset \bar{H}^{(1)}_1(h_2)) = 1$; and (iv) $I^a_0(b + \tau^{\beta}_{0,b}(\pi^*_{0,b}), \pi^*_{0,b})$ need not satisfy a coverage requirement (i.e., $P(h_1 \in I^a_0(b + \tau^{\beta}_{0,b}(\pi^*_{0,b}), \pi^*_{0,b}) \geq 1 - a)$. The proof of Theorem Bonf-Adj in McCloskey (2012) still goes through with this modification of Assumption CS. Condition (i) is satisfied by the consistency of $(\widehat{\zeta}_n, \widehat{\delta}_n)$ and the uniform consistency of $\widehat{\Sigma}_n(\cdot)$ under any $\{\gamma_n\} \in \Gamma(\gamma_0)$. The former holds by Corollary Est and Assumption FD while the latter holds by Assumptions V1 and V2 of AC12. For condition (ii), $I^a_0(\cdot)$ is clearly continuous and compact-valued. Note that $\mathcal{P}(\widehat{\zeta}_n, \widehat{\delta}_n)$ and $\mathcal{P}(\zeta_0, \delta_0)$ are equal to $\bar{H}^{(1)}(\widehat{h}_{n,2})$ and $\bar{H}^{(1)}(h_2)$ in the notation of McCloskey (2012) so that condition (iii) holds by construction.

Seventh, note that rather than using a quantile adjustment function ($a^{(j)}(\cdot)$ in the notation of McCloskey, 2012), we are fixing the quantile at level $1-\alpha$ and adding a size-correction function $\eta(\cdot)$ to it. The proof of Theorem Bonf-Adj of McCloskey (2012) can be easily adjusted to this modification. Rather than requiring the quantile adjustment function to be continuous, the proof requires $\eta(\cdot)$ to be continuous. That is, Assumption a(i) of McCloskey (2012) may be replaced by the analogous assumption: $\eta(\cdot)$ is continuous. In practice, $\eta(\cdot)$ is only evaluated at the point $(\widehat{\zeta}_n, \widehat{\delta}_n, \widehat{\Sigma}_n(\cdot))$, which is consistent with this assumption. Due to the replacement of quantile adjustment by additive size-correction, Assumption a(ii) of McCloskey (2012) should also be replaced by the analogous assumption: $P(\widetilde{W}(b,\gamma_0) \geq \sup_{\lambda \in \Lambda^{\alpha}_{\gamma_0}(b,\gamma_0)\cap\Lambda(v)} c_{1-\alpha}(\lambda) + \eta(\zeta_0, \delta_0, \Sigma_0(\cdot))) \leq \alpha$ for all $(b,\gamma_0) \in \Lambda_0 \cap \Lambda(v)$. This assumption holds by the construction of $\eta(\widehat{\zeta}_n, \widehat{\delta}_n, \widehat{\Sigma}_n(\cdot))$ and the (uniform) consistency of $(\widehat{\zeta}_n, \widehat{\delta}_n, \widehat{\Sigma}_n(\cdot))$.

Finally, Assumption Inf-Adj of McCloskey (2012) holds vacuously since $\bar{H}^{(1),c} = \emptyset$ and Assumption LB-Adj of that paper is imposed by Assumption DF2. ■

13 Appendix B: Assumptions and Notation for Vector β Case

For the vector β case, reparameterize β as $(\|\beta\|, \omega)$, where $\omega = \beta/\|\beta\|$ if $\beta \neq 0$ and define $\omega = 1_{d_\beta}/\|1_{d_\beta}\|$ if $\beta = 0$. Correspondingly, reparameterize $\widetilde{\theta}$ as $\widetilde{\theta}^+ = (\|\beta\|, \omega, \zeta, \widetilde{\pi})$. Let $\widehat{\widetilde{\theta}}{}^+_n$ and $\widetilde{\theta}^+_0$ be the correspondingly reparameterized versions of $\widehat{\widetilde{\theta}}_n$ and $\widetilde{\theta}_0$. Let $\widetilde{\Sigma}_0(\theta^+) \equiv \widetilde{\Sigma}(\theta^+;\gamma_0)$ for $\theta^+ \in \Theta^+ \equiv \{\theta^+ : \theta \in \Theta\}$ be a nonstochastic $d_\theta \times d_\theta$ matrix-valued function. Let

$$\widetilde{\Sigma}_0(\theta^+) = \begin{pmatrix} \Sigma_0(\theta^+) & \Sigma_0(\theta^+)\begin{pmatrix} 0_{d_\psi \times d_{\pi^1}} \\ \partial \pi_0^{1+}(\theta^+)'/\partial \pi \end{pmatrix} \\ (0_{d_{\pi^1} \times d_\psi} : \partial \pi_0^{1+}(\theta^+)/\partial \pi')\Sigma_0(\theta^+) & (0_{d_{\pi^1} \times d_\psi} : \partial \pi_0^{1+}(\theta^+)/\partial \pi')\Sigma_0(\theta^+)\begin{pmatrix} 0_{d_\psi \times d_{\pi^1}} \\ \partial \pi_0^{1+}(\theta^+)'/\partial \pi \end{pmatrix} \end{pmatrix}$$

$$\equiv \begin{pmatrix} \Sigma_0(\theta^+) & \Sigma_0^{12}(\theta^+) \\ \Sigma_0^{12}(\theta^+)' & \Sigma_0^{22}(\theta^+), \end{pmatrix}$$

$$\widetilde{\Sigma}_0(\widetilde{\pi}, \omega) \equiv \widetilde{\Sigma}(\widetilde{\pi}, \omega; \gamma_0) = \widetilde{\Sigma}(\|\beta_0\|, \omega, \zeta_0, \pi; \gamma_0),$$

where $\pi_0^{1+}(\theta^+) = \pi_0^1(\theta)$ and $\Sigma_0(\theta^+) \equiv \Sigma(\theta^+;\gamma_0)$ is defined in (8.1) of Andrews and Cheng (2012b).

Assumption Var1. *(Vector β) (i) $\widehat{\Sigma}_n^{12} = \widehat{\Sigma}_n^{12}(\widehat{\theta}_n^+)$ and $\widehat{\Sigma}_n^{22} = \widehat{\Sigma}_n^{22}(\widehat{\theta}_n^+)$ for some (stochastic) $d_\theta \times d_{\pi^1}$ and $d_{\pi^1} \times d_{\pi^1}$ matrix-valued functions $\widehat{\Sigma}_n^{12}(\cdot)$ and $\widehat{\Sigma}_n^{22}(\cdot)$ on Θ^+ that satisfy $\sup_{\theta^+ \in \Theta^+} \|\widehat{\Sigma}_n^{12}(\theta^+) - \Sigma_0^{12}(\theta^+)\| \xrightarrow{p} 0$ and $\sup_{\theta^+ \in \Theta^+} \|\widehat{\Sigma}_n^{22}(\theta^+) - \Sigma_0^{22}(\theta^+)\| \xrightarrow{p} 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$.*

(ii) $\Sigma_0^{12}(\theta^+)$ and $\Sigma_0^{22}(\theta^+)$ are continuous on $\theta^+ \in \Theta^+$ for all $\gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(iii) $\lambda_{\min}(\Sigma_0(\widetilde{\pi},\omega)) > 0$ *and* $\lambda_{\max}(\widetilde{\Sigma}_0(\widetilde{\pi},\omega)) < \infty$ *for all* $\widetilde{\pi}$ *such that there is some* $\widetilde{\theta} = (\psi,\widetilde{\pi}) \in \widetilde{\Theta}$, $\omega \in \mathbb{R}^{d_\beta}$ *with* $\|\omega\| = 1$ *and* $\gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

(iv) $P(\tau^{\beta}_{0,b}(\pi^*(\gamma_0, b)) = 0) = 0$ *for all* $\gamma_0 \in \Gamma$ *with* $\beta_0 = 0$ *and* b *with* $\|b\| < \infty$.

14 Appendix C: Derivations for Threshold-Crossing Example

Beginning with the Gaussian processes, note that

$$\begin{aligned} D_\psi Q_n(\theta) &= -\sum_{y,d,z=0,1} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{ydz}(W_i) D_\psi \log p_{yd,z}(\theta, \widehat{\pi}^1_n(\theta)) \\ D_\psi Q(\theta;\gamma_n) &= -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n} D_\psi \log p_{yd,z}(\theta, \pi^1_n(\theta)) \end{aligned}$$

so that

$$\begin{aligned} &\sqrt{n}(D_\psi Q_n(\theta) - D_\psi Q(\theta;\gamma_n)) \\ &= -\sum_{y,d,z=0,1}\left[n^{-1/2}\sum_{i=1}^n (\mathbf{1}_{ydz}(W_i) - p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n})\right] D_\psi \log p_{yd,z}(\theta,\widehat{\pi}^1_n(\theta)) \qquad (14.1) \\ &\quad + \sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n}\sqrt{n}\left[D_\psi \log p_{yd,z}(\theta,\pi^1_n(\theta)) - D_\psi \log p_{yd,z}(\theta,\widehat{\pi}^1_n(\theta))\right]. \qquad (14.2) \end{aligned}$$

Note that a CLT for triangular arrays provides that under $\{\gamma_n\} \in \Gamma(\gamma_0)$,

$$n^{-1/2}\sum_{i=1}^n \begin{pmatrix} \mathbf{1}_{000}(W_i) - p_{00,0}(\widetilde{\theta}_n)\phi_{0,n} \\ \vdots \\ \mathbf{1}_{111}(W_i) - p_{11,1}(\widetilde{\theta}_n)\phi_{1,n} \end{pmatrix} \xrightarrow{d} \widetilde{Z} = \begin{pmatrix} \widetilde{Z}_{000} \\ \vdots \\ \widetilde{Z}_{111} \end{pmatrix} \overset{d}{\sim} \mathcal{N}(0,\widetilde{\mathcal{V}}_0) \qquad (14.3)$$

with $\widetilde{\mathcal{V}}_0$ defined such that

$$\begin{aligned} Var(\widetilde{Z}_{ydz}) &= p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}(1 - p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0}) \\ Cov(\widetilde{Z}_{ydz}, \widetilde{Z}_{y'd'z'}) &= -p_{yd,z}(\widetilde{\theta}_0)p_{y'd',z'}(\widetilde{\theta}_0)\phi_{z,0}\phi_{z',0}. \end{aligned}$$

On the other hand, a (uniform in θ) mean value expansion, along with the results of Lemma Conc1 (see Appendix A) provides that

$$D_\psi \log p_{yd,z}(\theta,\widehat{\pi}^1_n(\theta)) = D_\psi \log p_{yd,z}(\theta,\pi^1_n(\theta)) + D_{\psi\pi^1}\log p_{yd,z}(\theta,\pi^{1\dagger}_n(\theta))(\widehat{\pi}^1_n(\theta) - \pi^1_n(\theta)),$$

where $\sup_{\theta\in\Theta}\|\pi_n^{1\dagger}(\theta)-\pi_n^1(\theta)\| \xrightarrow{p} 0$. Hence, (14.2) is equal to

$$
\begin{aligned}
&-\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n}D_{\psi\pi^1}\log p_{yd,z}(\theta,\pi_n^{1\dagger}(\theta))\sqrt{n}(\widehat{\pi}_n^1(\theta)-\pi_n^1(\theta))\\
&=\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n}D_{\psi\pi^1}\log p_{yd,z}(\theta,\pi_n^{1\dagger}(\theta))[D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1}\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))+o_{p\theta}(1)
\end{aligned}
\tag{14.4}
$$

since

$$
\sqrt{n}(\widehat{\pi}_n^1(\theta)-\pi_n^1(\theta)) = -[D_{\pi^1\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))]^{-1}\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))+o_{p\theta}(1)
$$

by taking the first order condition with respect to π^1 in the quadratic expansion in Assumption FSReg3. Now, note that

$$
\begin{aligned}
\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta)) &= -\sum_{y,d,z=0,1} n^{-1/2}\sum_{i=1}^n \mathbf{1}_{ydz}(W_i)D_{\pi^1}\log p_{yd,z}(\theta,\pi_n^1(\theta))\\
&= -\sum_{y,d,z=0,1} n^{-1/2}\sum_{i=1}^n \left[\mathbf{1}_{ydz}(W_i)-p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n}\right]D_{\pi^1}\log p_{yd,z}(\theta,\pi_n^1(\theta)),
\end{aligned}
$$

where the second equality follows from the first order condition with respect to π^1 of $\widetilde{Q}(\widetilde{\theta};\gamma_n)$:

$$
D_{\pi^1}\widetilde{Q}(\widetilde{\theta};\gamma_n) = -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n}D_{\pi^1}\log p_{yd,z}(\theta,\pi_n^1(\theta)) = 0.
$$

Invoking (14.3), the results of Lemma Conc1 the and the uniform CLT, we have the following joint weak convergence result for (14.1) and $\sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))$:

$$
\begin{aligned}
&\begin{pmatrix} -\sum_{y,d,z=0,1}\left[n^{-1/2}\sum_{i=1}^n(\mathbf{1}_{ydz}(W_i)-p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n})\right]D_\psi\log p_{yd,z}(\theta,\widehat{\pi}_n^1(\theta))\\ \sqrt{n}D_{\pi^1}\widetilde{Q}_n(\theta,\pi_n^1(\theta))\end{pmatrix}\\
&\Rightarrow \begin{pmatrix} -\sum_{y,d,z=0,1}\widetilde{Z}_{ydz}D_\psi\log p_{yd,z}(\theta,\pi_0^1(\theta))\\ -\sum_{y,d,z=0,1}\widetilde{Z}_{ydz}D_{\pi^1}\log p_{yd,z}(\theta,\pi_0^1(\theta))\end{pmatrix}.
\end{aligned}
$$

By (14.1), (14.2), (14.4), the results of Lemma Conc1 and the uniform CLT, this result implies

$$
\sqrt{n}\begin{pmatrix} D_{\pi^1}\widetilde{Q}_n(\cdot,\pi_n^1(\cdot))\\ D_\psi Q_n(\cdot)-D_\psi Q(\cdot;\gamma_n)\end{pmatrix} \Rightarrow \begin{pmatrix}\widetilde{G}_0(\cdot)\\ \bar{G}_0(\cdot)\end{pmatrix}, \tag{14.5}
$$

where

$$\widetilde{G}_0(\theta) = -\sum_{y,d,z=0,1} \widetilde{Z}_{ydz} D_{\pi^1} \log p_{yd,z}(\theta, \pi_0^1(\theta))$$

$$\begin{aligned}\bar{G}_0(\theta) = &\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0} D_{\psi\pi^1} \log p_{yd,z}(\theta, \pi_0^1(\theta))[\widetilde{H}_{0,\pi^1\pi^1}(\theta)]^{-1}\widetilde{G}_0(\theta) \\ &- \sum_{y,d,z=0,1} \widetilde{Z}_{ydz} D_{\psi} \log p_{yd,z}(\theta, \pi_0^1(\theta))\end{aligned}$$

with

$$\widetilde{H}_{0,\pi^1\pi^1}(\theta) = D_{\pi^1\pi^1}\widetilde{Q}_0(\theta, \pi_0^1(\theta)) = -\sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0} D_{\pi^1\pi^1} \log p_{yd,z}(\theta, \pi_0^1(\theta)).$$

Finally, by (14.5), $G_0(\pi) = \bar{G}_0(0, \zeta_0, \pi)$.

For the other two deterministic functions of interest, note that by Lemma Conc1 and the uniform CLT, $\sup_{\theta\in\Theta} |D_{\psi\psi}Q_n(\theta) - \bar{H}_0(\theta)| \xrightarrow{p} 0$ for

$$\bar{H}_0(\theta) = -\sum_{y,d,z} p_{yd,z}(\widetilde{\theta}_0)\phi_{z,0} D_{\psi\psi} \log p_{yd,z}(\theta, \pi_0^1(\theta)).$$

Hence, $H_0(\pi) = \bar{H}_0(0, \zeta_0, \pi)$. Finally, let $K_n(\theta; \gamma^*) = \partial E_{\gamma^*} D_\psi Q_n(\theta)/\partial\beta^*$ and note that for non stochastic sequences $\{\bar{\psi}_n\}$ and $\{\gamma_n\}$ such that $\gamma_n \in \Gamma$, $\gamma_n \to \gamma_0 = (0, \zeta_0, \pi_0, \pi_0^1, \widetilde{\phi}_0)$ for some $\gamma_0 \in \Gamma$, $(\bar{\psi}_n, \pi) \in \Theta$ and $\bar{\psi}_n \to \psi_0 = (0, \zeta_0)$,

$$\begin{aligned}K_n(\bar{\psi}_n, \pi; \gamma_n) &= \frac{\partial}{\partial\widetilde{\beta}_n} E_{\gamma_n} \frac{1}{n}\sum_{i=1}^n \{-\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) D_\psi \log p_{yd,z}(\bar{\psi}_n, \pi, \widehat{\pi}_n^1(\bar{\psi}_n, \pi))\} \\ &= \frac{\partial}{\partial\beta_n} E_{\gamma_n} \frac{1}{n}\sum_{i=1}^n \{-\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) D_\psi \log p_{yd,z}(\bar{\psi}_n, \pi, \pi_n^1(\bar{\psi}_n, \pi))\} \qquad (14.6) \\ &\quad + \frac{\partial}{\partial\beta_n} E_{\gamma_n} \frac{1}{n}\sum_{i=1}^n \{-\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{\partial \log p_{yd,z}(\bar{\psi}_n, \pi, \pi_n^{1*}(\bar{\psi}_n, \pi))}{\partial\pi^1} [\widehat{\pi}_n^1(\bar{\psi}_n, \pi) - \pi_n^1(\bar{\psi}_n, \pi)]\}, \\ &\qquad (14.7)\end{aligned}$$

by the mean value theorem, where each element of $\pi_n^{1*}(\bar{\psi}_n, \pi)$ lies in between the corresponding elements of $\widehat{\pi}_n^1(\bar{\psi}_n, \pi)$ and $\pi_n^1(\bar{\psi}_n, \pi)$. The quantity (14.6) is equal to

$$-\frac{\partial}{\partial\beta_n} \sum_{y,d,z=0,1} p_{yd,z}(\widetilde{\theta}_n)\phi_{z,n} D_\psi \log p_{yd,z}(\bar{\psi}_n, \pi, \pi_n^1(\bar{\psi}_n, \pi))\}$$

$$\longrightarrow -\sum_{y,d,z=0,1} D_\psi \log p_{yd,z}(\psi_0,\pi,\pi_0^1(\psi_0,\pi))[D_{\beta_0} p_{yd,z}(\widetilde{\theta}_0)]'\phi_{z,0} = K_0(\psi_0,\pi).$$

For the term (14.7), note that for all $n \geq 1$,

$$\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}\{-\sum_{y,d,z=0,1}\mathbf{1}_{ydz}(W_i)\frac{\partial \log p_{yd,z}(\bar{\psi}_n,\pi,\pi_n^{1*}(\bar{\psi}_n,\pi))}{\partial \pi^1}[\widehat{\pi}_n^1(\bar{\psi}_n,\pi)-\pi_n^1(\bar{\psi}_n,\pi)]\}\right| \tag{14.8}$$
$$\leq 2\sum_{y,d,z=0,1}\sup_{\widetilde{\theta}\in\widetilde{\Theta}}\left|\frac{\partial \log p_{yd,z}(\widetilde{\theta})}{\partial \pi^1}\right|\sup_{\theta\in\Theta}\sup_{\pi^1\in\Pi^1(\theta)}\pi^1 < \infty$$

almost surely under P_{γ_n} due to the continuity of $\partial \log p_{yd,z}(\widetilde{\theta})/\partial\pi^1$ in $\widetilde{\theta}$ and the compactness of the parameter spaces $\widetilde{\Theta}$, Θ and $\Pi^1(\theta)$ so that (14.8) is uniformly integrable under P_{γ_n} for all $\pi \in \Pi$. In conjunction with the fact that

$$\frac{1}{n}\sum_{i=1}^{n}\{-\sum_{y,d,z=0,1}\mathbf{1}_{ydz}(W_i)\frac{\partial \log p_{yd,z}(\bar{\psi}_n,\pi,\pi_n^{1*}(\bar{\psi}_n,\pi))}{\partial \pi^1}[\widehat{\pi}_n^1(\bar{\psi}_n,\pi)-\pi_n^1(\bar{\psi}_n,\pi)]\}$$
$$= -[\widehat{\pi}_n^1(\bar{\psi}_n,\pi)-\pi_n^1(\bar{\psi}_n,\pi)]\sum_{y,d,z=0,1}\frac{\partial \log p_{yd,z}(\bar{\psi}_n,\pi,\pi_n^{1*}(\bar{\psi}_n,\pi))}{\partial \pi^1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{ydz}(W_i)\xrightarrow{p} 0$$

uniformly over $\pi \in \Pi$ by Theorem Conc, this implies that $K_n(\bar{\psi}_n,\pi;\gamma_n) \to K_0(\psi_0,\pi) = K_0(\pi)$ uniformly over $\pi \in \Pi$.

References

Andrews, D. W. K., 1987. Asymptotic results for generalized Wald tests. Econometric Theory 3, 348–358.

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.

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

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

Andrews, D. W. K., Cheng, X., 2014. GMM estimation and uniform subvector inference with possible identification failure. Econometric Theory 30, 287–333.

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

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

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

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

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

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

Dovonon, P., Renault, E., 2013. Testing for common conditionally heteroskedastic factors. Econometrica 81, 2561–2586.

Escanciano, J. C., Zhu, L., 2013. Set inferences and sensitivity analysis in semiparametric conditionally identified models, Unpublished Manuscript, Indiana University, Department of Economics.

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.

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

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

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

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

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

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

Phillips, P. C. B., 2015. Inference in near singular regression, Cowles Foundation Discussion Paper No. 2009.

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.

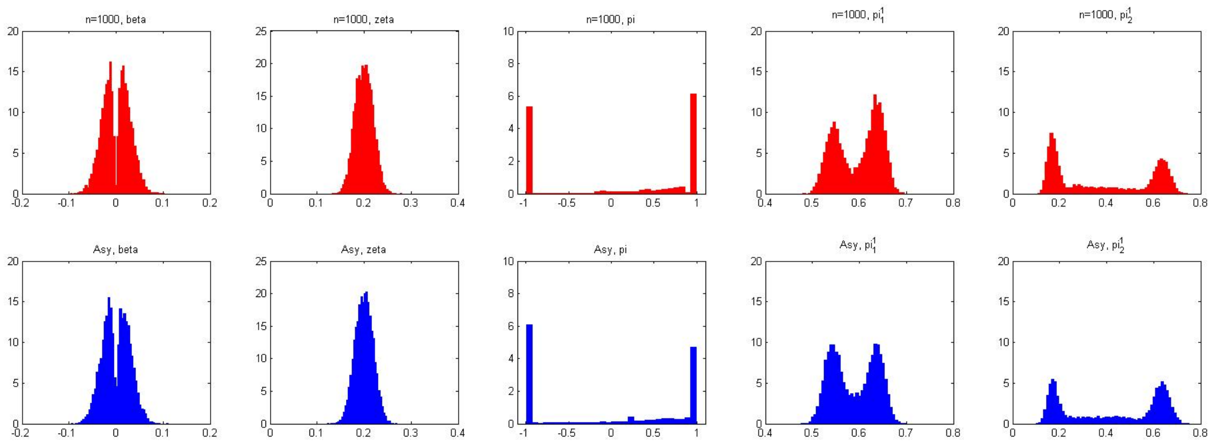
Rothenberg, T. J., 1971. Identification in parametric models. Econometrica 39, 577–591.

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

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

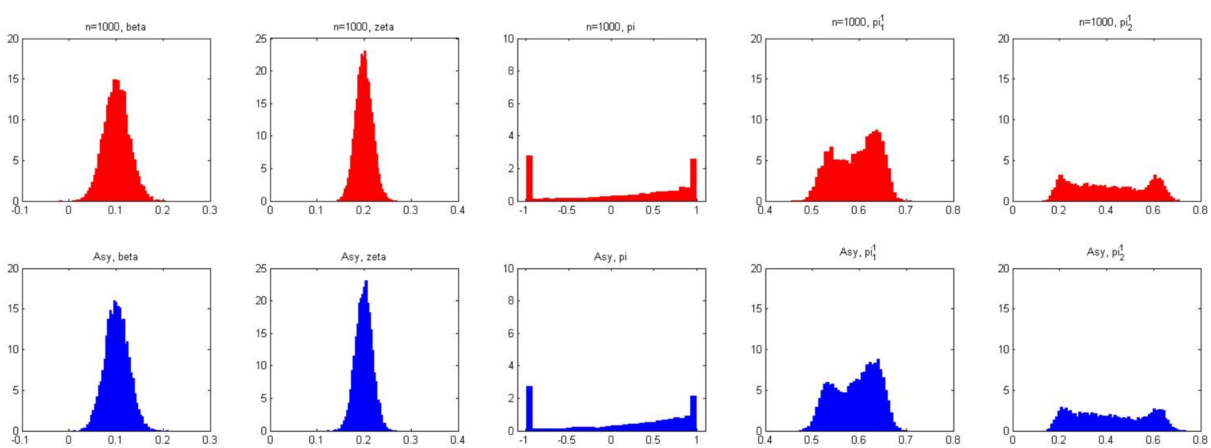
Stock, J. H., Wright, J. H., 2000. GMM with weak identification. Econometrica 68, 1055–1096.

Figure 1: Threshold Crossing Model Parameter Estimator Densities when $b = 0$



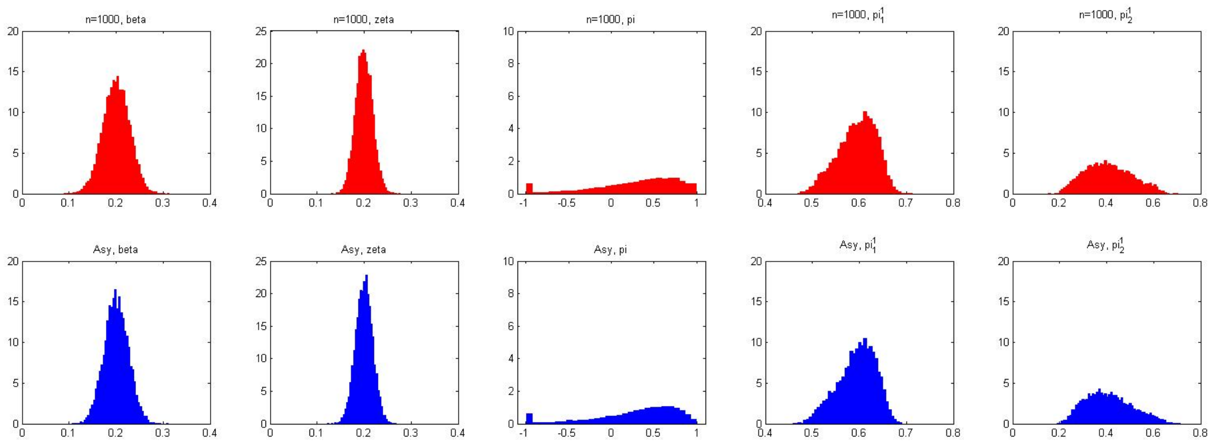
Asymptotic and finite-sample ($n = 1000$) densities of the estimators of β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$.

Figure 2: Threshold Crossing Model Parameter Estimator Densities when $b = \sqrt{n}0.1$



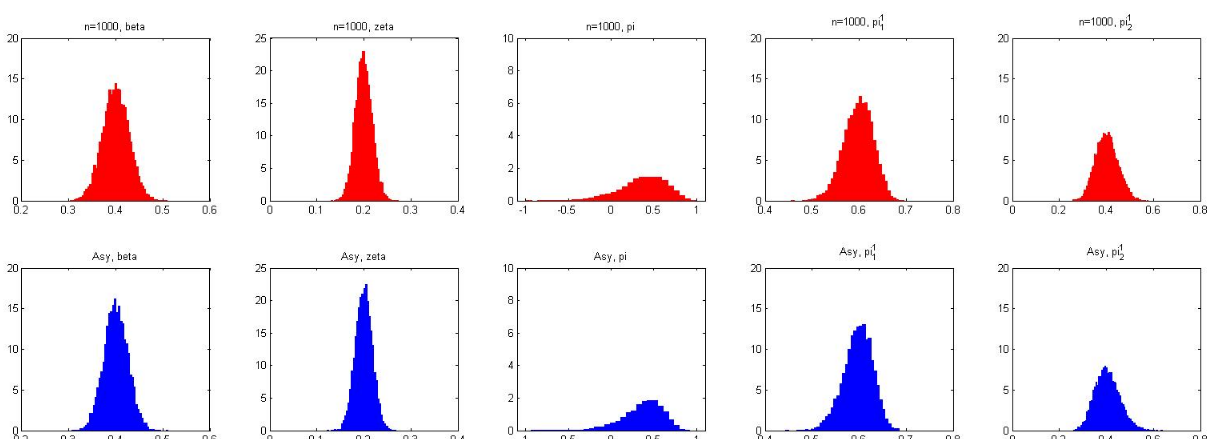
Asymptotic and finite-sample ($n = 1000$) densities of the estimators of β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$.

Figure 3: Threshold Crossing Model Parameter Estimator Densities when $b = \sqrt{n}0.2$



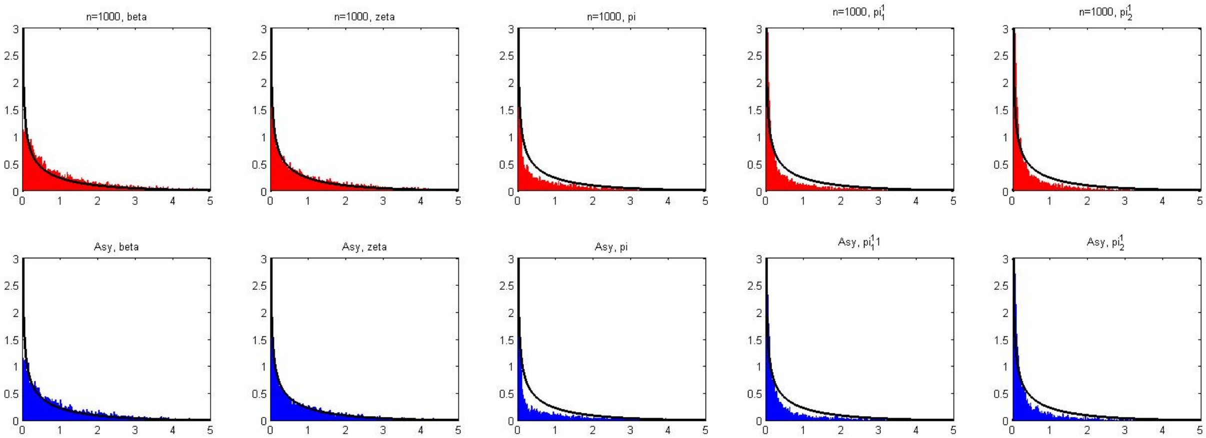
Asymptotic and finite-sample ($n = 1000$) densities of the estimators of β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$.

Figure 4: Threshold Crossing Model Parameter Estimator Densities when $b = \sqrt{n}0.4$



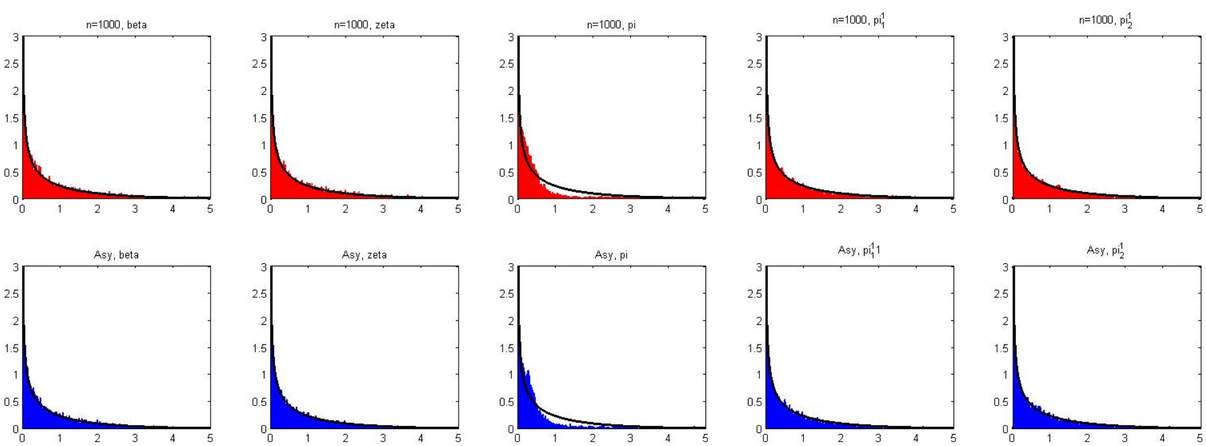
Asymptotic and finite-sample ($n = 1000$) densities of the estimators of β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$.

Figure 5: Wald Statistic Densities for the Threshold Crossing Model when $b = 0$



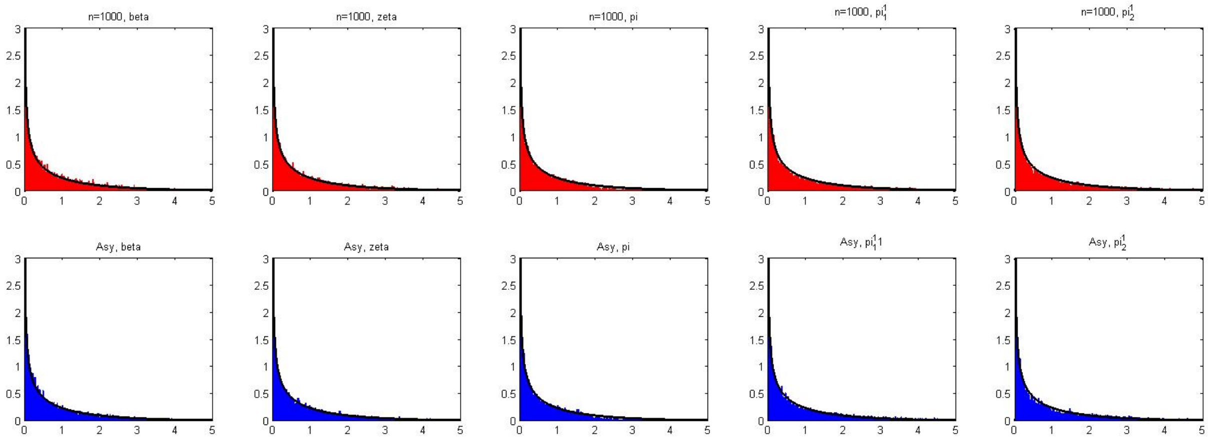
Asymptotic and finite-sample ($n = 1000$) densities of the Wald statistic for the parameters β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$, with a χ_1^2 density overlay (black line).

Figure 6: Wald Statistic Densities for the Threshold Crossing Model when $b = \sqrt{n}0.1$



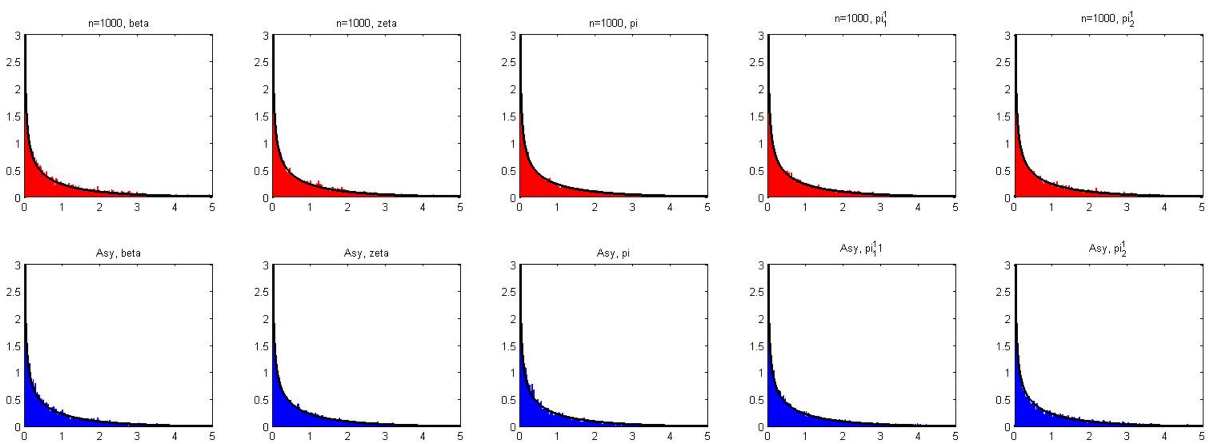
Asymptotic and finite-sample ($n = 1000$) densities of the Wald statistic for the parameters β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$, with a χ_1^2 density overlay (black line).

Figure 7: Wald Statistic Densities for the Threshold Crossing Model when $b = \sqrt{n}0.2$



Asymptotic and finite-sample ($n = 1000$) densities of the Wald statistic for the parameters β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$, with a χ_1^2 density overlay (black line).

Figure 8: Wald Statistic Densities for the Threshold Crossing Model when $b = \sqrt{n}0.4$



Asymptotic and finite-sample ($n = 1000$) densities of the Wald statistic for the parameters β, ζ, π, π_1^1 and π_2^1 in the Threshold-Crossing model when $\zeta_0 = 0.2$, $\pi_0 = 0.4$, $\pi_{1,0}^1 = 0.6$ and $\pi_{2,0}^1 = 0.4$, with a χ_1^2 density overlay (black line).